CHAPTER 12
Introduction to Time Series Regression and Forecasting

Time series data—data collected for a single entity at multiple points in time—can be used to answer quantitative questions for which cross-sectional data are inadequate. One such question is, what is the causal effect on a variable of interest, $Y$, of a change in another variable, $X$, over time? In other words, what is the dynamic causal effect on $Y$ of a change in $X$? For example, what is the effect on traffic fatalities of a law requiring passengers to wear seatbelts, both initially and subsequently as drivers adjust to the law? Another such question is, what is your best forecast of the value of some variable at a future date? For example, what is your best forecast of next month's rate of inflation, interest rates, or stock prices? Both of these questions—one about dynamic causal effects, the other about economic forecasting—can be answered using time series data. But time series data pose special challenges, and overcoming those challenges requires some new techniques.

Chapters 12–14 introduce techniques for the econometric analysis of time series data and apply these techniques to the problems of forecasting and estimating dynamic causal effects. Chapter 12 introduces the basic concepts and tools of regression with time series data and applies them to economic forecasting. In Chapter 13, the concepts and tools developed in Chapter 12 are applied to the problem of estimating dynamic causal effects using time series data. Chapter 14 takes up some more advanced topics in time series analysis, including forecasting multiple time series and modeling changes in volatility over time.

The empirical problem studied in this chapter is forecasting the rate of inflation, that is, the percentage increase in overall prices. While in a sense forecasting is just an application of regression analysis, forecasting is quite
different from the estimation of causal effects, the focus of this book until now. As discussed in Section 12.1, models that are useful for forecasting need not have a causal interpretation: if you see pedestrians carrying umbrellas you might forecast rain, even though carrying an umbrella does not cause it to rain. Section 12.2 introduces some basic concepts of time series analysis and presents some examples of economic time series data. Section 12.3 presents time series regression models in which the regressors are past values of the dependent variable; these "autoregressive" models use the history of inflation to forecast its future. Often, forecasts based on autoregressions can be improved by adding additional predictor variables and their past values, or "lags," as regressors, and these so-called autoregressive distributed lag models are introduced in Section 12.4. For example, we find that inflation forecasts made using lagged values of the rate of unemployment in addition to lagged inflation—that is, forecasts based on an empirical Phillips curve—improve upon the autoregressive inflation forecasts. A practical issue is deciding how many past values to include in autoregressions and autoregressive distributed lag models, and Section 12.5 describes methods for making this decision.

The assumption that the future will be like the past is an important one in time series regression, sufficiently so that it is given its own name, "stationarity." Time series variables can fail to be stationary in various ways, but two are especially relevant for regression analysis of economic time series data: (1) the series can have persistent, long-run movements, that is, the series can have trends; and (2) the population regression can be unstable over time, that is, the population regression can have breaks. These departures from stationarity jeopardize forecasts and inferences based on time series regression. Fortunately, there are statistical procedures for detecting trends and breaks and, once detected, for adjusting the model specification. These procedures are presented in Sections 12.6 and 12.7.

### 12.1 Using Regression Models for Forecasting

Chapter 4 began by considering the problem of a school superintendent who wants to know how much test scores would increase if she reduces class sizes in her school district; that is, the superintendent wants to know the causal effect on test scores of a change in class size. Accordingly, Parts II and III focused on using regression analysis to estimate causal effects using observational data.

Now consider a different problem, that of a parent moving to a metropolitan area and choosing a town within that area in part based on the local school system. The parent would like to know how the different school districts perform on standardized tests. Suppose, however, that test score data are unavailable (perhaps they are confidential) but data on class sizes are. The parent thus must guess at how well the different districts perform on standardized tests based on a limited amount of information. That is, the parent's problem is to forecast average test scores in a given district based on information related to test scores, such as class size.

The superintendent's problem and the parent's problem are conceptually very different. Multiple regression is a powerful tool for both, but because the problems are different, the criteria used to assess the suitability of a particular regression model is different as well. To obtain the credible estimates of causal effects desired by the superintendent, we must worry about the issues raised in Chapter 7: omitted variable bias, selection, simultaneous causality, and so forth. In contrast, to obtain the reliable forecast desired by the parent, it is important that the estimated regression have good explanatory power, that its coefficients be estimated precisely, and that it is stable in the sense that the regression estimated on one set of data can be reliably used to make forecasts using other data.

For example, recall the regression of test scores on the student-teacher ratio (STR) from Chapter 4:

\[
\text{TestScore} = 698.9 - 2.28 \times STR. \tag{12.1}
\]

We concluded that this regression is not useful for the superintendent: the OLS estimator of the slope is biased because of omitted variables such as the composition of the student body and their other learning opportunities outside school. The superintendent cannot change the district's average income or the fraction of non-English speakers, both of which affect test scores. Because these variables are also correlated with class size, there is omitted variable bias. Thus the regression
of test scores on the student-teacher ratio yields a biased estimator of the effect on test scores of a change in the student-teacher ratio, and Equation (12.1) does not answer the superintendent's question.

Nevertheless, Equation (12.1) could be useful to the parent trying to choose a district. To be sure, class size is not the only determinant of test performance, but from the parent's perspective what matters is whether it is a reliable predictor of test performance. The parent interested in forecasting test scores does not care whether the coefficient in Equation (12.1) estimates the causal effect on test scores of class size. Rather, the parent simply wants the regression to explain much of the variation in test scores across districts and to be stable, that is, to apply to the districts to which the parent is considering moving. Although omitted variable bias makes Equation (12.1) useless for answering the causal question, it still can be useful for forecasting.

The applications in this chapter are different than the test score/class size prediction problem because this chapter focuses on using time series data to forecast future events. Yet time series forecasting is similar conceptually to the parent's problem: the task is to use the known values of some variables (current and past values of the rate of price inflation instead of class size) to forecast the value of another variable (future inflation instead of test scores). As in the parent's problem, regression models can produce reliable forecasts, even if their coefficients have no causal interpretation. In Chapter 13, we return to problems like that faced by the school superintendent and discuss the estimation of causal effects using time series variables.

2 Introduction to Time Series Data and Serial Correlation

This section introduces some basic concepts and terminology that arise in time series econometrics. A good place to start any analysis of time series data is by plotting the data, so that is where we begin.

The Rates of Inflation and Unemployment in the United States

Figure 12.1a plots the U.S. rate of inflation—the annual percentage change in prices in the United States, as measured by the Consumer Price Index (CPI)—from 1960 to 1999 (the data are described in Appendix 12.1). The inflation rate...
Lags, First Differences, Logarithms, and Growth Rates

The observation on the time series variable \( Y \) made at date \( t \) is denoted \( Y_t \), and the total number of observations is denoted \( T \). The interval between observations, that is, the period of time between observation \( t \) and observation \( t+1 \), is some unit of time such as weeks, months, quarters (three-month units), or years. For example, the inflation data studied in this chapter are quarterly, so the unit of time (a "period") is a quarter of a year.

Special terminology and notation are used to indicate future and past values of \( Y \). The value of \( Y \) in the previous period is called its first lagged value or, more simply, its first lag, and is denoted \( Y_{t-1} \). Its \( j \)th lagged value (or simply its \( j \)th lag) is its \( j \) periods ago, which is \( Y_{t-j} \). Similarly \( Y_{t+1} \) denotes the value of \( Y \) one period into the future.

The change in the value of \( Y \) between period \( t-1 \) and period \( t \) is \( Y_t - Y_{t-1} \); this change is called the first difference in the variable \( Y \). In time series data, "\( \Delta \)" is used to represent the first difference, so that \( \Delta Y_t = Y_t - Y_{t-1} \).

Economic time series are often analyzed after computing their logarithms or the changes in their logarithms. One reason for this is that many economic series, such as gross domestic product (GDP), exhibit growth that is approximately exponential, that is, over the long run the series tends to grow by a certain percentage per year on average; if so, the logarithm of the series grows approximately linearly. Another reason is that the standard deviation of many economic time series is approximately proportional to its level, that is, the standard deviation is well expressed as a percentage of the level of the series; if so, then the standard deviation of the logarithm of the series is approximately constant. In either case, it is useful to transform the series so that changes in the transformed series are proportional (or percentage) changes in the original series, and this is achieved by taking the logarithm of the series.¹

Lags, first differences, and growth rates are summarized in Key Concept 12.1.

Lags, changes, and percentage changes are illustrated using the U.S. inflation rate in Table 12.1. The first column shows the date, or period, where the first quarter of 1999 is denoted 1999:Q1, the second quarter of 1999 is denoted 1999:Q2, and so forth. The second column shows the value of the CPI in that quarter, and the third column shows the rate of inflation. For example, from the first to the second quarter of 1999, the index increased from 164.9 to 166.0, a percentage increase of \( 100 \times (166.03 - 164.87) / 164.87 = 0.704\% \). This is the percentage increase from one quarter to the next. It is conventional to report rates of inflation (and other growth rates in macroeconomic time series) on an annual basis, which is the percentage increase in prices that would occur over a year, if the series were to continue to increase at the same rate. Because there are four quarters a year, the annualized rate of inflation in 1999:II is \( 0.704 \times 4 = 2.82 \), or 2.8% per year after rounding.

This percentage change can also be computed using the differences-of-logarithms approximation in Key Concept 12.1. The difference in the logarithm of the CPI from 1999:Q1 to 1999:II is \(\ln(166.03) - \ln(164.87) = 0.00701\), yielding

¹ Recall from Section 6.2 that the change of the logarithm of a variable is approximately equal to the proportional change of that variable; that is, \( \ln(X + \Delta X) \approx \ln(X) + \Delta \ln(X) \). This approximation works best when \( \Delta X \) is small. Now, replace \( X \) with \( Y_{t-1} \), \( \Delta X \) with \( \Delta Y_t \), and note that \( \Delta Y_t = Y_t - Y_{t-1} \). This means that the proportional change in the series \( Y_t \) between periods \( t-1 \) and \( t \) is approximately \( \ln(Y_t) - \ln(Y_{t-1}) \approx \ln(Y_{t-1}) + \Delta Y_t - \ln(Y_{t-1}) = \Delta \ln(Y_t) \). The expression \( \Delta \ln(Y_t) \) is the first difference of \( \ln(Y_t) \), \( \Delta \ln(Y_t) \). Thus \( \Delta \ln(Y_t) \approx \Delta Y_t / Y_{t-1} \). The percentage change is 100 times the fractional change, so the percentage change in the series \( Y_t \) is approximately \( 100 \Delta \ln(Y_t) \).
The approximate quarterly percentage difference $100 \times 0.00701 = 0.701\%$. On an annualized basis, this is $0.701 \times 4 = 2.80$, or 2.8\% after rounding, the same as obtained by directly computing the percentage growth. These calculations can be summarized as

$$\text{annualized rate of inflation} = \text{Inf}_t = 400 \ln(CPI_t) - \ln(CPI_{t-4}) = 400 \Delta \ln(CPI_t),$$

where $CPI_t$ is the value of the Consumer Price Index at date $t$. The factor of 400 arises from converting fractional change to percentages (multiplying by 100) and converting quarterly percentage change to an equivalent annual rate (multiplying by 4).

The final two columns of Table 12.1 illustrate lags and changes. The first lag of inflation in 1999:II is 1.6\%, the inflation rate in 1999:II was 1.6\% - 1.6\% = 1.2\%.

### Autocorrelation

In time series data, the value of $Y$ in one period typically is correlated with its value in the next period. The correlation of a series with its own lagged values is called autocorrelation or serial correlation. The first autocorrelation (or autocorrelation coefficient) is the correlation between $Y_t$ and $Y_{t-1}$, that is, the correlation between values of $Y$ at two adjacent dates. The second autocorrelation is the correlation between $Y_t$ and $Y_{t-2}$, and the $j^{th}$ autocorrelation is the correlation between $Y_t$ and $Y_{t-j}$. Similarly, the $j^{th}$ autocovariance is the covariance between $Y_t$ and $Y_{t-j}$. Autocorrelation and autocovariance are summarized in Key Concept 12.2.

The $j^{th}$ population autocovariances and autocorrelations in Key Concept 12.2 can be estimated by the $j^{th}$ sample autocovariances and autocorrelations, $	ext{cov}(Y_t, Y_{t-j})$ and $\hat{\rho}_j$:

$$\text{cov}(Y_t, Y_{t-j}) = \frac{1}{T-j} \sum_{t=j+1}^{T} (Y_t - \overline{Y}_{1:T})(Y_{t-j} - \overline{Y}_{1:T})$$

$$\hat{\rho}_j = \frac{\text{cov}(Y_t, Y_{t-j})}{\text{var}(Y_t)}$$

where $\overline{Y}_{1:T}$ denotes the sample average of $Y_t$ computed over the observations $t = j + 1, \ldots, T$ and where $\text{var}(Y_t)$ is the sample variance of $Y$. (Equation (12.6) uses the assumption that $\text{var}(Y_t)$ and $\text{var}(Y_{t-j})$ are the same, an implication of the assumption that $Y$ is stationary, which is discussed in Section 12.4.)

The first four sample autocovariances of the inflation rate and of the change in the inflation rate are listed in Table 12.2. These entries show that inflation is strongly positively autocorrelated: the first autocorrelation is 0.85. The sample autocorrelation declines as the lag increases, but it remains large even at a lag of four quarters. The change in inflation is negatively autocorrelated: an increase in
the rate of inflation in one quarter tends to be associated with a decrease in the next quarter.

At first, it might seem contradictory that the level of inflation is strongly positively correlated but its change is negatively correlated. These two autocorrelations, however, measure different things. The strong positive autocorrelation in inflation reflects the long-term trends in inflation evident in Figure 12.1: inflation was low in the first quarter of 1965 and again in the second; it was high in the first quarter of 1981 and again in the second. In contrast, the negative autocorrelation of the change of inflation means that, on average, an increase in inflation in one quarter is associated with a decrease in inflation in the next.

Other Examples of Economic Time Series

Economic time series differ greatly. Four examples of economic time series are plotted in Figure 12.2: the U.S. Federal Funds interest rate; the rate of exchange between the dollar and the British pound; the logarithm of real Japanese gross domestic product; and the daily return on the Standard and Poor’s 500 (S&P 500) stock market index.

The U.S. Federal Funds rate (Figure 12.2a) is the interest rate that banks pay to each other to borrow funds overnight. This rate is important because it is controlled by the Federal Reserve and is the Fed’s primary monetary policy instrument. If you compare the plots of the Federal Funds rate and the rates of unemployment and inflation in Figure 12.1, you will see that sharp increases in the Federal Funds rate often have been associated with subsequent recessions.

The dollar/pound exchange rate (Figure 12.2b) is the price of a British pound (£) in U.S. dollars. Before 1972, the developed economies ran a system of fixed exchange rates—called the “Bretton Woods” system—under which governments worked to keep exchange rates from fluctuating. In 1972, inflationary pressures led to the breakdown of this system; thereafter, the major currencies were allowed to “float,” that is, their values were determined by the supply and demand for currencies in the market for foreign exchange. Prior to 1972, the exchange rate was approximately constant, with the exception of a single devaluation in 1968 in which the official value of the pound, relative to the dollar, was decreased to $2.40. Since 1972 the exchange rate has fluctuated over a very wide range.
3 Autoregressions

What will be the rate of price inflation—the percentage increase in overall prices—be next year? Wall Street investors rely on forecasts of inflation when deciding how much to pay for bonds. Economists at central banks, like the U.S. Federal Reserve Bank, use inflation forecasts when they set monetary policy. Firms use inflation forecasts when they forecast sales of their product, and local governments use inflation forecasts when they develop their budget for the upcoming year. In this section, we consider forecasts made using an autoregression, a regression model that relates a time series variable to its past values.

The First Order Autoregressive Model

If you want to predict the future of a time series, a good place to start is in the immediate past. For example, if you want to forecast the change in inflation from this quarter to the next, you might see whether inflation rose or fell last quarter. A systematic way to forecast the change in inflation, $\Delta \text{Inf}_t$, using the previous quarter’s change, $\Delta \text{Inf}_{t-1}$, is to estimate an OLS regression of $\Delta \text{Inf}_t$ on $\Delta \text{Inf}_{t-1}$. Estimated using data from 1962–1999, this regression is

$$
\Delta \text{Inf}_t = 0.02 - 0.211 \Delta \text{Inf}_{t-1},
$$

(12.7)

where, as usual, standard errors are given in parentheses under the estimated coefficients, and $\Delta \text{Inf}_t$ is the predicted value of $\Delta \text{Inf}_t$ based on the estimated regression line. The model in Equation (12.7) is called a first order autoregression: an autoregression because it is a regression of the series onto its own lag, $\Delta \text{Inf}_{t-1}$, and first order because only one lag is used as a regressor. The coefficient in Equation (12.7) is negative, so an increase in the inflation rate in one quarter is associated with a decline in the inflation rate in the next quarter.

A first order autoregression is abbreviated by AR(1), where the “1” indicates that it is first order. The population AR(1) model for the series $Y_t$ is

$$
Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t,
$$

(12.8)

where $u_t$ is an error term.

Forecasts and forecast errors. Suppose you have historical data on $Y_t$ and you want to forecast its future value. If $Y_t$ follows the AR(1) model in Equation (12.8) and if $\beta_0$ and $\beta_1$ are known, then the forecast of $Y_t$ based on $Y_{t-1}$ is $\hat{Y}_{t|t-1} = \beta_0 + \beta_1 Y_{t-1}$.

In practice, $\beta_0$ and $\beta_1$ are unknown, so forecasts must be based on estimates of $\beta_0$ and $\beta_1$. We will use the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ which are constructed using historical data. In general, $\hat{Y}_{t|t-1}$ will denote the forecast of $Y_t$ based on information through period $t-1$ using a model estimated with data through period $t-1$. Accordingly, the forecast based on the AR(1) model in Equation (12.8) is

$$
\hat{Y}_{t|t-1} = \hat{\beta}_0 + \hat{\beta}_1 Y_{t-1},
$$

(12.9)

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimated using historical data through time $t-1$.

The forecast error is the mistake made by the forecast; this is the difference between the value of $Y_t$ that actually occurred and its forecasted value based on $Y_{t-1}$:

$$
\text{forecast error} = Y_t - \hat{Y}_{t|t-1}.
$$

(12.10)
Forecasts vs. predicted values. The forecast is not an OLS predicted value, and the forecast error is not an OLS residual. OLS predicted values are calculated for the observations in the sample used to estimate the regression. In contrast, the forecast is made for some date beyond the data set used to estimate the regression, so the data on the actual value of the forecasted dependent variable are not in the sample used to estimate the regression. Similarly, the OLS residual is the difference between the actual value of Y and its predicted value for observations in the sample, whereas the forecast error is the difference between the future value of Y, which is not contained in the estimation sample, and the forecast of that future value. Said differently, forecasts and forecast errors pertain to "out-of-sample" observations, whereas predicted values and residuals pertain to "in-sample" observations.

Root mean squared forecast error. The root mean squared forecast error (RMSFE) is a measure of the size of the forecast error, that is, of the magnitude of a typical mistake made using a forecasting model. The RMSFE is the square root of the mean squared forecast error:

\[
\text{RMSFE} = \sqrt{\text{E}[(Y_t - \hat{Y}_{t|t-1})^2]}.
\] (12.11)

The RMSFE has two sources of error: the error arising because future values of \( u_t \) are unknown, and the error in estimating the coefficients \( \beta_0 \) and \( \beta_1 \). If the first source of error is much larger than the second, as it can be if the sample size is large, then the RMSFE is approximately \( \sqrt{\text{var}(u_t)} \), the standard deviation of the error \( u_t \) in the population autoregression (Equation (12.8)). The standard deviation of \( u_t \) is in turn estimated by the standard error of the regression (SER, see Section 5.10). Thus, if uncertainty arising from estimating the regression coefficients is small enough to be ignored, the RMSFE can be estimated by the standard error of the regression. Estimation of the RMSFE including both sources of forecast error is taken up in Section 12.4.

Application to inflation. What is the forecast of inflation in the first quarter of 2000 (2000:1) that a forecaster would have made in 1999:IV, based on the estimated AR(1) model in Equation (12.7) (which was estimated using data through 1999:IV)? From Table 12.1, the inflation rate in 1999:IV was 3.2% (so \( \text{Inf}_{1999:IV} = 3.2\% \)), an increase of 0.4 percentage points from 1999:III (so \( \Delta \text{Inf}_{1999:IV} = 0.4 \)). Plugging these values into Equation (12.7), the forecast of the change in inflation from 1999:IV to 2000:1 is \( \Delta \text{Inf}_{2000:1} = 0.02 - 0.211 \times \Delta \text{Inf}_{1999:IV} = 0.02 - 0.211 \times 0.4 = -0.06 = -0.1 \) (rounded to the nearest tenth). The predicted rate of inflation is the past rate of inflation plus its predicted change:

\[
\Delta \text{Inf}_{2000:1} = \Delta \text{Inf}_{1999:IV} = \Delta \text{Inf}_{1999:IV} + \Delta \text{Inf}_{1999:IV} = 3.2\% - 0.1\% = 3.1\%.
\] (12.12)

Because \( \text{Inf}_{1999:IV} = 3.2\% \) and the predicted change in the inflation rate from 1999:IV to 2000:1 is -0.1, the predicted rate of inflation in 2000:1 is \( \text{Inf}_{2000:1} = \text{Inf}_{1999:IV} + \Delta \text{Inf}_{1999:IV} = 3.2\% - 0.1\% = 3.1\% \). Thus, the AR(1) model forecasts that inflation will drop slightly from 3.2% in 1999:IV to 3.1% in 2000:1.

How accurate was this AR(1) forecast? From Table 12.1, the actual value of inflation in 2000:1 was 4.1%, so the AR(1) forecast is low by a full percentage point; that is, the forecast error is 1.0%. The \( R^2 \) of the AR(1) model in Equation (12.7) is only 0.04, so the lagged change of inflation explains a very small fraction of the variation in inflation in the sample used to fit the autoregression. This low \( R^2 \) is consistent with the poor forecast of inflation in 2000:1 produced using Equation (12.7). More generally, the low \( R^2 \) suggests that this AR(1) model will forecast only a small amount of the variation in the change of inflation.

The standard error of the regression in Equation (12.7) is 1.67; ignoring uncertainty arising from estimation of the coefficients, our estimate of the RMSFE for forecasts based on Equation (12.7) therefore is 1.67 percentage points.

The \( p \)th Order Autoregressive Model
The AR(1) model uses \( Y_{t-1} \) to forecast \( Y_t \), but doing so ignores potentially useful information in the more distant past. One way to incorporate this information is to include additional lags in the AR(1) model; this yields the \( p \)th order autoregressive, or AR(\( p \)), model.

The \( p \)th order autoregressive model (the AR(\( p \)) model) represents \( Y_t \) as a linear function of \( p \) of its lagged values; that is, in the AR(\( p \)) model, the regressors are \( Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p} \) plus an intercept. The number of lags, \( p \), included in an AR(\( p \)) model is called the order, or lag length, of the autoregression.

For example, an AR(4) model of the change in inflation uses four lags of the change in inflation as regressors. Estimated by OLS over the period 1962-1999, the AR(4) model is

\[
\Delta \text{Inf}_{t} = 0.02 - 0.21 \Delta \text{Inf}_{t-1} - 0.32 \Delta \text{Inf}_{t-2} + 0.19 \Delta \text{Inf}_{t-3} - 0.04 \Delta \text{Inf}_{t-4} \quad (12.13)
\]

The coefficients on the final three additional lags in Equation (12.13) are jointly significantly different from zero at the 5% significance level: the F-statistic is 6.43
of the four quarters of 1999 into Equation (12.13): \( \Delta \text{Inf}_{2000:1-1999:IV} = 0.02 - 0.21 \Delta \text{Inf}_{1999:IV} - 0.32 \Delta \text{Inf}_{1999:III} + 0.19 \Delta \text{Inf}_{1999:II} - 0.04 \Delta \text{Inf}_{1999:I} = 0.02 - 0.21 \times 0.4 - 0.32 \times 0.0 + 0.19 \times 1.1 - 0.04 \times (-0.4) = 0.2 \), where the 1999 values for the change of inflation are taken from the final column of Table 12.1.

The corresponding forecast of inflation in 2000:1 is the value of inflation in 1999:IV, plus the forecasted change, that is, 3.2% + 0.2% = 3.4%. The forecast error is the actual value, 4.1%, minus the forecast, or 4.1% - 3.4% = 0.7%, slightly smaller than the AR(1) forecast error of 1.0%.

### 12.4 Time Series Regression with Additional Predictors and the Autoregressive Distributed Lag Model

Economic theory often suggests other variables that could help to forecast the variable of interest. These other variables, or predictors, can be added to an autoregression to produce a time series regression model with multiple predictors. When other variables and their lags are added to an autoregression, the result is an autoregressive distributed lag model.

#### Forecasting Changes in the Inflation Rate Using Past Unemployment Rates

A high value of the unemployment rate tends to be associated with a future decline in the rate of inflation. This negative relationship, known as the short-run Phillips curve, is evident in the scatterplot of Figure 12.3, in which year-to-year changes in the rate of price inflation are plotted against the rate of unemployment in the previous year. For example, in 1982 the unemployment rate averaged 9.9%, and during the next year the rate of inflation fell by 2.9%. Overall, the correlation in Figure 12.3 is -0.40.

The scatterplot in Figure 12.3 suggests that past values of the unemployment rate might contain information about the future course of inflation that is not already contained in past changes of inflation. This conjecture is readily checked by augmenting the AR(4) model in Equation (12.13) to include the first lag of the unemployment rate:

\[
\Delta \text{Inf}_t = 1.42 - 0.26 \Delta \text{Inf}_{t-1} - 0.40 \Delta \text{Inf}_{t-2} + 0.11 \Delta \text{Inf}_{t-3} - 0.09 \Delta \text{Inf}_{t-4} - 0.23 \text{Unemp}_{t-1} - 0.15 \text{Unemp}_{t-2} - 0.07 \text{Unemp}_{t-3} - 0.03 \text{Unemp}_{t-4}.
\]
The \( t \)-statistic on \( \text{Unemp}_{t-1} \) is \(-2.33\), so this term is significant at the 5% level. The \( R^2 \) of this regression is 0.22, a small improvement over the AR(4) \( R^2 \) of 0.21.

The forecast of the change in inflation in 2000:I is obtained by substituting the 1999 values of the change of inflation into Equation (12.16), along with the value of the unemployment rate in 1999:IV (which is 4.1%); the resulting forecast is \( \Delta \text{Inf}_{2000:1|1999:IV} = 0.5 \). Thus the forecast of inflation in 2000:I is 3.2% + 0.5% = 3.7%, and the forecast error is 0.4%. This forecast is closer to actual 2000:I inflation than was the AR(4) forecast.

If one lag of the unemployment rate is helpful for forecasting inflation, several lags might be even more helpful; adding three more lags of the unemployment rate yields

\[
\Delta \text{Inf}_t = 1.32 - 0.36 \Delta \text{Inf}_{t-1} - 0.34 \Delta \text{Inf}_{t-2} + 0.07 \Delta \text{Inf}_{t-3} - 0.03 \Delta \text{Inf}_{t-4} \\
(0.47) \quad (0.09) \quad (0.10) \quad (0.08) \quad (0.09)
\]

\[
-2.68 \text{Unemp}_{t-1} + 3.43 \text{Unemp}_{t-2} - 1.04 \text{Unemp}_{t-3} + 0.07 \text{Unemp}_{t-4} \\
(0.47) \quad (0.89) \quad (0.89) \quad (0.44)
\] (12.17)

The \( F \)-statistic testing the joint significance of the second through fourth lags of the unemployment rate is 4.93 (\( p \)-value = 0.003), so they are jointly significant.

The \( R^2 \) of the regression in Equation (12.17) is 0.35, a solid improvement over 0.22 for Equation (12.16). The \( F \)-statistic on all the unemployment coefficients is 8.51 (\( p \)-value < 0.001), indicating that this model represents a statistically significant improvement over the AR(4) model of Section 12.3 (Equation (12.13)). The standard error of the regression in Equation (12.17) is 1.37, a substantial improvement over the SER of 1.53 for the AR(4).

The forecast change in inflation from 1999:IV to 2000:I using Equation (12.17) is computed by substituting the values of the variables into the equation. The unemployment rate was 4.3% in 1999:I and 4.2% in 1999:II, and 4.1% in 1999:IV. The forecast of the change in inflation from 1999:IV to 2000:I, based on Equation (12.17), is

\[
\Delta \text{Inf}_{2000:1|1999:IV} = 1.32 - 0.36 \times 0.4 - 0.34 \times 0.0 + 0.07 \times 1.1 - 0.03 \times (-0.4) - 2.68 \times 4.1 + 3.43 \times 4.2 - 1.04 \times 4.3 + 0.07 \times 4.3 = 0.5
\] (12.18)

Thus the forecast of inflation in 2000:I is 3.2% + 0.5% = 3.7%. The forecast error is small, 0.4. Adding multiple lags of the unemployment rate appears to improve inflation forecasts beyond those of an AR(4).

The autoregressive distributed lag model. The models in Equations (12.16) and (12.17) are autoregressive distributed lag (ADL) models; “autoregressive” because lagged values of the dependent variable are included as regressors, as in an autoregression, and “distributed lag” because the regression also includes multiple lags (a “distributed lag”) of an additional predictor. In general, an autoregressive distributed lag model with \( p \) lags of the dependent variable \( Y \) and \( q \) lags of an additional predictor \( X \), is called an ADL\((p,q)\) model. In this notation, the model in Equation (12.16) is an ADL\((4,1)\) model and the model in Equation (12.17) is an ADL\((4,4)\) model.

The autoregressive distributed lag model is summarized in Key Concept 12.4. With all these regressors, the notation in Equation (12.19) is somewhat cumbersome, and alternative optional notation, based on the so-called lag operator, is presented in Appendix 12.3.

The assumption that the errors in the ADL model have a conditional mean of zero given all past values of \( Y \) and \( X \), that is, \( E(\epsilon_t | Y_{t-1}, Y_{t-2}, \ldots, X_{t-1}, X_{t-2}, \ldots) = 0 \), implies that no additional lags of either \( Y \) or \( X \) belong in the ADL model. In other words, the lag lengths \( p \) and \( q \) are the true lag lengths and the coefficients on additional lags are zero.

The ADL model contains lags of the dependent variable (the autoregressive component) and a distributed lag of a single additional predictor, \( X \). In general,
The Autoregressive Distributed Lag Model

The autoregressive distributed lag model with $p$ lags of $Y_t$ and $q$ lags of $X_t$, denoted $\text{ADL}(p,q)$, is

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p} + \delta_1 X_{t-1} + \delta_2 X_{t-2} + \ldots + \delta_q X_{t-q} + u_t,$$

where $\beta_0, \beta_1, \ldots, \beta_p, \delta_1, \ldots, \delta_q$ are unknown coefficients and $u_t$ is the error term with $E(u_t|Y_{t-1}, Y_{t-2}, \ldots, X_{t-p}, X_{t-q}, \ldots) = 0$.

however, forecasts can be improved by using multiple predictors. But before turning to the general time series regression model with multiple predictors, we first introduce the concept of stationarity, which will be used in that discussion.

Stationarity

Regression analysis of time series data necessarily uses data from the past to quantify historical relationships. If the future is like the past, then these historical relationships can be used to forecast the future. But if the future differs fundamentally from the past, then those historical relationships might not be reliable guides to the future.

In the context of time series regression, the idea that historical relationships can be generalized to the future is formalized by the concept of stationarity. The precise definition of stationarity, given in Key Concept 12.5, is that the distribution of the time series variable does not change over time.

Time Series Regression with Multiple Predictors

The general time series regression model with multiple predictors extends the ADL model to include multiple predictors and their lags. The model is summarized in Key Concept 12.6. The presence of multiple predictors and their lags leads to double subscripting of the regression coefficients and regressors.

The time series regression model assumptions. The assumptions in Key Concept 12.6 modify the four least squares assumptions of the multiple regression model for cross-sectional data (Key Concept 5.4) for time series data.

A time series $Y_t$ is stationary if its probability distribution does not change over time, that is, if the joint distribution of $(Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p})$ does not depend on $t$; otherwise, $Y_t$ is said to be nonstationary. A pair of time series, $X_t$ and $Y_t$, are said to be jointly stationary if the joint distribution of $(X_{t-1}, Y_{t-1}, X_{t-2}, Y_{t-2}, \ldots, X_{t-p}, Y_{t-p})$ does not depend on $t$. Stationarity requires the future to be like the past, at least in a probabilistic sense.

The first assumption is that $u_t$ has conditional mean zero, given all the regressors and the additional lags of the regressors beyond the lags included in the regression. This assumption extends the assumption used in the AR and ADL models and implies that the best forecast of $Y_t$ using all past values of $Y$ and the $X$'s is given by the regression in Equation (12.20).

The second least squares assumption for cross-sectional data (Key Concept 5.4) is that $(X_{ti}, \ldots, X_{tk}, Y_t), i = 1, \ldots, n,$ are independently and identically distributed (i.i.d.). The second assumption for time series regression replaces the i.i.d. assumption by a more appropriate one with two parts. Part (a) is that the data are drawn from a stationary distribution, so that the distribution of the data today is the same as its distribution in the past. This assumption is a time series version of the "identically distributed" part of the i.i.d. assumption: the cross-sectional requirement of each draw being identically distributed is replaced by the time series requirement that the joint distribution of the variables, including lags, does not change over time. In practice, many economic time series appear to be nonstationary, which means that this assumption can fail to hold in applications. If the time series variables are nonstationary, then one or more problems can arise in time series regression: the forecast can be biased, the forecast can be inefficient (there can be alternative forecasts based on the same data with lower variance), or conventional OLS-based statistical inferences (for example performing a hypothesis test by comparing the OLS t-statistic to ±1.96) can be misleading. Precisely which of these problems occurs, and its remedy, depends on the source of the nonstationarity. In Sections 12.6 and 12.7, we study the problems posed by tests for, and solutions to two empirically important types of nonstationarity in economic time series, trends and breaks. For now, however, we simply assume that the series are jointly stationary and accordingly focus on regression with stationary variables.
Time Series Regression with Multiple Predictors

The general time series regression model allows for \( k \) additional predictors, where \( q_1 \) lags of the first predictor are included, \( q_2 \) lags of the second predictor are included, and so forth:

\[ y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_p y_{t-p} + \delta_1 x_{t-1} + \delta_2 x_{t-2} + \cdots + \delta_{k_1} x_{t-k_1} + \cdots + \delta_{k_2} x_{t-k_2} + \cdots + \delta_{kr} x_{t-k_r} + \epsilon_t, \tag{12.20} \]

where

1. \( E(\epsilon_t | y_{t-1}, y_{t-2}, \ldots, x_{t-k_1}, x_{t-k_2}, \ldots) = 0; \)
2. (a) The random variables \( (y_t, x_{t-j} ; j = 1, 2, \ldots, k) \) have a stationary distribution, and \( b) (y_t, x_{t-j} ; j = 1, 2, \ldots) \) become independent as \( j \) gets large;
3. \( x_{t-1}, \ldots, x_t \) and \( y_t \) have nonzero, finite fourth moments; and
4. There is no perfect multicollinearity.

Part (b) of the second assumption requires that the random variables become independently distributed when the amount of time separating them becomes large. This replaces the cross-sectional requirement that the variables be independently distributed from one observation to the next with the time series requirement that they be independently distributed when they are separated by long periods of time. This assumption is sometimes referred to as weak dependence, and it ensures that in large samples there is sufficient randomness in the data for the law of large numbers and the central limit theorem to hold. We do not provide a precise mathematical statement of the weak dependence condition, rather, the reader is referred to Hayashi (2000, Chapter 2).

The third assumption, which is the same as the third least squares assumption for cross-sectional data, is that all the variables have nonzero finite fourth moments.

Finally, the fourth assumption, which is also the same as for cross-sectional data, is that the regressors are not perfectly multicollinear.

Statistical inference and the Granger causality test. Under the assumptions of Key Concept 12.6, inference on the regression coefficients using OLS proceeds in the same way as it usually does using cross-sectional data.

Granger Causality Tests (Tests of Predictive Content)

The Granger causality statistic is the F-statistic testing the hypothesis that the coefficients on all the values of one of the variables in Equation (12.20) (for example, the coefficients on \( x_{t-1}, x_{t-2}, \ldots, x_{t-k_r} \)) are zero. This null hypothesis implies that these regressors have no predictive content for \( Y \) beyond that contained in the other regressors, and the test of this null hypothesis is called the Granger causality test.

One useful application of the F-statistic in time series forecasting is to test whether the lags of one of the included regressors has useful predictive content, above and beyond the other regressors in the model. The claim that a variable has no predictive content corresponds to the null hypothesis that the coefficients on all lags of that variable are zero. The F-statistic testing this null hypothesis is called the Granger causality statistic, and the associated test is called a Granger causality test (Granger (1969)). This test is summarized in Key Concept 12.7.

Granger causality has little to do with causality in the sense that it is used elsewhere in this book. In Chapter 1, causality was defined in terms of an ideal randomized controlled experiment, in which different values of \( X \) are applied experimentally and we observe the subsequent effect on \( Y \). In contrast, Granger causality means that if \( X \) Granger-causes \( Y \), then \( X \) is a useful predictor of \( Y \), given the other variables in the regression. While "Granger predictability" is a more accurate term than "Granger causality," the latter has become part of the jargon of econometrics.

As an example, consider the relationship between the change in the inflation rate and its past values and past values of the unemployment rate. Based on the OLS estimates in Equation (12.17), the F-statistic testing the null hypothesis that the coefficients on all four lags of the unemployment rate are zero is 8.51 (\( p < 0.001 \); in the jargon of Key Concept 12.7, we can conclude (at the 1% significance level) that the unemployment rate Granger-causes changes in the inflation rate. This does not necessarily mean that a change in the unemployment rate will cause—in the sense of Chapter 1—a subsequent change in the inflation rate. It does mean that the past values of the unemployment rate appear to contain information that is useful for forecasting changes in the inflation rate, beyond that contained in past values of the inflation rate.
Forecast Uncertainty and Forecast Intervals

In any estimation problem, it is good practice to report a measure of the uncertainty of that estimate, and forecasting is no exception. One measure of the uncertainty of a forecast is its root mean square forecast error. Under the additional assumption that the errors \( u_t \) are normally distributed, the RMSFE can be used to construct a forecast interval, that is, an interval that contains the future value of the variable with a certain probability.

**Forecast uncertainty.** The forecast error consists of two components: uncertainty arising from estimation of the regression coefficients, and uncertainty associated with the future unknown value of \( u_t \). For regression with few coefficients and many observations, the uncertainty arising from future \( u_t \) can be much larger than the uncertainty associated with estimation of the parameters. In general, however, both sources of uncertainty are important, so we now develop an expression for the RMSFE that incorporates these two sources of uncertainty.

To keep the notation simple, consider forecasts of \( Y_{T+1} \) based on an ADL(1,1) model with a single predictor, that is, \( Y_t = \beta_0 + \beta_1 X_{t-1} + \delta Y_{t-1} + u_t \), and suppose that \( u_t \) is homoskedastic. The forecast is \( \hat{Y}_{T+1|T} = \hat{\beta}_0 + \hat{\beta}_1 Y_T + \hat{\delta}_1 Y_T \) and the forecast error is

\[
Y_{T+1} - \hat{Y}_{T+1|T} = u_{T+1} - [(\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1) Y_T + (\hat{\delta}_1 - \delta_1) Y_T].
\]  
(12.21)

Because \( u_{T+1} \) has conditional mean zero and is homoskedastic, \( u_{T+1} \) has variance \( \sigma^2_u \) and is uncorrelated with the final expression in brackets in Equation (12.21). Thus the mean squared forecast error (MSFE) is

\[
\text{MSFE} = \mathbb{E}[(Y_{T+1} - \hat{Y}_{T+1|T})^2] = \sigma^2_u + \text{var}(\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1) \text{var}(Y_T) + (\hat{\delta}_1 - \delta_1) \text{var}(Y_T),
\]

and the RMSFE is the square root of the MSFE.

Estimation of the MSFE entails estimation of the two parts in Equation (12.22). The first term, \( \sigma^2_u \), can be estimated by the square of the standard error of the regression, as discussed in Section 12.3. The second term requires estimating the variance of a weighted average of the regression coefficients, and methods for doing so were discussed in Section 6.1 (see the discussion following Equation (6.7)).

An alternative method for estimating the MSFE is to use the variance of pseudo out-of-sample forecasts, a procedure discussed in Section 12.7.

**Forecast intervals.** A forecast interval is like a confidence interval, except that it pertains to a forecast. That is, a 95\% forecast interval is an interval that contains the future value of the series in 95\% of repeated applications.

One important difference between a forecast interval and a confidence interval is that the usual formula for a 95\% confidence interval (the estimator \( \pm 1.96 \) standard errors) is justified by the central limit theorem and therefore holds for a wide range of distributions of the error term. In contrast, because the forecast error in Equation (12.21) includes the future value of the error \( u_{T+1} \), to compute a forecast interval requires either estimating the distribution of the error term or making some assumption about that distribution.

In practice, it is convenient to assume that \( u_{T+1} \) is normally distributed. If so, Equation (12.21) and the central limit theorem applied to \( \hat{\beta}_0, \hat{\beta}_1 \), and \( \hat{\delta}_1 \) imply that the forecast error is the sum of two independent, normally distributed terms, so that the forecast error is itself normally distributed with variance equaling the MSFE. It follows that a 95\% confidence interval is given by \( \hat{Y}_{T+1|T} \pm 1.96 \text{SE}(Y_{T+1} - \hat{Y}_{T+1|T}) \), where \( \text{SE}(Y_{T+1} - \hat{Y}_{T+1|T}) \) is an estimator of the RMSFE.

This discussion has focused on the case that the error term, \( u_{T+1} \), is homoskedastic. If instead \( u_{T+1} \) is heteroskedastic, then one needs to develop a model of the heteroskedasticity so that the term \( \sigma^2_u \) in Equation (12.22) can be estimated, given the most recent values of \( Y \) and \( X \), and methods for modeling this conditional heteroskedasticity are presented in Section 14.5.

Because of uncertainty about future events—that is, uncertainty about \( u_{T+1} \)—95\% forecast intervals can be so wide that they have limited use in decision making. Professional forecasters therefore often report forecast intervals that are tighter than 95\%, for example, one standard error forecast intervals (which are 68\% forecast intervals if the errors are normally distributed). Alternatively, some forecasters report multiple forecast intervals, as is done by the economists at the Bank of England when they publish their inflation forecasts (see the River of Blood box on the following page).

12.5 Lag Length Selection Using Information Criteria

The estimated inflation regressions in Sections 12.3 and 12.4 have either one or four lags of the predictors. One lag makes some sense, but why four? More generally, how many lags should be included in a time series regression? This section discusses statistical methods for choosing the number of lags, first in an autoregression, then in a time series regression model with multiple predictors.
The River of Blood

As part of its efforts to inform the public about monetary policy decisions, the Bank of England regularly publishes forecasts of inflation. These forecasts combine output from econometric models maintained by professional econometricians at the bank with the expert judgment of the members of the bank’s senior staff and Monetary Policy Committee. The forecasts are presented as a set of forecast intervals designed to reflect what these economists consider to be the range of probable paths that inflation might take. In its Inflation Report, the bank prints these ranges in red, with the darkest red reserved for the central band. Although the bank prosaically refers to this as the “fan chart,” the press has called these spreading shades of red the “river of blood.”

The river of blood for February 2001 is shown in Figure 12.4 (in this figure the blood is green, not red, so you will need to use your imagination). This chart shows that, as of February 2001, the bank’s economists expected inflation to fall over the next year from approximately 3% to just over 2%, but then to increase. There is considerable uncertainty about this forecast, however. In their written discussion, they cited in particular the possibility of a further slowdown in the United States—which in fact became the recession of 2001—that could lead to lower inflation in the United Kingdom. As it happened, their forecast was a good one: in the fourth quarter of 2001, the rate of inflation was 2.0%.

The Bank of England has been a pioneer in the movement towards greater openness by central banks, and other central banks now also publish inflation forecasts. The decisions made by monetary policymakers are difficult ones and affect the lives—and wallets—of many of their fellow citizens. In a democracy in the information age, reasoned the economists at the Bank of England, it is particularly important for citizens to understand the bank’s economic outlook and the reasoning behind its difficult decisions.

To see the River of Blood in its original red hue, visit the Bank of England’s website at www.bankofengland.co.uk/inflationreport.

Determining the Order of an Autoregression

In practice, choosing the order \( p \) of an autoregression requires balancing the benefit of including more lags against the cost of additional estimation uncertainty. On the one hand, if the order of an estimated autoregression is too low, you will omit potentially valuable information contained in the more distant lagged values. On the other hand, if it is too high, you will be estimating more coefficients than necessary, which in turn introduces additional estimation error into your forecasts.

The \( F \)-statistic approach. One approach to choosing \( p \) is to start with a model with many lags and to perform hypothesis tests on the final lag. For example, you might start by estimating an AR(6) and test whether the coefficient on the sixth lag is significant at the 5% level; if not, drop it and estimate an AR(5), test the coefficient on the fifth lag, and so forth. The drawback of this method is that it will produce too large a model, at least some of the time: even if the true AR order is five, so the sixth coefficient is zero, a 5% test using the \( F \)-statistic will incorrectly reject this null hypothesis 5% of the time just by chance. Thus, when the true value of \( p \) is five, this method will estimate \( p \) to be six 5% of the time.

The BIC. A way around this problem is to estimate \( p \) by minimizing an “information criterion.” One such information criterion is the Bayes information criterion (BIC), also called the Schwarz information criterion (SIC), which is

\[
\text{BIC}(p) = \ln\left(\frac{\text{SSR}(p)}{T}\right) + (p + 1)\ln\frac{T}{T},
\]  

(12.23)
where SSR(p) is the sum of squared residuals of the estimated AR(p). The BIC estimator of \( p \), \( \hat{p} \), is the value that minimizes BIC(p) among the possible choices \( p = 0, 1, \ldots, p_{\text{max}} \) where \( p_{\text{max}} \) is the largest value of \( p \) considered.

The formula for the BIC might look a bit mysterious at first, but it has an intuitive appeal. Consider the first term in Equation (12.23). Because the regression coefficients are estimated by OLS, the sum of squared residuals necessarily decreases (or at least does not increase) when you add a lag. In contrast, the second term is the number of estimated regression coefficients (the number of lags, \( p \), plus one for the intercept) times the factor \((\ln T)/T\). This second term increases when you add a lag. The BIC trades off these two forces so that the number of lags that minimizes the BIC is a consistent estimator of the true lag length. The mathematics of this argument is given in Appendix 12.5.

As an example, consider estimating the AR order for an autoregression of the change in the inflation rate. The various steps in the calculation of the BIC are carried out in Table 12.3 for autoregressions of maximum order six \( (p_{\text{max}} = 6) \). For example, for the AR(1) model in Equation (12.7), \( \text{SSR}(1)/T = 2.726 \), so \( \ln(\text{SSR}(1)/T) = 1.003 \). Because \( T = 152 \) (38 years, four quarters per year), \( \ln(T)/T = 0.033 \) and \( (p + 1)\ln(T)/T = 2 \times 0.033 = 0.066 \). Thus BIC(1) = 1.003 + 0.066 = 1.069.

The BIC is smallest when \( p = 3 \) in Table 12.3. Thus the BIC estimate of the lag length is 3. As can be seen in Table 12.3, as the number of lags increases the \( R^2 \) increases and the SSR decreases. The increase in the \( R^2 \) is large from one to two lags, smaller from two to three, and quite small from three to four. The BIC helps decide precisely how large the increase in the \( R^2 \) must be to justify including the additional lag.

The AIC. The BIC is not the only information criterion; another is the Akaike information criterion, or AIC:

\[
\text{AIC}(p) = \ln\left(\frac{\text{SSR}(p)}{T}\right) + (p + 1)\frac{2}{T}.
\]

The difference between the AIC and the BIC is that the term \( \ln(T) \) in the BIC is replaced by \( \ln(2) \) in the AIC, so the second term in the AIC is smaller. For example, for the 152 observations used to estimate the inflation autoregressions, \( \ln(T) = \ln(152) = 5.02 \), so that the second term for the BIC is more than twice as large as the term in AIC. Thus a smaller decrease in the SSR is needed in the AIC to justify including another lag. As a matter of theory, the second term in the AIC is not large enough to ensure that the correct lag length is chosen, even in large samples, so the AIC estimator of \( p \) is not consistent. As is discussed in Appendix 12.5, in large samples the AIC will overestimate \( p \) with nonzero probability.

Despite this theoretical blemish, the AIC is widely used in practice. If you are concerned that the BIC might yield a model with too few lags, the AIC provides a reasonable alternative.

### A note on calculating information criteria.
How well two estimated regressions fit the data is best assessed when they are estimated using the same data sets. Because the BIC and AIC are formal methods for making this comparison, the autoregressions under consideration should be estimated using the same observations. For example, in Table 12.3 all the regressions were estimated using data from 1962:I–1999:IV, for a total of 152 observations. Because the autoregressions involve lags of the change in inflation, this means that earlier values of the change in inflation (values before 1962:I) were used as regressors for the preliminary observations. Said differently, the regressions examined in Table 12.3 each include observations on \( \Delta \text{Inf}_t, \Delta \text{Inf}_{t-1}, \ldots, \Delta \text{Inf}_{t-p} \), for \( t = 1962:I, \ldots, 1999:IV \), corresponding to 152 observations on the dependent variable and regressors, so \( T = 152 \) in Equations (12.23) and (12.24).
Lag Length Selection in Time Series Regression with Multiple Predictors

The tradeoff involved with lag length choice in the general time series regression model with multiple predictors (Equation (12.20)) is similar to that in an autoregression: using too few lags can decrease forecast accuracy because valuable information is lost, but adding lags increases estimation uncertainty. The choice of lags must balance the benefit of using additional information against the cost of estimating the additional coefficients.

The F-statistic approach. As in the univariate autoregression, one way to determine the number of lags to include is to use F-statistics to test joint hypotheses that sets of coefficients equal zero. For example, in the discussion of Equation (12.17), we tested the hypothesis that the coefficients on the second through fourth lag of the unemployment rate equal zero against the alternative that they are nonzero; this hypothesis was rejected at the 1% significance level, lending support to the longer-lag specification. If the number of models being compared is small, then this F-statistic method is easy to use. In general, however, the F-statistic method can produce models that are too large, in the sense that the true lag order is overestimated.

Information criteria. As in an autoregression, the BIC and AIC can be used to estimate the number of lags and variables in the time series regression model with multiple predictors. If the regression model has $K$ coefficients (including the intercept), the BIC is

$$BIC(K) = \ln \left( \frac{SSR(K)}{T} \right) + K \ln T. \tag{12.25}$$

The AIC is defined in the same way, but with 2 replacing $\ln T$ in Equation (12.25). For each candidate model, the BIC (or AIC) can be evaluated, and the model with the lowest value of the BIC (or AIC) is the preferred model, based on the information criterion.

There are two important practical considerations when using an information criterion to estimate the lag lengths. First, as is the case for the autoregression, all the candidate models must be estimated over the same sample; in the notation of Equation (12.25), the number of observations used to estimate the model, $T$, must be the same for all models. Second, when there are multiple predictors, this approach is computationally demanding because it requires computing many different models (many combinations of the lag parameters). In practice, a convenient shortcut is to require all the regressors to have the same number of lags, that is, to require that $p = q_1 = \cdots = q_o$ so that only $p_{\text{max}} + 1$ models need to be compared (corresponding to $p = 0, 1, \ldots, p_{\text{max}}$).

12.6 Nonstationarity I: Trends

In Key Concept 12.6, it was assumed that the dependent variable and the regressors are stationary. If this is not the case, that is, if the dependent variable and/or regressors are nonstationary, then conventional hypothesis tests, confidence intervals, and forecasts can be unreliable. The precise problem created by nonstationarity, and the solution to that problem, depends on the nature of that nonstationarity.

In this and the next section, we examine two of the most important types of nonstationarity in economic time series data: trends and breaks. In each section, we first describe the nature of the nonstationarity, then discuss the consequences for time series regression if this type of nonstationarity is present but is ignored. We next present tests for nonstationarity and discuss remedies for, or solutions to, the problems caused by that particular type of nonstationarity. We begin by discussing trends.

What Is a Trend?

A trend is a persistent long-term movement of a variable over time. A time series variable fluctuates around its trend.

Inspection of Figure 12.1a suggests that the U.S. inflation rate has a trend consisting of a general upward tendency through 1982 and a downward tendency thereafter. The series in Figures 12.2a, b, and c also have trends, but their trends are quite different. The trend in the U.S. Federal Funds interest rate is similar to the trend in the U.S. inflation rate. The $/$ exchange rate clearly had a prolonged downward trend after the collapse of the fixed exchange rate system in 1972. The logarithm of Japanese real GDP has a complicated trend: fast growth at first, then moderate growth, and finally slow growth.

Deterministic and stochastic trends. There are two types of trends seen in time series data, deterministic and stochastic. A deterministic trend is a nonrandom function of time. For example, a deterministic trend might be linear in time; if inflation had a deterministic linear trend so that it increased by 0.1 percentage point per quarter, this trend could be written as $0.1t$, where $t$ is measured in quarters. In contrast, a stochastic trend is random and varies over
include an adjustment for the tendency of the series to increase. This adjustment leads to an extension of the random walk model to include a tendency to move, or “drift” in one direction or the other. This extension is referred to as a random walk with drift:

$$Y_t = \beta_0 + Y_{t-1} + \epsilon_t$$  \hspace{1cm} (12.27)

where $E(\epsilon_t | Y_{t-1}, Y_{t-2}, \ldots) = 0$ and $\beta_0$ is the “drift” in the random walk. If $\beta_0$ is positive, then $Y_t$ increases on average. In the random walk with drift model, the best forecast of the series tomorrow is the value of the series today, plus the drift $\beta_0$.

The random walk model (with drift as appropriate) is simple yet versatile, and it is the primary model for trends used in this book.

A random walk is nonstationary. If $Y_t$ follows a random walk, then it is not stationary: the variance of a random walk increases over time so the distribution of $Y_t$ changes over time. One way to see this is to recognize that, because $\epsilon_t$ is serially uncorrelated in Equation (12.26), $\text{var}(Y_t) = \text{var}(Y_{t-1}) + \text{var}(\epsilon_t)$; for $Y_t$ to be stationary, $\text{var}(Y_t)$ cannot depend on time, so in particular $\text{var}(Y_t) = \text{var}(Y_{t-1})$ must hold, but this can only happen if $\text{var}(\epsilon_t) = 0$. Another way to see this is to imagine that $Y_t$ starts out at zero, that is, $Y_0 = 0$. Then $Y_1 = \epsilon_1$, $Y_2 = \epsilon_1 + \epsilon_2$, and so forth, so that $Y_t = \epsilon_1 + \epsilon_2 + \cdots + \epsilon_t$. Because $\epsilon_t$ is serially uncorrelated, $\text{var}(Y_t) = \text{var}(\epsilon_1 + \epsilon_2 + \cdots + \epsilon_t) = t\sigma^2$. Thus the variance of $Y_t$ depends on $t$; in fact, it increases as $t$ increases. Because the variance of $Y_t$ depends on $t$, its distribution depends on $t$, that is, it is nonstationary.

Because the variance of a random walk increases without bound, its population autocorrelations are not defined (the first autocovariance and variance are infinite and the ratio of the two is not well defined). However, a feature of a random walk is that its sample autocorrelations tend to be very close to one, in fact, the $j$th sample autocorrelation of a random walk converges to one in probability.

Stochastic trends, autoregressive models, and a unit root. The random walk model is a special case of the AR(1) model (Equation (12.8)) in which $\beta_1 = 1$. In other words, if $Y_t$ follows an AR(1) with $\beta_1 = 1$, then $Y_t$ contains a stochastic trend and is nonstationary. If, however, $|\beta_1| < 1$ and $\epsilon_t$ is stationary, then the joint distribution of $Y_t$ and its lags does not depend on $t$ (a result shown in Appendix 12.2) so $Y_t$ is stationary as long as $\epsilon_t$ is stationary.

The analogous condition for an AR($p$) to be stationary is more complicated than the condition $|\beta_1| < 1$ for an AR(1). Its formal statement involves the roots...
of the polynomial, \(1 - \beta_1 x - \beta_2 x^2 - \beta_3 x^3 - \cdots - \beta_p x^p\). (The roots of this polynomial are the solutions to the equation \(1 - \beta_1 x - \beta_2 x^2 - \beta_3 x^3 - \cdots - \beta_p x^p = 0\).) For an AR(\(p\)) to be stationary, the roots of this polynomial must all be greater than one in absolute value. In the special case of an AR(1), the root is the value of \(z\) that solves \(1 - \beta_1 z = 0\), so its root is \(z = 1/\beta_1\). Thus the statement that the root be greater than one in absolute value is equivalent to \(|\beta_1| < 1\).

If an AR(\(p\)) has a root that equals one, the series is said to have a unit autoregressive root or, more simply, a unit root. If \(Y_t\) has a unit root, then it contains a stochastic trend. If \(Y_t\) is stationary (and thus does not have a unit root), it does not contain a stochastic trend. For this reason, we will use the terms "stochastic trend" and "unit root" interchangeably.

**Problems Caused by Stochastic Trends**

If a regressor has a stochastic trend (has a unit root), then the OLS estimator of its coefficient and its OLS \(t\)-statistic can have nonstandard (that is, nonnormal) distributions, even in large samples. We discuss three specific aspects of this problem: first, the estimator of the autoregressive coefficient in an AR(1) is biased towards zero if its true value is one; second, \(t\)-statistics on regressors with a stochastic trends can have a nonnormal distribution, even in large samples; and third, an extreme example of the risks posed by stochastic trends is that two series that are independent will, with high probability, misleadingly appear to be related if they both have stochastic trends, a situation known as spurious regression.

**Problem #1: Autoregressive coefficients that are biased towards zero.** Suppose that \(Y_t\) follows the random walk in Equation (12.26) but this is unknown to the econometrician, who instead estimates the AR(1) model in Equation (12.8). Because \(Y_t\) is nonstationary, the least squares assumptions for time series regression in Key Concept 12.6 do not hold, so as a general matter we cannot rely on estimators and test statistics having their usual large-sample normal distributions. In fact, in this example the OLS estimator of the autoregressive coefficient, \(\hat{\beta}_1\), is consistent, but it has a nonnormal distribution, even in large samples: the asymptotic distribution of \(\hat{\beta}_1\) is shifted towards zero. The expected value of \(\hat{\beta}_1\) is approximately \(E(\hat{\beta}_1) = 1 - 5.3/T\). This results in a large bias in sample sizes typically encountered in economic applications. For example, 20 years of quarterly data contain 80 observations, in which case the expected value of \(\hat{\beta}_1\) is \(E(\hat{\beta}_1) = 1 - 5.3/80 = 0.934\). Moreover, this distribution has a long left tail: the 5% percentile of \(\hat{\beta}_1\) is approximately \(1 - 14.1/T\) which, for \(T = 80\), corresponds to 0.824, so that 5% of the time \(\hat{\beta}_1 < 0.824\).

One implication of this bias towards zero is that, if \(Y_t\) follows a random walk, then forecasts based on the AR(1) model can perform substantially worse than those based on the random walk model, which imposes the true value \(\beta_1 = 1\). This conclusion also applies to higher order autoregressions, in which there are forecasting gains from imposing a unit root (that is, from estimating the autoregression in first differences instead of in levels) when in fact the series contains a unit root.

**Problem #2: Nonnormal distributions of \(t\)-statistics.** If a regressor has a stochastic trend, then its usual OLS \(t\)-statistic can have a nonnormal distribution under the null hypothesis, even in large samples. This nonnormal distribution means that conventional confidence intervals are not valid and hypothesis tests cannot be conducted as usual. In general, the distribution of this \(t\)-statistic is not readily tabulated because the distribution depends on the relationship between the regressor in question and the other regressors. One important case in which it is possible to tabulate this distribution is in the context of an autoregression with a unit root, and we return to this special case when we take up the problem of testing whether a time series contains a stochastic trend.

**Problem #3: Spurious regression.** Stochastic trends can lead two time series to appear related when they are not, a problem called spurious regression.

For example, U.S. inflation was steadily rising from the mid-1960s through the early 1980s, and at the same time Japanese GDP was steadily rising. These two trends conspire to produce a regression that appears to be "significant" using conventional measures. Estimated by OLS using data from 1965 through 1981, this regression is

\[
\text{U.S. Inflation}_t = -2.84 + 0.18 \text{Japanese GDP}_t, \quad R^2 = 0.56. \quad (12.28)
\]

The \(t\)- statistic on the slope coefficient exceeds 9, which by our usual standards indicates a strong positive relationship between the two series, and the \(R^2\) is high. However, running this regression using data from 1982 through 1999 yields

\[
\text{U.S. Inflation}_t = 6.25 - 0.03 \text{Japanese GDP}_t, \quad R^2 = 0.07. \quad (12.29)
\]

The regressions in Equation (12.28) and (12.29) could hardly be more different. Interpreted literally, Equation (12.28) indicates a strong positive relationship, while Equation (12.29) indicates a weak negative relationship.
The source of these conflicting results is that both series have stochastic trends. These trends happened to align from 1965 through 1981, but did not align from 1982 through 1999. There is, in fact, no compelling economic or political reason to think that the trends in these two series are related. In short, these regressions are spurious.

The regressions in Equations (12.28) and (12.29) illustrate empirically the theoretical point that OLS can be misleading when the series contain stochastic trends (see Exercise 12.6 for a computer simulation that demonstrates this result). One special case in which certain regression-based methods are reliable is when the trend component of the two series is the same, that is, when the series contain a common stochastic trend; if so, the series are said to be cointegrated. Econometric methods for detecting and analyzing cointegrated economic time series are discussed in Section 14.4.

**Detecting Stochastic Trends: Testing for a Unit AR Root**

Trends in time series data can be detected by informal and formal methods. The informal methods involve inspecting a time series plot of the data and computing the autocorrelation coefficients, as we did in Section 12.2. Because the first autocorrelation coefficient will be near one if the series has a stochastic trend, at least in large samples, a small first autocorrelation coefficient combined with a time series plot that has no apparent trend suggests that the series does not have a trend. If doubt remains, however, there are formal statistical procedures that can be used to test the hypothesis that there is a stochastic trend in the series against the alternative that there is no trend.

In this section, we use the Dickey-Fuller test (named after its inventors David Dickey and Wayne Fuller (1979)) to test for a stochastic trend. Although the Dickey-Fuller test is not the only test for a stochastic trend (another test is discussed in Section 14.3), it is the most commonly used test in practice and is one of the most reliable.

**The Dickey-Fuller test in the AR(1) model.** The starting point for the Dickey-Fuller test is the autoregressive model. As discussed earlier, the random walk in Equation (12.27) is a special case of the AR(1) model with $\beta_1 = 1$. If $\beta_1 = 1$, $Y_t$ is nonstationary and contains a (stochastic) trend. Thus, within the AR(1) model, the hypothesis that $Y_t$ has a trend can be tested by testing

$$H_0: \beta_1 = 1 \text{ vs. } H_1: \beta_1 < 1 \text{ in } Y_t = \beta_0 + \beta_1 Y_{t-1} + \epsilon_t$$

(12.30)

If $\beta_1 = 1$, the AR(1) has an autoregressive root of one, so the null hypothesis in Equation (12.30) is that the AR(1) has a unit root, and the alternative is that it is stationary.

This test is most easily implemented by estimating a modified version of Equation (12.30) obtained by subtracting $Y_{t-1}$ from both sides. Let $\delta = \beta_1 - 1$; then Equation (12.30) becomes

$$Y_t = \beta_0 + \delta Y_{t-1} + \epsilon_t$$

(12.31)

The OLS $t$-statistic testing $\delta = 0$ in Equation (12.31) is called the Dickey-Fuller statistic. The formulation in Equation (12.31) is convenient because regression software automatically prints out the $t$-statistic testing $\delta = 0$. Note that the Dickey-Fuller test is one-sided, because the relevant alternative is that $Y_t$ is stationary so $\beta_1 < 1$ or, equivalently, $\delta < 0$. The Dickey-Fuller statistic is computed using "nonrobust" standard errors, that is, the "homoskedasticity-only" standard errors presented in Appendix 4.4 (Equation (4.62) for the case of a single regressor and in Section 16.4 for the multiple regression model).2

**The Dickey-Fuller test in the AR(p) model.** The Dickey-Fuller statistic presented in the context of Equation (12.31) applies only to an AR(1). As discussed in Section 12.3, for some series the AR(1) model does not capture all the serial correlation in $Y_t$, in which case a higher order autoregression is more appropriate.

The extension of the Dickey-Fuller test to the AR(p) model is summarized in Key Concept 12.8. Under the null hypothesis, $\delta = 0$ and $\Delta Y_t$ is a stationary AR($p$). Under the alternative hypothesis, $\delta < 0$ so that $Y_t$ is stationary. Because the regression used to compute this version of the Dickey-Fuller statistic is augmented by lags of $\Delta Y_t$, the resulting $t$-statistic is referred to as the augmented Dickey-Fuller (ADF) statistic.

In general, the lag length $p$ is unknown, but it can be estimated using an information criterion applied to regressions of the form (12.32) for various values of $p$. Studies of the ADF statistic suggest that it is better to have too many lags than too few, so it is recommended to use the AIC instead of the BIC to estimate $p$ for the ADF statistic.3

---

2Under the null hypothesis of a unit root the usual "nonrobust" standard errors produce a $t$-statistic that is in fact robust to heteroskedasticity, a surprising and special result.

3See Stock (1994) for a review of simulation studies of the finite-sample properties of the Dickey-Fuller and other unit root tests.
The Augmented Dickey-Fuller Test for a Unit Autoregressive Root

The augmented Dickey-Fuller (ADF) test for a unit autoregressive root tests the null hypothesis $H_0: \delta = 0$ against the one-sided alternative $H_1: \delta < 0$ in the regression

$$\Delta Y_t = \beta_0 + \delta Y_{t-1} + \gamma_1 \Delta Y_{t-1} + \gamma_2 \Delta Y_{t-2} + \cdots + \gamma_p \Delta Y_{t-p} + u_t$$ (12.32)

Under the null hypothesis, $Y_t$ has a stochastic trend; under the alternative hypothesis, $Y_t$ is stationary. The ADF statistic is the OLS $t$-statistic testing $\delta = 0$ in Equation (12.32).

If instead the alternative hypothesis is that $Y_t$ is stationary around a deterministic linear time trend, then this trend, “$t$” (the observation number), must be added as an additional regressor, in which case the Dickey-Fuller regression becomes

$$\Delta Y_t = \beta_0 + \alpha t + \delta Y_{t-1} + \gamma_1 \Delta Y_{t-1} + \gamma_2 \Delta Y_{t-2} + \cdots + \gamma_p \Delta Y_{t-p} + u_t$$ (12.33)

where $\alpha$ is an unknown coefficient and the ADF statistic is the OLS $t$-statistic testing $\delta = 0$ in Equation (12.33).

The lag length $p$ can be estimated using the BIC or AIC. The ADF statistic does not have a normal distribution, even in large samples. Critical values for the one-sided ADF test depend on whether the test is based on Equation (12.32) or (12.33) and are given in Table 12.4.

Testing against the alternative of stationarity around a linear deterministic time trend. The discussion so far has considered the null hypothesis that the series has a unit root and the alternative hypothesis that it is stationary. This alternative hypothesis of stationarity is appropriate for series, like the rate of inflation, that do not exhibit long-term growth. But other economic time series, like Japanese GDP (Figure 12.2c), exhibit long-run growth, and for such series the alternative of stationarity without a trend is inappropriate. Instead, a commonly used alternative is that the series are stationary around a deterministic time trend, that is, a trend that is a deterministic function of time.

One specific formulation of this alternative hypothesis is that the time trend is linear, that is, the trend is a linear function of $t$; thus, the null hypothesis is that the series has a unit root and the alternative is that it does not have a unit root but does have a deterministic time trend. The Dickey-Fuller regression must be modified to test the null hypothesis of a unit root against the alternative that it is stationary around a linear time trend. As summarized in Equation (12.33) in Key Concept 12.8, this is accomplished by adding a time trend (the regressor $X_t = t$) to the regression.

A linear time trend is not the only way to specify a deterministic time trend; for example, the deterministic time trend could be quadratic, or it could be linear but have breaks (that is, be linear with slopes that differ in two parts of the sample). The use of alternatives like these with nonlinear deterministic trends should be motivated by economic theory. For a discussion of unit root tests against stationarity around nonlinear deterministic trends, see Maddala and Kim (1998, Chapter 13).

Critical values for the ADF statistic. Under the null hypothesis of a unit root, the ADF statistic does not have a normal distribution, even in large samples. Because its distribution is nonstandard, the usual critical values from the normal distribution cannot be used when using the ADF statistic to test for a unit root; a special set of critical values, based on the distribution of the ADF statistic under the null hypothesis, must be used instead.

The critical values for the ADF test are given in Table 12.4. Because the alternative hypothesis of stationarity implies that $\delta < 0$ in Equations (12.32) and (12.33), the ADF test is one-sided. For example, if the regression does not include a time trend, then the hypothesis of a unit root is rejected at the 5% significance level if the ADF statistic is less than $-2.86$. If a time trend is included in the regression, the critical value is instead $-3.41$.

The critical values in Table 12.4 are substantially larger (more negative) than the one-sided critical values of $-1.28$ (at the 10% level) and $-1.645$ (at the 5% level) from the standard normal distribution. The nonstandard distribution of the ADF statistic is an example of how OLS $t$-statistics for regressors with stochastic trends can have nonnormal distributions. Why the large-sample distribution of the ADF statistic is nonstandard is discussed further in Section 14.3.

<table>
<thead>
<tr>
<th>Deterministic Regressors</th>
<th>10%</th>
<th>5%</th>
<th>1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept only</td>
<td>-2.57</td>
<td>-2.86</td>
<td>-3.43</td>
</tr>
<tr>
<td>Intercept and time trend</td>
<td>-3.12</td>
<td>-3.41</td>
<td>-3.96</td>
</tr>
</tbody>
</table>
**Does U.S. inflation have a stochastic trend?** The null hypothesis that inflation has a stochastic trend can be tested against the alternative that it is stationary by performing the ADF test for a unit autoregressive root. The ADF regression with four lags of $\Delta \text{Inf}_t$ is

$$\Delta \text{Inf}_t = 0.53 - 0.11 \text{Inf}_{t-1} - 0.14 \Delta \text{Inf}_{t-1} - 0.25 \Delta \text{Inf}_{t-2} + 0.24 \Delta \text{Inf}_{t-3} + 0.014 \text{Inf}_{t-4}$$

The ADF $t$-statistic is the $t$-statistic testing the hypothesis that the coefficient on $\text{Inf}_{t-1}$ is zero; this is $t = -2.60$. From Table 12.4, the 5% critical value is $-2.86$. Because the ADF statistic of $-2.60$ is less negative than $-2.86$, the test does not reject at the 5% significance level. Based on the regression in Equation (12.34), we therefore cannot reject (at the 5% significance level) the null hypothesis that inflation has a unit autoregressive root, that is, that inflation contains a stochastic trend, against the alternative that it is stationary.

The ADF regression in Equation (12.34) includes four lags of $\Delta \text{Inf}_t$ to compute the ADF statistic. When the number of lags is estimated using the AIC, where $0 \leq p \leq 6$, the AIC estimator of the lag length is, however, three. When three lags are used (that is, when $\Delta \text{Inf}_{t-1}, \Delta \text{Inf}_{t-2},$ and $\Delta \text{Inf}_{t-3}$ are included as regressors), the ADF statistic is $-2.65$, which is less negative than $-2.86$. Thus, when the number of lags in the ADF regression is chosen by AIC, the hypothesis that inflation contains a stochastic trend is not rejected at the 5% significance level.

These tests were performed at the 5% significance level. At the 10% significance level, however, the tests reject the null hypothesis of a unit root: the ADF statistics of $-2.60$ (four lags) and $-2.65$ (three lags) are slightly more negative than the 10% critical value of $-2.57$. Thus the ADF statistics paint a rather ambiguous picture, and the forecasters must make an informed judgment about whether or not to model inflation as having a stochastic trend. Clearly, inflation in Figure 12.1a exhibits long-run swings, consistent with the stochastic trend model. Moreover, in practice, many forecasters treat U.S. inflation as having a stochastic trend, and we follow that strategy here.

**Avoiding the Problems Caused by Stochastic Trends**

The most reliable way to handle a trend in a series is to transform the series so that it does not have the trend. If the series has a stochastic trend, that is, if the series has a unit root, then the first difference of the series does not have a trend.

For example, if $Y_t$ follows a random walk so $Y_t = \beta_0 + Y_{t-1} + u_t$, then $\Delta Y_t = \beta_0 + u_t$ is stationary. Thus using first differences eliminates random walk trends in a series. In practice, you can rarely be sure whether a series has a stochastic trend or not. Recall that, as a general point, failure to reject the null hypothesis does not necessarily mean that the null hypothesis is true; rather, it simply means that you have insufficient evidence to conclude that it is false. Thus, failure to reject the null hypothesis of a unit root using the ADF test does not mean that the series actually has a unit root. For example, in an AR(1) the true coefficient $\beta_1$ might be very close to one, say 0.98, in which case the ADF test would have low power, that is, a low probability of correctly rejecting the null hypothesis in samples of size of our inflation series. Even though failure to reject the null hypothesis of a unit root does not mean the series has a unit root, it still can be reasonable to approximate the true autoregressive root as equaling one and therefore to use differences of the series rather than its levels.

### 12.7 Nonstationarity II: Breaks

A second type of nonstationarity arises when the population regression function changes over the course of the sample. In economics, this can occur for a variety of reasons, such as changes in economic policy, changes in the structure of the economy, or an invention that changes a specific industry. If such changes, or "breaks," occur, then a regression model that neglects those changes can provide a misleading basis for inference and forecasting.

This section presents two strategies for checking for breaks in a time series regression function over time. The first strategy looks for potential breaks from the perspective of hypothesis testing, and entails testing for changes in the regression coefficients using F-statistics. The second strategy looks for potential breaks from the perspective of forecasting; you pretend that your sample ends sooner than it actually does and evaluate the forecasts you would have made had this been so. Breaks are detected when the forecasting performance is substantially poorer than expected.

**What Is a Break?**

Breaks can arise either from a discrete change in the population regression coefficients at a distinct date or from a gradual evolution of the coefficients over a longer period of time.

---

1For additional discussion of stochastic trends in economic time series variables and of the problems they pose for regression analysis, see Stock and Watson (1988).
One source of discrete breaks in macroeconomic data is a major change in macroeconomic policy. For example, the breakdown of the Bretton Woods system of fixed exchange rates in 1972 produced the break in the time series behavior of the $\$/C exchange rate that is evident in Figure 12.2b. Prior to 1972, the exchange rate was essentially constant, with the exception of a single devaluation in 1968 in which the official value of the pound, relative to the dollar, was decreased. In contrast, since 1972 the exchange rate has fluctuated over a very wide range.

Breaks also can occur more slowly as the population regression evolves over time. For example, such changes can arise because of slow evolution of economic policy and ongoing changes in the structure of the economy. The methods for detecting breaks described in this section can detect both types of breaks, distinct changes and slow evolution.

Problems caused by breaks. If a break occurs in the population regression function during the sample, then the OLS regression estimates over the full sample will estimate a relationship that holds “on average,” in the sense that the estimate combines the two different periods. Depending on the location and the size of the break, the “average” regression function can be quite different than the true regression function at the end of the sample, and this leads to poor forecasts.

Testing for Breaks

One way to detect breaks is to test for discrete changes, or breaks, in the regression coefficients. How this is done depends on whether the date of the suspected break (the break date) is known or not.

Testing for a break at a known date. In some applications you might suspect that there is a break at a known date. For example, if you are studying international trade relationships using data from the 1970s, you might hypothesize that there is a break in the population regression function of interest in 1972 when the Bretton Woods system of fixed exchange rates was abandoned in favor of floating exchange rates.

If the date of the hypothesized break in the coefficients is known, then the null hypothesis of no break can be tested using a binary variable interaction regression of the type discussed in Chapter 6 (Key Concept 6.4). To keep things simple, consider an ADL(1,1) model, so there is an intercept, a single lag of $Y_t$, and a single lag of $X_t$. Let $\tau$ denote the hypothesized break date and let $D_t(\tau)$ be a binary variable that equals zero before the break date and one after, so $D_t(\tau) = 0$ if $t < \tau$ and $D_t(\tau) = 1$ if $t \geq \tau$. Then the regression including the binary break indicator and all interaction terms is

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 X_{t-1} + \gamma_0 D_t(\tau) + \gamma_1 (D_t(\tau) \times Y_{t-1}) + \gamma_2 (D_t(\tau) \times X_{t-1}) + \epsilon_t.$$  

(12.35)

If there is not a break, then the population regression function is the same over both parts of the sample so the terms involving the break binary variable $D_t(\tau)$ do not enter Equation (12.35). That is, under the null hypothesis of no break, $\gamma_0 = \gamma_1 = \gamma_2 = 0$. Under the alternative hypothesis that there is a break, then the population regression function is different before and after the break date $\tau$, in which case at least one of the $\gamma$’s is nonzero. Thus the hypothesis of a break can be tested using the F-statistic that tests the hypothesis that $\gamma_0 = \gamma_1 = \gamma_2 = 0$ against the hypothesis that at least one of the $\gamma$’s is nonzero. This is often called a Chow test for a break at a known break date, named for its inventor, Gregory Chow (1960).

If there are multiple predictors or more lags, then this test can be extended by constructing binary variable interaction variables for all the regressors and testing the hypothesis that all the coefficients on terms involving $D_t(\tau)$ are zero.

This approach can be modified to check for a break in a subset of the coefficients by including only the binary variable interactions for that subset of regressors of interest.

Testing for a break at an unknown break date Often the date of a possible break is unknown or known only within a range. Suppose, for example, you suspect that a break occurred sometime between two dates, $\tau_0$ and $\tau_1$. The Chow test can be modified to handle this by testing for breaks at all possible dates $\tau$ in between $\tau_0$ and $\tau_1$, then using the largest of the resulting F-statistics to test for a break at an unknown date. This modified Chow test is variously called the Quandt likelihood ratio (QLR) statistic (Quandt, 1960) (the term we shall use) or, more obscurely, the sup-Wald statistic.

Because the QLR statistic is the largest of many F-statistics, its distribution is not the same as an individual F-statistic. Instead, the critical values for the QLR statistic must be obtained from a special distribution. Like the F-statistic, this distribution depends on the number of restrictions being tested, $q$, that is, the number of coefficients (including the intercept) that are being allowed to break, or change, under the alternative hypothesis. The distribution of the QLR statistic also depends on $\tau_1/\tau_0$ and $\tau_0/\tau_1$, that is, on the endpoints, $\tau_0$ and $\tau_1$, of the sub-sample over which the F-statistics are computed, expressed as a fraction of the total sample size.
For the large-sample approximation to the distribution of the QLR statistic to be a good one, the subsample endpoints, $t_0$ and $t_1$, cannot be too close to the end of the sample. For this reason, in practice the QLR statistic is computed over a "trimmed" range, or subset, of the sample. A common choice is to use 15% trimming, that is, to set for $t_0 = 0.15T$ and $t_1 = 0.85T$ (rounded to the nearest integer). With 15% trimming, the $F$-statistic is computed for break dates in the central 70% of the sample.

The critical values for the QLR statistic, computed with 15% trimming, are given in Table 12.5. Comparing these critical values with those of the $F_{nm}$ distribution (Appendix Table 4) shows that the critical values for the QLR statistics are larger. This reflects the fact that the QLR statistic looks at the largest of many individual $F$-statistics. By examining $F$-statistics at many possible break dates, the QLR statistic has many opportunities to reject, leading to QLR critical values that are larger than the individual $F$-statistic critical values.

Like the Chow test, the QLR test can be used to focus on the possibility that there are breaks in only some of the regression coefficients. This is done by first computing the Chow tests at different break dates using binary variable interactions only for the variables with the suspect coefficients, then computing the maximum of those Chow tests over the range $t_0 \leq t \leq t_1$. The critical values for this version of the QLR test are also taken from Table 12.5, where the number of restrictions ($q$) is the number of restrictions tested by the constituent $F$-tests.

If there is a discrete break at a date within the range tested, then the QLR statistic will reject with high probability in large samples. Moreover, the date at which the constituent $F$-statistic is at its maximum, $\hat{t}$, is an estimate of the break date $t$. This estimate is a good one in the sense that, under certain technical conditions, $\hat{t}/T \xrightarrow{d} t/T$, that is, the fraction of the way through the sample at which the break occurs is estimated consistently.

The QLR statistic also rejects with high probability in large samples when there are multiple discrete breaks or when the break comes in the form of a slow evolution of the regression function. This means that the QLR statistic detects forms of instability other than a single discrete break. As a result, if the QLR statistic rejects, it can mean that there is a single discrete break, that there are multiple discrete breaks, or that there is slow evolution of the regression function.

The QLR statistic is summarized in Key Concept 12.9.

**Warning:** You probably don’t know the break date even if you think you do. Sometimes an expert might believe that he or she knows the date of a possible break, so that the Chow test can be used instead of the QLR test. But if this knowledge is based on the expert’s knowledge of the series being analyzed, then in fact this date was estimated using the data, albeit in an informal way. Preliminary estimation of the break date means that the usual $F$ critical values cannot be used for the Chow test for a break at that date. Thus it remains appropriate to use the QLR statistic in this circumstance.

<table>
<thead>
<tr>
<th>Number of Restrictions ($q$)</th>
<th>10%</th>
<th>5%</th>
<th>1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7.12</td>
<td>8.68</td>
<td>12.16</td>
</tr>
<tr>
<td>2</td>
<td>5.00</td>
<td>5.86</td>
<td>7.78</td>
</tr>
<tr>
<td>3</td>
<td>4.09</td>
<td>4.71</td>
<td>6.02</td>
</tr>
<tr>
<td>4</td>
<td>5.59</td>
<td>4.09</td>
<td>5.12</td>
</tr>
<tr>
<td>5</td>
<td>3.26</td>
<td>3.66</td>
<td>4.53</td>
</tr>
<tr>
<td>6</td>
<td>3.02</td>
<td>3.37</td>
<td>4.12</td>
</tr>
<tr>
<td>7</td>
<td>2.84</td>
<td>3.15</td>
<td>3.82</td>
</tr>
<tr>
<td>8</td>
<td>2.69</td>
<td>2.98</td>
<td>3.57</td>
</tr>
<tr>
<td>9</td>
<td>2.58</td>
<td>2.84</td>
<td>3.38</td>
</tr>
<tr>
<td>10</td>
<td>2.48</td>
<td>2.71</td>
<td>3.23</td>
</tr>
<tr>
<td>11</td>
<td>2.40</td>
<td>2.62</td>
<td>3.09</td>
</tr>
<tr>
<td>12</td>
<td>2.33</td>
<td>2.54</td>
<td>2.97</td>
</tr>
<tr>
<td>13</td>
<td>2.27</td>
<td>2.46</td>
<td>2.87</td>
</tr>
<tr>
<td>14</td>
<td>2.21</td>
<td>2.40</td>
<td>2.78</td>
</tr>
<tr>
<td>15</td>
<td>2.16</td>
<td>2.34</td>
<td>2.71</td>
</tr>
<tr>
<td>16</td>
<td>2.12</td>
<td>2.29</td>
<td>2.64</td>
</tr>
<tr>
<td>17</td>
<td>2.08</td>
<td>2.25</td>
<td>2.58</td>
</tr>
<tr>
<td>18</td>
<td>2.05</td>
<td>2.20</td>
<td>2.53</td>
</tr>
<tr>
<td>19</td>
<td>2.01</td>
<td>2.17</td>
<td>2.48</td>
</tr>
<tr>
<td>20</td>
<td>1.99</td>
<td>2.13</td>
<td>2.43</td>
</tr>
</tbody>
</table>

These critical values apply when $t_0 = 0.15T$ and $t_1 = 0.85T$ (rounded to the nearest integer), so that the $F$-statistic is computed for all potential break dates in the central 70% of the sample. The number of restrictions $q$ is the number of restrictions tested by each individual $F$-statistic. This table was kindly provided to us by Donald Andrews, and supersedes Table 1 in Andrews (1993).
The QLR Test for Coefficient Stability

Let \( F(t) \) denote the \( F \)-statistic testing the hypothesis of a break in the regression coefficients at date \( t \); in the regression in Equation (12.35), for example, this is the \( F \)-statistic testing the null hypothesis that \( \gamma_0 = \gamma_1 = \gamma_2 = 0 \). The QLR (or Sup-Wald) test is the largest of statistics in the range \( t_0 \leq t \leq t_1 \):

\[
\text{QLR} = \max\{F(t_0), F(t_0 + 1), \ldots, F(t_1)\}.
\]  

(12.36)

1. Like the \( F \)-statistic, the QLR statistic can be used to test for a break in all or just some of the regression coefficients.

2. In large samples, the distribution of the QLR statistic under the null hypothesis depends on the number of restrictions being tested, \( q \), and on the endpoints \( t_0 \) and \( t_1 \) as a fraction of \( T \). Critical values are given in Table 12.5 for 15% trimming \((t_0 = 0.15T \text{ and } t_1 = 0.85T, \text{ rounded to the nearest integer})\).

3. The QLR test can detect a single discrete break, multiple discrete breaks, and/or slow evolution of the regression function.

4. If there is a distinct break in the regression function, the date at which the largest Chow statistic occurs is an estimator of the break date.

Application: Has the Phillips curve been stable? The QLR test provides a way to check whether the Phillips curve has been stable from 1962 to 1999. Specifically, we focus on whether there have been changes in the coefficients on the lagged values of the unemployment rate and the intercept in the ADL(4,4) specification in Equation (12.17) containing four lags each of \( \Delta \text{InP} \) and \( \text{Unemp} \).

The Chow \( F \)-statistics testing the hypothesis that the intercept and the coefficients on \( \text{Unemp}_{-4}, \ldots, \text{Unemp}_{-1} \) in Equation (12.17) are constant against the alternative that they break at a given date are plotted in Figure 12.5 for breaks in the central 70% of the sample. For example, the \( F \)-statistic testing for a break in 1980:1 is 2.26, the value plotted at that date in the figure. Each \( F \)-statistic tests five restrictions (no change in the intercept and in the four coefficients on lags of the unemployment rate), so \( q = 5 \). The largest of these \( F \)-statistics is 3.53, which occurs in 1982:II; this is the QLR statistic. Comparing 3.53 to the critical values for \( q = 5 \) in Table 12.5 indicates that the hypothesis that these coefficients are stable is rejected at the 10% significance level (the critical value is 3.26), but not 5% significance level (the critical value is 3.66). Thus there is some evidence that at least one of these five coefficients has changed over the sample, but the evidence is not especially strong.

Pseudo Out-of-Sample Forecasting

The ultimate test of a forecasting model is its out-of-sample performance, that is, its forecasting performance in "real time," after the model has been estimated. Pseudo out-of-sample forecasting is a method for simulating the real-time performance of a forecasting model. The idea of pseudo out-of-sample forecasting is simple: pick a date near the end of the sample, estimate your forecasting model using data up to that date, then use that estimated model to make a forecast. Performing this exercise for multiple dates near the end of your sample yields a series of pseudo forecasts and thus pseudo forecast errors. The pseudo forecast errors can then be examined to see if they are representative of what you would expect if the forecasting relationship were stationary.
Pseudo Out-of-Sample Forecasts

Pseudo out-of-sample forecasts are computed using the following steps:

1. Choose a number of observations, \( P \), for which you will generate pseudo out-of-sample forecasts; for example, \( P \) might be 10% or 15% of the sample size. Let \( s = T - P \).
2. Estimate the forecasting regression using the shortened data set for \( t = 1, \ldots, s \).
3. Compute the forecast for the first period beyond this shortened sample, \( s + 1 \); call this \( \hat{Y}_{s+1} \).
4. Compute the forecast error, \( \tilde{u}_{s+1} = \hat{Y}_{s+1} - \hat{Y}_{s+1} \).
5. Repeat steps 2-4 for the remaining dates, \( s = T - P + 1 \) to \( T - 1 \) (re-estimate the regression at each date). The pseudo out-of-sample forecasts are \( \{\hat{Y}_{s+1} \mid s = T - P, \ldots, T - 1 \} \) and the pseudo out-of-sample forecast errors are \( \{\tilde{u}_{s+1} \mid s = T - P, \ldots, T - 1 \} \).

The reason this is called "pseudo" out-of-sample forecasting is that it is not true out-of-sample forecasting. True out-of-sample forecasting occurs in real time, that is, you make your forecast without the benefit of knowing the future values of the series. In pseudo out-of-sample forecasting, you simulate real time forecasting using your model, but you have the "future" data against which to assess those simulated, or pseudo, forecasts. Pseudo out-of-sample forecasting mimics the forecasting process that would occur in real time, but without having to wait for new data to arrive.

Pseudo out-of-sample forecasting gives a forecaster a sense of how well the model has been forecasting at the end of the sample. This can provide valuable information, either bolstering confidence that the model has been forecasting well or suggesting that the model has gone off track in the recent past. The methodology of pseudo out-of-sample forecasting is summarized in Key Concept 12.10.

Other uses of pseudo out-of-sample forecasting. A second use of pseudo out-of-sample forecasting is to estimate the RMSFE. Because the pseudo out-of-sample forecasts are computed using only data prior to the forecast date, the pseudo out-of-sample forecast errors reflect both the uncertainty associated with future values of the error term and the uncertainty arising because the regression coefficients were estimated; that is, the pseudo out-of-sample forecast error include both sources of error in Equation (12.21). Thus the sample standard deviation of the pseudo out-of-sample forecast errors is an estimator of the RMSFE. As discussed in Section 12.4, this estimator of the RMSFE can be used to quantify forecast uncertainty and to construct forecast intervals.

A third use of pseudo out-of-sample forecasting is to compare two or more candidate forecasting models. Two models that appear to fit the data equally well can perform quite differently in a pseudo out-of-sample forecasting exercise. When the models are different, for example, when they include different predictors, pseudo out-of-sample forecasting provides a convenient way to compare the two models that focuses on their potential to provide reliable forecasts.

Application: Did the Phillips curve change during the 1990s? If the coefficients of the Phillips curve changed during the 1990s, then pseudo out-of-sample forecasts computed over that period should deteriorate. The pseudo out-of-sample forecasts of inflation for the period 1994:I to 1999:IV, computed using the four-lag Phillips curve, are plotted in Figure 12.6 along with the actual values of inflation. For example, the forecast of inflation for 1994:I was computed by regressing \( \Delta \text{Inf}_{t} \) on \( \Delta \text{Inf}_{t-1}, \ldots, \Delta \text{Inf}_{t-4}, \text{Unemp}_{t-1}, \ldots, \text{Unemp}_{t-4} \) with an intercept using the data through 1993:IV, then computing the forecast \( \hat{\Delta \text{Inf}}_{1994:1} \) using these estimated coefficients and the data through 1993:IV. The inflation forecast for 1994:I is then \( \hat{\text{Inf}}_{1994:1} = \hat{\Delta \text{Inf}}_{1994:1} + \hat{\text{Inf}}_{1993:4} \). This entire procedure was repeated using data through 1994:I to compute the forecast \( \hat{\text{Inf}}_{1994:4} \). Doing this for all 24 quarters from 1994:I–1999:IV creates 24 pseudo out-of-sample forecasts, which are plotted in Figure 12.6. The pseudo out-of-sample forecast errors are the differences between actual inflation and its pseudo out-of-sample forecast, that is, the differences between the two lines in Figure 12.6. For example, in the third quarter of 1997, the inflation rate rose by 0.9 percentage points, but the pseudo out-of-sample forecast of \( \Delta \text{Inf}_{1997:III} \) was 1.9 percentage points, so the pseudo out-of-sample forecast error was \( \Delta \text{Inf}_{1997:III} - \hat{\Delta \text{Inf}}_{1997:III} = 0.9 - 1.9 = -1.0 \) percentage points. In other words, a forecaster using the ADL(4,4) model of the Phillips curve, estimated through 1997:II, would have forecasted that inflation would increase by 1.9 percentage points in 1997:III, whereas in reality it only increased by 0.9 percentage points.

How do the mean and standard deviation of the pseudo out-of-sample forecast errors compare with the in-sample fit of the model? The standard error of the regression of the four-lag Phillips curve fit using data through 1993:IV is 1.47, so based on the in-sample fit we would expect the out-of-sample forecast errors to have mean zero and root mean square forecast error of 1.47. In fact, the aver-
The pseudo out-of-sample forecasts made using a four-lag Phillips curve of the form in Equation (12.17) generally track actual inflation, but on average the forecasts are higher than actual inflation. This upward bias in the forecasts may have been caused by a decline in the natural rate of unemployment, which would appear as a shift in the intercept of the Phillips curve.

The average forecast error is $-0.37$ and the sample RMSFE is $0.75$. Thus the RMSFE of the pseudo out-of-sample forecasts is less than predicted by the in-sample fit of the regression. However, the average forecast error is negative rather than zero, that is, on average the forecasts predicted larger increases in inflation (and thus higher inflation) than actually occurred. In fact, the $t$-statistic testing the hypothesis that the mean out-of-sample forecast error is zero is $t = -2.71$, so the hypothesis that the mean is zero is rejected at the 1% significance level. This suggests that the forecasts were biased over this period, systematically forecasting higher inflation than actually occurred. The finding that the pseudo out-of-sample forecasts are biased is reflected in Figure 12.6: forecasted inflation typically exceeds actual inflation so the average forecast error is negative.

These biased forecasts suggest that the Phillips curve regression was unstable towards the end of this sample, and that this instability led to forecasts of the change in inflation that were systematically too high. Before using this model for real-time forecasting, it would be important to try to identify the source of this shift and to incorporate it into a modified version of the Phillips curve model.

Taken together, this bias in the pseudo out-of-sample forecasts and the rejection of stability by the QLR statistic (at the 10% level) suggest that the four-lag Phillips curve has been unstable. This instability was a matter of considerable interest during the 1990s and early 2000s because economic forecasters recognized that, as seen in Figure 12.6, inflation forecasts based on the Phillips curve were too high. Some macroeconomists think that the source of this instability was a decline in the natural rate of unemployment during the 1990s, which would translate into a negative shift in the intercept in the regressions examined here. Other macroeconomists think that this breakdown is more complete, however, and that the entire concept of the Phillips curve—a link between the pressures of excess demand and overall price inflation—is just an antiquated feature of the pre-information age economy. If you are interested in reading more on this debate, see the symposium on the Phillips curve in the Winter 1997 issue of the *Journal of Economic Perspectives*.

### Avoiding the Problems Caused by Breaks

The best way to adjust for a break in the population regression function depends on the presence of a break. If a distinct break occurs at a specific date, this break will be detected with high probability by the QLR statistic, and the break date can be estimated. Thus the regression function can be estimated using a binary variable indicating the two subsamples associated with this break, interacted with the other regressors as needed. If all the coefficients break, then this regression takes the form of Equation (12.35), where $t$ is replaced by the estimated break date, $\hat{t}$, while if only some of the coefficients break, then only the relevant interaction terms appear in the regression. If there is in fact a distinct break, then inference on the regression coefficients can proceed as usual, for example using the usual normal critical values for hypothesis tests based on $t$-statistics. In addition, forecasts can be produced using the estimated regression function that applies to the end of the sample.

If the break is not distinct but rather arises from a slow, ongoing change in the parameters, the remedy is more difficult, and goes beyond the scope of this book.\(^3\)

\(^3\)For additional discussion of estimation and testing in the presence of discrete breaks, see Hansen (2001). For an advanced discussion of estimation and forecasting when there are slowly evolving coefficients, see Hamilton (1994, Chapter 13).
12.8 Conclusion

In time series data, a variable generally is correlated from one observation, or date, to the next. A consequence of this correlation is that linear regression can be used to forecast future values of a time series based on its current and past values. The starting point for time series regression is an autoregression, in which the regressors are lagged values of the dependent variable. If additional predictors are available, then their lags can be added to the regression.

This chapter has considered several technical issues that arise when estimating and using regressions with time series data. One such issue is determining the number of lags to include in the regressions. As discussed in Section 12.5, if the number of lags is chosen to minimize the BIC, then the estimated lag length is consistent for the true lag length.

Another of these issues concerns whether or not the series being analyzed are stationary. If the series are stationary, then the usual methods of statistical inference (such as comparing t-statistics to normal critical values) can be used, and, because the population regression function is stable over time, regressions estimated using historical data can be used reliably for forecasting. If, however, the series are nonstationary, then things become more complicated, where the specific complication depends on the nature of the nonstationarity. For example, if the series is nonstationary because it has a stochastic trend, then the OLS estimator and t-statistic can have nonstandard (nonnormal) distributions, even in large samples, and forecast performance can be improved by specifying the regression in first differences. A test for detecting this type of nonstationarity—the augmented Dickey–Fuller test for a unit root—was introduced in Section 12.6. Alternatively, if the population regression function has a break, then neglecting this break results in estimating an average version of the population regression function that in turn can lead to biased and/or imprecise forecasts. Procedures for detecting a break in the population regression function were introduced in Section 12.7.

In this chapter, the methods of time series regression were applied to economic forecasting, and the coefficients in these forecasting models were not given a causal interpretation. You do not need a causal relationship to forecast, and ignoring causal interpretations liberates the quest for good forecasts. In some applications, however, the task is not to develop a forecasting model but rather to estimate causal relationships among time series variables, that is, to estimate the dynamic causal effect on $Y$ over time of a change in $X$. Under the right conditions, the methods of this chapter, or closely related methods, can be used to estimate dynamic causal effects, and that is the topic of the next chapter.

Summary

1. Regression models used for forecasting need not have a causal interpretation.
2. A time series variable generally is correlated with one or more of its lagged values; that is, it is serially correlated.
3. An autoregression of order $p$ is a linear multiple regression model in which the regressors are the first $p$ lags of the dependent variable. The coefficients of an AR($p$) can be estimated by OLS, and the estimated regression function can be used for forecasting. The lag order $p$ can be estimated using an information criterion such as the BIC.
4. Adding other variables and their lags to an autoregression can improve forecasting performance. Under the least squares assumptions for time series regression (Key Concept 12.6), the OLS estimators have normal distributions in large samples and statistical inference proceeds the same way as for cross-sectional data.
5. Forecast intervals are one way to quantify forecast uncertainty. If the errors are normally distributed, an approximate 68% forecast interval can be constructed as the forecast ± an estimate of the root mean squared forecast error.
6. A series that contains a stochastic trend is nonstationary, violating the second least squares assumption in Key Concept 12.6. The OLS estimator and t-statistic for the coefficient of a regressor with a stochastic trend can have a nonstandard distribution, potentially leading to biased estimators, inefficient forecasts, and misleading inferences. The ADF statistic can be used to test for a stochastic trend. A random walk stochastic trend can be eliminated by using first differences of the series.
7. If the population regression function changes over time, then OLS estimates neglecting this instability are unreliable for statistical inference or forecasting. The QLR statistic can be used to test for a break and, if a discrete break is found, the regression function can be re-estimated in a way that allows for the break.
8. Pseudo out-of-sample forecasts can be used to assess model stability towards the end of the sample, to estimate the root mean squared forecast error, and to compare different forecasting models.
CHAPTER 12 Introduction to Time Series Regression and Forecasting

APPENDIX 12.1 Time Series Data Used in Chapter 12

Macroeconomic time series data for the United States are collected and published by various government agencies. The U.S. Consumer Price Index is measured using household surveys and is compiled by the Bureau of Labor Statistics (BLS). The unemployment rate is computed from the BLS’s Current Population Survey (see Appendix 3.1). The quarterly data used here were computed by averaging the monthly values. The Federal Funds rate data are the monthly average of daily rates as reported by the Federal Reserve and the dollar-exchange rate data are the monthly average of daily rates; both are for the final month in the quarter. Japanese real GDP data were obtained from the OECD. The daily percentage change in the NYSE Composite Index was computed as 100Δln(NYSEt), where NYSEt is the value of the index at the daily close of the New York Stock Exchange; because the stock exchange is not open on weekends and holidays, the time period of analysis is a business day. These and thousands of other economic time series are freely available on the websites maintained by various data collecting agencies.

APPENDIX 12.2 Stationarity in the AR(1) Model

This appendix shows that, if $|\beta_1|<1$ and $u_t$ is stationary, then $y_t$ is stationary. Recall from Key Concept 12.5 that the time series variable $y_t$ is stationary if the joint distribution of $(y_t, \ldots, y_{t+r})$ does not depend on $t$. To streamline the argument, we show this formally for $T=2$ under the simplifying assumptions that $\beta_0=0$ and $(u_t)$ are i.i.d. $N(0, \sigma^2_u)$.

The first step is deriving an expression for $y_t$ in terms of the $u_t$'s. Because $\beta_0=0$, Equation (12.8) implies that $y_t = \beta_1 y_{t-1} + u_t$. Substituting $y_{t-1} = \beta_1 y_{t-2} + u_{t-1}$ into this expression yields $y_t = \beta_1 (\beta_1 y_{t-2} + u_{t-1}) + u_t = \beta_1^2 y_{t-2} + \beta_1 u_{t-1} + u_t$. Continuing this substitution another step yields $y_t = \beta_1^2 y_{t-2} + \beta_1^3 u_{t-2} + \beta_1^4 u_{t-3} + u_t$ and continuing indefinitely yields

$$y_t = u_t + \beta_1 u_{t-1} + \beta_1^2 u_{t-2} + \beta_1^3 u_{t-3} + \ldots = \sum_{i=0}^{\infty} \beta_1^i u_{t-i}$$  (12.37)

Thus $y_t$ is a weighted average of current and past $u_t$'s. Because the $u_t$'s are normally distributed and because the weighted average of normal random variables is normal (Section 2.6), $y_{t+1}$ and $y_{t+2}$ have a bivariate normal distribution. Recall from Section 2.6 that the bivariate normal distribution is completely determined by the means of the two variables, their variances, and their covariance. Thus, to show that $y_t$ is stationary, we need to show that the means, variances, and covariances of $(y_{t+1}, y_{t+2})$ do not depend on $t$. An extension of the argument used below can be used to show that the distribution of $(y_{t+1}, y_{t+2}, \ldots, y_{t+r})$ does not depend on $t$.

The means and variances of $y_{t+1}$ and $y_{t+2}$ can be computed using Equation (12.37), with the subscripts $s+1$ and $s+2$ replacing $t$. First, because $E(u_t) = 0$ for all $t$, $E(y_t) = E(\sum_{i=0}^{\infty} \beta_1^i u_{t-i}) = \sum_{i=0}^{\infty} \beta_1^i E(u_{t-i}) = 0$, so the mean of $y_{t+1}$ and $y_{t+2}$ are both zero and in particular do not depend on $t$. Second, $\text{var}(y_t) = \text{var}(\sum_{i=0}^{\infty} \beta_1^i u_{t-i}) = \sum_{i=0}^{\infty} \text{var}(u_{t-i}) = \sigma_u^2 (1 - \beta_1^2)$, where the final equality follows from the fact that, if $|\beta_1|<1$, $\sum_{i=0}^{\infty} 1/(1-\beta_1^2) = \sigma_u^2 (1 - \beta_1^2)$, which does not depend on $s$ as long as $|\beta_1|<1$. Finally, because $y_{t+1} = y_{t+2} = \beta_1 y_{t+3} + u_{t+2}$, $\text{cov}(y_{t+1}, y_{t+2}) = E(y_{t+1} y_{t+2}) = E(y_{t+1} \beta_1 y_{t+3} + u_{t+2}) = \beta_1 \text{var}(y_{t+1}) + \text{cov}(y_{t+1}, u_{t+2}) = \beta_1 \sigma_u^2 (1 - \beta_1^2)$. The covariance does not depend on $s$, so $y_{t+1}$ and $y_{t+2}$ have a joint probability distribution that does not depend on $s$, that is, their joint distribution is stationary. If $|\beta_1|\geq 1$, this calculation breaks down because the infinite sum in Equation (12.37) does not converge and the variance of $y_t$ is infinite. Thus $y_t$ is stationary if $|\beta_1|<1$, but not if $|\beta_1|\geq 1$.

The preceding argument was made under the assumptions that $\beta_0=0$ and $u_t$ is normally distributed. If $\beta_0 \neq 0$, the argument is similar except that the means of $y_{t+1}$ and $y_{t+2}$ are $\beta_0/(1-\beta_1)$ and Equation (12.37) must be modified for this nonzero mean. The assumption that $u_t$ is i.i.d. normal can be replaced with the assumption that $u_t$ is stationary with a finite variance because, by Equation (12.37), $y_t$ can still be expressed as a function of current and past $u_t$'s, so the distribution of $y_t$ is stationary as long as the distribution of $u_t$ is stationary and the infinite sum expression in Equation (12.37) is meaningful in the sense that it converges, which requires $|\beta_1|<1$.

APPENDIX 12.3 Lag Operator Notation

The notation in this and the next two chapters is streamlined considerably by adopting what is known as lag operator notation. Let $L$ denote the lag operator, which has the property that it transforms a variable into its lag. That is, the lag operator $L$ has the
consistency of the BIC lag length estimator

Appendix 12.3, let \( u_t = \beta(L)e_t \), where \( e_t \) is a serially uncorrelated, unobserved random variable, and \( \beta(L) \) is a lag polynomial of degree \( q \) with \( \beta_0 = 1 \). Then the ARMA\((p,q)\) model is

\[
a(L)y_t = \beta_0 + \beta(L)e_t
\]

where \( a(L) \) is a lag polynomial of degree \( p \) with \( a_0 = 1 \).

Both AR and ARMA models can be thought of as ways to approximate the autocovariances of \( Y_t \). The reason for this is that any stationary time series \( Y_t \) with a finite variance can be written either as an AR or as a MA with a serially uncorrelated error term, although the AR or MA models might need to have an infinite order. The second of these results, that a stationary process can be written in moving average form, is known as the Wold decomposition theorem, and is one of the fundamental results underlying the theory of stationary time series analysis.

As a theoretical matter, the families of AR, MA, and ARMA models are equally rich, as long as the lag polynomials have a sufficiently high degree. Still, in some cases the autocovariances can be better approximated using an ARMA\((p,q)\) model with small \( p \) and \( q \) than by a pure AR model with only a few lags. As a practical matter, however, the estimation of ARMA models is more difficult than the estimation of AR models, and ARMA models are more difficult to extend to additional regressors than AR models.

**APPENDIX 12.5** Consistency of the BIC Lag Length Estimator

This appendix summarizes the argument that the BIC estimator of the lag length, \( \hat{p} \), in an autoregression is correct in large samples, that is, \( \Pr(\hat{p} = p) \rightarrow 1 \). This is not true for the AIC estimator, which can overestimate \( p \) even in large samples.

**BIC**

First consider the special case that the BIC is used to choose among autoregressions with zero, one, or two lags, when the true lag length is one. It is shown below that (i) \( \Pr(\hat{p} = 0) \rightarrow 0 \), and (ii) \( \Pr(\hat{p} = 2) \rightarrow 0 \), from which it follows that \( \Pr(\hat{p} = 1) \rightarrow 1 \). The extension of this argument to the general case of searching over \( 0 \leq p \leq p_{\text{max}} \) entails showing that \( \Pr(\hat{p} < p) \rightarrow 0 \) and \( \Pr(\hat{p} > p) \rightarrow 0 \); the strategy for showing these is the same as used in (i) and (ii) below.

**APPENDIX 12.4 ARMA Models**

The autoregressive-moving average (ARMA) model extends the autoregressive model by modeling \( u_t \) as serially correlated, specifically, as being a distributed lag (or "moving average") of another unobserved error term. That is, in the lag operator notation of the autoregressive model, the moving average is added to the autoregressive part.
Proof of (i) and (ii)

Proof of (i). To choose \( \hat{p} = 0 \) it must be the case that \( \text{BIC}(0) < \text{BIC}(1) \); that is, \( \text{BIC}(0) - \text{BIC}(1) < 0 \). Now \( \text{BIC}(0) - \text{BIC}(1) = [\ln(\text{SSR}(0)/T) + (\ln T)/T] - [\ln(\text{SSR}(1)/T) + 2(\ln T)/T] = \ln(\text{SSR}(0)/T) - \ln(\text{SSR}(1)/T) - (\ln T)/T. \) Now \( \text{SSR}(0)/T = [(T-1)/T]\hat{\beta}_{2}^{2} \rightarrow \sigma_{1}^{2}, \text{SSR}(1)/T \rightarrow \sigma_{2}^{2}, \text{and} \ (\ln T)/T \rightarrow 0; \) putting these pieces together, \( \text{BIC}(0) - \text{BIC}(1) \rightarrow \ln \sigma_{1}^{2} - \ln \sigma_{2}^{2} > 0 \) because \( \sigma_{1}^{2} > \sigma_{2}^{2}. \) It follows that \( \Pr[\text{BIC}(0) < \text{BIC}(1)] \rightarrow 0, \) so that \( \Pr(\hat{p} = 0) \rightarrow 0. \)

Proof of (ii). To choose \( \hat{p} = 2 \) it must be the case that \( \text{BIC}(2) < \text{BIC}(1), \) or \( \text{BIC}(2) - \text{BIC}(1) < 0. \) Now \( T[\text{BIC}(2) - \text{BIC}(1)] = T[\ln(\text{SSR}(2)/T) + 3(\ln T)/T] - [\ln(\text{SSR}(1)/T) + 2(\ln T)/T]] = T[\ln(\text{SSR}(2)/\text{SSR}(1))] + \ln T = -T[\ln(1 + F/(T-2)] + \ln T, \) where \( F = [\text{SSR}(1) - \text{SSR}(2)]/[\text{SSR}(2)/(T-2)] \) is the "rule of thumb" \( F \)-statistic (Appendix 5.3) testing the null hypothesis that \( \beta_{2} = 0 \) in the AR(2). If \( a_{t} \) is homoskedastic, \( F \) has a \( \chi^{2} \) asymptotic distribution; if not, it has some other asymptotic distribution. Thus \( \Pr[\text{BIC}(2) - \text{BIC}(1) < 0] = \Pr[T[\text{BIC}(2) - \text{BIC}(1)] < 0 \] = \Pr(-T[\ln(1 + F/(T-2)] + \ln T) < 0 = \Pr(T[\ln(1 + F/(T-2)] > \ln T). \) As \( T \) increases, \( \ln(1 + F/(T-2)] \rightarrow \ln a, \) which becomes exact as \( a \rightarrow 0. \) Thus \( \Pr[\text{BIC}(2) - \text{BIC}(1) < 0] \rightarrow \Pr(F > \ln T) \rightarrow 0, \) so that \( \Pr(\hat{p} = 2) \rightarrow 0. \)

AIC

In the special case of an AR(1) when zero, one, or two lags are considered, (i) applies to the AIC, where the term \( \ln T \) is replaced by 2, so \( \Pr(\hat{p} = 0) \rightarrow 0. \) All the steps in the proof of (ii) for the BIC also apply to the AIC, with the modification that \( \ln T \) is replaced by 2; thus \( \Pr(\text{AIC}(2) - \text{AIC}(1) < 0) \rightarrow \Pr(F > 2) > 0. \) If \( a_{t} \) is homoskedastic, \( \Pr(F > 2) \rightarrow \Pr(\hat{p} = 2) = 0.16. \) In general, when \( \hat{p} \) is chosen using the AIC, \( \Pr(\hat{p} = p) \rightarrow 0 \) but \( \Pr(\hat{p} > p) \) tends to a positive number, so \( \Pr(\hat{p} = p) \) does not tend to 1.

CHAPTER 13

Estimation of Dynamic Causal Effects

In the 1983 movie *Trading Places*, the characters played by Dan Aykroyd and Eddie Murphy used inside information on how well Florida oranges had fared the winter to make millions in the orange juice concentrate futures market, a market for contracts to buy or sell large quantities of orange juice concentrate at a specified price on a future date. In real life, traders in orange juice futures in fact do pay close attention to the weather in Florida: freezes in Florida kill Florida oranges, the source of almost all frozen orange juice concentrate made in the United States, so its supply falls and the price rises. But precisely how much does the price rise when the weather in Florida turns sour? Does the price rise all at once, or are there delays; if so, for how long? These are questions that real life traders in orange juice futures need to answer if they want to succeed.

This chapter takes up the problem of estimating the effect on \( Y \) now and in the future of a change in \( X, \) that is, the dynamic causal effect on \( Y \) of a change in \( X. \) What, for example, is the effect on the path of orange juice prices over time of a freezing spell in Florida? The starting point for modeling and estimating dynamic causal effects is the so-called distributed lag regression model, in which \( Y_{t} \) is expressed as a function of current and past values of \( X_{t}. \) Section 13.1 introduces the distributed lag model in the context of estimating the effect of cold weather in Florida on the price of orange juice concentrate over time. Section 13.2 takes a closer look at what, precisely, is meant by a dynamic causal effect.

One way to estimate dynamic causal effects is to estimate the coefficients of the distributed lag regression model using OLS. As discussed in Section 13.3, this estimator is consistent if the regression error has a conditional mean of zero given current and past values of \( X, \) a condition that (as in Chapter 10) is
The New Econometrics of Structural Change: Dating Breaks in U.S. Labor Productivity

Bruce E. Hansen

Applied time series analysis and forecasting is based on the assumption of stationarity—the constancy of parameters like the mean, variance and trend over time. But what happens if the parameters change? A particularly common assertion is that United States labor productivity experienced a “slowdown” around 1973 and a “speedup” in the second half of the 1990s. While one can certainly pick groups of years in the 1970s and 1980s where the average annual productivity growth is lower than it was in the 1960s and the late 1990s, this ad hoc selection of convenient time periods hardly qualifies as serious analysis.

Structural change is a statement about parameters, which only have meaning in the context of a model. To focus our discussion, we will discuss structural change in the simplest dynamic model, the first-order autoregression:

\[ y_t = \alpha + \rho y_{t-1} + \epsilon_t \]

\[ E\epsilon_t = \sigma^2, \]

where \( \epsilon_t \) is a time series of serially uncorrelated shocks. The parameters are (\( \alpha, \rho, \sigma^2 \)). The assumption of stationarity implies that these parameters are constant over time. We say that a structural break has occurred if at least one of these parameters has changed at some date—the breakdate—in the sample period. While it may seem unlikely that a structural break could be immediate and might seem more reasonable to allow a structural change to take a period of time to take effect, we most often focus on the simple case of an immediate structural break for simplicity and parsimony.

A structural break may affect any or all of the model parameters, and these cases have different implications. Changes in the autoregressive parameter \( \rho \) reflect changes in the serial correlation in \( y_t \). The intercept \( \alpha \) controls the mean of \( y_t \) through the relationship \( E(y_t) = \mu = \alpha / (1 - \rho) \). Since \( \mu \) is the growth rate in labor productivity, changes in \( \mu \) are identical to changes in the trend and are probably the issue of primary interest. Finally, changes in \( \sigma^2 \) imply changes in the volatility of labor productivity.

The econometrics of structural change looks for systematic methods to identify structural breaks. In the past 15 years, the most important contributions to this literature include the following three innovations: 1) Tests for a structural break of unknown timing; 2) Estimation of the timing of a structural break; and 3) Tests to distinguish between a random walk and broken time trends. These three innovations have dramatically altered the face of applied time series econometrics. We discuss these three topics in turn and use U.S. labor productivity data to illustrate their applicability.

Testing for Structural Change of Unknown Timing

The classical test for structural change is typically attributed to Chow (1960). His famous testing procedure splits the sample into two subperiods, estimates the parameters for each subperiod, and then tests the equality of the two sets of parameters using a classic \( F \) statistic. This test was popular for many years and was extended to cover most econometric models of interest. For a recent treatment, see Andrews and Fair (1988).

However, an important limitation of the Chow test is that the breakdate must be known a priori. A researcher has only two choices: to pick an arbitrary candidate breakdate or to pick a breakdate based on some known feature of the data. In the first case, the Chow test may be uninformative, as the true breakdate can be missed. In the second case, the Chow test can be misleading, as the candidate breakdate is endogenous—it is correlated with the data—and the test is likely to indicate a break falsely when none in fact exists. Furthermore, since the results can be highly sensitive to these arbitrary choices, different researchers can easily reach quite distinct conclusions—hardly an example of sound scientific practice.

To illustrate this point, let’s take United States labor productivity in the manufacturing/durables sector. We measure this as the growth rate of the ratio of the Industrial Production Index for manufacturing/durables to average weekly labor hours, a monthly time series available from February 1947 to April 2001 (yielding 601 observations).\(^1\) If we compute a Chow statistic using 1973 as the

---

\(^1\) Average weekly hours is the number of employees multiplied by average weekly hours, as measured by the Bureau of Labor Statistics.
breakdate, we obtain a value of 5.9. The 5 percent critical value from the chi-square distribution is 6.0, so we fail to find evidence of a structural break. If we instead compute a Chow statistic using 1975 as the breakdate—just two years forward—we obtain a value of 7.1. As this exceeds the 5 percent critical value, it appears to provide evidence of a structural break. But how can we be confident, as the two similar breakdates give such different answers?

The necessary solution is to treat the breakdate as unknown. This idea—and solution—goes back to Quandt (1960), who proposed taking the largest Chow statistic over all possible breakdates. For illustration, imagine a devilish trickster who is willing to go to any extreme to prove the existence of a structural break in labor productivity. This devilish trickster performs every possible Chow test imaginable, searching across possible breakdates. Through this search, this trickster finds the worst-case Chow statistic, the breakdate where the test is largest. This is Quandt’s statistic.

One way to see the construction of this statistic is to plot the sequence of Chow statistics as a function of candidate breakdates. I have done so in Figure 1 for our labor productivity example. The candidate breakdates are along the x-axis; the values of the Chow statistics on the y-axis. To compute these Chow statistics for a particular breakdate, you split the sample at that breakdate and estimate the model parameters separately on each subsample, as well as their covariance matrices. If the true parameters are constant, the subsample estimates should be (roughly) constant across candidate breakdates—subject to estimation error. On the other hand, if there is a structural break, then the subsample estimates will vary systematically across candidate breakdates, and this will be reflected in the Chow test sequence.

In Figure 1, we see considerable variation of the Chow test sequence across candidate breakdates, reaching a high of 20.2 in May 1991. This value—20.2—is the Quandt statistic.

If the breakdate is known a priori, then the chi-square distribution can be used to assess statistical significance. We sketch this 5 percent critical value with dashed lines in Figure 1. For breakdates where the Chow test sequence lies below this critical value, the test appears to be “insignificant,” and conversely for breakdates where the Chow test lies above the critical value.

However, if the breakdate is unknown a priori, then the chi-square critical values are inappropriate. What critical values should be used instead? For many years, this question remained unanswered, and the Quandt statistic had no practical application. In the early 1990s, the problem was solved simultaneously by several sets of authors, with the most elegant and general statements given by Andrews (1993) and Andrews and Ploberger (1994). These authors provide tables of critical values, and Hansen (1997) provides a method to calculate p-values.

These asymptotic critical values are considerably larger than the comparable chi-square critical values, depending on the number of parameters in the model and other factors. In our example (which has two parameters), the Andrews 5 percent critical value is 12.9, just over twice the chi-square critical value. We have sketched this critical value as well (dotted line) in Figure 1. A visual way to assess significance is to see if the Chow test sequence breaks above the critical value (as this is equivalent to the maximum of the sequence exceeding the critical value). We can see that the maximum (20.2) easily exceeds the Andrews critical value (the p-value is 0.0016), so we are easily able to reject the hypothesis of no structural break. We are therefore quite confident that this time series has a structural break.

If we find evidence of one structural break, could there be more than one? Bai and Perron (1998) develop tests for multiple structural changes. Their method is sequential, starting by testing for a single structural break. If the test rejects the null hypothesis that there is no structural break, the sample is split in two (based on the breakdate estimate presented in the next section) and the test is reapplied to each.

---

5 I use the Wald form of the statistic computed with a heteroskedasticity-consistent covariance matrix.
6 This is the likelihood ratio test under normality.
7 Actually, we do not consider literally “all” possible breakdates. We cannot consider breakdates too close to the beginning or end of the sample, as there are not enough observations to identify the subsample parameters. The conventional solution is to consider all breakdates in the interior ε percent to (1 − ε) percent of the sample, where the trimming parameter ε is typically between 5 percent and 15 percent. In the examples presented here, I use 5 percent trimming.
8 Andrews and Ploberger (1994) show that improved power can be obtained by taking exponential averages of the Chow test sequence.
9 Hansen (2000) shows that Andrews’ critical values are not robust to structural change in the marginal distribution of the regression, which is undesirable in tests focusing on conditional relationships. He shows how to simulate robust critical values on a case-by-case basis. In this example, the robust p-value (calculated with 10,000 bootstrap replications) is 0.0044, yielding the same conclusion.
subsample. This sequence continues until each subsample test fails to find evidence of a break.

The Quandt-Andrews and Andrews-Ploberger family of statistics have essentially replaced the Chow statistic in recent econometric practice. One comprehensive application is Stock and Watson (1996), who apply the tests systematically to 76 monthly time series using both univariate and bivariate regressions. They reject stability at the 10 percent level for over half of their models.

Another interesting application is Ben-David and Papell (1998), who look for evidence of "slowdowns" (a decrease in the trend function) in the Summers-Heston GDP data from 74 countries. They find statistically significant evidence of a slowdown in 46 countries. In 21 of these cases, the postbreak trend function is actually negative.

A final example getting considerable recent attention is McConnell and Perez-Quiros (2000). They test for the stability of the volatility of United States GDP growth rates and find overwhelming evidence of a substantial decrease in volatility around 1984.

### Estimating the Timing of Structural Change

In many applications, it is useful to know when the structural change occurred. Treating the date of structural change—the breakdate—as an unknown parameter, the issues are how to estimate the breakdate and how to obtain confidence intervals for the breakdate.

An obvious candidate for a breakdate estimate is the date that yields the largest value of the Chow test sequence (in our labor productivity example, May 1991). It turns out that this is known to be a good estimate only in one special case—in linear regressions when the Chow test is constructed with the "homoskedastic" form of the covariance matrix.7

In regression models, an appropriate method to estimate the parameters—including the breakdate—is least squares. Operationally, the sample is split at each possible breakdate, the other parameters estimated by ordinary least squares and the sum of squared errors calculated and stored. The least squares breakdate estimate is the date that minimizes the full-sample sum of squared errors (equivalently, maximizes the residual variance).

A theory of least squares estimation has been developed in a sequence of papers by Jushan Bai, both alone and with coauthors. Bai (1994, 1997a) derives the asymptotic distribution of the breakdate estimator and shows how to construct confidence intervals for the breakdate.8 These confidence intervals are easy to calculate and hence are very useful in applications, as they indicate the degree of estimation accuracy. Bai, Lucasdaine and Stock (1998) extend this analysis to multiple time series with simultaneous structural breaks. They show that using multiple time series improves estimation precision. Bai and Perron (1998) discuss simultaneous estimation of multiple breakdates.

Chong (1995) and Bai (1997b) show how to estimate multiple breakdates sequentially. The key insight is that when there are multiple structural breaks, the sum of squared errors (as a function of the breakdate) can have a local minimum near each breakdate. Thus, the global minimum can be used as a breakdate estimator, and the other local minima can be viewed (cautiously) as candidate breakdate estimators. The sample is then split at the breakdate estimate, and analysis continues on the subsamples. Bai (1997b) shows that important improvements are obtained by iterative refinements: reestimation of breakdates based on refined samples.

These methods can be best illustrated through our application to labor productivity. In Figure 2, we plot the residual variance (sum of squared errors divided by sample size) as a function of a single breakdate. The breakdates are on the x-axis and the residual variance on the y-axis. The sample is split at each breakdate and the regression parameters estimated separately on each subsample. The sum of squared errors is calculated (for the entire sample) and the residual variance plotted. If the true parameters are constant, the subsample estimates (and hence, sum of squared errors) will vary randomly and erratically across candidate breakdates. On the other hand, if there is a structural break, then the subsample estimates will vary systematically across candidate breakdates, and the sum of squared errors will have a well-defined minimum near the true breakdate.

The plot in Figure 2 is partially erratic, but we can discern three well-defined minima: a global minimum in January 1982 and two local minima in July 1962 and September 1993. The visual evidence suggests that two or three structural breaks are possible in this sample period.

We break the sample at the estimated breakdate (January 1982) and test for structural breaks on the two subsamples. We find no evidence for a break in the period (1947, 1982), but find evidence for a break in the period (1982, 2001). For this latter period, the residual variance has a strong V-shape as a function of the breakdate, indicating good identification, and the minimum is obtained in December 1994. Now we split the sample in December 1994 and reestimate on the sample period (1947, 1994). The Quandt test rejects the hypothesis of parameter constancy at the 5 percent level, indicating a structural break, and the least squares estimate of the breakdate is December 1965, with a second local minimum in January 1982. Now taking the sample (1964, 1994), the Quandt test fails to find evidence for a structural break. The point estimate of the breakdate again is January 1982. Finally,

---

7 When the Wald test is calculated using the homoskedastic covariance matrix, it is algebraically a transformation of the sum of squared residuals. Thus, the date that maximizes this Wald test sequence is algebraically identical to the date that minimizes the sum of squared residuals—the least squares estimator.

8 This builds on earlier work by Picard (1985).
Choosing Between Random Walk and Structural Change

Time series are often usefully described as being composed of a trend and a cycle. Before the work of Nelson and Plosser (1982), it was commonplace to assume that the trend was linear. Nelson and Plosser challenged this assumption by providing evidence that for many widely used aggregate macroeconomic time series, the trend could be characterized as a random walk. That is, instead of being a fixed trend to which the time series would revert over the business cycle, the trend would be moved by random shocks—and then would stay at the new level until disturbed by another random shock.

While this result prompted many countereffects, the most constructive was mounted by Perron (1989). Perron argued that the movement of the trend could be explained by a parsimonious single structural break in an otherwise constant linear trend. This explanation is plausible, since a trend break produces serial correlation properties that are similar to those of a random walk.

Perron (1989) showed how to test the random walk hypothesis against the trend-break model. This is achieved by estimating a linear autoregression augmented with dummy interactions to capture the desired broken trend specification. The hypothesis of a random walk trend implies that the sum of the autoregressive coefficients equals one (that is, a "unit root" in the autoregressive polynomial), so this can be easily assessed with a $t$-ratio statistic. The distribution of the $t$-ratio is non-normal, but Perron provided a distribution theory and critical values.

Perron (1989) applied this test to the Nelson-Plosser macroeconomic time series, specifying the breakdate as 1929 for the annual series and 1973 for the postwar quarterly series. He was able to reject the random walk model for most of the series at the 5 percent significance level, suggesting that the series were stationary after accounting for structural change in the trend.

However, Perron’s (1989) analysis was disputed by a collection of papers, most notably Christiano (1992), Zivot and Andrews (1992), Banerjee, Lumdaine and Stock (1998) and even Perron and Vogelsang (1992). These authors argue that it is inappropriate to specify the breakdate as known, as it is not reasonable to believe that the choice has been made independently of the data. These authors collectively suggest that an appropriate procedure is to select the breakdate that provides the most evidence against the random walk hypothesis—the breakdate that produces the largest $t$-ratio.

This procedure for choosing a breakdate can produce the same numerical value for the test statistic, as Perron’s choices of 1929 and 1973 were quite judicious. However, the test is constructed using a different procedure, and hence it has a different sampling distribution. The critical values for the modified test are much larger, making it harder to reject the null hypothesis of a random walk. Armed

\[ It \text{ is also different from the Dickey-Fuller (1979) distribution.} \]

\[ \text{The authors also found that asymptotic critical values should only be used as a crude guide and that finite-sample bootstrap critical values are much preferred.} \]
with this new theory, the evidence against the hypothesis of a random walk had evaporated.

Will longer samples of data settle the debate? Perron (1997) revisited the issue, extending the sample to 1991 III and using different methods to select the autoregressive lag order. He found slightly stronger evidence against the random walk model, but the evidence was not conclusive. The key problem is that the trend functions specified earlier in Perron (1989) do not predict well out of sample. For the annual series, the trend function severely overpredicts the 1970s and 1980s. For the quarterly series, it underpredicts for the period 1987–2000. This is consistent with a random walk trend.

The contributions to this family of tests continue to grow. Of particular note, Lundstaine and Papell (1997) allow two break dates rather than one and find that the case against the random walk is strengthened. However, the need for two structural breaks also reduces the distinction between the trend-break and random walk models.11

The Perron (1989) idea has had a large and well-deserved impact on empirical analysis and has focused attention on the time series properties of the trend. As we now understand, the distinction between a random walk and a trend break largely concerns the frequency of permanent shocks to the trend. In a random walk process, such shocks occur frequently, while in a trend-break process, they occur infrequently (once or twice in a sample). Future work may attempt to find alternative ways to narrow the difference between these models.

Perron’s (1989) idea and its variants have seen many applications. One creative example of this work is Fernandez (1997). His focus is on the question of whether changes in money help to forecast output, even after conditioning on lagged output. An earlier literature had shown that the results depend on whether or not the interest rates are included in the regression and whether a trend is included to detrend the series linearly. Fernandez uses tests from Perron (1997) to argue that output is well represented by a stationary process about a trend with a single trend break, and he uses the estimated broken trend function to detrend output. Fernandez finds that when the sample period is confined to pre-1985, this produces very robust results; however, he is unable to produce robust results when data after 1985 is included.

Another application is Papell, Murray and Ghiblai (2000). These authors are concerned with hysteresis in unemployment rates in 16 OECD countries. Hysteresis is the theory that a one-time change in unemployment can have permanent effects; thus, it is closely related to the idea that trend unemployment can be described as a random walk. Using the Perron-Vogelsang (1992) tests, the authors are able to reject the random walk hypothesis for ten of the 16 countries in favor of a one-time break in time trend. This finding suggests a very different economic interpretation about hysteresis.

U.S. Productivity

We now return to our empirical investigation of labor productivity in the U.S. manufacturing/durables sector. To review, we found strong evidence of a structural break sometime between 1992 and 1996, and weaker evidence of a structural break in the 1950s and the early 1980s. What is the nature of these changes in labor productivity? Our subsample estimates of mean growth rates (in annualized units) are 3.4 percent for 1947–1984, 2.5 percent for 1984–1992, 4.2 percent for 1992–1995 and 7.7 percent for 1995–2001. Clearly, the growth rate in the final period is quite large relative to previous history.

As there are many possible measures of labor productivity, it is natural to ask if our results are robust to alternative choices. We consider the quarterly labor productivity index issued by the Bureau of Labor Statistics, which measures output per hour of all persons for the manufacturing/durables sector. Applying the same model and methods to this series (1947 through 2000), we again find strong evidence for a break in the 1990s, with a point estimate of 1997. However, with this series, there is no evidence of a second break.

To investigate further, we disaggregate the manufacturing/durables sector by two-digit SIC industry group, and monthly series are available at this level of disaggregation for both the industrial production index and weekly labor hours. The ten industries are the following: Lumber (SIC 24); Furniture (SIC 25); Stone, clay and glass products (SIC 32); Primary metal industries (SIC 33); Fabricated metal products (SIC 34); Industrial machinery and equipment (SIC 35); Electronic equipment (SIC 36); Transportation equipment (SIC 37); Instruments (SIC 38); and Miscellaneous (SIC 39).

We apply the same empirical methods as described earlier in this paper for the manufacturing/durables sector. There is evidence of structural change in the regression function for seven of the ten industry groups, but only two—Industrial machinery and Electronic equipment—show evidence of a structural break in the mean growth rate. For the other five industry groups—Furniture, Primary Metals, Fabricated Metals, Instruments and Miscellaneous—there is a statistically significant change in the autoregressive parameter, but not in the mean growth rate. This means that the response of labor productivity to shocks has changed, but the longrun impact is unaltered.

As we stated above, the industrial machinery sector has a statistically significant break in the mean growth rate. The breakdate estimate is February 1992, and the Bai 90 percent confidence interval is October 1990 to June 1994. The model implies a mean growth rate (in annualized units) of 3.3 percent before 1992 and 7.8 percent after the break. There is no evidence of a second structural break in this series.
The other series with a break in the mean growth rate is electronic equipment. The
breakdate estimate is December 1993, and the Bai 50 percent confidence
interval is June 1998 to January 1995. The estimated mean growth rate increased
from 5.8 percent before 1995 to 17.8 percent after 1993.

There is a remarkable coincidence of the breakdates for the manufacturing/
durable sector as a whole and the industrial machinery and electronic equipment
industry groups. Apparently, the structural break in the sector is due to roughly
simultaneous breaks in these two industry groups.

This investigation raises many questions: Is this break permanent or transitory?
Can further disaggregation identify the sources of the productivity break? Is the
timing of the structural break simultaneous across sectors? Why are there no
spillover effects into other industries? Can the labor productivity gain be explained
by increased utilization of other factors? Are these increases unique to the United
States, or have they occurred in other countries as well? I expect that these and
related questions will be exciting avenues for future research.

Conclusion

Structural change is pervasive in economic time series relationships, and it can
be quite perilous to ignore. Inferences about economic relationships can go astray,
forecasts can be inaccurate, and policy recommendations can be misleading or
worse. The new tools developed in the past few years are useful aids in econometric
model specification, analysis and evaluation.

I thank the editors—Alan Krueger, Brad De Long, Michael Waldman and Timothy
Tayler—for detailed and helpful comments on a previous draft, which greatly improved the
paper. I also thank John Kneser and Ken West for insightful suggestions about labor
productivity. This research was supported by grant from the National Science Foundation.
Any opinions, findings and conclusions or recommendations expressed in this material are
those of the author, and they do not necessarily reflect the views of the NSF. The data and
Gauss programs used in this paper are available on the author’s webpage, (http://www.ssc.
wisc.edu/~bhamsem).

References

Andrews, Donald W.K. 1993. "Tests for Parameter Instability and Structural Change with Un-


Andrews, Donald W.K. and Werner Ploberger. 1994. "Optimal Tests When a Nuisance Parame-


Bai, Jushan. 1997b. "Estimating Multiple Breaks One at a Time." Econometric Theory, June,

Bai, Jushan and Pierre Perron. 1998. "Estim-
ing and Testing Linear Models with Multiple
Structural Changes." Econometrica, January, 66:3,
pp. 47–78.

Bai, Jushan, Robin L. Luminesdine and James H.
Stock. 1998. "Testing for and Dating Com-
mon Breaks in Multivariate Time Series." Review

Banerjee, Awanj, Robin L. Luminesdine and
Tests of the Unit-Root and Trend-Break Hypoth-
eses: Theory and International Evidence." Jour-
nal of Business and Economic Statistics, July, 10:3,
pp. 297–317.

Ben-David, Dan and David H. Papell. 1998.
"Stochastic and Meltowns: Postwar Growth Evi-
dence from 74 Countries." Review of Economics

Cheung, Yung-Wai and Edward Lo. 1995. "Partial
Parameter Consistency in a Mispecified Structural
Change Model." Econometrica, October, 49:4,
pp. 1051–76.

Mean Reversion in Stationary and Nonstationary
Time Series." Journal of Business and Economic

Break in GNP." Journal of Business and Economic

"Distribution of the Estimators for Autoregres-
sive Time Series with a Unit Root." Journal of the
American Statistical Association, June, 74:366,
pp. 621–36.

Elliott, Graham, Thomas J. Rothenberg and
Autoregressive Unit Root." Econometrica, July,
64:4, pp. 813–36.

Money-Output Correlation." Review of Economic

Hansen, Bruce E. 1997. "Approximate Asym-
ptotic P-Values for Structural Change Tests." Jour-
nal of Business and Economic Statistics, January,

Change in Conditional Models." Journal of Eco-

"Multiple Trend Breaks and the Unit-Root
Hypothesis." Review of Economics and Statistics,

McConnell, Margaret M. and Gabriel Peres-
quiros. 2001. "Output Fluctuations in the
United States: What Has Changed Since the
Early 1980s?" American Economic Review, December,
90:5, pp. 1460–74.

"Trends and Random Walks in Macroecon-
omic Time Series." Journal of Monetary Eco-
nomics, September, 10:2, pp. 139–66.

Papell, David H., Christian J. Mayre and Hala
Ghiliani. 2000. "The Structure of Unemploy-
ment." Review of Economics and Statistics, May,
82:2, pp. 509–15.

Perron, Pierre. 1989. "The Great Crash, the
Oil-Price Shock, and the Unit-Root Hypothesis.
"Econometrica, November, 57(6), pp. 1691–701.

Breaking Trends Functions in Macroeconomic
Variables." Journal of Economicometrics, October, 83:2,
pp. 355–85.

Perron, Pierre and Timothy J. Vogelsang.
1992. "Nonstationarity and Level Shifts with an
Application to Purchasing Power Parity." Journal
of Business and Economic Statistics, July, 10:3,
pp. 391–408.

Change-Points in Time Series." Advances in Ap-
plied Probability, 17:2, pp. 581–84.

esis that a Linear Regression Obers Two Sepa-
rate Regimes." Journal of the American Statistical
Association, 55, pp. 343–347.

Salkovskis, Pentti and Heimito Luttaspehl.
with a Level Shift at Unknown Time." Contrib-
uted paper, Economic Society World Congre.
Econometric Theory, Forthcoming.

Stock, James H. and Mark W. Watson.
1996. "Evidence on Structural Instability in
Macroeconomic Time Series Relations." Journal of Business

Evidence on the Great Crash, the Oil-Price
Shock, and the Unit-Root Hypotheses," Journal of
Business and Economic Statistics, 10, pp. 251–70.
The Reality of Macroeconomic Structures

All are keeping a sharp look-out in front, but none suspects
that the danger may be creeping up from behind. This shows
how real the island was.

J. M. Barrie, Peter Pan

David Hume, the macroeconomist and historian, talked freely about
causes linking macroeconomic and social variables and events; David
Hume, the philosopher of causality, analyzed almost exclusively physical
causes. In the preceding chapters, we have followed in Hume's footsteps,
concentrating on physical and biological examples. It is now time to
inquire, as Hume never did, into the connection between the practical
causal reasoning of macroeconomics and the general causal analysis
developed for other problems. We address two questions. First, in this
chapter, we ask, is macroeconomics a suitable subject for a structural
causal account? That is, are there macroeconomic structures analogous
to physical structures? In the next chapter, we ask, what constraints does
the nature of macroeconomics place on the details of a suitable causal
account?

The subject of these chapters (and, for that matter, of the book as a
whole) has been left up to now implicit. It might be useful to start with
a working definition. To begin at the beginning, we define the key terms.
"Macroeconomics" is sometimes thought to be the economics of broad
aggregates, and "microeconomics" the economics of individual economic
actions. Although he did not use the terms microeconomics and macro-
economics, John Maynard Keynes (1936, pp. 292–293) drew a related dis-
tinction: Microeconomics is the theory of the individual industry or firm;
macroeconomics is the theory of output and employment as a whole. As
Maarten Janssen (1993, ch. 1) shows, these alternative definitions cut in
somewhat different ways. They are, however, similar enough for present
purposes, since any quantification of output and employment as a whole
is bound to involve broad aggregates. Economics is a social science; but
macroeconomics is remarkably dehumanized. People are at best referred
to as representatives of a class; and, mostly, the talk is about variables.
Macroeconomics is thus that area of economics that studies GDP, unem-
ployment, interest rates, the flow of financial resources, exchange rates,
and so forth.

The structural account of causality presupposes a realist ontology.
Causes are real properties of variables in structures. It seems unpro-
blematic to think of a billiard ball or a piston and cylinder or a disease as
a real structure. It is somewhat harder to think of GDP or the general
price level as structures or parts of structures. They seem to be derived
from something more basic, and not to be real things at all, but creations
of our minds—mere summaries for the genuinely real behaviors of indi-
vidual economic actors. The recent history of macroeconomics, with the
emphasis on the microfoundations of macroeconomics, reinforces these
doubts.

What is meant by realism? Uskali Mäki (1996) offers a careful dis-
cussion of realism in relation to economics that might help to define the
term. Mäki distinguishes between ontological realism, which raises ques-
tions about what there is, and semantic realism, which raises questions
about the connection between language and what there is. Semantic
realism can be analyzed further, but the central issue here is ontological:
Do macroeconomic aggregates and causal relations between them exist?
More precisely, and again using Mäki's terminology, do macroeconomic
aggregates exist externally (i.e., independently of any individual human
mind) and objectively (i.e., unconstituted by the representations of
macroeconomic theory). Advocates of what might be called the strong
program of microfoundations would reject realism with respect to
macroeconomic entities in this sense and, implicitly, reject macro-
economics as a suitable subject of structural causal analysis. If it could
be established that macroeconomics was a suitable subject for a realist
causal account, then one of the central rationales for the program of
microfoundations for macroeconomics would be eliminated.

5.1 THE PROGRAM OF MICROFOUNDATIONS

Before the twentieth century, the most common definition of economics
was epitomized by Alfred Marshall (1920, p. 1): "[A] study of mankind in

1 This chapter is a revised and slightly expanded version of Hoover (1995c). Copyright ©
the ordinary business of life; [economics] examines that part of individual and social action which is most closely connected with the attainment and with the use of the material requisites of well-being." This definition, which is reasonably hospitable to macroeconomics, has now largely been supplanted by that of Lionel Robbins (1935, p. 16): "Economics is the science which studies human behaviour as a relationship between ends and scarce means which have alternative uses." On Robbins’s definition, economics must be fundamentally about the individual.

Modern macroeconomics developed in the wake of Keynes’s General Theory of Employment Interest and Money (1936). Typical elements of Keynes’s analysis were the consumption function, which related aggregate consumption to aggregate national income, the investment function, which related aggregate investment to the general rate of interest, and the liquidity preference function, which related the aggregate stock of money to aggregate national income and the general rate of interest. It is easy to understand that in a profession committed to Robbins’s definition of economics, such aggregate relationships were at best rather unappealing way stations on the path to an individualist economics. The program of microfoundations, as it has developed over the past 60 years, aims to explain all macroeconomic properties of the economy – in principle, at least – by reference to the behavior of rational economic actors such as postulated by microeconomics.³

Approval of the program of microfoundations is almost universal among economists. Those economists who have reflected at all deeply on the matter typically associate microfoundations with methodological individualism (e.g., Janssen 1993, pp. 26–29 passim; Boland 1979, chs. 2 and 5). Mark Blaug (1992, p. 44) defines methodological individualism as the principle that "asserts that explanations of social, political, or economic phenomena can only be regarded as adequate if they run in terms of the beliefs, attitudes, and decisions of individuals."

Methodological individualism is a doctrine about explanation. Despite lip-service to it, it is not widely practiced by economists. The reason is what I have elsewhere labeled the "Cournot problem" after its lucid, early formulation by Augustine Cournot (1838/1927, p. 127), the nineteenth-century mathematician and economist: There are too many individuals (firms and consumers) and too many goods to be handled by direct modelling.² Blaug (1992, p. 46) observes that few explanations of macroeconomic phenomena have been successfully reduced to their microfoundations, so that insistence on microfoundations would eliminate explanations of macroeconomic phenomena tout court. Even Lucas (1987, pp. 107–108), an important advocate of the program of microfoundations, holds up only the hope of the elimination of distinction between microeconomics and macroeconomics.⁶

The commitment of economists to methodological individualism is thus not grounded in successful applications. Rather it appears to be based on an instinctive belief in ontological individualism: the doctrine that all that exists fundamentally for the economy are individual economic actors. Lucas and his fellow new classical economists have promoted representative-agent models, a class of models in which the mathematical methods of microeconomic optimal choice are applied to a single individual who takes national income as his budget constraint and whose choices are taken to represent the aggregate choices of the economy, because they appear to achieve the reduction of macroeconomics to microeconomics as required by the program of microfoundations for macroeconomics. A. P. Kirman (1992) severely criticizes the representative-agent model, not because it aspires to methodological individualism, but because it fails to fulfill the necessary conditions for perfect aggregation, so that the representative agent in the models fails to represent actual individuals successfully.² Methodological individualism remains the goal. Similarly, David Levy (1985) argues that complete methodological individualism is impossible, because, given imperfect information, individual economic actors must make reference to collective entities as part of their own decision-making processes. Nevertheless, Levy (1985, p. 107) writes: "These collectives have no real existence but are simply the product of theories." While defending macroeconomics against the strong claims of the program of microfoundations, Blaug (1992, p. 45) nevertheless writes: "ontological individualism is trivially true."

It is important to understand that there are some senses in which neither the methodological nor the ontological individualist denies the existences of aggregates, collectives, or wholes. No one denies that GDP calculations can be made and reported, or even that GDP may have some locally stable relationship to unemployment or average interest rates. No one denies the existence of

³ See Weintraub (1979) and Janssen (1993) for general discussions of microfoundations for macroeconomics.

⁴ I must be careful not to leave a false impression: Both Janssen and Boland are critics of the program of microfoundations.

⁵ Hoover (1988a, p. 135). Also see Hoover (1988a, ch. 9, esp. sec. 2; ch. 10, pp. 241–244); Friedman (1955).

⁶ It is because he regards the success of the program of microfoundations as a hope rather than an inevitable triumph that I have previously characterized Lucas as calling for the euthanasia of macroeconomics – a pleasant and, no doubt, incremental death – rather than for its extermination or liquidation (Hoover 1988a, p. 87).

Harley (1997) provides a detailed examination of the representative-agent model and related aggregation issues.
social organizations such as governments or firms (in the sense that talk of governments and firms conveys meaning). What is typically denied is that such aggregates or organizations are among the fundamental units from which economic reality is constructed.

Hayek (1979, ch. 4) argues that such entities are secondary, and that the role of a social science is “compositive” – that is, that it must explain these entities as arising from the fundamental individual components. Hayek (1979, ch. 6) denies that the wholes that social science explains through compositive methods are subject to scientific laws. He holds up the attempt to connect them through laws as an example of Whitehead’s “fallacy of misplaced concreteness.” He writes, “the wholes about which we speak exist only if, and to the extent to which, the theory is correct which we have formed about the connection of the parts which they imply, and which we can explicitly state only in the form of a model built from those relationships” (Hayek 1979, p. 98). Hayek thus argues that aggregates exist, but derivatively rather than fundamentally, and that in Mäki’s terminology they do not exist objectively (i.e., unconstituted by the representations of theory). Still even Hayek does not endorse practicable methodological individualism, stressing the importance of a reduction to microfoundations in principle and himself citing the Cournot problem (Hayek 1979, pp. 74–75, esp. fn. 8).

It may come as a surprise to most economists to discover that an essentially similar debate over reductionism rages in physics. Some condensed-matter physicists are for the emergence of properties that are not reducible without loss to the behavior of the particles that constitute their substance. Some nuclear and particle physicists insist that such a reduction is fundamental. (See Cat 1998.) My position is analogous to that of the condensed-matter physicists.

5.2 IS MACROECONOMICS ONTOLOGICALLY PROBLEMATIC?

One might concede the main point of the last section – namely, that the drive for microfoundations is driven by ontological individualism – yet not believe that any interesting metaphysical issue is involved, because the ontology of economics is too well understood by common sense to

---

While not denying that aggregates such as GDP or the general price level can be calculated, Mises (1949/1966, p. 217) goes farther than Hayek in arguing that they are quite devoid of meaning (also see Lachmann 1976, p. 96).

The terminology of “fundamental” or “derivative” is fraught with difficulties. It is beyond my purpose to try to sort such matters out here. It is enough for the point at hand, however, to note that Hayek does not believe that economic aggregates can be causes in their own right. They might serve as some sort of shorthand, but he argues that there is always an adequate causal mechanism independent of that shorthand.

---

pose any serious puzzle. Mäki (1996), for example, contrasts folk economics with scientific economics, arguing that scientific economics merely presents modifications of the “ontic furniture” of the general folk views of “man” and society.” He lists some types of possible modifications: selection, abstraction, idealization, exaggeration, projection, and aggregation. But he maintains that none of these modifications or combinations thereof “accomplishes a major departure from the ontic furniture of the ordinary realm. No new kinds of entities or properties are introduced” (Mäki 1996). Mäki illustrates his point with a discussion of the business firm in standard neoclassical analysis, and concludes that “folk economics and neoclassical economics have real business firms as their shared referent even though they represent these firms differently.”

Mäki’s case for the “ontological commonsense realism” of economics in which the ontic furniture poses no special challenges to the understanding is more persuasive for some parts of economics than others. It is not particularly cogent for macroeconomics: Macroeconomic aggregates do not share a referent with folk notions. The case can be illustrated with reference to the related notions of “real GDP” (or “real income” or “real output,” these terms being used almost interchangeably) and the “general price level.”

On any interpretation, macroeconomics takes a larger view of the economy and deals with aggregates, which are, in turn, constructed from characteristics of individual economic actors. It is helpful to distinguish two types of aggregates.

First are what we might call natural aggregates: simple sums or averages. Examples of natural aggregates are the level of total employment or the average rate of interest on six-month commercial paper. These are called natural because they are measured in the same units (i.e., they have the same dimensionality) as the individual components that they comprise and, therefore, preserve a close analogy with those individual components. Employment, for instance, is measured by the number of workers or the number of labor hours at both the level of a single individual and in aggregate. The rate of interest on a bond and the average rate of interest on a group of similar maturity bonds are both expressed as a percentage yield per unit time.

A second type of aggregate, what we might call synthetic aggregates, are important for macroeconomics. These are called synthetic because they are fabricated out of components in a way that alters the structure of the components, so that they are dimensionally distinct from the components and so that there is no close analogy (despite their sometimes sharing a common name) with the components. The nature of the synthesis is well illustrated by the general price level. The notion of a general price level aims to capture a prescientific insight: “A dollar just ain’t what
it used to be!"; "when I was a lad, a penny would buy what a quarter does these days." To capture this insight, one would like to have some notion of the average level of prices. A simple average will not of course work: \((10\text{¢}/\text{orange} + 20\text{¢}/\text{apple} + \$27,948/\text{Volvo station wagon})/3\) does not convey any useful information. One cannot add apples and Volvos, as they say.

It might be argued that any sort of an average is altogether the wrong way to start. What is really wanted is an estimate of the price of money itself, and not the average price of goods. The price of money would, like the price of oranges, be a single number. The general price level could be defined to be its inverse \((1/p_m)\). Since relative prices of goods change because of changes in the conditions of demand and supply, there would be, at best, a rough proportionality between individual prices and the general price level. Indeed, it would permit one to isolate which changes in individual prices were the result of "real" factors and which of monetary factors.

This approach, however, does not do justice to the prescientific insight, for it does not provide us with a concept or a measurement of prices that is independent of highly particular and highly inadequate theoretical models. To see this, consider how one would actually determine the value of \(p_m\). One might, for instance, write down a complete Walrasian general equilibrium model in which commodities were expressed in natural quantities, prices in terms of money, and all assets denominated in money were valued using \(p_m\). Aside from the impracticality of formulating and solving such a model for an actual economy (the Cournot problem again), it is well known that \(p_m\) might not be determinable in such a model, or if it is, might not be unique.\(^{10}\) The essence of the problem is that the real quantity of money (i.e., the useful services that it provides), unlike the real quantities of apples or Volvos, depends fundamentally on the price of other goods. In adjusting the prices of various goods, including money, unique convergence may not be possible, because each time prices adjust to remove an excess demand or supply, the quantity of money changes—possibly in a discontinuous manner—which can increase rather than diminish some of the excess demands or supplies.

This problem in the foundations of monetary theory has yet to be satisfactorily resolved. But what if it had been? It would tie the notion of the general price level extremely closely to a particular theoretical analysis. The measurement of \(p_m\) would be derived rather than fundamental (see Ellis 1966, ch. 8). The prescientific notion is not tied to such a derivation. That would not pose any special problem if the general price level derived in this way correlated closely with numerous other theoretical methods of deriving it, which in turn correlated reasonably with the prescientific sense of a general rise in prices.\(^{11}\) Temperature provides an example of what is wanted (Ellis 1966, ch. 6). The prescientific notion of hotter and colder is vague. The first attempts to provide quantification relied on some presuppositions—e.g., the linearity of the expansion of the various fluids used in thermometers—but were not tied to particular theories. Temperature measures can now be derived from particular theoretical understandings—e.g., from the kinetic theory of gases. Such derived measures show considerable—if necessarily imperfect—consistency with the prescientific notions of hotter and colder and with the theoretical measurement systems. They permit the extension of temperature scales beyond ordinary experience—e.g., to the measurement of the temperature of the sun—but retain their independence from particular theories to the degree that there is consonance of measurements derived from different theoretical starting points. In contrast, the measurement of the general price level remains at the atheoretical stage, in which the makers of price index numbers, Laspeyres, Paasche, and Fisher, are the economists' Fahrenheit and Celsius.

The point of raising these difficulties in measuring the general price level is not that the existence of aggregates is tied to their measurement. Rather it is that the difficulties in measuring them help to expose what problematic entities they are and undermine the appeal of seeing them as close analogues of their components (particular prices, particular goods, and so forth). The disanalogies can be made clearer through a more detailed examination of the general price level.

The fundamental insight of the index number is that one can avoid some of the dimensional nonsense of averaging disparate prices by averaging percentage rates of change instead. A simple average, however, does not capture the common-sense feeling for the degree of price change. A change in the price of gasoline should count for more than, say, a change in the price of caviar in measuring the change in the general price level. How to weight various price changes turns out to have an irreducible degree of arbitrariness.

In general, the percentage change in the general price level, indicated as \(\Delta p\) (where \(p\) is the logarithm of the general price level), is related to the individual underlying prices as

\[
\Delta p = f(\Delta p_1, \Delta p_2, \ldots, \Delta p_n)
\]

\(^{10}\) Hahn 1965; Samuelson and Sato 1984; see Hoover 1988a, ch. 5, sec. 1 for a simple exposition.

\(^{11}\) Avogadro's number, for example, can be computed to take the same value from numerous theoretically independent methods (see the discussion in Hacking 1983, pp. 54-55).
where $p_j$ is the price of good $j$ for $j = 1, 2, \ldots, n$. Now, the properties that theoretically restrict the functional form of $f(.)$ are very weak:

\[
\Delta p \leq \max\{\Delta p_1, \Delta p_2, \ldots, \Delta p_n\}, \quad (5.2)
\]

\[
\Delta p \geq \min\{\Delta p_1, \Delta p_2, \ldots, \Delta p_n\}. \quad (5.3)
\]

Conditions (5.2) and (5.3) together imply two obvious corollaries:

\[
\text{if } \forall j \Delta p_j \equiv 0, \text{ then } \Delta p \equiv 0. \quad (5.4)
\]

\[
\text{if } \Delta p_1 = \Delta p_2 = \ldots = \Delta p_n, \text{ then } \Delta p = \Delta p_j \text{ for any } j. \quad (5.5)
\]

Conditions (5.2) and (5.3) say that the general price level cannot increase by more than the largest nor decrease by less than the smallest individual price change. Condition (5.4) says that if each price increases (decreases) the general price level must itself increase (decrease), and that the general price level cannot change if no individual price changes. Condition (5.5) says that if every individual price changes equiproportionally so must the general price level. An infinite number of functions fulfill these conditions, and the range of consistent changes in the general price level, given a fixed set of underlying price changes, is wide. In practice, price indices are generally linear:

\[
\Delta p = \sum_{j=1}^{m} w_j \Delta p_j, \quad \text{where } \sum_{j=1}^{m} w_j = 1. \quad (5.6)
\]

For $m < n$, this formulation recognizes the practical fact that price indices are based on samples of selected goods. The weights $w_j$ in these indices are chosen in practice to capture the prescientific sense of the amount of a price rise. Two common weighting schemes with rationales in microeconomic consumer theory are the Laspeyres index, which chooses the weights to reflect the share of each good in base period consumption, so that the general price level effectively measures the changing cost of a fixed bundle of goods, and the Paasche index, which chooses weights to reflect the share of each good in current period consumption, so that some compensation is made for substitution from relatively more expensive to relatively cheaper goods in the face of changes in relative prices.

Neither index is "correct"; there are an infinite number of indices lying between the two, and economists have from time to time argued the case for other indices with different weighting schemes.\(^{12}\) The non-uniquness of the price index is important for the point of this chapter. It is a fundamental property. A price index is an attempt to quantify the prescientific insight that the value of money changes. The different admissible price indices are not, however, approximations to some true underlying general price level. The general price level is, in some fundamental sense, nonscalar, although there is no currently acceptable scientific refinement that captures that fact.\(^{13}\) No similar indefiniteness attaches to any of the prices of the underlying individual goods.

The change in the general price level, $\Delta p$, may be integrated over time to generate a price level

\[
(p_{t+1} - p_t = \int_t^{t+1} \Delta p dt = p_{t+1} + c).\]

The constant of integration $c$ permits us to choose an arbitrary base. The base is usually $P_0 = 1$ or $P_t = 100$, where $P = \exp(p)$, for some desired base period $t$ (other base values are less common, but not unknown). $P$ differs from the particular price of, say, a Volvo, not just in its intrinsic indefiniteness, but also in its dimensions. The dimensions of the price of a Volvo are dollars/Volvo; the dimensions of $P$ are period-$t$ dollars/base-period dollars. The dimensions of $P$ are not the dimensions of the price of any good. They appear to be the inverse dimensions of the price of money, taking base-period money to be the numéraire. Given the indefiniteness of the price index, however, it is evident that the price of money is unlike the price of other goods, and represents a substantial departure from pretheoretic notions of price. The price index is used to normalize the price of particular goods, thereby to decompose individual price changes into a common or general element and a real or relative (to the index) element. That this operation is not obvious to common sense will be evident to anyone who has taught elementary economics or read policy analysis by noneconomists.

Real GDP is another important example of a synthetic aggregate. Considered as national income, nominal GDP adds up the incomes of each individual in the economy and is an obvious extension of the accounting framework for business or personal income. In a major innovation in economic analysis, the national accounting framework since the 1930s establishes the three-way identity between the sum of all incomes, the value added in production, and the value of all final goods and services. That these other methods of computing GDP have obvious common-sense analogues is less clear. If final goods (i.e., goods that are not inputs into other production processes) are indicated by $Q$, then nominal GDP is

\[
\text{(5.7)}
\]

\[\text{This may be an area in which the theory of fuzzy sets would be helpful. The use of scalar indices may account for some portion of the apparently irreducible randomness in estimated macroeconomic relations.}\]
\[ Y^N = \sum_{j=1}^{n} P_j Q_j, \quad j = 1, 2, \ldots, n. \] (5.7)

The dimension of income is dollars/unit time. Money provides the common unit that is essential if disparate goods are to be added.

If some or all of the prices of individual goods increase, it is obvious that nominal GDP could increase without any of the individual quantities changing. In the utilitarian framework that underlies economics, this is an undesirable characteristic, because the measure of income has changed without the underlying utility, which is assumed to be generated by the quantities of the goods themselves, changing. It is clearly desirable to correct for changing prices. The usual way to do this is to compute real GDP as

\[ Y^R = Y^N / P. \] (5.8)

Real GDP is often treated as the analogue of an individual good. It does not, however, have the dimensions of a real good. Rather its dimension is base-period dollars (not dollars/unit good). Real GDP is a derivative measurement. One gets a different measurement for it for every different admissible price index. It inherits the fundamental fuzziness of the general price level.

The analogy of real GDP to an individual good is suggested to some by the possibility of perfect aggregation. If, for example, relative prices are constant (i.e., \( p_j / p_k \) is constant for all \( j \) and \( k \)), then \( \sum_{j=1}^{n} P_j Q_j / k \) (where the \( i \) in the subscript indicates the base time period \( t \)) can be normalized by choosing the units for the \( Q_j \), so that each \( P_j = 1 \). Then nominal GDP at time \( n \) can be written

\[ \sum_{j=1}^{n} P_{j,n} Q_{j,n} = P_{n,n} \sum_{j=1}^{n} Q_{j,n}. \] (5.9)

In this case, conditions (5.2) and (5.3) above ensure that \( P \) is unique. Some conclude therefore that in this limited case, one can treat the summation in the right-hand side of equation (5.9) as a natural aggregate quantity analogous to an individual quantity. The conditions for constant relative prices are almost certainly never fulfilled, but even if they were, the summation is not analogous to an individual quantity. The general price level \( P \) in equation (5.9) still has the dimension period-\( n \) dollars/period-\( t \) (i.e., current period/base period) dollars. To sum heterogeneous goods, they must still be converted to a common denominator, and in this case, the summation still has the dimensions of period-\( t \) dollars. This would be more perspicuous if (5.9) were written as

\[ \sum_{j=1}^{n} P_{j,n} Q_{j,n} = P_{n,n} \sum_{j=1}^{n} Q_{j,n}. \] (5.10)

where the subscripted numeral 1 is a placeholder for the dimensional conversion.

The general price level and real GDP are the most important aggregates in macroeconomics. There are many others. Each mixes the characteristics of simple and synthetic aggregates to different degrees. Average interest rates were cited above as an example of a simple aggregate, but when averaging is across nonhomogeneous maturities and risk classes, interest rates, too, are complicated by the fundamental problems of index numbers. Aggregation of employment across skill or quality levels faces similar considerations. There are other derivative quantities as well. The real rate of interest is defined to be the market interest rate less the percentage change in the general level of prices (\( \Delta p \)). Like real GDP, the real rate of interest inherits the fundamental fuzziness of the general price level.

The history of quantitative economics demonstrates that even the use of simple averages represented a difficult conceptual leap. On the best interpretation, what is accepted to common sense is relative. To treat synthetic aggregates as mere extensions of common-sense notions appears in comparison to make a category mistake. Despite their deceptively related names, there is no simple analogy between the general price level and individual prices or between quantities of individual goods and real GDP.

### 5.3 THE SUPERVENIENCE OF MACROECONOMICS ON MICROECONOMICS

Synthetic aggregates, at least, are not direct extensions of folk ontology. It is clear, however, that, if their independent reality is to be demonstrated in a sense more fundamental than that one can always calculate them according to some algorithm, we must first show that such aggregates cannot be reduced to properties of individual economic actors. Aggregates are in fact calculated; they clearly do not exist in a separate Platonic realm; and ontological individualism has immediate appeal, because we all have first-hand experience as economic actors. Any account of the autonomy of macroeconomic aggregates must account, therefore, for the relationship of the individual to the aggregate. To deny an essential connection between individual economic behaviors and macroeconomic aggregates would be absurd. The thing to be demonstrated is rather that economic reality is necessarily characterized differently at the microeconomic and the macroeconomic levels and that
one level is not reducible to the other. To sustain this claim we appeal to the philosophical literature on supervenience, to which we give an anti-reductionist twist.

Macroeconomic aggregates *supervene* upon microeconomic reality. What this means is that even though macroeconomics cannot be reduced to microeconomics, if two parallel worlds possessed exactly the same configuration of microeconomic or individual economic elements, they would also possess exactly the same configuration of macroeconomic elements. It is not the case, however, that the same configuration of macroeconomic elements implies the same configuration of microeconomic elements.

The relationship of the ideal gas laws to the kinetic theory of gases provides a simple illustration of supervenience. The kinetic theory describes gases as a collection of elastic particles (molecules) with randomly distributed momenta. Pressure corresponds to the average force that those particles impart to the walls of an enclosing container. Temperature corresponds to the average translational kinetic energy per molecule. The micro level of the molecules and the macro level of pressure and temperature operate under different descriptions. Pressure and temperature are not distinct or well defined at the micro level. A single speeding molecule has a momentum and kinetic energy, but it does not have a pressure or temperature, which are properties of a collection large enough to exemplify a consistent statistical character. But temperature and pressure supervene on the micro level in the sense that any particular arrangement of molecules and their momenta corresponds to a particular pressure and temperature, and, if that arrangement is ever repeated, the same pressure and temperature would be registered. There are, however, infinitely many arrangements of molecules and their momenta that register the same pressure and temperature. The gas laws provide a classic case of supervenience. Although the kinetic-theoretic account is sometimes presented as a clear case of a successful reduction of the macro to the micro, in fact, something is lost in the reduction. On the one hand, there is a shift of fundamental concepts between the levels, even though there is a mapping: Temperature and pressure are not defined at the micro level. And, in order to make the reduction, the kinetic theory has to appeal to an irrevocably statistical (and, therefore, macro) postulate: Momenta are assumed to be distributed so that there is an equal probability that any molecule is moving in any direction.

A second illustration is drawn from the life sciences. Biology provides both analogies and disanalogies for economics. Alexander Rosenberg

---

14 An account of this relationship is standard in most introductory physics textbooks. See, e.g., Weidner and Sells (1973, ch. 19, secs. 1–3).

15 The account of Kim’s analysis here omits most of the technical details (these are also reproduced in Rosenberg 1985, pp. 113–116), and changes his notation. The identification of the distinct realms of properties as “micro” and “macro” is my addition – literally ad hoc – and does not significantly affect Kim’s analysis.

16 Kim goes on to show that in some cases the implication in relation (5.11) can be strengthened to a biconditional.
Using Kim's analysis, Rosenberg argues against the autonomy of Mendelian genetics. The conceptual scheme of Mendelian genetics (the macro level) does not map easily into the conceptual scheme of molecular biology (the micro level). Mendelian genetics permits explanation of phenomena not easily explainable directly from the molecular level. However, Mendelian genetics fails to account for some phenomena within its scope. Rosenberg argues that Mendelian genetics supervenes on molecular biology, and that molecular biology is the more scientifically advanced, more fundamental and autonomous theory. Mendelian genetics is reducible in principle (that is the upshot of Kim's analysis in relation (5.11) above), but it retains heuristic power, because something like the Cournot problem prevents the practicable application of molecular biology to some phenomena in which Mendelian genetics is relatively successful. Indeed, one might argue that supervenience fails if the supervenience base is only the DNA. It holds only if the supervenience base includes the full developmental environment. In that case, the Cournot problem applies with great viciousness.

Even if supervenience implies reduction in the case of genetics, the analogous argument does not work in economics. Rosenberg (1992, p. 129) himself has argued that the intentional character of microeconomics limits its scientific development: Microeconomic "theory's prediction and explanation of the choices of individuals [cannot] exceed the precision and accuracy of commonsense explanations and predictions with which we have all been familiar since prehistory." In fact, macroeconomic explanation and prediction is not only often better, but may have more scope for improvability. An electric supplier could not say when Mary Smith will switch on her oven, but it may know pretty precisely how many kilowatts it must supply at a given time, based on an aggregate analysis of past behavior. Insurance companies know that whether an individual is, say, a smoker or obese matters probabilistically to his chances of dying. But the company would go broke trying to predict individuals' precise dates of death.17

It is important to remember that it is not macroeconomic theory that supervenes on microeconomic theory, but macroeconomic reality that supervenes on microeconomic reality. The disabilities of microeconomic theory thus prove, at most, that there can be no automatic presumption that microeconomics is more basic, because it is more successful, and that macroeconomics is merely heuristic. The critical relationship is the reducibility in principle suggested by relation (5.11) above. To begin to undermine reducibility in the case of macroeconomics, it helps to note a crucial disanalogy with biology. Reduction appeals to biologists because it removes scientifically suspect teleological explanation common in evolutionary biology and other functional accounts. The aim of reduction in economics, however, is precisely the opposite: Macroeconomics appears mechanical and dehumanized, and the point of the program of microfoundations is to reintroduce human decision making as an explanatory element. The point is to recover intentionality.

Kim's analysis posits two levels of properties that are (semantically at least) distinct and then investigates how they must be related if one set is supervenient on the other. Intentionality at the macroeconomic level undermines the distinctness of microeconomic properties from macroeconomic properties. Levy's argument (see Section 5.1) that individual economic actors will invariably make reference to social wholes and aggregates is even more fundamental than he imagines. In evaluating the future, individuals must form expectations about real prices and real quantities. Independently of the uncertainty of the future, the Cournot problem implies that it is impracticable to solve good-by-good, price-by-price, period-by-period planning problems in all their fine detail. Practically, the best that one is able to do is to work with aggregates. The information on which these are based is fundamentally monetary. Economic actors must use estimates and expectations of the general price level and real interest rates to form any practical assessment of their situations. Hayek (1979, p. 62) writes:

in the social sciences it is necessary to draw a distinction between those ideas which are constitutive of the phenomena we want to explain and the ideas which either we ourselves or the very people whose actions we have to explain may have formed about these phenomena and which are not the cause of, but theories about, the social structures.

What Levy's argument demonstrates is that Hayek is mistaken, that how people theorize about the economy is constitutive of macroeconomic phenomena.18 Since people cannot theorize about certain sorts of phenomena without appealing to macroeconomic categories - that are not themselves reducible to microeconomic categories - the Cournot problem introduces analytical constraints, not only in practice, but in principle as well. The distinctiveness of the properties at the microeconomic and macroeconomic levels is breached, undermining Kim's analysis, because complete characterizations of the microeconomic must include characterizations of the macroeconomic on the part of individual agents.

17 Both these examples are repeated verbatim from Hoover (1995a).

18 In contrast to Hayek, his fellow Austrian-school economists, Mises (1943, p. 252) argues that knowledge of economic theory can prevent the mistaken investments that fuel the business cycle.
To challenge the applicability to economics of the reductionism implicit in Kim's analysis and in Rosenberg's application of it to biology, does not challenge the notion that macroeconomics supervenes on microeconomics. Kim's analysis is epistemological; it argues that there must be laws that would permit us to draw consequences between the micro and macro levels. The point here is ontological: Even though macroeconomics cannot be reduced to microeconomics as the program of microfoundations suggests, the elements of macroeconomics could not exist without the substrate of microeconomic individuals.

5.4 TWO ARGUMENTS FOR THE REALITY OF MACROECONOMICS

So far we have argued that the ontological status of macroeconomic entities is problematic in the sense that, like other entities posited by scientific theories, they are not part of our common-sense ontic furniture. Furthermore, the nature of the relationship through which the elements of macroeconomics supervene on the elements of microeconomics precludes direct reduction of the macroeconomic to the microeconomic, even in principle. If macroeconomic entities exist, they cannot be said therefore to exist only derivatively, despite their supervenience on microeconomic entities. It remains to argue directly for the existence of macroeconomic entities.

The first argument is based on the argument from manipulability championed by Ian Hacking (1983, esp. pp. 22–24): "If you can spray them, then they are real." Hacking argues that convincing evidence of the reality of the electron is found in experiments aimed at detecting the existence of free quarks, in which niobium balls are charged by "spraying" them with electrons. The general point is that an entity defined by a scientific theory has real existence when procedures used to manipulate parts of the world outside the domain of the theory are best understood as procedures in which the entity is an instrument or tool. This argument is similar to the "no-miracles" argument for the reality of scientific entities. The best explanation of why theories are predictively successful, including successfully predicting the consequences of using them to design experimental or engineering manipulations of the world is that the entities posited by them in fact exist - anything else would be an inexplicable miracle.

It is common to denigrate the empirical success of economics (see, e.g., Rosenberg 1992, pp. 18, 56, 112, 238, passim). It is true that economics does not have the precision of physics or chemistry, although it is less clearly inferior to meteorology, geology, climatology, parts of biology, and cosmology - to name just a few of the less exact, but nevertheless scientific disciplines. The reputation of economics for predicting poorly arises partly because people seek unconditional forecasts ("what will happen tomorrow?") while economic theories typically predict only conditionally ("tomorrow X will happen if Y happens"). Quantified economic relations are at best locally stable: The precise estimate of the price elasticity of demand for Volvos changes with changes in the range of alternative brands and models, with changes in the proportion of academics to the total population, and with changes in other background conditions. Nevertheless, qualitatively stable relations are well established; for example, demand curves slope down (i.e., when the price of Volvos rises, sales of Volvos fall). And there is often enough local stability that useful quantitative assessments are possible. Can irreducible macroeconomic aggregates be manipulated as well?

The answer seems to be clearly yes. Consider the following example. Almost every macroeconomic theory predicts that sufficiently large expansions of government expenditure will change (probably increase) nominal GDP and the general level of prices. Different theories differ in their precise understanding of the mechanisms. Similarly, no macroeconomic theory disputes the ability of the Federal Reserve to use its ability to supply or remove reserves from the banking system to set the level of the federal funds rate (the rate at which one commercial bank borrows from another overnight). The empirical evidence in support of these propositions is also overwhelming. Now consider two irreducibly macroeconomic aggregate entities: the real rate of interest (i.e., the market rate of interest less the percentage change in the aggregate price level, \( \Delta p \)) and the yield curve (an aggregate relation portrayed as a graph of market interest rates against time to maturity of the associated bonds).

Both the real rate of interest and the yield curve are synthetic aggregates, and both are entities with causal powers in some economic theories. Every macroeconomic theory that I know predicts that actions that increase the general price level or the federal funds rate will shift the yield curve upwards in the short run. And, at least if the changes are unanticipated, increases in the general price level will reduce the level of the real interest rate. The empirical evidence for these effects is overwhelming, and indeed are easily confirmed by anyone willing to read the Wall Street Journal regularly for a month. Just like the electron, some macroeconomic aggregates can not only be controlled, but can be used to manipulate other macroeconomic aggregates.

39 The caveat "almost" and the ambiguity over the direction both hinge on the financial market. If the interest elasticity of money demand were zero (empirically a false supposition), there would be no change in prices. If the increase in the demand for money induced by an increase in government bonds financing an increase in government expenditure were large enough (again unlikely), the price level could fall.
Applied to causality, Hacking's and other such arguments from manipulability face the same charge of circularity leveled at analyses of causality that refer to direct control (Chapter 2, Section 3). The defense is also the same. The appeal to manipulability is not intended to provide a reduction of causality or reality to something else. Manipulability provides evidence for the reality of entities (physical or macroeconomic) or of causal relations only in a context of contingently indubitable knowledge. The argument is an abductive or bootstrap argument, the conclusions of which are corrigible and fallible if one is willing to challenge maintained assumptions about the context.

The second argument is related to the first. Leszek Nowak (1980) and others have argued that the principal method of constructing scientific theories is idealization. Nowak's (1980, p. 29) paradigm idealization statement is:

\[
\text{If } G(x) \text{ and } p_1(x) = 0 \text{ and } \ldots \text{ and } p_{k-1}(x) = 0 \text{ and } p_k(x) = 0, \\
\text{then } F(x) = f_k(H_1(x), \ldots, H_n(x)),
\]  

(5.12)

where \( G(x) \) is the complete theory, \( F(x) \) is the idealized theory, \( H_i \) \((i = 1, \ldots, n)\) denote primary factors, and the \( p_j \) \((j = 1, \ldots, k)\) denote secondary factors. An idealized theory is one that picks out the primary factors by setting the secondary factors to extreme values: zero or \( \pm \infty \), represented here, without loss of generality, as \( p_j = 0 \).

Were \( G(x) \) a known and exhaustively complete theory of the phenomenon within its explanatory range such that one could accurately specify each of the secondary factors that were set aside, then the distinction between primary and secondary factors would in fact be unclear, because our complete knowledge of \( G(x) \) would allow us for example to replace (5.12) with

\[
\text{If } G(x) \text{ and } H_1(x) = 0 \text{ and } \ldots \text{ and } H_{k-1}(x) = 0 \text{ and } H_k(x) = 0, \\
\text{then } F(x) = f_k(p_1(x), \ldots, p_k(x)).
\]  

(5.13)

In the case of either (5.12) or (5.13), releasing the idealizing conditions \((p_k(x) = 0 \text{ or } H_k(x) = 0)\) allows us to recover the complete theory, \( G(x) \). Idealization has been reduced to a fancy name for an arbitrary selection of ceteris paribus conditions or to a formal nesting relationship for theories.

Hoover (1994b) argues that in an empirical context, the method of idealization has power only if we recognize that not all of the idealizing conditions can be explicitly stated. The claim to distinguish between primary and secondary factors is then a claim that the primary factors are the essence of the matter. Idealized theories thus aim to identify, isolate, and relate the real essences or causally effective capacities of eco-

nomic reality. The success of such an idealized theory then amounts to an ontological claim for its primary factors.

That Keynesian macroeconomics could be cast as an idealization that employs macroeconomic aggregates essentially is beyond doubt. The major competitor to Keynesian macroeconomics today, new classical macroeconomics, trades on an explicitly microfoundational approach. There is, however, less here than meets the eye. Currently, the most popular new classical macroeconomic theory is embodied in the real-business-cycle model (see Hartley, Hoover, and Salyer 1997; 1998, ch. 1). The proponents of this representative-agent model would like it to be interpreted as an idealization from a complete Walrasian general equilibrium model of the economy in which distributional issues are idealized out of the model so that what remains is a one-agent, one-good, one-price representation of the economy. This would work if the analogue for \( G(x) \) in Nowak's schema, the complete Walrasian model (i.e., the “fantasy” model of Chapter 4, Section 4), were both true and known in detail. At least the second condition is false, which undermines the evidential basis for the first condition.

Empirically, far from isolating a microeconomic core, real-business-cycle models, as with other representative-agent models, use macroeconomic aggregates for their testing and estimation. Thus, to the degree that such models are successful in explaining empirical phenomena, it is because they capture the relationships among aggregates, not among individuals. They point to the ontological centrality of macroeconomic, not microeconomic, entities. The appeal to the methods of microeconomics does not in this case amount to the successful implementation of the program of microfoundations, for they are but the simulacrum of microeconomics. The relationship between models that are macroeconomic in form and their macroeconomic empirical implementation is metaphorical. The nature of metaphorical connection deserves further exploration. It is enough for the present purpose to understand that, at

---

20 Cartwright (1983, 1989) argues for realism with respect to causal capacities, but for an instrumentalist interpretation of scientific laws. Laws are either literally false (“the laws of physics lie”) – to quote the title of Cartwright's (1983) earlier book – or are merely phenomenal – i.e., atheoretic regularities. Hoover (1994b) argues that if idealized models represent essences, then phenomenal laws are necessary bridges to take the place of those secondary factors that cannot in fact be identified explicitly. Mäki (1992) argues that Nowak conflates idealization with isolation, which comprises idealization, omission, and other techniques. To apply Mäki's account we would have to say that the omission of secondary factors amounts to a claim that retained primary factors are the essence of the matter.

21 Hartley et al. (1997, 1998) provide a fuller argument for the ways in which the real-business-cycle models fails to provide genuine microfoundations.
5.5 MACROECONOMIC STRUCTURES

The structural account of causality presupposes that there are in fact real entities that may be related in causal structures. We are all familiar with such structures from Hume’s examples of billiard balls on the table to the machines, chemical processes, and living things that we encounter in our day-to-day lives. The argument of this chapter is that macroeconomics aggregates provide the raw materials for causal structures analogous to these familiar ones. Maki (1996, sec. 6) argues that the no-miracles argument and other arguments from manipulability cannot be applied successfully to economics, even if they apply to physical sciences. (There may, of course, be other arguments — and Maki supplies some — for existential beliefs about economic entities.) The conclusion of this chapter is that these reservations cannot be applied generally to economics: Macroeconomics shares characteristics with physical sciences that microeconomics may not. In consequence, while there may be differences in detail about the application of the structural account of causality to macroeconomics, there is no objection in principle. The differences in detail between macroeconomics and the physical and biological sciences are considered next.

Causality and Macroeconomic Constraints

Effect, n. The second of two phenomena which always appear together in the same order. The first, called a Cause, is said to generate the other — which is no more sensible than it would be for one who has never seen a dog except in pursuit of a rabbit to declare the rabbit the cause of the dog.

— Ambrose Bierce, The Devil’s Dictionary

In Chapter 5, I argued that macroeconomics is a suitable subject for a structural causal account. In this chapter, I ask, what constraints does the nature of macroeconomics place on that account?1

6.1 THE NATURE OF MACROECONOMICS

As we saw in Chapter 5, macroeconomics deals in aggregates. These aggregates are composed of the behaviors of individuals. Consumption as reported in the national income accounts, for example, is just the summation of the purchases of a nation’s citizens. It is tempting then to see economic agents as human molecules and the relations postulated in macroeconomic theory or measured in macroeconomic metrics as the analogues of the ideal gas laws or other macrophysical relations.2 But there is a crucial and obvious difference: Molecules do not make choices, people do; and they do so with reference not just to the immediate past and their immediate surroundings, but also with reference to future goals and to global or macro relations (e.g., people use aggregate price level to calculate their real wages in striking wage bargains; firms

---

1 This chapter is a substantially revised and expanded version of Hoover (1993), “Causality and Temporal Order in Macroeconomics or Why Even Economists Don’t Know How to Get Causes from Probabilities,” British Journal for the Philosophy of Science 44 (4), 693–710, by permission of Oxford University Press.

2 See Nelson (1992), Koopmans (in Hendry and Morgan 1995, p. 515), using the analogy of the gas laws, observes that science is possible without microfoundations, but suggests that economics has an advantage over physics, because we are the “molecules” of the economy and we know molecular behavior from direct acquaintance.
Macroeconomics and Methodology

Christopher A. Sims

This essay begins with a sketch of some ways I find it useful to think about science and its uses. Following that, the essay applies the framework it has sketched to discussion of several aspects of the recent history of macroeconomics. It considers skeptically the effort by some economists in the real business cycle school to define a quantitative methodology that stands in opposition to, or at least ignores, econometrics "in the modern (narrow) sense of the term." It connects this effort to the concurrent tendency across much of social science for scholars to question the value of statistical rigor and increasingly see their disciplines as searches for persuasive arguments rather than as searches for objective truth. The essay points to lines of substantive progress in macroeconomics that apparently flout the methodological prescriptions of the real business cycle school purists, yet are producing advances in understanding at least as important as what purist research has in fact achieved.

Science as Data Reduction

Advances in the natural sciences are discoveries of ways to compress data concerning the natural world—both data that already exists and potential data—with minimal loss of information. For example, Tycho Brahe accumulated large amounts of reliable data on the movements of the planets. Kepler observed that they are all on elliptical orbits with the sun at a focus, thereby accomplishing a sharp data compression.\(^1\) Newton found the inverse-square law, allowing still further compression\(^2\) and also allowing the same formula to organize existing data and predict new experimental or practical data in areas remote from the study of planetary motion.

Economics aims to accomplish the same sort of thing in relation to data on the economy, but is less successful. Whatever theory economists use to characterize data, the actual data always contain substantial variation that is not captured in the theory. The quality of the theory’s characterization of the data tends to deteriorate as we extend it to data remote in time, location or circumstances from the data from which the theory was initially developed.

This view, treating science as data reduction, may sound oversimplified, but it is in fact a flexible metaphor that should not be controversial. The contentious issues should concern what "data" are to be characterized and what constitutes a "compression."

It was once common for economists to think of the scientific enterprise as formulating testable hypotheses and confronting them with data. True hypotheses would survive the tests, while false ones would be eliminated. The science-as-data-compression view lets us see the limits of this hypothesis-testing view. The latter is dependent on the idea that there are true and false theories, when in fact the degree to which theories succeed in reducing data can be a continuum. The theory that planetary orbits are ellipses is only approximate if measurements are made carefully enough. It does not seem helpful to say that therefore it is false and should be rejected. Furthermore, "theories" can be so complex that they do not actually allow important data reduction, even though a naive hypothesis-testing approach might accept them as "true." More commonly, theories can differ less in whether they pass tests of match with the data than in the degree to which the theories are themselves simple. Planetary motions could be predicted quite accurately before Kepler; Kepler nonetheless had a better theory.

A good theory must not only display order in the data (which is the same thing as compressing it), it must do so in a way that is convincing and understandable to the target audience for the theory. But this does not mean that a successful scientific theory is understandable by many people. In fact, the most successful scientific theories are fully understood by very few people. They are successful because of institutions and conventions that support the recognition of specialized expertise and its perpetuation by rigorous training.

So though an effective theory must be persuasive, its persuasiveness cannot be determined entirely by examining the theory itself. One has to look also at who the accepted experts are, what kinds of arguments they are trained to understand and

---

\(^1\) Kepler's theory allowed data on the position of a planet that before required four coordinates (three spatial, one temporal) for each of \(N\) observed points to be reproduced with very high accuracy (from an economist's point of view, at least) with two coordinates (arc-length along the ellipse, time) for each of the \(N\) data points, plus the five numbers necessary to characterize the ellipse in 3 space.

\(^2\) Newton's theory allowed nearly the same accuracy with a single coordinate, time, together with the location and velocity vectors of the planet relative to the sun at some base time.
approve. And it is part of the continuing task of the discipline to assess what arguments its members ought to be trained to understand and approve.

Priesthoods and guilds—organizations of people with acknowledged expertise, training programs and hierarchical structure—are the imperfect social mechanisms by which bodies of knowledge are perpetuated. Modern science and economics are special cases. In understanding methodological disputes, it helps to bear in mind that the discussion is part of the workings of such an institution.

**Limits of the Analogy Between Economics and Physical Sciences**

Most natural sciences give a much less important role to probability-based formal inference than does economics. Since economics seems closer to natural sciences than are the other social sciences, in that economics makes more use of mathematically sophisticated theory and has more abundant data, why should it not also be less in need of statistical methodology? Examining the differences among sciences in a little more detail, we can see that probability-based inference is unavoidable in economics and that in this economics resembles related sciences, whether social or natural.

Economists can do very little experimentation to produce crucial data. This is particularly true of macroeconomics. Important policy questions demand opinions from economic experts from month to month, regardless of whether professional consensus has emerged on the questions. As a result, economists normally find themselves considering many theories and models with legitimate claims to matching the data and predicting the effects of policy. We have to deliver recommendations or accurate description of the nature of the uncertainty about the consequences of alternative policies, despite the lack of a single accepted theory. Because macroeconomists often favor one policy or another based on their own interests, or prefer economic advice that pretends to certainty, there is an incentive for economists to become contending advocates of theories, rather than cool assessors of the state of knowledge.

There are natural sciences that share some of these characteristics. Astronomers can’t do experiments, but they have more data than we do. Cosmology is short of relevant data and has contending theories, but is not pressed into service on policy decisions. Epidemiology is policy relevant and has limits on experimentation, but some kinds of experimentation are open to it—particularly use of animal models. Atmospheric science has limited experimental capacity, but in weather forecasting has more data than we do and less demand to predict the effects of policy. In modeling the effects of pollution and global warming, though, atmospheric science begins to be close to economics, with competing models that give different policy-relevant answers. But in this area atmospheric science does not have methodological lessons to teach us; I would say if anything the reverse is true.

Axiomatic arguments can produce the conclusion that anyone making decisions under uncertainty must act as if that agent has a probability distribution over the uncertainty, updating the probability distribution by Bayes’ rule as new evidence accumulates. People making decisions whose results depend on which of a set of scientific theories is correct should therefore be interested in probabilistic characterizations of the state of evidence. Yet in most physical sciences, such probabilistic characterizations of evidence are rare. Scientists understand the concept of standard error, but it seldom plays a central role in their discussion of results. In experimental sciences, this is due to the possibility of constructing an experiment in such a way, or continuing it to such a length, that standard errors of measurement are negligible. When this is possible, it certainly makes sense to do it.

In nonexperimental sciences with a great deal of data, like some branches of astronomy or atmospheric science, data may be plentiful but not suited to resolve some important outstanding theoretical issue. An interesting example is the narrative in Lindzen (1990, sec. 9.1) of the development of the theory of atmospheric tides—diurnal variations of barometric pressure. For a long time in this field, theory and data collection leapfrogged each other, with theory postulating mechanisms on which little data were available, the data becoming available and contradicting the theory and new theory then emerging. Because the amount of data was large and it was error ridden, something like what economists call reduced-form modeling went on continually in order to extract patterns from the noisy data. Even at the time Lindzen wrote, the best theory could not account for important features of the data. The gaps were well documented, and Lindzen’s narrative closes with suggestions for how they might be accounted for. There is no formal statistical comparison of models in the narrative, but also no account of any use of the models in decision making. If they had to be used to extrapolate the effects of interventions (pollution regulations, say) on atmospheric tides, and if the consequences were important, there would be no way to avoid making assumptions on, or even explicitly modeling, the variation the theories could not account for: it would have to be treated as random error.

In clinical medicine and epidemiology, statistical assessment of evidence is as pervasive as it is in economics. A treatment for a disease is a kind of theory, and when one is compared to another in a clinical trial, the comparison is nearly always statistical. If clinical trials were cheap, and if there were not ethical problems, they could be run at such a scale that, as in experimental science, the uncertainty in the results would become negligible. In fact, though, trials are expensive and patients cannot be given apparently worse treatments once the better therapy has acquired

---

3 This is of course an oversimplification, for rhetorical effect. For a more nuanced discussion of how modern science relates to and emerged from priesthoods and guilds, see Ben-David (1971).

4 See, for example, the first two chapters of Ferguson (1967) or chapters 2 and 6 of Robert (1994).

For an extended analysis of why some natural sciences in practice consider only "objective" probabilistics and make little formal use of probabilistic inference despite the validity of the axiomatic foundations of Bayesian decision theory, see Burks (1977).
a high probability, even though near certainty is not yet available. Epidemiology therefore often must work with nonexperimental data that produce difficulties in causal interpretation much like those facing economists. The debate over the evidence linking smoking and cancer has strong parallels with debates over macroeconomic policy issues, and it was inevitably statistical. Biological experiments not involving human subjects were possible in that case, though, and for macroeconomic policy questions, there is seldom anything comparable.

In other social sciences, there has recently been a reaction against formal statistical methodology. Many sociologists, for example, argue that insistence on quantitative evidence and formal statistical inference forces field research into a rigid pattern. Close observation and narrative description, like what has been common in anthropology, is advocated instead (for example, Bryman, 1988). A few economists also take this point of view. Bewley (1994) has undertaken work on wage and employment adjustment, using interviews with individual firms, that is close to the spirit of the new style in sociology.6

The coincident timing of attacks on statistical methods across disparate social sciences is probably not an accident. But the common element in these attacks is not a unified alternative approach—those advocating anthropological-style field research are criticizing statistical method from an almost precisely opposite point of view to that of purist real business cycle theorists. Instead, the popularity of the critiques probably arises from the excesses of enthusiasts of statistical methods. Pioneering statistical studies can be followed by mechanical imitations. Important formal inference techniques can be elaborated beyond what is useful for, or even at the expense of, their actual application. Indeed, insistence on elaborate statistical method can stifle the emergence of new ideas. Hence a turning away from statistical method can in some contexts play a constructive role. Anthropological method in field research in economics seems promising at a stage (as in the theory of price and wage rigidity in economics) where there are few theories, or only abstract and unconvincing theories, available and informal exploration in search of new patterns and generalizations is important. A focus on solving and calibrating models, rather than carefully fitting them to data, is reasonable at a stage where solving the models is by itself a major research task. When plausible theories have been advanced, though, and when decisions depend on evaluating them, more systematic collection and comparison of evidence cannot be avoided.

The pattern of variation across disciplines in the role of formal statistical inference reflects two principles. First, formal statistical inference is not important when the data are so abundant that they allow the available theories to be clearly ranked. This is typical of experimental natural sciences. Second, formal statistical inference is not necessary when there is no need to choose among competing theories among which the data do not distinguish decisively. But if the data do not make the choice of theory obvious, and if decisions depend on the choice, experts can report and discuss their conclusions reasonably only using notions of probability.

All the argument of this section is Bayesian—that it, it treats uncertainty across theories as no different conceptually from stochastic elements of the theories themselves. It is only from this perspective that the claim that decision making under uncertainty must be probabilistic can be supported. It is also only from this perspective that the typical inference problem in macroeconomics—where a single set of historically given time series must be used to sort out which of a variety of theoretical interpretations are likely—makes sense (Sims, 1982).7

The Rhetoric of Economics

Any economist who uses "rhetoric" in an article these days usually is reflecting at least implicitly the influence of McCloskey's (1983) antithetical methodological essay and subsequent related writing. This work in part reflected, in part instigated, an impatience with demands for technical rigor that emerged not only in the attitudes of the real business cycle school purists, but also in some macroeconomists of quite disparate substantive views. McCloskey wanted economists to recognize that in their professional writing, even at its most academic or scientific, they were engaged in persuasion. The essay identified and analyzed some of the rhetorical tools specific to economic argument, as well as the way economists use more universal tools. My own viewpoint as laid out above is consistent with McCloskey's in a number of respects. Both recognize that theories are not "true" or "false" and are not "tested" in single decisive confrontations with data. Both recognize that one can legitimately prefer one theory to another even when both fit the data to the same degree. Both reflect a suspicion of orthodoxy, hierarchy and methodological prescriptions as potential tools of priestly resistance to change. But McCloskey's enthusiasm for identifying rhetorical devices in economic argument and encouraging rhetorical skill among economists risks making us soft on quackery. For example, a simple theory is preferable to a complicated one if both accord equally well with the data, making the simpler one a more thorough data compression. Thus I agree with McCloskey that a naïve hypothesis-testing model of how theories are evaluated is a mistake. But a simple theory may gain adherents for

---

6 Another source of skepticism about econometrics may be the side effects of the Lucas critique of econometric policy evaluation. Nothing in the explicit logic of that critique suggests that probabilistic inference is in itself invalid or problematic. It criticizes a particular way of modeling macroeconomic policy interventions in a particular class of models. But the crude summary of its message---"econometric models are useless for policy evaluation"---no doubt contributed to the broader tendency of economists to question econometric method. In Sims (1987), I argue that the original formulation of the Lucas critique is itself logically flawed.

7 It should be noted that this point of view implies a critical stance toward some recent developments in econometric theory, particularly the literature on hypothesis testing in the presence of possible nonstationarity and cointegration, and is in this respect aligned with real business cycle purists (Sims, 1980; Sims and Uhlig, 1991).
other reasons—it may appeal to people with less training, who want to believe that a theory accessible to them is correct, or the evidence of its poorer fit may not be understandable without requiring rare technical skills; or the simple theory may fit the political or aesthetic tastes of many people. Convincing people that a simple theory is better than a more complicated one by appeal to something like these latter sources of support can be rhetorically effective in that it persuades people, and it may be done with admirable skill. But it is bad economics. Indeed, while I agree with McCloskey that recognizing rhetorical devices in economic discourse and analyzing their effectiveness is worthwhile, my first reaction on recognizing a persuasive type of argument is not enthusiasm, but wariness.

Economics is not physics. Science in general does not consist of formulating theories, testing them against data and accepting or rejecting them. But we can recognize these points without losing sight of the qualitative difference between modern science and classical or medieval natural philosophy: modern science has successfully created agreement that in scientific discourse certain types of apparently persuasive arguments are not legitimate. The only kind of argument that modern science treats as legitimate concerns the match of theory to data generated by experiment and observation. This means that sometimes badly written, difficult papers presenting theories that are aesthetically, politically or religiously displeasing are more persuasive to scientists than clearly written, easily understood papers that present theories that many people find inherently attractive. The fact that economics is not physics does not mean that we should not aim to apply the same fundamental standards for what constitutes legitimate argument; we can insist that the ultimate criterion for judging economic ideas is the degree to which they help us order and summarize data, that it is not legitimate to try to protect attractive theories from the data.

We can insist on such standards, but it is not at all inevitable that we will do so. Because economics, like other social sciences, does not achieve the clean successes and consensuses of the natural sciences, there can be disagreement not only about which theories are best, but about which modes of argument are legitimate. The standard that theories need to confront data, not be protected from it, is itself in constant need of defense, and its implications are regularly in dispute among economists who believe themselves committed to it. While McCloskey’s argument and analysis can in part be seen as a useful part of this process of defending scientific standards in economics, some of the influence of McCloskey’s writing has been malign.

Though it does not show up often explicitly in written literature, I encounter with increasing frequency in one-on-one professional argument an attitude I think of as rhetorical cynicism. A bit of it appears in the original McCloskey essay. For example, McCloskey (p. 489) cites with admiration the Friedman and Schwartz _Monetary History_ in a paragraph that ends with, “what was telling in the debate was the sheer bulk of the book—the richness and intelligence of its arguments, however irrelevant most of the arguments were to the main point.” It was perhaps not as clear in 1983 when McCloskey wrote as it is now that monetarism was on its way to the same macroeconomic limbo as Keynesianism, but even so, did McCloskey mean to suggest that it is good that economists were persuaded by irrelevant arguments? Or that we should admire rhetorical techniques that succeed in persuading the profession even if we ourselves can recognize that they should not persuade? I think many economists now see themselves as experts in persuasion as much as experts in substantive knowledge. They are willing to use arguments they know are flawed without explaining the flaws or to cite evidence they know could be shown to be misleading, for the sake of rhetorical effectiveness. There have always been economists who became, sincerely or cynically, uncritical apologists for particular viewpoints. The recent phenomenon is somewhat different. Economists seem to be telling themselves a story like this: a bold article, free of technical detail beyond a respectable minimum, is more likely to be cited frequently than a more cautious one that carefully defines the limits of its results. Leaving it to someone else to write carping technical critiques generates more citations and is not irresponsible if one thinks of one’s role as like that of a lawyer in adversarial court proceedings.

As is probably apparent, my own opinion is that whatever the value of viewing economics as rhetoric, that view of economics should remain secondary, with the view of economics as science, in the sense that it is an enterprise that holds theory accountable to data, remaining primary. It then follows that if economists are to communicate about the central questions of the discipline, they will need the language of statistical inference.

### The Real Business Cycle School

One way to characterize macroeconomics is as that branch of economics that makes drastic simplifications for the sake of studying phenomena—determination of the price level, the business cycle, economic growth—that inherently require analysis of general equilibrium. It is therefore natural and promising that macroeconomists, as computational power expands, are exploring methods for using previously intractable dynamic, stochastic, general equilibrium (DSGE) models. This phase of research, in which people examine which kinds of models are manageable and interesting and sharpen methods of numerical analysis, shares some characteristics with "normal science" as Kuhn describes it: textbooks are written (Sargent’s (1987) *Dynamic Macroeconomic Theory*, Stokey, Lucas and Prescott’s (1989) *Recursive Methods in Economic Dynamics*), researchers pose and solve puzzles, and there is a general sense of powerful methods being extended to cover new areas of application.

This activity has critics in the profession. The models are still too stylized and too remote from fitting the data to provide reliable guides to policy. Since considerable intellectual energy is going in to exploring them nonetheless, economists with strong interests in current policy, or high rates of time discount, or a reluctance to invest in learning the newly fashionable analytic methods, are ready to argue
that research of this kind should not be supported or taken seriously. Not surprisingly, the people who go ahead into this area of work despite the arguments against it develop some intellectual armor against such attacks. Much of this armor is visible more in informal interactions than in published writing. It is therefore valuable to have the armor displayed, even if in a form more rigid than most economists working in the area would probably be comfortable in, in the Kydland and Prescott essay in this issue.

The argument seems to be that what dynamic, stochastic, general equilibrium models in economics are doing not only resembles Kuhn’s normal science, it is normal science. Macroeconomists are said to have available a “well-tested,” or “standard” theory. They do (computational) “experiments.” These experiments usually result in “established theory becoming stronger,” but occasionally discover an extension of the existing theory that is useful, and thereby “established theory” is “improved.”

But these analogies with established physical sciences are strained. The neoclassical stochastic growth model that Kydland and Prescott put forth as the foundation of dynamic, stochastic, general equilibrium modeling is legitimately labeled accepted theory in one limited sense. There is an interacting group of researchers working out the implications of models built on this base; within this group the theory is accepted as a working hypothesis. But even within this group there is no illusion that the theory is noncontroversial in the profession at large. Most in the group would not even assert confidently that it is clear that theory of this type will deliver on its promise, any more than did Keynesian simultaneous equations models or natural rate rational expectations models.

What Kydland and Prescott call computational experiments are computations, not experiments. In economics, unlike experimental sciences, we cannot create observations designed to resolve our uncertainties about theories; no amount of computation can change that.

Dynamic, stochastic, general equilibrium modeling has delivered little empirical payoff so far. Macroeconomists have developed a variety of approaches to compressing the time series data using only informal theoretical ideas. The business cycle stage charts of the early NBER business cycle analysts were among the first of these reduced form modeling approaches, but multivariate spectral analysis, distributed lag regression, turning point timing analysis, cross-correlation functions, distributed lag regression, principle component analysis, vector autoregression impulse response analysis and dynamic factor analysis have all seen use. Certain patterns turned up by these analyses have the status of stylized facts: for example, Okun’s law (the tendency of employment to lag output), the strong predictive value of interest rate innovations for output and prices, the smoothness of aggregate price and wage movements, the tendency of productivity to fluctuate procyclically, and the strong correlation of monetary aggregates with nominal income and their (Granger) causal priority to it. I think it is fair to say that most real business cycle research has ignored most of the known facts about the business cycle in ascertaining the match between DSGE models and the facts. Kydland and Prescott rightly point out that all theories (at least in macroeconomics) are false and that therefore it does not make sense to discard a theory if it fails to fit perfectly. But if a theory fits much worse than alternative theories, that is a strike against it. We may still be interested in a poorly fitting theory if the theory offers an especially dramatic data compression (that is, is very simple relative to the data it confronts) or if it is a type of theory that promises to fit better with further work. But there can be no argument for deliberately shying away from documenting the ways in which the theory does and does not match the data.

This issue is distinct from the question of whether we should employ formal methods of statistical inference. Here the issue is only whether the data is going to be confronted as it exists, in all its density. When Mark Watson (1993), eschewing “formal statistical inference,” used an extension of standard tools for examining fit of a time series model by frequency in a Fourier analysis, he allowed us to see that the neoclassical stochastic growth model at the core of the real business cycle approach is very far from accounting well even for what we think of as business cycle variation in output itself. Watson’s analysis does not imply that the real business cycle approach should be abandoned. It does suggest that frequency domain analysis or other standard methods of orthogonal decomposition of macroeconomic time series data (like vector autoregression impulse responses) ought to be a standard part of evaluating real business cycle models, with the aim of getting better fit than what Watson found.

Kydland and Prescott appear to argue strongly against using econometric tools in the modern sense of the term. In part this stems from their caricature of formal statistical inference as “statistical hypothesis testing” that will certainly reject any (necessarily false) theory when given enough data. Bayesian critiques of classical hypothesis testing have long made the same kind of point, without rejecting formal statistical inference. Kydland and Prescott also claim, “Searching within some parametric class of economies for the one that best fits a set of aggregate time series makes little sense, because it isn’t likely to answer an interesting question.” Yet they put forth as an interesting type of question, “How much the U.S. postwar economy would have fluctuated if technology shocks had been the only source of fluctuations?” Surely one approach to such a question would be to construct a parametric class of dynamic, stochastic, general-equilibrium models in which the parameter indexed the contribution of technology shocks to fluctuations and to examine the behavior of model fit as a function of this parameter. Of course if it might turn out that model fit was insensitive to the parameter— the model was weakly identified in this dimension—but it might be instead that some sources of impulse response could be ruled out as unlikely, because they implied a poor fit. Showing this would not amount to simply finding the fit-maximizing parameter value, of course, but instead to characterizing the shape of the likelihood. If Kydland and Prescott are objecting only to the idea of ending inference with the picking of the parameter that gives the best fit, they are taking the position of Bayesian or likelihood
principle-based inference.\* It seems likely, though, that they intend a broader objection to probabilistic inference, in which case they seem to contradict some of their own position on what are interesting research questions.

Kydland and Prescott do approve some probability-based inference. They argue that it is reasonable to look at the theoretical probability distribution that is implied by a model for a set of statistics and to compare this to the corresponding statistics computed from the actual data. But a stochastic model produces a distribution, not a statistic. How are we meant to "compare" a distribution to a statistic? What conclusions might we draw from such a comparison? In this paper there is little guidance as to how we should make or interpret such a comparison.\* It is perhaps not surprising in the light of Kydland and Prescott's inference aversion that there is little guidance, because these questions are the root out of which all of statistical inference grows.

We can guess what use of such a comparison Kydland and Prescott intend by looking at their discussion of what constitutes "well-tested theory." They consider the neoclassical growth framework a well-tested theory. They say that it gives us confidence in the theory that it implies that when a model economy is subjected to realistic shocks, "it should display business cycle fluctuations of a quantitative nature similar to those actually observed." Another way of putting this, apparently, is that stochastic models based on the neoclassical growth framework produce "normal-looking" fluctuations. For some purposes this kind of language may suffice, but when we need to consider which of two or more models or theories with different policy implications is more reliable, it does not take us very far to be told that we should be more confident in the one whose simulated data is more "normal-looking" or is of a "quantitative nature" more "similar" to the actual data. My view, for example, is that Watson (1993) showed us that the stochastic neoclassical growth model as usually applied in the real business cycle literature produces simulated data that is drastically dissimilar to the actual data. If Kydland, Prescott and I are to have a reasoned discussion about this point, we will have to start talking about what we mean by "similar" and about what alternative models or theories are the standard against which the match of this one to the data is to be judged. Then we will be engaged in statistical inference.

If hyperbolic claims for the research accomplishments of dynamic, stochastic, general equilibrium models and for immunity of such models to criticism for na"{i}ve econometric method were necessary to sustain the enthusiasm of the participants, Kydland and Prescott might be justified in a bit of forensic exaggeration. But the field is interesting enough to attract attention from researchers without this, and the air of dogma and ideological rigidity these claims attach to the field are an unnecessary burden on it.

\* The likelihood principle is an implication of a Bayesian approach to inference, but can be argued for on other grounds as well. See Berger and Wolpert (1988).

\* Actually, in an earlier version there was a little more guidance, and it sounded a lot like simple statistical inference.

---

**Progress in Quantitative Macroeconomics**

There is work that is based on dynamic, stochastic, general equilibrium models and that makes serious use of formal methods of probability-based inference. McGrattan, Rogerson and Wright (1993), for example, estimate a standard real business cycle model using maximum likelihood, which could be a first step toward comparing its fit to other types of DSGE models or naive reduced form models. Leeper and Sims (1994) fit a DSGE model that adds to what Kydland and Prescott lay out as the standard neoclassical model a fluctuating relative price of consumption and capital goods and an articulated monetary and fiscal policy sector. The Leeper-Sims model comes close to the fit of a first-order reduced form vector autoregression model on a set of three variables. This is apparently much better than the fit of the versions of the neoclassical model that appear throughout the real business cycle literature, though because the fit of those models is seldom examined in a careful, standardized way, it is difficult to be certain of this. In any case, it is clear that it is becoming quite feasible to produce likelihood surfaces and one-step-ahead prediction residuals for DSGE models, thus providing a basis for comparison of alternative models meant to explain the same data series. There is considerable interest in making such comparisons, and it is essential before the models can become the basis for quantitative policy analysis. It will be happening even if Kydland and Prescott cannot be persuaded to assist in the process.

Other streams of research in macroeconomics are making as much progress, from an empirical point of view, as real business cycle modeling has yet achieved. One is modern policy modeling in the tradition of simultaneous equations modeling. These models have had limited attention from academic macroeconomists because of the concentration of research interest on building equilibrium models. But dynamic, stochastic, general equilibrium models have not been produced at a scale, level of detail and fit that allows them to be used in the actual process of monetary and fiscal policy formation. The result is that the field of policy modeling has been left to economists close to the policy process outside of academia, together with a few academic economists with strong policy interests. The models have developed to include rational expectations and expanded to include international linkages. This has come at a cost, however. Because implementing rational expectations at the scale of these models has been seen as computationally difficult, the stochastic structure of the models has become more stylized and unbelievable than when I wrote about the then-existing models in Sims (1980b).

A good example of work on this line is Taylor (1993). The larger of the two models in that book uses consumption and investment functions loosely based on dynamic optimizing theories and incorporating expectations of future income explicitly. It has wage adjustments that are sluggish and depend on expectations of the future. It has international linkages that use expectation interest rate parity conditions. In these respects it is an advance beyond the state of the art in 1980. But because the optimizing theory is used only informally, the model contains some structural anomalies. Further, equations of the model are estimated one at a time,
in some cases using statistical methods that are incompatible with the model's stochastic specification. The work is nonetheless promising, because it maintains the simultaneous equation modeling tradition of dense contact with the data and because it seems on the verge of substantial further progress. Recent developments in computing power and standardization of solution methods for linear rational expectations models make it seem likely that Taylor could apply to the full-scale model in the latter part of his book the more internally consistent statistical methodology he uses on a smaller example model in the first chapter. For the same reason, he should be able to connect his model more completely to a theory of dynamic optimizing behavior. It would thus more closely approach the real business cycle school's level of internal consistency and intertemperability, retain or improve upon the standards of statistical fit of the original simultaneous equation modeling tradition and still preserve the scale and detail needed for application to policy analysis.

Another distinct stream of research, which may look more significant to me than it should because of my own involvement with it, uses weakly identified time series models to isolate the effects of monetary policy. This style of work, in contrast with standard simultaneous equation modeling, begins with careful multivariate time series modeling of the data, developing evidence on prominent regularities. Restrictions based on substantive economic reasoning are imposed only as necessary to interpret the data and always with an eye to avoiding distortion of the model's fit.

Work in this style began with Friedman and Schwartz, who challenged conventional Keynesian thinking by displaying evidence of the strong correlation between monetary aggregates and income and of timing relationships, both in time series data and in particular historical episodes, that suggested causality running from money to income. I showed that beyond correlation and timing, there was a one-way predictive relationship between money and income (Sims, 1972). This implied that money was causally prior to nominal income according to the same definition of a causal ordering that underlies putting the "causally prior" variable on the right-hand side of a regression.

From this point onward, most of this literature focused on using reduced-form models that summarized the data by showing how all variables in the system are predicted to change following a surprise change in any one variable. This in effect breaks the variation in the data into mutually uncorrelated pieces, helping to isolate the major regularities. The surprise change are called "innovations," and the predicted patterns of change are called "impulse responses." I showed in Sims (1980a), following work by Y. P. Mehta (1978), that short-term interest rate innovations absorbed most of the predictive power of money innovations for output when the interest rates were added to a multivariate system. This undermined the monetarist claim that exogenous disturbances to the money stock were generated by policy and a dominant source of business cycle fluctuations, but left open the question of how the interest rate innovations should be interpreted. Starting in the mid-'80s with work by Bernanke (1986), Blanchard and Watson (1986) and Sims (1986), the informal identification arguments that had been used in this literature were supplemented with formal ones, adaptations to this weakly identified context of the methods used in the simultaneous equations literature. Bernanke and Blinder (1992) argued for identifying federal funds rate innovations as policy shocks, supporting their time series analysis with institutional detail. I showed (1992) that the stylized facts that this literature was codifying and interpreting were stable across a range of economies outside the United States, but that in certain respects these facts did not accord well with the interpretation of interest rate innovations as policy disturbances. I had already showed that the model anomaly—the "price puzzle" in which inflation rises after an apparent monetary contraction—would disappear under a particular set of identifying assumptions for U.S. data (Sims, 1986). Other researchers joined in trying to delineate the range of identifying assumptions consistent with the data (Christiano, Eichenbaum and Evans, 1994; Gordon and Leeper, 1995; Sims and Zha, 1995). Other recent work has extended the models to open economy environments with interesting results (Eichenbaum and Evans, 1993; Soyoung Kim, 1994; Soyoung Kim and Roubini, 1995; Guhan and Zha, 1994).

This literature has advanced knowledge in several ways. It has established firmly that most of the observed variation in monetary policy instruments—interest rates and monetary aggregates—cannot be treated as exogenously generated by random shifts in policy. (Incredibly, real business cycle school attempts to introduce monetary variables into dynamic, stochastic, general equilibrium models still often contradict this by now elementary business cycle fact.) It has given us a clearer quantitative picture of the size and dynamics of the effects of monetary policy. It has shown us that our knowledge about the size of these effects is still uncertain and that, because monetary contraction is rarely a spontaneous policy decision, the apparently eloquent fact that monetary contractions are followed by recession is hard to interpret.

The work suffers from some rhetorical handicaps. It cannot be understood without some familiarity with time series and simultaneous equations modeling ideas. Though the interpretations that it puts forward are influenced by rational expectations arguments, they find no need to make formal use of them. The work has proceeded incrementally, with no single paper or idea having dramatically changed people's thinking. The conclusions the work leads to tend more strongly to undermine naively confident interpretations of the data than to provide technical support for any simple policy position.

Some macroeconomists seem to have the impression that because this literature has not used dynamic optimization theory or rational expectations explicitly, and because it has found a version of simultaneous equations modeling essential, it is part of a tradition that we now know to be obsolete. In fact, this literature aims at minimizing reliance on ad hoc modeling conventions of both the traditional simultaneous equations style and the new dynamic, stochastic, general equilibrium style, in order to focus cleanly on the central issue of identifying the distinction between variation generated by deliberate policy action and variation generated by
disturbances outside of the policy process.\textsuperscript{40} The literature is probably now reaching a level of maturity at which it will pay to open up connections to DSGE models (by examining a range of more tightly restricted identifications) and to large-scale policy models (by considering larger and internationally linked versions of the models). Soyoung Kim (1994) is a step in the latter direction, and Sims (1989), Jinni Kim (1995) and Leeper and Sims (1994) are steps in the former direction.

Conclusion

Empirical macroeconomists are engaged in several promising lines of work. They are also engaged in making strained analogies between their work and the natural sciences and in classifying work in styles other than their own as outdated or mistaken based on its methods, not its substance. Since there is also a tendency in the profession to turn away from all technically demanding forms of theorizing and data analysis, it does not make sense for those of us who persist in such theorizing and data analysis to focus a lot of negative energy on each other. All the lines of work described in the previous section, including recent business cycle modeling, are potentially useful, and the lines of work show some tendency to converge. We would be better off if we spent more time in reading each others’ work and less in thinking up grand excuses for ignoring it.

References


Besse, James, and Joseph E. Stiglitz, "A Field Study on Downward Wage Rigidity," processed, Yale University, 1994.


Christiano, Lawrence, Martin Eichenbaum, and C. Evans, "The Effects of Monetary Policy


Kim, Soyoung, "Do Monetary Policy Shocks Matter in the G-7 Countries?" processed, Yale University, 1994, 81–117.


Chapter 13

Methods for Stationary Time-Series Data

13.1 Introduction

Time-series data have special features that often require the use of specialized econometric techniques. We have already dealt with some of these. For example, we discussed methods for dealing with serial correlation in Sections 7.6 through 7.9 and in Section 10.7, and we discussed heteroskedasticity and autocorrelation consistent (HAC) covariance matrices in Section 9.3. In the chapter and the next, we discuss a variety of techniques that are commonly used to model, and test hypotheses about, economic time series.

A first point concerns notation. In the time-series literature, it is usual to refer to a variable, series, or process by its typical element. For instance, one may speak of a variable $y_t$ or a set of variables $Y_t$, rather than defining a vector $y$ or a matrix $Y$. We will make free use of this convention in our discussion of time series.

The methods we will discuss fall naturally into two groups. Some of them are intended for use with stationary time series, and others are intended for use with nonstationary time series. We defined stationarity in Section 7.6. Recall that a random process for a time series $y_t$ is said to be covariance stationary if the unconditional expectation and variance of $y_t$, and the unconditional covariance between $y_t$ and $y_{t-j}$, for any lag $j$, are the same for all $t$. In this chapter, we restrict our attention to time series that are covariance stationary. Nonstationary time series and techniques for dealing with them will be discussed in Chapter 14.

Section 13.2 discusses stochastic processes that can be used to model the way in which the conditional mean of a single time series evolves over time. These are based on the autoregressive and moving-average processes that were introduced in Section 7.6. Section 13.3 discusses methods for estimating this sort of univariate time-series model. Section 13.4 then discusses single-equation dynamic regression models, which provide richer ways to model the relationships among time-series variables than do static regression models. Section 13.5 deals with seasonality and seasonal adjustment. Section 13.6 discusses autoregressive conditional heteroskedasticity, which provides a way to model the evolution of the conditional variance of a time series. Finally, Section 13.7 deals with vector autoregressions, which are a particularly simple and commonly used way to model multivariate time series.

13.2 Autoregressive and Moving-Average Processes

In Section 7.6, we introduced the concept of a stochastic process and briefly discussed autoregressive and moving-average processes. Our purpose there was to provide methods for modeling serial dependence in the error terms of a regression model. But these processes can also be used directly to model the dynamic evolution of an economic time series. When they are used for this purpose, it is common to add a constant term, because most economic time series do not have mean zero.

**Autoregressive Processes**

In Section 7.6, we discussed the $p^{th}$ order autoregressive, or AR($p$), process. If we add a constant term, such a process can be written, with slightly different notation, as

$$y_t = \gamma + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \ldots + \rho_p y_{t-p} + \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \sigma^2). \quad (13.01)$$

According to this specification, the $\varepsilon_t$ are homoskedastic and uncorrelated innovations. The process for $\varepsilon_t$ is often referred to as white noise, by a peculiar and metaphorical, of long standing, which cheerfully mixes a visual and auditory image. Throughout this chapter, the notation $\varepsilon_t$ refers to a white noise process with variance $\sigma^2$.

Note that the constant term $\gamma$ in equation (13.01) is not the unconditional mean of $y_t$. We assume throughout this chapter that the processes we consider are covariance stationary, in the sense that was given to that term in Section 7.6. This implies that $\mu = \mathbb{E}(y_t)$ does not depend on $t$. Thus, by equating the expectations of both sides of (13.01), we find that

$$\mu = \gamma + \mu \sum_{i=1}^p \rho_i.$$  

Solving this equation for $\mu$ yields the result that

$$\mu = \frac{\gamma}{1 - \sum_{i=1}^p \rho_i}. \quad (13.02)$$

If we define $u_t = y_t - \mu$, it is then easy to see that

$$u_t = \sum_{i=1}^p \rho_i u_{t-i} + \varepsilon_t, \quad (13.03)$$

which is exactly the definition (7.33) of an AR($p$) process given in Section 7.6. The lag operator notation we introduced in that section, equation (13.03)...
can also be written as

\[ u_t = \rho(L)u_t + \varepsilon_t, \quad \text{or as} \quad (1 - \rho(L))u_t = \varepsilon_t, \]

where the polynomial \( \rho \) is defined by equation (7.35), that is, \( \rho(z) = \rho_1 z + \rho_2 z^2 + \ldots + \rho_p z^p \). Similarly, the expression for the unconditional mean \( \mu \) in equation (13.02) can be written as \( \gamma/(1 - \rho(1)) \).

For \( u_t \), an AR(1) process, the \textbf{autocovariance matrix} was given in Section 7.6 by equation (7.32). The elements of this matrix are called the \textbf{autocovariances} of the AR(1) process. If the matrix is multiplied by a scalar chosen to make the diagonal elements equal to unity, the result is the \textbf{autocorrelation matrix}. For an AR(\( p \)) process, the autocovariances and the corresponding autocorrelations can be computed by using a set of equations called the \textbf{Yule-Walker equations}. We discuss these equations in detail for an AR(2) process; the generalization to the AR(\( p \)) case is straightforward but algebraically more complicated.

An AR(2) process without a constant term is defined by the equation

\[ u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \varepsilon_t. \]  

(13.04)

Let \( \nu_0 \) denote the unconditional variance of \( u_t \), and let \( \nu_{1} \) denote the covariance of \( u_t \) and \( u_{t-1} \), for \( i = 1, 2, \ldots \). Because the process is stationary, the \( \nu_i \), which are by definition the autocovariances of the AR(2) process, do not depend on \( t \). Multiplying equation (13.04) by \( u_t \) and taking expectations of both sides, we find that

\[ \nu_0 = \rho_1 \nu_1 + \rho_2 \nu_2 + \sigma^2. \]  

(13.05)

Because \( \nu_{t-1} \) and \( \nu_{t-2} \) are uncorrelated with the innovation \( \varepsilon_t \), the last term on the right-hand side here is \( E(u_t \varepsilon_t) = E(\varepsilon^2_t) = \sigma^2 \). Similarly, multiplying equation (13.04) by \( u_{t-1} \) and \( u_{t-2} \) and taking expectations, we find that

\[ \nu_1 = \rho_1 \nu_0 + \rho_2 \nu_1 \quad \text{and} \quad \nu_2 = \rho_1 \nu_1 + \rho_2 \nu_0. \]  

(13.06)

Equations (13.05) and (13.06) can be rewritten as a set of three simultaneous linear equations for \( \nu_0, \nu_1, \) and \( \nu_2 \):

\[ \begin{align*}
\nu_0 - \rho_1 \nu_1 - \rho_2 \nu_2 &= \sigma^2 \\
\rho_1 \nu_0 + (\rho_2 - 1)\nu_1 &= 0 \\
\rho_2 \nu_0 + \rho_1 \nu_1 - \nu_2 &= 0.
\end{align*} \]

(13.07)

These equations are the first three Yule-Walker equations for the AR(2) process. As readers are asked to show in Exercise 13.1, their solution is

\[ \begin{align*}
\nu_0 &= \frac{\sigma^2}{D} (1 - \rho_2), \quad \nu_1 = \frac{\sigma^2}{D} \rho_1, \quad \nu_2 = \frac{\sigma^2}{D} (\rho_1^2 + \rho_2 (1 - \rho_2)),
\end{align*} \]

(13.08)

where \( D \equiv (1 + \rho_2)(1 + \rho_1 - \rho_2)(1 - \rho_1 - \rho_2) \).

The result (13.08) makes it clear that \( \rho_1 \) and \( \rho_2 \) are not the autocorrelations of an AR(2) process. Recall that, for an AR(1) process, the same \( \rho \) that appears in the defining equation \( u_t = \rho u_{t-1} + \varepsilon_t \) is also the correlation of \( u_t \) and \( u_{t-1} \).

This simple result does not generalize to higher-order processes. Similarly, the autocovariances and autocorrelations of \( u_t \) and \( u_{t-i} \) for \( i > 2 \) have a more complicated form for AR processes of order greater than 1. They can, however, be determined readily enough by using the Yule-Walker equations. Thus, if we multiply both sides of equation (13.04) by \( u_{t-i} \) for any \( i \geq 2 \), and take expectations, we obtain the equation

\[ v_i = \rho_1 v_{i-1} + \rho_2 v_{i-2}. \]

(13.09)

Since \( \nu_0, \nu_1, \) and \( \nu_2 \) are given by equations (13.08), this equation allows us to solve recursively for any \( v_i \) with \( i > 2 \).

Necessary conditions for the stationarity of the AR(2) process follow directly from equations (13.08). The \( 3 \times 3 \) covariance matrix

\[ \begin{bmatrix}
\nu_0 & \nu_1 & \nu_2 \\
\nu_1 & \nu_0 & \nu_1 \\
\nu_2 & \nu_1 & \nu_0 \\
\end{bmatrix} \]

is positive definite only if the determinant

\[ D \equiv (1 + \rho_2)(1 + \rho_1 - \rho_2)(1 - \rho_1 - \rho_2) \]

of any three consecutive elements of an AR(2) process must be a positive definite matrix. Otherwise, the solution (13.08) to the first three Yule-Walker equations, based on the hypothesis of stationarity, would make no sense. The Yule-Walker equations are shown in Figure 13.1. The stationarity triangle is shown in Figure 13.1.
Moving-Average Processes

A $q$th order moving-average, or MA($q$), process with a constant term can be written as

$$y_t = \mu + \alpha_0 \varepsilon_t + \alpha_1 \varepsilon_{t-1} + \ldots + \alpha_q \varepsilon_{t-q},$$

(13.10)

where the $\varepsilon_t$ are white noise, and the coefficient $\alpha_0$ is generally normalized to 1 for purposes of identification. The expectation of the $y_t$ is readily seen to be $\mu$, and so we can write

$$u_t = y_t - \mu = \varepsilon_t + \sum_{j=1}^{q} \alpha_j \varepsilon_{t-j} = (1 + \alpha(L)) \varepsilon_t,$$

where the polynomial $\alpha$ is defined by $\alpha(z) = \sum_{j=1}^{q} \alpha_j z^j$.

The autocovariances of an MA process are much easier to calculate than those of an AR process. Since the $\varepsilon_t$ are white noise, and hence uncorrelated, the variance of the $u_t$ is seen to be

$$\text{Var}(u_t) = \text{E}(u_t^2) = \sigma^2 \left(1 + \sum_{j=1}^{q} \alpha_j^2\right).$$

Similarly, the $j$th order autocovariance is, for $j > 0$,

$$\text{E}(u_t u_{t-j}) = \begin{cases} 
\sigma^2 \left(\alpha_j + \sum_{i=1}^{q-j} \alpha_j \alpha_{i} \right) & \text{for } j < q, \\
\sigma^2 \alpha_j & \text{for } j = q, \text{ and} \\
0 & \text{for } j > q.
\end{cases}$$

(13.12)

Using (13.12) and (13.11), we can calculate the autocorrelation $\rho(j)$ between $y_t$ and $y_{t-j}$ for $j > 0$. We find that

$$\rho(j) = \frac{\alpha_j + \sum_{i=1}^{q-j} \alpha_j \alpha_{i}}{1 + \sum_{i=1}^{q} \alpha_i^2} \text{ for } j \leq q, \quad \rho(j) = 0 \text{ otherwise},$$

(13.13)

where it is understood that, for $j = q$, the numerator is just $\alpha_j$. The fact that all of the autocorrelations are equal to 0 for $j > q$ is sometimes convenient, but it suggests that $q$ may often have to be large if an MA($q$) model is to be satisfactory. Expression (13.13) also implies that $q$ must be large if an MA($q$) model is to display any autocorrelation coefficients that are big in absolute value. Recall from Section 7.6 that, for an MA(1) model, the largest possible absolute value of $\rho(1)$ is only 0.5.

1 The notation $\rho$ is unfortunately in common use both for the parameters of an AR process and for the autocorrelations of an AR or MA process. We shall distinguish between the parameter $\rho_1$ and the autocorrelation $\rho(j)$.
We will not consider general methods for inverting a polynomial in the lag operator; see Hamilton (1994) or Hayashi (2000), among many others. In this particular case, though, the solution turns out to be

$$(1 - \rho_1 L)^{-1} = 1 + \rho_1 L + \rho_1^2 L^2 + \ldots.$$  \hspace{1cm} (13.17)

To see this, note that $\rho_1 L$ times the right-hand side of equation (13.17) is the same series without the first term of 1. Thus, as required,

$$(1 - \rho_1 L)^{-1} - \rho_1 L (1 - \rho_1 L)^{-1} = (1 - \rho_1 L)(1 - \rho_1 L)^{-1} = 1.$$  \hspace{1cm} \text{(13.17)}

We can now use this result to solve equation (13.16). We find that

$$u_t = \varepsilon_t + \rho_1 \varepsilon_{t-1} + \rho_1^2 \varepsilon_{t-2} + \ldots.$$  \hspace{1cm} (13.18)

It is clear that (13.18) is a special case of the MA(\infty) process (13.14), with $\alpha_t = \rho_1^i$ for $i = 0, \ldots, \infty$. Square summability of the $\alpha_t$ is easy to check provided that $|\rho_1| < 1$.

In general, if we can write a stationary AR(p) process as

$$(1 - \rho(L))u_t = \varepsilon_t,$$  \hspace{1cm} (13.19)

where $\rho(L)$ is a polynomial of degree $p$ in the lag operator, then there exists an MA(\infty) process

$$u_t = (1 + \alpha(L))\varepsilon_t,$$  \hspace{1cm} (13.20)

where $\alpha(L)$ is an infinite series in $L$ such that $(1 - \rho(L))(1 + \alpha(L)) = 1$. This result provides an alternative way to the Yule-Walker equations to calculate the variance, autocovariances, and autocorrelations of an AR(p) process by using equations (13.11), (13.12), and (13.13), after we have solved for \alpha(L). However, these methods make use of the theory of functions of a complex variable, and so they are not elementary.

The close relationship between AR and MA processes goes both ways. If (13.20) is an MA(q) process that is invertible, then there exists a stationary AR(\infty) process of the form (13.19) with

$$(1 - \rho(L))(1 + \alpha(L)) = 1.$$  \hspace{1cm} \text{(13.20)}

The condition for a moving-average process to be invertible is formally the same as the condition for an autoregressive process to be stationary; see the discussion around equation (7.36). We require that all the roots of the polynomial equation $1 + \alpha(z) = 0$ must lie outside the unit circle. For an MA(1) process, the invertibility condition is simply that $|\alpha_1| < 1$.

\textbf{ARMA Processes}

If our objective is to model the evolution of a time series as parsimoniously as possible, it may well be desirable to employ a stochastic process that has both autoregressive and moving-average components. This is the autoregressive moving-average process, or ARMA process. In general, we can write an ARMA(p, q) process with nonzero mean as

$$(1 - \rho(L))y_t = \gamma + (1 + \alpha(L))\varepsilon_t,$$  \hspace{1cm} (13.21)

and a process with zero mean as

$$(1 - \rho(L))u_t = (1 + \alpha(L))\varepsilon_t,$$  \hspace{1cm} (13.22)

where $\rho(L)$ and $\alpha(L)$ are, respectively, a $p^{th}$ order and a $q^{th}$ order polynomial in the lag operator, neither of which includes a constant term. If the process is stationary, the expectation of $y_t$ given by (13.21) is $\mu = \gamma/(1 - \rho(1))$, just as for the AR(p) process (13.01). Provided the autoregressive part is stationary and the moving-average part is invertible, an ARMA(p, q) process can always be represented as either an MA(\infty) or an AR(\infty) process.

The most commonly encountered ARMA process is the ARMA(1,1) process, which, when there is no constant term, has the form

$$u_t = \rho_1 u_{t-1} + \varepsilon_t + \alpha_1 \varepsilon_{t-1}.$$  \hspace{1cm} (13.23)

This process has one autoregressive and one moving-average parameter.

The Yule-Walker method can be extended to compute the autocovariances of an ARMA process. We illustrate this for the ARMA(1,1) case and invite readers to generalize the procedure in Exercise 13.6. As before, we denote the $i^{th}$ autocovariance by $\gamma_i$ and let $E(u_t \varepsilon_{t-i}) = w_{i}$, for $i = 0, 1, \ldots$. Note that $E(u_t \varepsilon_t) = 0$ for all $s > t$. If we multiply (13.23) by $\varepsilon_t$ and take expectations, we see that $u_0 = \gamma_0$. If we then multiply (13.23) by $\varepsilon_{t-1}$ and repeat the process, we find that $u_1 = \rho_1 w_0 + \alpha_1 \gamma_2$, from which we conclude that $w_1 = \gamma_2 (\rho_1 + \alpha_1)$. Although we do not need them at present, we note that the $w_i$ for $i > 1$ can be found by multiplying (13.23) by $\varepsilon_{t-i}$, which gives us the recursion $w_i = \rho_1 w_{i-1}$, with solution $w_i = \alpha_1 \gamma_2 (\rho_1 + \alpha_1)$.

We proceed by the way in which the Yule-Walker equations are set up for an AR process. Multiplying equation (13.23) first by $u_t$ and then by $u_{t-1}$, and subsequently taking expectations, gives

$$u_0 = \rho_1 u_1 + w_0 + \alpha_1 w_1 = \rho_1 v_1 + \gamma_2 (1 + \alpha_1 \rho_1 + \alpha_1),$$

$$u_1 = \rho_1 u_0 + w_0 + \alpha_1 \gamma_2 = \rho_1 v_0 + \alpha_1 \gamma_2,$$

where $\gamma_2 = \alpha_1 \gamma_2 (\rho_1 + \alpha_1)$.

We have used the expressions for $w_0$ and $w_1$ given in the previous
paragraph. When these two equations are solved for \( v_0 \) and \( v_1 \), they yield

\[
v_0 = \sigma^2 \frac{1 + 2\rho_1 \alpha_1 + \alpha_1^2}{1 - \rho_1^2}, \quad \text{and} \quad v_1 = \sigma^2 \frac{\rho_1 + \rho_1^2 \alpha_1 + \rho_1 \alpha_1^2 + \alpha_1}{1 - \rho_1^2}. \tag{13.24}
\]

Finally, multiplying equation (13.23) by \( u_{t-i} \) for \( i > 1 \) and taking expectations gives \( u_i = \rho_1 u_{i-1} \), from which we conclude that

\[
v_i = \sigma^2 \rho_{1}^{i-1}(\rho_1 + \rho_1^2 \alpha_1 + \rho_1 \alpha_1^2 + \alpha_1) \tag{13.25}
\]

Equation (13.25) provides all the autocovariances of an ARMA(1, 1) process. Using it and the first of equations (13.24), we can derive the autocorrelations.

**Autocorrelation Functions**

As we have seen, the autocorrelation between \( u_i \) and \( u_{i-j} \) can be calculated theoretically for any known stationary ARMA process. The autocorrelation function, or ACF, expresses the autocorrelation as a function of the lag \( k \) for \( j = 1, 2, \ldots \). If we have a sample \( y_t \), \( t = 1, \ldots, n \), from an ARMA process of possibly unknown order, then the \( j^{th} \) order autocorrelation \( \rho(j) \) can be estimated by using the formula

\[
\hat{\rho}(j) = \frac{\text{Cov}(y_t, y_{t-j})}{\text{Var}(y_t)}, \tag{13.26}
\]

where

\[
\text{Cov}(y_t, y_{t-j}) = \frac{1}{n-j-1} \sum_{t=j+1}^{n} (y_t - \bar{y})(y_{t-j} - \bar{y}), \tag{13.27}
\]

and

\[
\text{Var}(y_t) = \frac{1}{n-1} \sum_{t=1}^{n} (y_t - \bar{y})^2. \tag{13.28}
\]

In equations (13.27) and (13.28), \( \bar{y} \) is the mean of the \( y_t \). Of course, (13.28) is just the special case of (13.27) in which \( j = 0 \). It may seem odd to divide by \( n-1 \) rather than by \( n-j-1 \) in (13.27). However, if we did not use the same denominator for every \( j \), the estimated autocorrelation matrix would not necessarily be positive definite. Because the denominator is the same, the factors of \( 1/(n-1) \) cancel in the formula (13.26).

The empirical ACF, or sample ACF, expresses the \( \hat{\rho}(j) \), defined in equation (13.26), as a function of the lag \( j \). Graphing the sample ACF provides a convenient way to see what the pattern of serial dependence in any observed time series looks like, and it may help to suggest what sort of stochastic process would provide a good way to model the data. For example, if the data were generated by an MA(1) process, we would expect that \( \hat{\rho}(1) \) would be an estimate of \( \alpha_1 \) and all the other \( \hat{\rho}(j) \) would be approximately equal to zero. If the data were generated by a MA(2) process with \( \alpha_2 > 0 \), we would expect that \( \hat{\rho}(1) \) would be an estimate of \( \alpha_1 \), that \( \hat{\rho}(2) \) would be relatively large, the next few \( \hat{\rho}(j) \) would be progressively smaller, and the ones for large \( j \) would be approximately equal to zero. A graph of the sample ACF is sometimes called a correlogram; see Exercise 13.15.

The partial autocorrelation function, or PACF, is another way to characterize the relationship between \( y_t \) and its lagged values. The partial autocorrelation coefficient of order \( j \) is defined as the true value of the coefficient \( \rho_j \) in the linear regression

\[
y_t = \gamma(j) + \rho_j^{(1)} y_{t-1} + \ldots + \rho_j^{(j)} y_{t-j} + \epsilon_t, \tag{13.29}
\]

or, equivalently, in the minimization problem

\[
\min_{\gamma(j), \rho_j} \sum_{i=1}^{J} \frac{E(y_t - \gamma(j) - \sum_{j=1}^{J} \rho_j^{(j)} y_{t-j})^2}. \tag{13.30}
\]

The superscript "(j)" appears on all the coefficients in regression (13.29) to make it plain that all the coefficients, not just the last one, are functions of \( j \), the number of lags. We can calculate the empirical PACF, or sample PACF, up to order \( J \) by running regression (13.29) for \( j = 1, \ldots, J \) and retaining only the estimate \( \rho_j^{(j)} \) for each \( j \). Just as a graph of the sample ACF may help to suggest what sort of stochastic process would provide a good way to model the data, so a graph of the sample PACF, interpreted properly, may do the same. For example, if the data were generated by an AR(2) process, we would expect the first two partial autocorrelations to be relatively large, and all the remaining ones to be insignificantly different from zero.

**13.3 Estimating AR, MA, and ARMA Models**

All of the time-series models that we have discussed so far are special cases of an ARMA\((p, q)\) model with a constant term, which can be written as

\[
y_t = \gamma + \sum_{i=1}^{p} \rho_i y_{t-i} + \epsilon_t + \sum_{j=1}^{q} \alpha_j \epsilon_{t-j}, \tag{13.31}
\]

where the \( \epsilon_t \) are assumed to be white noise. There are \( p+q+1 \) parameters to estimate in the model (13.31): the \( \rho_i \), for \( i = 1, \ldots, p \), the \( \alpha_j \), for \( j = 1, \ldots, q \), and \( \gamma \). Recall that \( \gamma \) is not the unconditional expectation of \( y_t \) unless all of the \( \rho_i \) are zero.

For our present purposes, it is perfectly convenient to work with models that allow \( y_t \) to depend on exogenous explanatory variables and are therefore even
more general than (13.31). Such models are sometimes referred to as ARMAX models. The 'X' indicates that \( y_t \) depends on a row vector \( X_t \) of exogenous variables as well as on its own lagged values. An ARMAX\((p,q)\) model takes the form
\[
y_t = X_t \beta + u_t, \quad u_t \sim \text{ARMA}(p,q), \quad E(u_t) = 0, \tag{13.32}
\]
where \( X_t \beta \) is the mean of \( y_t \) conditional on \( X_t \) but not conditional on lagged values of \( y_t \). The ARMA model (13.31) can evidently be recast in the form of the ARMAX model (13.32); see Exercise 13.13.

**Estimation of AR Models**

We have already studied a variety of ways of estimating the model (13.32) when \( u_t \) follows an AR(1) process. In Chapter 7, we discussed three estimation methods. The first was estimation by a nonlinear regression, in which the first observation is dropped from the sample. The second was estimation by feasible GLS, possibly iterated, in which the first observation can be taken into account. The third was estimation by the GNR that corresponds to the nonlinear regression with an extra artificial observation corresponding to the first observation. It turned out that estimation by iterated feasible GLS and by this extended artificial regression, both taking the first observation into account, yield the same estimates. Then, in Chapter 10, we discussed estimation by maximum likelihood, and, in Exercise 10.21, we showed how to extend the GNR by yet another artificial observation in such a way that it provides the ML estimates if convergence is achieved.

Similar estimation methods exist for models in which the error terms follow an AR\((p)\) process with \( p > 1 \). The easiest method is just to drop the first \( p \) observations and estimate the nonlinear regression model
\[
y_t = X_t \beta + \sum_{i=1}^{p} \rho_i (y_{t-i} - X_{t-i} \beta) + \epsilon_t
\]
by nonlinear least squares. If this is a pure time-series model for which \( X_t \beta = \beta \), then this is equivalent to OLS estimation of the model
\[
y_t = \gamma + \sum_{i=1}^{p} \rho_i y_{t-i} + \epsilon_t,
\]
where the relationship between \( \gamma \) and \( \beta \) is derived in Exercise 13.13. This approach is the simplest and most widely used for pure autoregressive model. It has the advantage that, although the \( \rho_i \) (but not their estimates) must satisfy the necessary condition for stationarity, the error terms \( u_t \) need not be stationary. This issue was mentioned in Section 7.8, in the context of the AR(1) model, where it was seen that the variance of the first error term \( u_t \) must satisfy a certain condition for \( u_t \) to be stationary.

**Maximum Likelihood Estimation**

If we are prepared to assume that \( u_t \) is indeed stationary, it is desirable not to lose the information in the first \( p \) observations. The most convenient way to achieve this goal is to use maximum likelihood under the assumption that the white noise process \( \epsilon_t \) is normal. In addition to using more information, maximum likelihood has the advantage that the estimates of the \( \rho_i \) are automatically constrained to satisfy the stationarity conditions.

For any ARMA\((p,q)\) process in the error terms \( u_t \), the assumption that the \( \epsilon_t \) are normally distributed implies that the \( u_t \) are normally distributed, and so the dependent variable \( y_t \), conditional on the explanatory variables. For any observed sample of size \( n \) from the ARMAX model (13.32), let \( y \) denote the \( n \)-vector of which the elements are \( y_1, \ldots, y_n \). The expectation of \( y \) conditional on the explanatory variables is \( X \beta \), where \( X \) is the \( n \times k \) matrix with typical row \( X_t \). Let \( \Omega \) denote the autocovariance matrix of the vector \( y \). This matrix can be written as
\[
\Omega = \begin{bmatrix}
v_0 & v_1 & v_2 & \cdots & v_{n-1} \\
v_1 & v_0 & v_1 & \cdots & v_{n-2} \\
v_2 & v_1 & v_0 & \cdots & v_{n-3} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
v_{n-1} & v_{n-2} & v_{n-3} & \cdots & v_0
\end{bmatrix}, \tag{13.33}
\]
where, as before, \( v_i \) is the stationary covariance of \( u_t \) and \( u_{t-i} \), and \( v_0 \) is the stationary variance of the \( u_t \). Then, using expression (12.122) for the multivariate normal density, we see that the log of the joint density of the observed sample is
\[
-\frac{1}{2} \log 2\pi - \frac{1}{2} \log |\Omega| - \frac{1}{2} (y - X \beta)^\top \Omega^{-1} (y - X \beta). \tag{13.34}
\]
In order to construct the loglikelihood function for the ARMAX model (13.32), the \( \rho_i \) must be expressed as functions of the parameters \( \rho_i \) and \( \alpha_j \) of the ARMA\((p,q)\) process that generates the error terms. Doing this allows us to replace \( \Omega \) in the log density (13.34) by a matrix function of these parameters. Unfortunately, a loglikelihood function in the form of (13.34) is difficult to work with, because of the presence of the \( n \times n \) matrix \( \Omega \). Most of the difficulty disappears if we can find an upper-triangular matrix \( \Psi \) such that \( \Psi \Psi^\top = \Omega^{-1} \), as was necessary when, in Section 7.8, we wished to estimate by feasible GLS a model like (13.32) with AR(1) errors. It then becomes possible to decompose expression (13.34) into a sum of contributions that are easier to work with than (13.34) itself.

If the errors are generated by an AR\((p)\) process, with no MA component, then such a matrix \( \Psi \) is relatively easy to find, as we will illustrate in a moment for the AR(2) case. However, if an MA component is present, matters are more difficult. Even for MA(1) errors, the algebra is quite complicated; see Hamilton (1994, Chapter 5) for a convincing demonstration of this fact. For
general ARMA($p, q$) processes, a technique called the Kalman filter can be used to evaluate the successive contributions to the loglikelihood for given parameter values, and can thus serve as the basis of an algorithm for maximizing the loglikelihood. This technique, to which Hamilton (1994, Chapter 13) provides an accessible introduction, is unfortunately beyond the scope of this book. A somewhat simpler approach, which solves for the parameters of $\Psi$ recursively, has been proposed by Galbraith and Zinde-Walsh (1992).

We now turn our attention to the case in which the errors follow an AR(2) process. In Section 7.8, we constructed a matrix $\Psi$ corresponding to the stationary covariance matrix of an AR(1) process by finding $n$ linear combinations of the error terms $u_t$ that were homoskedastic and serially uncorrelated. We perform a similar exercise for AR(2) errors here. This will show how to set about the necessary algebra for more general AR($p$) processes.

Errors generated by an AR(2) process satisfy equation (13.04). Therefore, for $t \geq 3$, we can solve for $\varepsilon_t$ to obtain

$$\varepsilon_t = u_t - \rho_1 u_{t-1} - \rho_2 u_{t-2}, \quad t = 3, \ldots, n.$$  \hspace{1cm} (13.35)

Under the normality assumption, the fact that the $\varepsilon_t$ are white noise means that they are mutually independent. Thus observations 3 through $n$ make contributions to the loglikelihood of the form

$$\ell_1(y^t, \beta, \rho_1, \rho_2, \sigma_\varepsilon) = -\frac{1}{2} \log 2\pi - \log \sigma_\varepsilon - \frac{1}{2\sigma_\varepsilon^2} \left( u_t(\beta) - \rho_1 u_{t-1}(\beta) - \rho_2 u_{t-2}(\beta) \right)^2,$$  \hspace{1cm} (13.36)

where $y^t$ is the vector that consists of $y_1$ through $y_t$, $u_t(\beta) = y_t - X_t \beta$, and $\sigma_\varepsilon^2$ is as usual the variance of the $\varepsilon_t$. The contribution (13.36) is analogous to the contribution (10.87) for the AR(1) case.

The variance of the first error term, $u_1$, is just the stationary variance $\sigma_0^2$ given by (13.08). We can therefore define $\varepsilon_1$ as $\sigma_\varepsilon u_1/\sqrt{\sigma_0^2}$, that is,

$$\varepsilon_1 = \left( \frac{D}{1 - \rho_2} \right)^{1/2} u_1,$$  \hspace{1cm} (13.37)

where $D$ was defined just after equations (13.08). By construction, $\varepsilon_1$ has the same variance $\sigma_\varepsilon^2$ as the $\varepsilon_t$ for $t \geq 3$. Since the $\varepsilon_t$ are innovations, it follows that, for $t > 1$, $\varepsilon_t$ is independent of $u_1$, and hence of $\varepsilon_1$. For the loglikelihood contribution from observation 1, we therefore take the log density of $\varepsilon_1$, plus a Jacobian term which is the log of the derivative of $\varepsilon_1$ with respect to $u_1$. The result is readily seen to be

$$\ell_1(y_1, \beta, \rho_1, \rho_2, \sigma_\varepsilon) = -\frac{1}{2} \log 2\pi - \log \sigma_\varepsilon - \frac{1}{2} \log \frac{D}{1 - \rho_2} - \frac{D}{2\sigma_\varepsilon^2 (1 - \rho_2)} u_1^2(\beta).$$  \hspace{1cm} (13.38)

13.3 Estimating AR, MA, and ARMA Models

Finding a suitable expression for $\varepsilon_2$ is a little trickier. What we seek is a linear combination of $u_1$ and $u_2$ that has variance $\sigma_\varepsilon^2$ and is independent of $u_1$. By construction, any such linear combination is independent of the $\varepsilon_t$ for $t > 2$.

A little algebra shows that the appropriate linear combination is

$$\sigma_\varepsilon \left( \frac{v_0}{v_0^2 - v_0^2} \right)^{1/2} \left( u_2 - \frac{v_1}{v_0} u_1 \right).$$

Use of the explicit expressions for $v_0$ and $v_1$ given in equations (13.08) then shows that

$$\varepsilon_2 = (1 - \rho_2^2)^{1/2} \left( u_2 - \frac{\rho_1}{1 - \rho_2} u_1 \right),$$  \hspace{1cm} (13.39)

as readers are invited to check in Exercise 13.9. The derivative of $\varepsilon_2$ with respect to $u_2$ is $(1 - \rho_2^2)^{1/2}$, and so the contribution to the loglikelihood from observation 2 can be written as

$$\ell_2(y^2, \beta, \rho_1, \rho_2, \sigma_\varepsilon) = -\frac{1}{2} \log 2\pi - \log \sigma_\varepsilon + \frac{1}{2} \log (1 - \rho_2^2) - \frac{1 - \rho_2^2}{2\sigma_\varepsilon^2} \left( u_2(\beta) - \frac{\rho_1}{1 - \rho_2} u_1(\beta) \right)^2.$$  \hspace{1cm} (13.40)

Summing the contributions (13.36), (13.38), and (13.40) gives the loglikelihood function for the entire sample. It may then be maximized with respect to $\beta, \rho_1, \rho_2$, and $\sigma_\varepsilon^2$ by standard numerical methods.

Exercise 13.10 asks readers to check that the $n \times n$ matrix $\Psi$ defined implicitly by the relation $\Psi \varepsilon^T = \varepsilon$, where the elements of $\varepsilon$ are defined by (13.35), (13.37), and (13.39), is indeed upper triangular and such that $\Psi \Psi^T$ is equal to $1/\sigma_\varepsilon^2$ times the inverse of the covariance matrix (13.33) for the $\varepsilon_t$ that correspond to an AR(2) process.

Estimation of MA and ARMA Models

Just why moving-average and ARMA models are more difficult to estimate than pure autoregressive models is apparent if we consider the MA(1) model

$$y_t = \mu + \varepsilon_t - \alpha_1 \varepsilon_{t-1},$$  \hspace{1cm} (13.41)

where for simplicity the only explanatory variable is a constant, and we have changed the sign of $\alpha_1$. For the first three observations, if we substitute recursively for $\varepsilon_{t-1}$, equation (13.41) can be written as

$$y_1 = \mu - \alpha_1 \varepsilon_0 + \varepsilon_1,$$

$$y_2 = (1 + \alpha_1) \mu - \alpha_1 y_1 - \alpha_1^2 \varepsilon_0 + \varepsilon_2,$$

$$y_3 = (1 + \alpha_1 + \alpha_1^2) \mu - \alpha_1 y_2 - \alpha_1^2 y_1 - \alpha_1^3 \varepsilon_0 + \varepsilon_3.$$
It is not difficult to see that, for arbitrary \( t \), this becomes

\[
y_t = \left( \sum_{s=0}^{t-1} \alpha_t^s \right) \mu - \sum_{s=1}^{t-1} \alpha_t^s y_{t-s} - \alpha_t^0 \varepsilon_0 + \varepsilon_t. \tag{13.42}
\]

Worse it not for the presence of the unobserved \( \varepsilon_0 \), equation (13.42) would be a nonlinear regression model, albeit a rather complicated one in which the form of the regression function depends explicitly on \( t \).

This fact can be used to develop tractable methods for estimating a model where the errors have a MA component without going to the trouble of setting up the complicated loglikelihood. The estimates are not equal to ML estimates, and are in general less efficient, although in some cases they are asymptotically equivalent. The simplest approach, which is sometimes rather misleadingly called conditional least squares, is just to assume that any unobserved pre-sample innovations, such as \( \varepsilon_0 \), are equal to 0, an assumption that is harmless asymptotically. A more sophisticated approach is to "backcast" the pre-sample innovations from initial estimates of the other parameters and then run the nonlinear regression (13.42) conditional on the backcasts, that is, the backward forecasts. Yet another approach is to treat the unobserved innovations as parameters to be estimated jointly by maximum likelihood with the parameters of the MA process and those of the regression function.

Alternative statistical packages use a number of different methods for estimating models with ARMA errors, and they may therefore yield different estimates; see Newbold, Agiakloglou, and Miller (1994) for a more detailed account. Moreover, even if they provide the same estimates, different packages may well provide different standard errors. In the case of ML estimation, for example, these may be based on the empirical Hessian estimator (10.42), the OPG estimator (10.44), or the sandwich estimator (10.45), among others. If the innovations are heteroskedastic, only the sandwich estimator is valid.

A more detailed discussion of standard methods for estimating AR, MA, and ARMA models is beyond the scope of this book. Detailed treatments may be found in Box, Jenkins, and Reinsel (1994, Chapter 7), Hamilton (1994, Chapter 5), and Fuller (1995, Chapter 8), among others.

### Indirect Inference

There is another approach to estimating ARMA models, which is unlikely to be used by statistical packages but is worthy of attention if the available sample is not too small. It is an application of the method of indirect inference, which was developed by Smith (1993) and Gouriéroux, Monfort, and Renault (1993). The idea is that, when a model is difficult to estimate, there may be an auxiliary model that is not too different from the model of interest but is much easier to estimate. For any such two models, there must exist so-called binding functions that relate the parameters of the model of interest to those of the auxiliary model. The idea of indirect inference is to estimate the parameters of interest from the parameter estimates of the auxiliary model by using the relationships given by the binding functions.

Because pure AR models are easy to estimate and can be used as auxiliary models, it is natural to use this approach with models that have an MA component. For simplicity, suppose the model of interest is the pure time-series MA(1) model (13.41), and the auxiliary model is the AR(1) model

\[
y_t = \gamma + \rho y_{t-1} + u_t, \tag{13.43}
\]

which we estimate by OLS to obtain estimates \( \hat{\gamma} \) and \( \hat{\rho} \). Let us define the elementary zero function \( u_t(\gamma, \rho) \) as \( y_t - \gamma - \rho y_{t-1} \). Then the estimating equations satisfied by \( \hat{\gamma} \) and \( \hat{\rho} \) are

\[
\sum_{t=2}^{n} u_t(\gamma, \rho) = 0 \quad \text{and} \quad \sum_{t=2}^{n} y_{t-1} u_t(\gamma, \rho) = 0. \tag{13.44}
\]

If \( y_t \) is indeed generated by (13.41) for particular values of \( \mu \) and \( \alpha_1 \), then we may define the pseudo-true values of the parameters \( \gamma \) and \( \rho \) of the auxiliary model (13.43) as those values for which the expectations of the left-hand sides of equations (13.44) are zero. These equations can thus be interpreted as correctly specified, albeit inefficient, estimating equations for the pseudo-true values. The theory of Section 9.5 then shows that \( \hat{\gamma} \) and \( \hat{\rho} \) are consistent for the pseudo-true values and asymptotically normal, with asymptotic covariance matrix given by a version of the sandwich matrix (9.67).

The pseudo-true values can be calculated as follows. Replacing \( y_t \) and \( y_{t-1} \) in the definition of \( u_t(\gamma, \rho) \) by the expressions given by (13.41), we see that

\[
u_t(\gamma, \rho) = (1 - \rho)u_t - \gamma + \varepsilon_t - (\alpha_1 + \rho)\varepsilon_{t-1} + \alpha_1 \rho \varepsilon_{t-2}. \tag{13.45}\]

The expectation of the right-hand side of this equation is just \((1 - \rho)\mu - \gamma\). Similarly, the expectation of \( y_{t-1} u_t(\gamma, \rho) \) can be seen to be

\[
\mu((1 - \rho)\mu - \gamma) - \sigma^2(\alpha_1 + \rho) - \sigma^2 \alpha_1 \rho.
\]

Equating these expectations to zero shows us that the pseudo-true values are

\[
\gamma = \frac{\mu(1 + \alpha_1 + \alpha_1^2)}{1 + \alpha_1^2} \quad \text{and} \quad \rho = \frac{-\alpha_1}{1 + \alpha_1^2}. \tag{13.46}
\]

In terms of the true parameters \( \mu \) and \( \alpha_1 \), Equations (13.46) express the binding functions that link the parameters of model (13.41) to those of the auxiliary model (13.43). The indirect estimates \( \hat{\mu} \) and \( \hat{\alpha_1} \) are obtained by solving these equations with \( \gamma \) and \( \rho \) replaced by \( \hat{\gamma} \) and \( \hat{\rho} \).
and $\hat{\beta}$. Note that, since the second equation of (13.46) is a quadratic equation for $\alpha_1$ in terms of $\rho$, there are in general two solutions for $\alpha_1$, which may be complex. See Exercise 13.11 for further elucidation of this point.

In order to estimate the covariance matrix of $\hat{\mu}$ and $\hat{\alpha}_1$, we must first estimate the covariance matrix of $\hat{\gamma}$ and $\hat{\beta}$. Let us define the $n \times 2$ matrix $Z$ as $[\gamma \ y_{-1}]$, that is, a matrix of which the first column is a vector of 1s and the second the vector of the $y_t$ lagged. Then, since the Jacobian of the zero functions $u_t(\gamma, \rho)$ is just $-Z$, it is easy to see that the covariance matrix (9.67) becomes

$$\lim_{n \to \infty} \frac{1}{n} (Z'Z)^{-1} Z' \Omega Z (Z'Z)^{-1},$$

where $\Omega$ is the covariance matrix of the error terms $u_t$, which are given by the $u_t(\gamma, \rho)$ evaluated at the pseudo-true values. If we drop the probability limit and the factor of $n^{-1}$ in expression (13.47) and replace $\Omega$ by a suitable estimate, we obtain an estimate of the covariance matrix of $\hat{\gamma}$ and $\hat{\beta}$. Instead of estimating $\Omega$ directly, it is convenient to employ a HAC estimator of the middle factor of expression (13.47). Since, as can be seen from equation (13.45), the $u_t$ have nonzero autocovariances only up to order 2, it is natural in this case to use the Hansen-White estimator (9.37) with lag truncation parameter set equal to 2. Finally, an estimate of the covariance matrix of $\hat{\mu}$ and $\hat{\alpha}_1$ can be obtained from the one for $\hat{\gamma}$ and $\hat{\beta}$ by the delta method (Section 5.6) using the relation (13.46) between the true and pseudo-true parameters.

In this example, indirect inference is particularly simple because the auxiliary model (13.43) has just as many parameters as the model of interest (13.41). However, this is rarely the case. We saw in Section 13.2 that a finite-order MA or ARMA process can always be represented by an AR($\infty$) process. This suggests that, when estimating an MA or ARMA model, we should use as an auxiliary model an AR($p$) model with $p$ substantially greater than the number of parameters in the model of interest. See Galbraith and Zinde-Walsh (1994, 1997) for implementations of this approach which solve for the binding functions analytically.

Clearly, indirect inference is impossible if the auxiliary model has fewer parameters than the model of interest. If, as is commonly the case, it has more, then the parameters of the model of interest are overidentified. This means that we cannot just solve for them from the estimates of the auxiliary model. Instead, we need to minimize a suitable criterion function, so as to make the estimates of the auxiliary model as close as possible, in the appropriate sense, to the values implied by the parameter estimates of the model of interest. In the next paragraph, we explain how to do this in a very general setting.

---

2 In this special case, an expression for $\Omega$ as a function of $\alpha$, $\rho$, and $\sigma^2_t$ can be obtained from equation (13.45), so that we can estimate $\Omega$ as a function of consistent estimates of those parameters. In most cases, however, it is necessary to use a HAC estimator.
for $t = 1, \ldots, n$, the simulated data are given by

$$u_t^* = \varepsilon_t + \sum_{j=1}^{q} \alpha_j \varepsilon_{t-j}.$$  

There is no need to worry about missing pre-sample innovations in the context of simulation, because they are simulated along with the other innovations. Simulating data from an AR($p$) process is not quite so easy, because of the initial observations. Recursive simulation can be used for all but the first $p$ observations, using the equation

$$u_t^* = \sum_{i=1}^{p} \rho_i u_{t-i}^* + \varepsilon_t. \quad (13.49)$$

For an AR(1) process, the first simulated observation $u_1^*$ can be drawn from the stationary distribution of the process, by which we mean the unconditional distribution of $u_t$. This distribution has mean zero and variance $\sigma^2 = \sigma^2 / (1 - \rho^2)$. The remaining observations are then generated recursively. When $p > 1$, the first $p$ observations must be drawn from the stationary distribution of $p$ consecutive elements of the AR(p) series. This distribution has mean vector zero and covariance matrix $\Omega$ given by expression (13.33) with $n = p$. Once the specific form of this covariance matrix has been determined, perhaps by solving the Yule-Walker equations, and $\Omega$ has been evaluated for the specific values of the $\rho_i$, a $p \times p$ lower-triangular matrix $A$ can be found such that $AA' = \Omega$; see the discussion of the multivariate normal distribution in Section 4.3. We then generate $\varepsilon_p$ as a $p$-vector of white noise innovations and construct the $p$-vector $u_0^*$ of the first $p$ observations as $u_0^* = A \varepsilon_p$. The remaining observations are then generated recursively.

Since it may take considerable effort to find $\Omega$, a simpler technique is often used. One starts the recursion (13.49) for a large negative value of $t$ with essentially arbitrary starting values, often zero. By making the starting value of $t$ far enough in the past, the joint distribution of $u_t^*$ through $u_{p-1}^*$ can be made arbitrarily close to the stationary distribution. The values of $u_t^*$ for nonpositive $t$ are then discarded.

Starting the recursion far in the past also works with an ARMA($p,q$) process. However, at least for simple models, we can exploit the covariances computed by the extension of the Yule-Walker method discussed in Section 13.2. The process (13.22) can be written explicitly as

$$u_t^* = \sum_{i=1}^{p} \rho_i u_{t-i}^* + \varepsilon_t + \sum_{j=1}^{q} \alpha_j \varepsilon_{t-j}. \quad (13.50)$$

In order to be able to compute the $u_t^*$ recursively, we need starting values for $u_1^*, \ldots, u_p^*$ and $\varepsilon_{p-q+1}, \ldots, \varepsilon_p$. Given these, we can compute $u_{p+1}^*$ by drawing the innovation $\varepsilon_{p+1}$ and using equation (13.50) for $t = p + 1, \ldots, n$. The starting values can be drawn from the joint stationary distribution characterized by the autocovariances $u_i$ and covariances $u_j$ discussed in the previous section. In Exercise 13.12, readers are asked to find this distribution for the relatively simple ARMA(1,1) case.

### 13.4 Single-Equation Dynamic Models

Economists often wish to model the relationship between the current value of a dependent variable $y_t$, the current and lagged values of one or more independent variables, and, quite possibly, lagged values of $y_t$ itself. This sort of model can be motivated in many ways. Perhaps it takes time for economic agents to perceive that the independent variables have changed, or perhaps it is costly for them to adjust their behavior. In this section, we briefly discuss a number of models of this type. For notational simplicity, we assume that there is only one independent variable, denoted $x_t$. In practice, of course, there is usually more than one such variable, but it will be obvious how to extend the models we discuss to handle this more general case.

#### Distributed Lag Models

When a dependent variable depends on current and lagged values of $x_t$, but not on lagged values of itself, we have what is called a distributed lag model. When there is only one independent variable, plus a constant term, such a model can be written as

$$y_t = \delta + \sum_{j=0}^{q} \beta_j x_{t-j} + u_t, \quad u_t \sim \text{IID}(0, \sigma^2), \quad (13.51)$$

in which $y_t$ depends on the current value of $x_t$ and on $q$ lagged values. The constant term $\delta$ and the coefficients $\beta_j$ are to be estimated.

In many cases, $x_t$ is positively correlated with some or all of the lagged values $x_{t-j}$ for $j \geq 1$. In consequence, the OLS estimates of the $\beta_j$ in equation (13.51) may be quite imprecise. However, this is generally not a problem if we are merely interested in the long-run impact of changes in the independent variable. This long-run impact is

$$\gamma \equiv \sum_{j=0}^{q} \beta_j = \sum_{j=0}^{q} \frac{\partial y_t}{\partial x_{t-j}}. \quad (13.52)$$

We can estimate (13.51) and then calculate the estimate $\hat{\gamma}$ using (13.52), or we can obtain $\hat{\gamma}$ directly by reparametrizing regression (13.51) as

$$y_t = \delta + \gamma x_t + \sum_{j=1}^{q} \beta_j (x_{t-j} - x_t) + u_t. \quad (13.53)$$
The advantage of this reparametrization is that the standard error of \( \hat{\gamma} \) is immediately available from the regression output.

In Section 3.4, we derived an expression for the variance of a weighted sum of parameter estimates. Expression (3.33), which can be written in a more intuitive fashion as (3.68), can be applied directly to \( \hat{\gamma} \), which is an unweighted sum. If we do so, we find that

\[
\text{Var}(\hat{\gamma}) = \mathbf{u}^T \text{Var}(\hat{\beta}) \mathbf{u} = \sum_{j=0}^{q} \text{Var}(\hat{\beta}_j) + 2 \sum_{j=1}^{q} \sum_{k=0}^{j-1} \text{Cov}(\hat{\beta}_j, \hat{\beta}_k),
\]

(13.54)

where the smallest value of \( j \) in the double summation is 1 rather than 0, because no valid value of \( k \) exists for \( j = 0 \). When \( x_{t-k} \) is positively correlated with \( x_{t-j} \) for all \( j \neq k \), the covariance terms in (13.54) are generally all negative. When the correlations are large, these covariance terms can often be large in absolute value, so much so that \( \text{Var}(\hat{\gamma}) \) may be smaller than the variance of \( \hat{\beta}_j \) for some or all \( j \). If we are interested in the long-run effects of \( x_1 \) on \( y_t \), it is therefore perfectly sensible just to estimate equation (13.33).

**The Partial Adjustment Model**

One popular alternative to distributed lag models like (13.51) is the **partial adjustment model**, which dates back at least to Nerlove (1958). Suppose that the desired level of an economic variable \( y_t \) is \( y_t^0 \). This desired level is assumed to depend on a vector of exogenous variables \( X_t \) according to

\[
y_t^0 = X_t \beta^0 + e_t, \quad e_t \sim \text{IID}(0, \sigma_e^2).
\]

(13.55)

Because of adjustment costs, \( y_t \) is not equal to \( y_t^0 \) in every period. Instead, it is assumed to adjust toward \( y_t^0 \) according to the equation

\[
y_t - y_{t-1} = (1 - \delta)(y_t^0 - y_{t-1}) + v_t, \quad v_t \sim \text{IID}(0, \sigma_v^2),
\]

(13.56)

where \( \delta \) is an adjustment parameter that is assumed to be positive and strictly less than 1. Solving (13.55) and (13.56) for \( y_t \), we find that

\[
y_t = y_{t-1} - (1 - \delta)y_t - (1 - \delta)X_t \beta^0 + (1 - \delta)e_t + v_t
\]

\[
= X_t \beta + \delta y_{t-1} + u_t,
\]

(13.57)

where \( \beta \equiv (1 - \delta)\beta^0 \) and \( u_t \equiv (1 - \delta)e_t + v_t \). Thus the partial adjustment model leads to a linear regression of \( y_t \) on \( X_t \) and \( y_{t-1} \). The coefficient of \( y_{t-1} \) is the adjustment parameter, and estimates of \( \beta^0 \) can be obtained from the OLS estimates of \( \beta \) and \( \delta \). This model does not make sense if \( \delta < 0 \) or if \( \delta \geq 1 \). Moreover, when \( \delta \) is close to 1, the implied speed of adjustment may be implausibly slow.

**Autoregressive Distributed Lag Models**

For simplicity of notation, we will continue to discuss only models with a single independent variable, \( x_t \). In this case, an **autoregressive distributed lag** or ADL model can be written as

\[
y_t = \beta_0 + \sum_{i=1}^{p} \beta_i y_{t-i} + \sum_{j=0}^{q} \gamma_j x_{t-j} + u_t, \quad u_t \sim \text{IID}(0, \sigma^2).
\]

(13.58)

Because there are \( p \) lags on \( y_t \) and \( q \) lags on \( x_t \), this is sometimes called an ADL(\( p, q \)) model.

A widely encountered special case of (13.58) is the **ADL(1, 1)** model

\[
y_t = \beta_0 + \beta_1 y_{t-1} + \gamma_0 x_t + \gamma_1 x_{t-1} + u_t.
\]

(13.59)

Because most results that are true for the ADL(1, 1) model are also true, with obvious modifications, for the more general ADL(\( p, q \)) model, we will largely confine our discussion to this special case.

Although the ADL(1, 1) model is quite simple, many commonly encountered models are special cases of it. When \( \beta_1 = \gamma_1 = 0 \), we have a static regression model with IID errors; when \( \gamma_0 = \gamma_1 = 0 \), we have a univariate AR(1) model; when \( \gamma_1 = 0 \), we have a partial adjustment model; when \( \gamma_1 = -\beta_1 \gamma_0 \), we have
Methods for Stationary Time-Series Data

a static regression model with AR(1) errors; and when $\beta_1 = 1$ and $\gamma_1 = -\gamma_0$, we have a model in first differences that can be written as

$$\Delta y_t = \beta_0 + \gamma_0 \Delta x_t + \nu_t.$$ 

Before we accept any of these special cases, it makes sense to test them against (13.59). This can be done by means of asymptotic $t$ or $F$ tests, which it may be wise to bootstrap when the sample size is not large.

It is usually desirable to impose the condition that $|\beta_1| < 1$ in (13.59). Strictly speaking, this is not a stationarity condition, since we cannot expect $y_t$ to be stationary without imposing further conditions on the explanatory variable $x_t$. However, it is easy to see that, if this condition is violated, the dependent variable $y_t$ exhibits explosive behavior. If the condition is satisfied, there may exist a long-run equilibrium relationship between $y_t$ and $x_t$, which can be used to develop a particularly interesting reparametrization of (13.59).

Suppose there exists an equilibrium value $x^\circ$ to which $x_t$ would converge as $t \to \infty$ in the absence of shocks. Then, in the absence of the error terms $\nu_t$, $y_t$ would converge to a steady-state long-run equilibrium value $y^\circ$ such that

$$y^\circ = \beta_0 + \beta_1 y^\circ + (\gamma_0 + \gamma_1) x^\circ.$$

Solving this equation for $y^\circ$ as a function of $x^\circ$ yields

$$y^\circ = \frac{\beta_0}{1 - \beta_1} + \frac{\gamma_0 + \gamma_1}{1 - \beta_1} x^\circ = \frac{\beta_0}{1 - \beta_1} + \lambda x^\circ,$$

where

$$\lambda = \frac{\gamma_0 + \gamma_1}{1 - \beta_1}.$$  (13.61)

This is the long-run derivative of $y^\circ$ with respect to $x^\circ$, and it is an elasticity if both series are in logarithms. An estimate of $\lambda$ can be computed directly from the estimates of the parameters of (13.59). Note that the result (13.60) and the definition (13.61) make sense only if the condition $|\beta_1| < 1$ is satisfied.

Because it is so general, the ADL($p,q$) model is a good place to start when attempting to specify a dynamic regression model. In many cases, setting $p = q = 1$ is sufficiently general, but with quarterly data it may be wise to start with $p = q = 4$. Of course, we very often want to impose restrictions on such a model. Depending on how we write the model, different restrictions may naturally suggest themselves. These can be tested in the usual way by means of asymptotic $F$ and $t$ tests, which may be bootstrapped to improve their finite-sample properties.

### 13.5 Seasonality

Error-Correction Models

It is a straightforward exercise to check that the ADL($1,1$) model of equation (13.59) can be rewritten as

$$\Delta y_t = \beta_0 + (\beta_1 - 1)(y_{t-1} - \lambda x_{t-1}) + \gamma_0 \Delta x_t + \nu_t,$$  (13.62)

where $\lambda$ was defined in (13.61). Equation (13.62) is called an error-correction model. It expresses the ADL($1,1$) model in terms of an error-correction mechanism; both the model and mechanism are often abbreviated to ECM. Although the model (13.62) appears to be nonlinear, it is really just a reparametrization of the linear model (13.59). If the latter is estimated by OLS, an appropriate GNR can be used to obtain the covariance matrix of the estimates of the parameters of (13.62). Alternatively, any good NLS package should do this for us if we start it at the OLS estimates.

The difference between $y_{t-1}$ and $\lambda x_{t-1}$ in the ECM (13.62) measures the extent to which the long-run equilibrium relationship between $x_t$ and $y_t$ is not satisfied. Consequently, the parameter $\beta_1 - 1$ can be interpreted as the proportion of the resulting disequilibrium that is reflected in the movement of $y_t$ in one period. In this respect, $\beta_1 - 1$ is essentially the same as the parameter $\delta - 1$ of the partial adjustment model. The term $(\beta_1 - 1)(y_{t-1} - \lambda x_{t-1})$ that appears in (13.62) is the error-correction term. Of course, many ADL models in addition to the ADL($1,1$) model can be rewritten as error-correction models. An important feature of error-correction models is that they can also be used with nonstationary data, as we will discuss in Chapter 14.

### 13.5 Seasonality

As we observed in Section 2.5, many economic time series display a regular pattern of seasonal variation over the course of every year. Seasonality, as such a pattern is called, may be caused by seasonal variation in the weather or by the timing of statutory holidays, school vacation periods, and so on. Many time series are observed quarterly, monthly, weekly, or daily, and display some form of seasonality, and this can have important implications for applied econometric work. Failing to account properly for seasonality can easily cause us to make incorrect inferences, especially in dynamic models.

There are two different ways to deal with seasonality in economic data. One approach is to try to model it explicitly. We might, for example, attempt to explain the seasonal variation in a dependent variable by the seasonal variation in some of the independent variables, perhaps including weather variables or, more commonly, seasonal dummy variables, which were discussed

---

3 Error-correction models were first used by Hendry and Anderson (1977) and Davidson, Hendry, Sraff, and Yoo (1978). See Banerjee, Dolado, Galbraith, and Hendry (1993) for a detailed treatment.
in Section 2.5. Alternatively, we can model the error terms as following a seasonal ARMA process, or we can explicitly estimate a seasonal ADL model. The second way to deal with seasonality is usually less satisfactory. It depends on the use of seasonally adjusted data, that is, data which have been filtered in such a way that they represent what the series would supposedly have been in the absence of seasonal variation. Indeed, many statistical agencies release only seasonally adjusted data for many time series, and economists often treat these data as if they were genuine. However, as we will see later in this section, using seasonally adjusted data can have unfortunate consequences.

Seasonal ARMA Processes

One way to deal with seasonality is to model the error terms of a regression model as following a seasonal ARMA process, that is, an ARMA process with nonzero coefficients only, or principally, at seasonal lags. In practice, purely autoregressive processes, with no moving-average component, are generally used. The simplest and most commonly encountered example is the simple AR(4) process

$$u_t = \rho_4 u_{t-4} + \varepsilon_t,$$  \hspace{1cm} (13.63)

where $\rho_4$ is a parameter to be estimated, and, as usual, $\varepsilon_t$ is white noise. Of course, this process makes sense only for quarterly data. Another purely seasonal AR process for quarterly data is the restricted AR(8) process

$$u_t = \rho_4 u_{t-4} + \rho_8 u_{t-8} + \varepsilon_t,$$  \hspace{1cm} (13.64)

which is analogous to an AR(2) process for nonseasonal data.

In many cases, error terms may exhibit both seasonal and nonseasonal serial correlation. This suggests combining a purely seasonal with a nonseasonal process. Suppose, for example, that we wish to combine an AR(1) process and a simple AR(4) process. The most natural approach is probably to combine them multiplicatively. Using lag-operator notation, we obtain

$$(1 - \rho_1 L)(1 - \rho_4 L^4)u_t = \varepsilon_t.$$  \hspace{1cm} (13.65)

This can be rewritten as

$$u_t = \rho_1 u_{t-1} + \rho_4 u_{t-4} - \rho_1 \rho_4 u_{t-5} + \varepsilon_t.$$  \hspace{1cm} (13.65)

Notice that the coefficient of $u_{t-5}$ in equation (13.65) is equal to the negative of the product of the coefficients of $u_{t-1}$ and $u_{t-4}$. This restriction can easily be tested. If it does not hold, then we should presumably consider more general ARMA processes with some coefficients at seasonal lags.

If adequate account of seasonality is not taken, there is often evidence of fourth-order serial correlation in a regression model. Thus testing for it often provides a useful diagnostic test. Moreover, seasonal autoregressive processes provide a parsimonious way to model seasonal variation that is not explained by the regressors. The simple AR(4) process (13.63) uses only one extra parameter, and the restricted AR(8) process (13.64) uses only two. However, just as evidence of first-order serial correlation does not mean that the error terms really follow an AR(1) process, evidence of fourth-order serial correlation does not mean that they really follow an AR(4) process.

By themselves, seasonal ARMA processes cannot capture one important feature of seasonality, namely, the fact that different seasons of the year have different characteristics: Summer is not just winter with a different label. However, an ARMA process makes no distinction among the dynamical processes associated with the different seasons. One simple way to alleviate this problem would be to use seasonal dummy variables as well as a seasonal ARMA process. Another potential difficulty is that the seasonal variation of many time series is not stationary, in which case a stationary ARMA process cannot adequately account for it. Trending seasonal variables may help to cope with nonstationary seasonality, as we will discuss shortly in the context of a specific example.

Seasonal ADL Models

Suppose we start with a static regression model in which $y_t$ equals $X_t \beta + u_t$ and then add three quarterly dummy variables, $s_1$ through $s_3$, assuming that there is a constant among the other explanatory variables. The dummies may be ordinary quarterly dummies, or else the modified dummies, defined in equations (2.49), that sum to zero over each year. We then allow the error term $u_t$ to follow the simple AR(4) process (13.63). Solving for $u_{t-4}$ yields the nonlinear regression model

$$y_t = \rho_4 y_{t-4} + X_t \beta - \rho_4 X_{t-4} \beta + \sum_{j=1}^{3} \delta_j s_{t-j} + \varepsilon_t.$$  \hspace{1cm} (13.66)

There are no lagged seasonal dummies in this model because they would be collinear with the existing regressors.

Equation (13.66) is a special case of the seasonal ADL model

$$y_t = \gamma_4 y_{t-4} + X_t \beta_1 + X_{t-4} \beta_2 + \sum_{j=1}^{3} \delta_j s_{t-j} + \varepsilon_t,$$  \hspace{1cm} (13.67)

which is just a linear regression model in which $y_t$ depends on $y_{t-4}$, the three seasonal dummies, $X_t$, and $X_{t-4}$. Before accepting the model (13.66), one would always want to test the common factor restrictions that it imposes on (13.67); this can readily be done by using asymptotic $F$ tests, as discussed in Section 7.9. One would almost certainly also want to estimate ADL models both more and less general than (13.67), especially if the common factor restrictions are rejected. For example, it would not be surprising if $y_{t-1}$ and at least some components of $X_{t-1}$ also belonged in the model, but it would also not be surprising if some components of $X_{t-4}$ did not belong.
Seasonally Adjusted Data

Instead of attempting to model seasonality, many economists prefer to avoid dealing with it entirely by using seasonally adjusted data. Although the idea of seasonally adjusting a time series is intuitively appealing, it is very hard to do so in practice without resorting to highly unrealistic assumptions. Seasonal adjustment of a series $y_t$ makes sense if, for all $t$, we can write $y_t = y_t^c + y_t^s$, where $y_t^c$ is a time series that contains no seasonal variation at all, and $y_t^s$ is a time series that contains nothing but seasonal variation. However, this is surely an extreme assumption, which would be false in almost any economic model of seasonal variation that could reasonably be imagined.

To make the discussion more concrete, consider Figure 13.2, which shows the logarithm of urban housing starts in Canada, quarterly, for the period 1966 to 2001. The solid line represents the actual data, and the dotted line represents a seasonally adjusted series.\(^4\) It is clear from the figure that housing starts in Canada are highly seasonal, with the first (winter) quarter usually having a much smaller number of starts than the other three quarters. There is also some indication that the magnitude of the seasonal variation may have become smaller in the latter part of the sample, perhaps because of changes in construction technology.

![Figure 13.2 Urban housing starts in Canada, 1966–2001](image)

\(^4\) These data come from Statistics Canada. The actual data, which start in 1948, are from CANSIM series J6001, and the adjusted data, which start in 1996, are from CANSIM series J9001.

### 13.5 Seasonality

#### Seasonal Adjustment by Regression

In Section 2.5, we discussed the use of seasonal dummy variables to construct seasonally adjusted data by regression. Although this approach is easy to implement and easy to analyze, it has a number of disadvantages, some of which we briefly discuss here. Because of these disadvantages, it is almost never used by official statistical agencies.

One problem with the simplest form of seasonal adjustment by regression is that it does not allow the pattern of seasonality to change over time. However, as Figure 13.2 illustrates, seasonal patterns often seem to do precisely that. A natural way to model this is to add additional seasonal dummy variables that have been interacted with powers of a time trend that increases annually. In the case of quarterly data, such a trend would be

$$t_q^T = [1 1 1 1 2 2 2 3 3 3 3 \ldots]. \quad (13.68)$$

The reason $t_q$ takes this rather odd form is that, when it is multiplied by the seasonal dummies, the resulting trending dummies always sum to zero over each year. If one simply multiplied seasonal dummies by an ordinary time trend, that would not be the case.

Let $S$ denote a matrix of seasonal dummies and seasonal dummies that have been interacted with powers of $t_q$ or, in the case of data at other than quarterly frequencies, whatever annually increasing trend term is appropriate. In the case of quarterly data, $S$ would normally have 3, 6, 9, or maybe 12 columns. In the case of monthly data, it would normally have 11, 22, or 33 columns. In all cases, every one of the variables in $S$ should sum to zero over each year. Then, if $y$ denotes the vector of observations on a series to be seasonally adjusted, we could run the regression

$$y = \beta_0 + S\delta + u \quad (13.69)$$

and estimate the seasonally adjusted series as $y' = y - S\hat{\delta}$. Unfortunately, although equations like (13.69) often provide a reasonable approximation to observed seasonal patterns, they frequently fail to do so, as readers will find when they answer Exercise 13.17.

Another problem with using seasonal dummies is that, as additional observations become available, the estimates from the dummy variable regression do not stay the same. It is inevitable that, as the sample size increases, the estimates of $\delta$ in equation (13.69) change, and so every element of $y'$ must change every time a new observation becomes available. This is clearly a most undesirable feature from the point of view of users of official statistics. Moreover, as the sample size gets larger, the number of trend terms may need to increase if a polynomial is to continue to provide an adequate approximation to changes in the pattern of seasonal variation.
Seasonal Adjustment and Linear Filters

The seasonal adjustment procedures that are actually used by statistical agencies tend to be very complicated. They attempt to deal with a host of practical problems, including changes in seasonal patterns over time, variations in the number of shopping days and the dates of holidays from year to year, and the fact that pre-sample and post-sample observations are not available. We will not attempt to discuss these methods at all.

Although official methods of seasonal adjustment are very complicated, they can often be approximated remarkably well by much simpler procedures based on what are called linear filters. Let \( y \) be an \( n \)-vector of observations (often in logarithms rather than levels) on a series that has not been seasonally adjusted. Then a linear filter consists of an \( n \times n \) matrix \( \Phi \), with rows that sum to 1, such that the seasonally adjusted series \( y' \) is equal to \( \Phi y \). Each row of the matrix \( \Phi \) consists of a vector of filter weights. Thus each element \( y'_i \) of the seasonally adjusted series is equal to a weighted average of current, leading, and lagged values of \( y \).

Let us consider a simple example for quarterly data. Suppose we first create three-term and eleven-term moving averages

\[
\bar{y}_t = \frac{1}{3}(y_{t-4} + y_t + y_{t+4}) \quad \text{and} \quad \bar{y}_t = \frac{1}{11} \sum_{j=-5}^{5} y_{t+j}.
\]

The difference between \( \bar{y}_t \) and \( \bar{y}_t \) is a rolling estimate of the amount by which the value of \( y \) for the current quarter tends to differ from its average value over the year. Thus one way to define a seasonally adjusted series would be

\[
y'_t = y_t - \bar{y}_t + \bar{y}_t = .0909y_{t-5} - .2424y_{t-4} + .0909y_{t-3} + .0909y_{t-2} + .0909y_{t-1} + .7576y_t + .0909y_{t+1} + .0909y_{t+2} + .0909y_{t+3} - .2424y_{t+4} + .0909y_{t+5}.
\]  

(13.70)

This example corresponds to a linear filter in which, for \( 5 < p < n-5 \), the \( p \)-th row of \( \Phi \) would consist first of \( p-6 \) zeros, followed by the eleven coefficients that appear in (13.70), followed by \( n-p-5 \) more zeros.

Although this example is very simple, the basic approach that it illustrates may be found, in various modified forms, in almost all official seasonal adjustment procedures. The latter generally do not actually employ linear filters, but they do employ a number of moving averages in a way similar to the example. These moving averages tend to be longer than the ones in the example, and they often give progressively less weight to observations farther from \( t \).

An important feature of almost all seasonally adjusted data is that, as in the example, the weight given to \( y_t \) is generally well below 1. For more on the relationship between official procedures and ones based on linear filters, see Buaridge and Wallis (1984) and Ghysels and Perron (1993).

We have claimed that official seasonal adjustment procedures in most cases have much the same properties as linear filters applied to either the levels or the logarithms of the raw data. This assertion can be checked empirically by regressing a seasonally adjusted series on a number of leads and lags of the corresponding seasonally unadjusted series. If the assertion is accurate, such a regression should fit well, and the coefficients should have a distinctive pattern. The coefficient of the current value of the raw series should be fairly large but less than 1; the coefficients of seasonal lags and leads should be negative, and the coefficients of other lags and leads should be small and positive. In other words, the coefficients should resemble those in equation (13.70). In Exercise 13.17, readers are asked to see whether a linear filter provides a good approximation to the method actually used for seasonally adjusting the housing starts data.

Consequences of Using Seasonally Adjusted Data

The consequences of using seasonally adjusted data depend on how the data were actually generated and the nature of the procedures used for seasonal adjustment. For simplicity, we will suppose that

\[
y = y_o + y_s \quad \text{and} \quad X = X_o + X_s,
\]

where \( y_s \) and \( X_s \) contain all the seasonal variation in \( y \) and \( X \), respectively, and \( y_o \) and \( X_o \) contain all other economically interesting variation. Suppose further that the DGP is

\[
y_o = X_o \beta_o + u, \quad u \sim \text{IID}(0, \sigma^2 I). \tag{13.71}
\]

(13.71)

Thus the economic relationship in which we are interested involves only the nonseasonal components of the data.

If the same linear filter is applied to every series, the seasonally adjusted data are \( \Phi y \) and \( \Phi X \), and the OLS estimator using those data is

\[
\hat{\beta}_s = (X^T \Phi^T \Phi X)^{-1} X^T \Phi^T \Phi y. \tag{13.72}
\]

(13.72)

This looks very much like a GLS estimator, with the matrix \( \Phi^T \Phi \) playing the role of the inverse covariance matrix.

The properties of the estimator \( \hat{\beta}_s \) defined in equation (13.72) depend on how the filter weights are chosen. Ideally, the filter would completely eliminate seasonality, so that

\[
\Phi y = \Phi y_o \quad \text{and} \quad \Phi X = \Phi X_o.
\]
In this ideal case, we see that
\[ \hat{\beta}_S = (X_0^\top \Phi^\top \Phi X_0)^{-1} X_0^\top \Phi^\top \Phi y_0 = \beta_0 + (X_0^\top \Phi^\top \Phi X_0)^{-1} X_0^\top \Phi^\top \Phi u. \] (13.73)

If every column of \( X \) is exogenous, and not merely predetermined, it is clear that the second term in the last line here has expectation zero, which implies that \( E(\hat{\beta}_S) = \beta_0 \). Thus we see that, under the exogeneity assumption, the OLS estimator that uses seasonally adjusted data is unbiased. But this is a very strong assumption for time-series data.

Moreover, this estimator is not efficient. If the elements of \( u \) are actually homoskedastic and serially independent, as we assumed in (13.71), then the Gauss-Markov Theorem implies that the efficient estimator would be obtained by an OLS regression of \( y_t \) on \( X_t \). Instead, \( \hat{\beta}_S \) is equivalent to the estimator from a certain GLS regression of \( y_t \) on \( X_t \). Of course, the efficient estimate is not feasible here, because we do not observe \( y_0 \) and \( X_0 \).

In many cases, we can prove consistency under much weaker assumptions that are needed to prove unbiasedness; see Sections 3.2 and 3.3. In particular, for OLS to be consistent, we usually just need the regressors to be predetermined. However, in the case of data that have been seasonally adjusted by means of a linear filter, this assumption is not sufficient. In fact, the exogeneity assumption that is needed in order to prove that \( \hat{\beta}_S \) is unbiased is also needed in order to prove that it is consistent. From (13.73) it follows that
\[ \operatorname{plim} \hat{\beta}_S = \beta_0 + \operatorname{plim} \left( \frac{1}{n} X_n^\top \Phi^\top \Phi X_n \right)^{-1} \operatorname{plim} \left( \frac{1}{n} X_n^\top \Phi^\top \Phi u \right), \]

provided we impose sufficient conditions for the probability limits to exist and be nonstochastic. The predeterminedness assumption (3.10) evidently does not allow us to claim that the second probability limit here is a zero vector. On the contrary, any correlation between error terms and regressors at leads and lags that are given nonzero weights by the filter generally causes it to be a nonzero vector. Therefore, the estimator \( \hat{\beta}_S \) is inconsistent if the regressors are merely predetermined.

Although the exogeneity assumption is always dubious in the case of time-series data, it is certainly false when the regressors include one or more lags of the dependent variable. There has been some work on the consequences of using seasonally adjusted data in this case; see Jaeger and Kunst (1990), Ghysels (1990), and Ghysels and Perron (1993), among others. It appears that, in models with a single lag of the dependent variable, estimates of the coefficient of the lagged dependent variable can be severely biased when seasonally adjusted data are used. This bias does not vanish as the sample size increases, and its magnitude can be substantial; see Davidson and MacKinnon (1993, Chapter 19) for an illustration.

### 13.6 Autoregressive Conditional Heteroskedasticity

Seasonally adjusted data are very commonly used in applied econometric work. Indeed, it is difficult to avoid doing so in many cases, either because the actual data are not available or because it is the seasonally adjusted series that are really of interest. However, the results we have just discussed suggest that, especially for dynamic models, the undesirable consequences of using seasonally adjusted data may be quite severe.

#### 13.6.1 ARCH Models

With time-series data, it is not uncommon for least-squares residuals to be quite small in absolute value for a number of successive periods of time, then much larger for a while, then smaller again, and so on. This phenomenon of time-varying volatility is often encountered in models for stock returns, foreign exchange rates, and other series that are determined in financial markets. Numerous models for dealing with this phenomenon have been proposed. One very popular approach is based on the concept of autoregressive conditional heteroskedasticity, or ARCH, that was introduced by Engle (1982). The basic idea of ARCH models is that the variance of the error term at time \( t \) depends on the realized values of the squared error terms in previous time periods.

If \( u_t \) denotes the error term adhering to a regression model, which may be linear or nonlinear, and \( \Omega_t \) denotes an information set that consists of data observed through period \( t - 1 \), then what is called an ARCH(\( q \)) process can be written as
\[ u_t = \sigma_t \varepsilon_t; \quad \sigma_t^2 = \text{E}(u_t^2 | \Omega_{t-1}) = \alpha_0 + \sum_{i=1}^{q} \alpha_i u_{t-i}^2, \] (13.74)

where \( \alpha_i > 0 \) for \( i = 0, 1, \ldots, q \), and \( \varepsilon_t \) is white noise with variance 1. Here and throughout this section, \( \sigma_t \) is understood to be the positive square root of \( \sigma_t^2 \). The skewness function for the ARCH(\( q \)) process is the rightmost expression in (13.74). Since this function depends on \( t \), the model is, as its name claims, heteroskedastic. The term "conditional" is due to the fact that, unlike the skewed functions we have so far encountered, the ARCH skew function is not exogenous, but merely predetermined. Thus the model prescribes the variance of \( u_t \) conditional on the past of the process.

Because the conditional variance of \( u_t \) is a function of \( u_{t-1} \), it is clear that \( u_t \) and \( u_{t-1} \) are not independent. They are, however, uncorrelated:
\[ \text{E}(u_t u_{t-1}) = \text{E}(E(u_t u_{t-1} | \Omega_t)) = \text{E}(u_{t-1} \sigma_t \varepsilon_t | \Omega_{t-1}) = 0, \]

where we have used the facts that \( \sigma_t \in \Omega_{t-1} \) and that \( \varepsilon_t \) is an innovation. Almost identical reasoning shows that \( \text{E}(u_t u_s) = 0 \) for all \( s < t \). Thus the ARCH process involves only heteroskedasticity, not serial correlation.
If an ARCH(\(q\)) process is covariance stationary, then \(\sigma^2\), the unconditional expectation of \(u_t^2\), exists and is independent of \(t\). Under the stationarity assumption, we may take the unconditional expectation of the second equation of (13.74), from which we find that

\[
\sigma^2 = \alpha_0 + \sigma^2 \sum_{i=1}^{\infty} \alpha_i.
\]

Therefore,

\[
\sigma^2 = \frac{\alpha_0}{1 - \sum_{i=1}^{\infty} \alpha_i}.
\] (13.75)

The condition \(\sum_{i=1}^{\infty} \alpha_i < 1\) is required for \(\sigma^2\) to be positive, and so it is also a necessary condition for stationarity. It is of course necessary that the conditional variances \(\sigma_t^2\) should be positive, and that is why we require that \(\alpha_i > 0\) for all \(i\). If that requirement were not satisfied, realizations of some of the \(\sigma_t^2\) could be negative.

Unfortunately, the ARCH(\(q\)) process has not proven to be very satisfactory in applied work. Many financial time series display time-varying volatility that is highly persistent, but the correlation between successive values of \(u_t^2\) is not very high; see Pagan (1996). In order to accommodate these two empirical regularities, \(q\) must be large. But if \(q\) is large, the ARCH(\(q\)) process has a lot of parameters to estimate, and the requirement that all the \(\alpha_i\) should be positive may not be satisfied if it is not explicitly imposed.

**GARCH Models**

The **generalized ARCH model**, which was proposed by Bollerslev (1986), is much more widely used than the original ARCH model. We may write a GARCH(\(p, q\)) process as

\[
u_t = \sigma_t \varepsilon_t; \quad \sigma_t^2 \equiv \mathbb{E}(u_t^2 | \Omega_{t-1}) = \alpha_0 + \sum_{i=1}^{q} \alpha_i u_{t-i}^2 + \sum_{j=1}^{p} \delta_j \sigma_{t-j}^2.
\] (13.76)

The conditional variance here can be written more compactly as

\[
\sigma_t^2 = \alpha_0 + \alpha(L) u_t^2 + \delta(L) \sigma_t^2,
\] (13.77)

where \(\alpha(L)\) and \(\delta(L)\) are polynomials in the lag operator \(L\), neither of which includes a constant term. All of the parameters in the infinite-order autoregressive representation

\[
(1 - \delta(L))^{-1} \alpha(L)
\]

must be nonnegative. Otherwise, as in the case of an ARCH(\(q\)) model with one or more of the \(\alpha_i < 0\), we could have negative conditional variances.

There is a strong resemblance between the GARCH(\(p, q\)) process (13.77) and the ARMA(\(p, q\)) process (13.21). In fact, if we let \(\delta(L) = \rho(L)\), \(\alpha_0 = \gamma, \sigma_t^2 = y_t,\) and \(u_t^2 = \varepsilon_t,\) we see that the former becomes formally the same as an ARMA(\(p, q\)) process in which the coefficient of \(\varepsilon_t\) equals 0. However, the formal similarity between the two processes masks some important differences. In a GARCH process, the \(\sigma_t^2\) are not observable, and \(\mathbb{E}(u_t^2) = \sigma_t^2 \neq 0\).

The simplest and by far the most popular GARCH model is the GARCH(1,1) process, for which the conditional variance can be written as

\[
\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \delta_1 \sigma_{t-1}^2.
\] (13.78)

Under the hypothesis of covariance stationarity, the unconditional variance \(\sigma^2\) can be found by taking the unconditional expectation of equation (13.78). We find that

\[
\sigma^2 = \alpha_0 + \alpha_1 \sigma^2 + \delta_1 \sigma^2.
\]

Solving this equation yields the result that

\[
\sigma^2 = \frac{\alpha_0}{1 - \alpha_1 - \delta_1}.
\] (13.79)

For this unconditional variance to exist, it must be the case that \(\alpha_1 + \delta_1 < 1\), and for it to be positive, we require that \(\alpha_0 > 0\).

The GARCH(1,1) process generally seems to work quite well in practice. In many cases, it cannot be rejected against any more general GARCH(\(p, q\)) process. An interesting empirical regularity is that the estimate \(\hat{\delta}_1\) is often small and positive, with the estimate \(\hat{\delta}_1\) much larger, and the sum of the coefficients, \(\hat{\alpha}_1 + \hat{\delta}_1\), between 0.9 and 1. These parameter values imply that the time-varying volatility is highly persistent.

**Testing for ARCH Errors**

It is easy to test a regression model for the presence of ARCH or GARCH errors. Imagine, for the moment, that we actually observe the \(u_t\). Then we can replace \(\sigma_t^2\) by \(u_t^2 - \varepsilon_t\), where \(\varepsilon_t\) is defined to be the difference between \(u_t^2\) and its conditional expectation. This allows us to rewrite the GARCH(\(p, q\)) model (13.76) as

\[
u_t^2 = \alpha_0 + \sum_{i=1}^{\max(p, q)} (\alpha_i + \delta_i) u_{t-i}^2 + \varepsilon_t - \sum_{j=1}^{p} \delta_j \varepsilon_{t-j}.
\] (13.80)

In this equation, we have replaced all of the \(\sigma_t^2\) by \(u_t^2 - \varepsilon_t\) and then grouped the two summations that involve the \(u_{t-j}^2\). Of course, if \(p \neq q\), either some of the \(\alpha_i\) or some of the \(\delta_j\) in the first summation are identically zero. Equation (13.80) can now be interpreted as a regression model with dependent variable \(u_t^2\) and MA(\(p\)) errors. If one were actually to estimate (13.80), the MA structure would yield estimates of the \(\delta_j\), and the estimated coefficients of the \(u_{t-j}^2\) would then allow the \(\alpha_i\) to be estimated.
Rather than estimating (13.80), it is easier to base a test on the Gauss-Newton regression that corresponds to (13.80), evaluated under the null hypothesis that \( \alpha_i = 0 \) for \( i = 1, \ldots, q \) and \( \delta_j = 0 \) for \( j = 1, \ldots, p \). Since equation (13.80) is linear with respect to the \( \alpha_i \) and the \( \delta_j \), the GNR is easy to derive. It is

\[
\max_{(p,q)} u_t^2 - \alpha_0 = b_0 + \sum_{i=1}^{\max(p,q)} b_i u_{t-i}^2 + \text{residual}. \tag{13.81}
\]

The artificial parameter \( b_0 \) here corresponds to the real parameter \( \alpha_0 \), and the \( b_i \), for \( i = 1, \ldots, \max(p,q) \), correspond to the sums \( \alpha_i + \delta_i \), because, under the null, the \( \alpha_i \) and \( \delta_i \) are not separately identifiable. In the regressand, \( \alpha_0 \) would normally be the error variance estimated under the null. If we are using equation (13.81) for testing, because there is a constant term on the right-hand side.

Under the alternative, the GNR should, strictly speaking, incorporate the MA structure of the error terms of (13.80). But, since these error terms are white noise under the null, a valid test can be constructed without taking account of the MA structure. The price to be paid for this simplification is that the \( \alpha_i \) and the \( \delta_i \) remain unidentified as separate parameters, which means that the test is the same for all GARCH\((p,q)\) alternatives with the same value of \( \max(p,q) \).

In practice, of course, we do not observe the \( u_t \). But, as for the GNR-based tests against other types of heteroskedasticity that we discussed in Section 7.5, it is asymptotically valid to replace the unobserved \( u_t \) by the least-squares residuals \( \hat{u}_t \). Thus the test regression is actually

\[
\hat{u}_t^2 = \hat{b}_0 + \sum_{i=1}^{\max(p,q)} \hat{b}_i \hat{u}_{t-i}^2 + \text{residual}, \tag{13.82}
\]

where we have arbitrarily set \( \hat{\alpha}_0 = 0 \). Because of the lags, this GNR would normally be run over the last \( n - \max(p,q) \) observations only. As usual, there are several possible test statistics. The easiest to compute is probably \( n \times \text{centered} \ R^2 \), which is asymptotically distributed as \( \chi^2(\max(p,q)) \) under the null. It is also asymptotically valid to use the standard \( F \) statistic for all of the slope coefficients to be 0, treating it as if it followed the \( F \) distribution with \( \max(p,q) \) and \( n - 2 \max(p,q) - 1 \) degrees of freedom. These tests are easily bootstrapped, and it is often wise to do so. We can use either parametric or a semiparametric bootstrap DGP.

Because it is very easy to compute a test statistic using regression (13.82), these tests are the most commonly used procedures to detect autoregressive conditional heteroskedasticity. However, other procedures may well perform better. In particular, Lee and King (1993) and Demos and Sentana (1998) have suggested various tests which take into account the fact that the alternative hypothesis is one-sided. These one-sided tests have better power than tests based on the Gauss-Newton regression (13.82).

The Stationary Distribution for ARCH and GARCH Processes

In the case of an ARMA process, the stationary, or unconditional, distribution of the \( u_t \) is normal whenever the innovations \( \varepsilon_t \) are normal white noise. However, this is not true for ARCH and GARCH processes, because the mapping from the \( \varepsilon_t \) to the \( u_t \) is nonlinear. As we will see, the stationary distribution is not normal, and it may not even have a fourth moment. For simplicity, we will confine our attention to the fourth moment of the ARCH(1) process. Other moments of this process, and moments of the GARCH\((1,1)\) process, are treated in the exercises.

For an ARCH\((1)\) process with normal white noise innovations, or indeed any such ARCH or GARCH process, the distribution of \( u_t \) is normal conditional on \( \Omega_{t-1} \). Since the variance of this distribution is \( \sigma_t^2 \), the fourth moment is \( 3\sigma_t^4 \), as we saw in Exercise 4.2. For an ARCH\((1)\) process, \( \sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 \). Therefore,

\[
\mathbb{E}(u_t^4 | \Omega_{t-1}) = 3(\alpha_0 + \alpha_1 u_{t-1}^2)^2 = 3\alpha_0^2 + 6\alpha_0\alpha_1 u_{t-1}^2 + 3\alpha_1^2 u_{t-1}^4.
\]

If we assume that the unconditional fourth moment exists and denote it by \( m_4 \), we can take the unconditional expectation of this relation to obtain

\[
m_4 = 3\alpha_0^2 + \frac{6\alpha_0\alpha_1}{1 - \alpha_1} + 3\alpha_1^2 m_4,
\]

where we have used the implication of equation (13.75) that the unconditional second moment is \( \alpha_0/(1 - \alpha_1) \). Solving this equation for \( m_4 \), we find that

\[
m_4 = \frac{3\alpha_0^2 (1 + \alpha_1)}{(1 - \alpha_1)(1 - 3\alpha_0^2)}. \tag{13.83}
\]

This result evidently cannot hold unless \( 3\alpha_0^2 < 1 \). In fact, if this condition fails, the fourth moment does not exist. From the result (13.83), we can see that \( m_4 > 3\sigma_t^4 = 3\alpha_0^2/(1 - \alpha_1)^2 \) whenever \( \alpha_1 > 0 \). Thus, whatever the stationary distribution of \( u_t \) might be, it certainly cannot be normal. At the time of writing there is, as far as the authors are aware, no explicit, analytical characterization of the stationary distribution for general ARCH and GARCH processes.

Estimating ARCH and GARCH Models

Since ARCH and GARCH processes induce heteroskedasticity, it might seem natural to estimate a regression model with ARCH or GARCH errors by using feasible GLS. The first step would be to estimate the underlying regression model by OLS or NLS in order to obtain consistent but inefficient estimates of the regression parameters, along with least-squares residuals \( \hat{u}_t \). The second step would be to estimate the parameters of the ARCH or GARCH process by treating the \( \hat{u}_t^2 \) as if they were actual squared error terms and estimating
a model with a specification something like (13.80), again by least squares. The final step would be to estimate the original regression model by feasible weighted least squares, using weights proportional to the inverse square roots of the fitted values from the model for the $\hat{u}_t^2$.

This approach is very rarely used, because it is not asymptotically efficient. The skedastic function, which would, for example, be the right-hand side of equation (13.78) in the case of a GARCH(1, 1) model, depends on the lagged squared residuals, which in turn depend on the estimates of the regression function. Because of this, estimating both functions together yields more efficient estimates than estimating each of them conditional on estimates of the other; see Engle (1982).

The most popular way to estimate models with GARCH errors is to assume that the error terms are normally distributed and use maximum likelihood: We can write a linear regression model with GARCH errors defined in terms of a normal innovation process as

$$ y_t - X_t \hat{\beta} = \varepsilon_t, \quad \varepsilon_t \sim N(0, 1), $$

(13.84)

where $y_t$ is the dependent variable, $X_t$ is a vector of exogenous or predetermined regressors, and $\hat{\beta}$ is a vector of regression parameters. The skedastic function $\sigma_t^2(\beta, \theta)$ is defined for some particular choice of $p$ and $q$ by equation (13.76) with $u_{t-1}$ replaced by $y_t - X_t \hat{\beta}$. It therefore depends on $\hat{\beta}$ as well as on the $\alpha_i$ and $\delta_j$ that appear in (13.76), which we denote collectively by $\theta$.

The density of $\varepsilon_t$ conditional on $\hat{\Omega}_{t-1}$ is then

$$ \frac{1}{\sigma_t(\hat{\beta}, \theta)} \phi \left( \frac{y_t - X_t \hat{\beta}}{\sigma_t(\hat{\beta}, \theta)} \right), $$

(13.85)

where $\phi(\cdot)$ denotes the standard normal density. The first factor in (13.85) is a Jacobian factor which reflects the fact that the derivative of $\varepsilon_t$ with respect to $y_t$ is $\sigma_t^{-1}(\hat{\beta}, \theta)$; see Section 10.8.

By taking the logarithm of expression (13.85), we find that the contribution to the loglikelihood function made by the $t$th observation is

$$ \ell_t(\beta, \theta) = -\frac{1}{2} \log 2\pi - \frac{1}{2} \log(\sigma_t^2(\beta, \theta)) - \frac{1}{2} \left( \frac{y_t - X_t \hat{\beta}}{\sigma_t(\hat{\beta}, \theta)} \right)^2. $$

(13.86)

Unfortunately, it is not entirely straightforward to evaluate this expression. The problem is the skedastic function $\sigma_t^2(\beta, \theta)$, which is defined implicitly by the recursion (13.77). This recursion does not constitute a complete definition because it does not provide starting values to initialize the recursion. In trying to find suitable starting values, we run into the difficulty, mentioned in the previous subsection, that there exists no closed-form expression for the stationary GARCH density.

If we are dealing with an ARCH(q) model, we can sidestep this problem by conditioning on the first q observations. Since, in this case, the skedastic function $\sigma_t^2(\beta, \theta)$ is determined completely by the q lags of the squared residuals, there is no missing information for observations $q + 1$ through $n$. We can therefore sum the contributions (13.86) for just those observations, and then maximize the result. This leads to ML estimates conditional on the first q observations. But such a procedure works only for models with pure ARCH errors, and these models are very rarely used in practice.

With a GARCH(p, q) model, p starting values of $\sigma_t^2$ are needed in addition to $\max(0, q - p)$ starting values of the squared residuals in order to initialize the recursion (13.77). It is therefore necessary to resort to some sort of ad hoc procedure to specify the starting values. A not very good idea is just to set all unknown pre-sample values of $\sigma_t^2$ to zero. A better idea is to replace them by an estimate of their common unconditional expectation. At least two different ways of doing this are in common use. The first is to replace the unconditional expectation by the appropriate function of the $\theta$ parameters, which would be given by the rightmost expression in equations (13.79) for GARCH(1, 1). The second is to use the sum of squared residuals from OLS estimation, divided by $n$.

Another approach, similar to one we discussed for models with MA errors, is to treat the unknown starting values as extra parameters, and to maximize the loglikelihood with respect to them, $\beta$, and $\theta$ jointly. In all but large samples, the choice of starting values can have a significant effect on the parameter estimates. Consequently, different programs for GARCH estimation can produce very different results. This unsatisfactory state of affairs, documented convincingly by Brooks, Burke, and Persand (2001), results from doing ML estimation conditional on different things.

For any choice of starting values, maximizing a loglikelihood function obtained by summing the contributions (13.86) is not particularly easy, especially in the case of GARCH models. Numerical difficulties seem to be quite common. It is vital to use analytical, rather than numerical, first derivatives, and for some algorithms it is highly desirable to use analytical second derivatives as well, these may be found in Fiorentini, Calzolari, and Panattoni (1996). Exercise 13.22 proposes an artificial regression, which makes use of first derivatives only. Not all software packages provide reliable estimates and standard errors; see McCullough and Renfro (1999) and Brooks, Burke, and Persand (2001). Therefore, we strongly recommend estimating this type of model more than once using different options and different computer programs.

Although GARCH models have error terms with thicker tails than those of the normal distribution, data from financial markets often have tails even thicker than those implied by a GARCH model with normal $\varepsilon_t$. It is therefore quite common to modify (13.84) by assuming that the $\varepsilon_t$ follow a distribution with thicker tails than the standard normal. One possibility is the Student’s $t$ distribution with a small number of degrees of freedom, which may be chosen
in advance or estimated. Maximum likelihood estimation then proceeds in the usual way.

We can use any of the estimators discussed in Section 10.4 to estimate the covariance matrix of the ML estimates. One of these, the information matrix estimator, can be computed by means of the artificial regression that is introduced in Exercise 13.22. If the error terms are not distributed according to the normal or whatever distribution we have assumed, the ML estimates are still consistent, but they are not asymptotically efficient. In this case, the sandwich covariance matrix estimator (10.45) is consistent, but covariance matrix estimators that rely on the information matrix equality generally are not. A variant of the sandwich estimator specifically adapted to GARCH models was derived by Bollerslev and Wooldridge (1992). These and other possible variants are discussed and compared by Fiorentini, Calzolari, and Panattoni (1996).  

Simulating ARCH and GARCH Models
ARCH and GARCH models can be simulated recursively in much the same way as ARMA models. The successive values of the \( \sigma_t^2 \) are computed on the basis of past realizations of the \( u_t^2 \) and \( \sigma_t^2 \) series, and the \( u_t \) are generated as \( \sigma_t \varepsilon_t \) for a white-noise series \( \varepsilon_t \), which is often but not always normal. However, the problem of finding suitable starting values for the recursion is much harder for ARCH and GARCH models than for ARMA ones, because we cannot simply draw them from the stationary distribution.

The easiest approach is the one already mentioned in the ARMA context, whereby one starts the recursion for some large negative \( t \) and discards the elements of the simulated series for nonpositive \( t \). It is natural to set the initial values of \( \sigma_0^2 \) in this recursion to the unconditional expectation of the \( u_t^2 \), or, in the bootstrap case, to an estimate of this unconditional expectation. However, this approach is not entirely satisfactory for bootstrapping, where we wish to condition on the observed data as far as possible. One possibility would be to condition on the first \( \max(p,q) \) observations, using the first \( q \) squared residuals as the initial values of \( u_t^2 \) and the first \( p \) squared residual as the initial values of \( \sigma_t^2 \). However, since much work remains to be done on bootstrapping ARCH and GARCH models, we cannot recommend this or any other approach at the present time.

Our discussion of autoregressive conditional heteroskedasticity has necessarily been quite superficial. There have been many extensions of the basic ARCH and GARCH models discussed here, among them the exponential GARCH model of Nelson (1991) and the absolute GARCH model of Hentschel (1995). These models are intended to explain empirical features of financial time series that the standard GARCH model cannot capture. More detailed treatments may be found in Bollerslev, Chou, and Kroner (1992), Bollerslev, Engle, and Nelson (1994), Hamilton (1994, Chapter 21), and Pagan (1996).

13.7 Vector Autoregressions
The dynamic models discussed in Section 13.4 were single-equation models. But we often want to model the dynamic relationships among several time-series variables. A simple way to do so without making many assumptions is to use what is called a vector autoregression, or VAR, model, which is the multivariate analog of an autoregressive model for a single time series.

Let the \( 1 \times g \) vector \( Y_t \) denote the \( t \)th observation on a set of \( g \) variables. Then a vector autoregressive model of order \( p \), sometimes referred to as a VAR(\( p \)) model, can be written as

\[
Y_t = \alpha + \sum_{j=1}^{p} Y_{t-j} \Phi_j + U_t, \quad U_t \sim \text{iid}(0, \Sigma),
\]

where \( U_t \) is a \( 1 \times g \) vector of error terms, \( \alpha \) is a \( 1 \times g \) vector of constant terms, and the \( \Phi_j \) for \( j = 1, \ldots, p \), are \( g \times g \) matrices of coefficients, all of which are to be estimated. If \( y_{ti} \) denotes the \( i \)th element of \( Y_t \) and \( \phi_{ji,k} \) denotes the \( k \)th element of \( \Phi_j \), then the \( i \)th column of (13.87) can be written as

\[
y_{ti} = \alpha_i + \sum_{j=1}^{p} \sum_{k=1}^{m} y_{t-j,k} \phi_{ji,k} + u_{ti}.
\]

This is just a linear regression, in which \( y_{ti} \) depends on a constant term and \( y_{t-k} \) through \( p \) of all of the \( g \) variables in the system. Thus we see that the VAR (13.87) has the form of a multivariate linear regression model, or SUR model, like the ones we discussed in Section 12.2.

To see this clearly, let us make the definitions

\[
X_t \equiv [1 \ Y_{t-1} \ \cdots \ Y_{t-p}] \quad \text{and} \quad \Pi \equiv \begin{bmatrix} \alpha \\ \Phi_1 \\ \vdots \\ \Phi_p \end{bmatrix}
\]

The row vector \( X_t \) has \( kp+1 \) elements, and the matrix \( \Pi \) is \( k \times g \). With these definitions, the VAR (13.87) becomes

\[
Y_t = X_t \Pi + U_t, \quad U_t \sim \text{iid}(0, \Sigma),
\]

(13.88)
which has the form of a multivariate regression model. In fact, if we stack
the rows, it has precisely the same form as (12.71), which is the unrestricted
reduced form for a linear simultaneous equations model. Thus a VAR can be
thought of as a set of reduced form linear equations relating the endogenous
variables in the vector $Y_t$ to the predetermined variables that are collected in
the vector $X_t$. Except for the constant term, these predetermined variables
are the first $p$ lags of all the endogenous variables themselves.

Estimating a vector autoregression is very easy. As we saw in Section 12.2,
it is appropriate to estimate a linear system like (13.88), in which the same
regressors appear in every equation, by ordinary least squares. In such a
case, OLS is both the efficient GLS estimator and the maximum likelihood
estimator under the assumption of multivariate normal errors. If $\hat{\Sigma}$
denotes the matrix of OLS estimates, it follows from (12.41) that the maximized value
of the loglikelihood function is

$$\frac{\eta n}{2} \left( \log 2\pi + 1 \right) - \frac{n}{2} \log |\hat{\Sigma}|,$$

where

$$\hat{\Sigma} = \frac{1}{n} (Y - X\hat{\Pi})' (Y - X\hat{\Pi}) = \frac{1}{n} \sum_{t=1}^{n} \hat{U}_t' \hat{U}_t,$$

is the ML estimate of the covariance matrix $\Sigma$. Here $Y$ is the $n \times g$
matrix with typical row $Y_t$, $X$ is the $n \times k$ matrix with typical row $X_t$, and $\hat{U}_t$ is
the row vector of OLS residuals for observation $t$. The estimate (13.90) is
often of considerable interest, because it captures the covariances between the
innovations in the various equations.

When specifying a VAR, it is important to determine how many lags need to
be included. If one wishes to test the null hypothesis that the longest lag in
the system is $p$ against the alternative that it is $p+1$, the easiest way to do
so is to compute the LR statistic

$$n \left( \log |\hat{\Sigma}(p)| - \log |\hat{\Sigma}(p+1)| \right),$$

where $\hat{\Sigma}(p)$ and $\hat{\Sigma}(p+1)$ denote the ML estimates of $\Sigma$ for systems with $p$
and $p+1$ lags, respectively; both of these may be computed using (13.90). This test statistic is asymptotically distributed as $\chi^2(g^2)$. However, unless the sample size $n$ is large relative to the number of parameters in the system ($g + pg^2$ under the null, and $g + (p+1)g^2$ under the alternative), the finite-sample distribution of the LR statistic (13.91) may differ substantially from
its asymptotic one. In consequence, this is a case in which it is often very
desirable to compute bootstrap rather than asymptotic $P$ values.

Since there is more than one way to generate bootstrap samples for a VAR, it
is worth saying a bit more about this. We suggest using (13.87) to generate
the data recursively, with OLS estimates under the null replacing the unknown
parameters. The bootstrap error terms are obtained by resampling the row
vectors $\hat{U}_t$, where $\hat{U}_t$ is equal to $(n/(n-1-gp))^{1/2}$ times the row vector
$U_t$ of OLS residuals, and actual pre-sample values of $Y_t$ are used to start
the recursive process of generating the bootstrap data. Limited simulation
evidence suggests that this procedure yields much more accurate $P$ values for
 tests based on (13.91) than using the $\chi^2(g^2)$ distribution.

If we wish to construct confidence intervals for, or test hypotheses about,
individual parameters in a VAR, we can use the OLS standard errors, which
are asymptotically valid. Similarly, if we wish to test hypotheses concerning
two or more parameters in a single equation, we can compute Wald tests in the
usual way based on the OLS covariance matrix for that equation. However, if
we wish to test hypotheses concerning coefficients in two or more equations, we
need the covariance matrix of the parameter estimates for the entire system.

We saw in Chapter 12 that the estimated covariance matrix for the feasible
GLS estimates of a multivariate regression model is given by expression
(12.19), and the one for the ML estimates is given by expression (12.38). These
covariance matrices differ only because they use different estimates of $\Sigma$. As in Section 12.2, we let $X_s \equiv I_g \otimes X$, which is a $gn \times gk$ matrix. Then,
if all the parameters are stacked into a vector of length $gk$, both covariance
matrices have the form

$$X_s' (\hat{\Sigma}^{-1} \otimes I_n) X_s .$$

Using the rules for manipulating Kronecker products given in equations
(12.88), we see that

$$X_s' (\hat{\Sigma}^{-1} \otimes I_n) X_s \sim (I_g \otimes X' \hat{\Sigma}^{-1} \otimes I_n) (I_g \otimes X)^{-1} = \hat{\Sigma} \otimes (X' X)^{-1}.$$

Thus the covariance matrix for all the coefficients of a VAR is easily computed
from $\hat{\Sigma}$, which is given in (13.90), and the inverse of the $X' X$ matrix.

The idea of using vector autoregressions instead of structural models to model
macroeconomic dynamics is often attributed to Sims (1980). Our treatment
has been very brief. For a more detailed introductory treatment, with many
references, see Lütkepohl (2001). For a review of macroeconomic applications
of VARs, see Stock and Watson (2001).

**Granger Causality**

One common use of vector autoregressions is to test the hypothesis that one
or more of the variables in a VAR do not "Granger cause" the others. The
concept of Granger causality was developed by Granger (1969). Other, closely
related, definitions of causality have been suggested, notably by Sims (1972).
Suppose we divide the variables in a VAR into two groups, $Y_{1t}$ and $Y_{2t}$, which
are row vectors of dimensions $g_1$ and $g_2$, respectively. Then we may say that
$Y_2$ does not Granger cause $Y_{1t}$ if the distribution of $Y_{1t}$, conditional on past

---

596 Methods for Stationary Time-Series Data

13.7 Vector Autoregressions

597
values of both $Y_{t1}$ and $Y_{t2}$, is the same as the distribution of $Y_{t1}$ conditional on its own past values.

In practice, it would be very difficult to test whether the entire distribution of $Y_{t1}$ depends on past values of $Y_{t2}$. Therefore, we always test ourselves with asking whether the conditional mean of $Y_{t1}$ depends on past values of $Y_{t2}$. In terms of the VAR (13.87), this is equivalent to imposing restrictions on the equations that correspond to $Y_{t1}$. We can rewrite the VAR as

$$[Y_{t1} \: Y_{t2}] = [\alpha_1 \: \alpha_2] + \sum_{j=1}^{p} [Y_{t-j,1} \: Y_{t-j,2}] \begin{bmatrix} \Phi_{j,11} & \Phi_{j,12} \\ \Phi_{j,21} & \Phi_{j,22} \end{bmatrix} + [U_{t1} \: U_{t2}],$$

where the matrices $\Phi_j$ have been partitioned to conform with the partition of $Y_t$ and its lags. If $Y_{t2}$ does not Granger cause $Y_{t1}$, then all of the $\Phi_{j,21}$ must be zero matrices. Similarly, if $Y_{t1}$ does not Granger cause $Y_{t2}$, then all of the $\Phi_{j,12}$ must be zero matrices.

Since the $\Phi_{j,21}$ appear only in the equations for $Y_{t1}$, it is easy to test the hypothesis that they are all zero. We obtain ML estimates of the two systems of equations

$$Y_{t1} = \alpha_1 + \sum_{j=1}^{p} Y_{t-j,1} \Phi_{j,11} + U_{t1}, \quad (13.92)$$

$$Y_{t2} = \alpha_1 + \sum_{j=1}^{p} (Y_{t-j,1} \Phi_{j,11} + Y_{t-j,2} \Phi_{j,21}) + U_{t1}, \quad (13.93)$$

which may be done using OLS for each equation, and then calculate the value of the loglikelihood function for each of the systems. As in (13.89), the loglikelihood depends only on the estimate of $\Sigma_{11}$, the $g_1 \times g_1$ upper left-hand block of $\Sigma$. This may easily be calculated using the OLS residuals, as in (13.90). We obtain the LR statistic

$$n(\log |\hat{\Sigma}_{11}| - \log |\Sigma_{11}|), \quad (13.94)$$

where $\hat{\Sigma}_{11}$ denotes the estimate of $\Sigma_{11}$ based on the OLS residuals from equations (13.92), and $\Sigma_{11}$ denotes the estimate of $\Sigma_{11}$ based on the OLS residuals from equations (13.93). The statistic (13.94) is asymptotically distributed as $\chi^2(p g_1 g_2)$, but more reliable inferences in finite samples can almost certainly be obtained by bootstrapping.

In practice, we are very commonly interested in testing Granger causality for a single dependent variable. In that case, equations (13.92) and (13.93) are univariate regressions. The restricted model, equation (13.92), becomes a regression of $y_{t1}$ on a constant and $p$ of its own lagged values. The unrestricted model, equation (13.93), adds $p$ lagged values of $y_{t2}$ additional variables to this regression. We can then perform an asymptotic $F$ test of the hypothesis that the $p g_2$ coefficients of the lags of all the additional variables are jointly equal to zero. For this test to be asymptotically valid, the error terms must be homoskedastic. If this assumption does not seem to be correct, we should instead perform a heteroskedasticity-robust test, as discussed in Section 6.8.

Our discussion of Granger causality has been quite brief. Hamilton (1994, Chapter 11) provides a much more detailed discussion of this topic. That book also discusses a number of other aspects of VAR models in more detail than we have done here.

## 13.8 Final Remarks

The analysis of time-series data has engaged the interest of a great many statisticians and econometricians and generated a massive literature. This chapter has provided only a superficial introduction to the subject. In particular, we have said nothing at all about frequency domain methods, because they are a bit too specialized for this book. See Brockwell and Davis (1991), Box, Jenkins, and Reinsel (1994, Chapter 2), Hamilton (1994, Chapter 6), and Fuller (1995), among many others.

This chapter has dealt only with stationary time series. A great many economic time series are, or at least appear to be, nonstationary. Therefore, in the next chapter, we turn our attention to methods for dealing with nonstationary time series. Such methods have been a subject of an enormous amount of research in econometrics during the past two decades.

## 13.9 Exercises

13.1 Show that the solution to the Yule-Walker equations (13.07) for the AR(2) process is given by equations (13.08).

13.2 Demonstrate that the first $p + 1$ Yule-Walker equations for the AR($p$) process $u_t = \sum_{i=1}^{p} \rho_i u_{t-i} + \varepsilon_t$ are

$$u_0 - \sum_{i=1}^{p} \rho_i v_i = \sigma^2, \quad \text{and} \quad \rho_1 u_0 - \rho_1 v_1 + \sum_{j=1,j \neq i}^{p} \rho_j v_{i-j} = 0, \quad i = 1, \ldots, p. \quad (13.95)$$

Then rewrite these equations using matrix notation.

13.3 Consider the AR(2) process

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \varepsilon_t,$$
4.6. Generalize Exercise 4.5 to deduce that if one adds together an AR(p) process with an MA(q) process and if these two processes are uncorrelated with each other at all leads and lags, then the result is an ARMA(p, p + q) process.

Chapter 4 References


5. Maximum Likelihood Estimation

5.1. Introduction

Consider an ARMA model of the form

\[ Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q}, \]  

with \( \epsilon \), white noise:

\[ E(\epsilon_t) = 0 \]  

\[ E(\epsilon_t \epsilon_{t-}\tau) = \begin{cases} \sigma^2 & \text{for } \tau = 0 \\ 0 & \text{otherwise}. \end{cases} \]

The previous chapters assumed that the population parameters \( (c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma^2) \) were known and showed how population moments such as \( E(Y_{t-1}) \) and linear forecasts \( E(Y_t | Y_{t-1}, Y_{t-2}, \ldots) \) could be calculated as functions of these population parameters. This chapter explores how to estimate the values of \( (c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma^2) \) on the basis of observations on \( Y \).

The primary principle on which estimation will be based is maximum likelihood. Let \( \theta = (c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma^2) \) denote the vector of population parameters. Suppose we have observed a sample of size \( T \) \( (y_1, y_2, \ldots, y_T) \). The approach will be to calculate the probability density

\[ f_{y_1, y_2, \ldots, y_T}(y_1, y_2, \ldots, y_T; \theta), \]

which might loosely be viewed as the probability of having observed this particular sample. The maximum likelihood estimate (MLE) of \( \theta \) is the value for which this sample is most likely to have been observed; that is, it is the value of \( \theta \) that maximizes \( f_{y_1, y_2, \ldots, y_T}(y_1, y_2, \ldots, y_T; \theta) \).

This approach requires specifying a particular distribution for the white noise process \( \epsilon \). Typically we will assume that \( \epsilon \) is Gaussian white noise:

\[ \epsilon_t \sim \text{i.i.d. } N(0, \sigma^2). \]

Although this assumption is strong, the estimates of \( \theta \) that result from it will often turn out to be sensible for non-Gaussian processes as well.

Finding maximum likelihood estimates conceptually involves two steps. First, the likelihood function \( f_{y_1, y_2, \ldots, y_T}(y_1, y_2, \ldots, y_T; \theta) \) must be calculated. Second, values of \( \theta \) must be found that maximize this function. This chapter is organized around these two steps. Sections 5.2 through 5.6 show how to calculate the likelihood function for different Gaussian ARMA specifications, while subsequent sections review general techniques for numerical optimization.
5.2. The Likelihood Function for a Gaussian AR(1) Process

Evaluating the Likelihood Function
A Gaussian AR(1) process takes the form
\[ Y_t = \phi Y_{t-1} + \epsilon_t, \quad [5.2.1] \]
with \( \epsilon_t \sim \text{i.i.d.} \ N(0, \sigma^2) \). For this case, the vector of population parameters to be estimated consists of \( \theta = (\phi, \sigma^2)^\top \).

Consider the probability distribution of \( Y_1 \), the first observation in the sample. From equations [3.4.3] and [3.4.4] this is a random variable with mean
\[ E(Y_1) = \mu = c(1 - \phi) \]
and variance
\[ E(Y_1 - \mu)^2 = \sigma^2(1 - \phi^2). \]

Since \( \{\epsilon_t\}_{t=0}^{\infty} \) is Gaussian, \( Y_1 \) is also Gaussian. Hence, the density of the first observation takes the form
\[ f_{Y_1}(y_1; \theta) = \frac{1}{\sqrt{2\pi\sigma^2(1 - \phi^2)}} \exp\left(\frac{-(y_1 - [c(1 - \phi)])^2}{2\sigma^2(1 - \phi^2)}\right). \] \[ [5.2.2] \]

Next consider the distribution of the second observation \( Y_2 \) conditional on observing \( Y_1 = y_1 \). From [5.2.1],
\[ Y_2 = c + \phi Y_1 + \epsilon_2. \] \[ [5.2.3] \]
Conditioning on \( Y_1 = y_1 \) means treating the random variable \( Y_1 \) as if it were the deterministic constant \( y_1 \). For this case, [5.2.3] gives \( Y_2 \) as the constant \( (c + \phi y_1) \) plus the \( N(0, \sigma^2) \) variable \( \epsilon_2 \). Hence,
\[ \{Y_2|Y_1 = y_1\} \sim N(c + \phi y_1, \sigma^2), \]
meaning
\[ f_{Y_2|Y_1}(y_2|y_1; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y_2 - [c + \phi y_1])^2}{2\sigma^2}\right). \] \[ [5.2.4] \]

The joint density of observations 1 and 2 is then just the product of [5.2.4] and [5.2.2]:
\[ f_{Y_1Y_2}(y_1, y_2; \theta) = f_{Y_2|Y_1}(y_2|y_1; \theta)f_{Y_1}(y_1; \theta). \]

Similarly, the distribution of the third observation conditional on the first two is
\[ f_{Y_3|Y_1Y_2}(y_3|y_1y_2; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y_3 - [c + \phi y_1])^2}{2\sigma^2}\right), \]
from which
\[ f_{Y_1Y_2Y_3}(y_1, y_2, y_3; \theta) = f_{Y_3|Y_1Y_2}(y_3|y_1y_2; \theta)f_{Y_2|Y_1}(y_2|y_1; \theta)f_{Y_1}(y_1; \theta). \]

In general, the values of \( Y_1, Y_2, \ldots, Y_{t-1} \) matter for \( Y_t \) only through the value of \( Y_{t-1} \), and the density of observation \( t \) conditional on the preceding \( t - 1 \) observations is given by
\[ f_{Y_t|Y_{t-1}, \ldots, Y_1}(y_t|y_{t-1}, \ldots, y_1; \theta) \]
\[ = f_{Y_t|Y_{t-1}}(y_t|y_{t-1}; \theta) \]
\[ = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y_t - [c + \phi y_{t-1}])^2}{2\sigma^2}\right). \] \[ [5.2.5] \]

The joint density of the first \( t \) observations is then
\[ f_{Y_1Y_2\ldots Y_t}(y_1, y_2, \ldots, y_t; \theta) = f_{Y_t|Y_{t-1}}(y_t|y_{t-1}; \theta)f_{Y_{t-1}|Y_{t-2}}(y_{t-1}|y_{t-2}; \theta)\ldots f_{Y_2|Y_1}(y_2|y_1; \theta) f_{Y_1}(y_1; \theta). \] \[ [5.2.6] \]

The likelihood of the complete sample can thus be calculated as
\[ f_{Y_1Y_2\ldots Y_T}(y_1, y_2, \ldots, y_T; \theta) = f_{Y_T|Y_{T-1}}(y_T|y_{T-1}; \theta)\prod_{t=2}^T f_{Y_t|Y_{t-1}}(y_t|y_{t-1}; \theta). \] \[ [5.2.7] \]

The log likelihood function (denoted \( \mathcal{L}(\theta) \)) can be found by taking logs of [5.2.7]:
\[ \mathcal{L}(\theta) = \log f_{Y_1}(y_1; \theta) + \sum_{t=2}^T \log f_{Y_t|Y_{t-1}}(y_t|y_{t-1}; \theta). \] \[ [5.2.8] \]

Clearly, the value of \( \theta \) that maximizes [5.2.8] is identical to the value that maximizes [5.2.7]. However, Section 5.8 presents a number of useful results that can be calculated as a by-product of the maximization if one always poses the problem as maximization of the log likelihood function [5.2.8] rather than the likelihood function [5.2.7].

Substituting [5.2.2] and [5.2.3] into [5.2.8], the log likelihood for a sample of size \( T \) from a Gaussian AR(1) process is seen to be
\[ \mathcal{L}(\theta) = -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2(1 - \phi^2)) \]
\[ - \frac{1}{2\sigma^2(1 - \phi^2)} \left( \frac{[(T - 1)/2] \log(2\pi)}{2\sigma^2(1 - \phi^2)} \right) \]
\[ - \frac{ [(T - 1)/2] \log(\sigma^2) - \frac{1}{2\sigma^2} \left[ (y_T - c - \phi y_{T-1})^2 \right] }{2\sigma^2}. \] \[ [5.2.9] \]

An Alternative Expression for the Likelihood Function
A different description of the likelihood function for a sample of size \( T \) from a Gaussian AR(1) process is sometimes useful. Collect the full set of observations in a \((T \times 1)\) vector,
\[ \mathbf{y} = (y_1, y_2, \ldots, y_T)^\top. \]
This vector could be viewed as a single realization from a \( T \)-dimensional Gaussian distribution. The mean of this \((T \times 1)\) vector is
\[ \mathbf{E(Y)} = \begin{bmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_T) \end{bmatrix} = \begin{bmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{bmatrix}, \] \[ [5.2.10] \]
where, as before, \( \mu = c(1 - \phi) \). In vector form, [5.2.10] could be written
\[ \mathbf{E(Y)} = \mu, \]
where \( \mu \) denotes the \((T \times 1)\) vector on the right side of [5.2.10]. The variance-covariance matrix of \( Y \) is given by
\[ \operatorname{E}[(Y - \mu)(Y - \mu)^\top] = \Omega, \] \[ [5.2.11] \]
where

\[
\Omega = \begin{bmatrix}
E(Y_1 - \mu)E(Y_1 - \mu) & \cdots & E(Y_1 - \mu)(Y_T - \mu) \\
E(Y_2 - \mu)(Y_1 - \mu) & \cdots & E(Y_2 - \mu)(Y_T - \mu) \\
\vdots & \ddots & \vdots \\
E(Y_T - \mu)(Y_1 - \mu) & \cdots & E(Y_T - \mu)(Y_T - \mu)
\end{bmatrix}
\]  

The elements of this matrix correspond to autocovariances of \(Y\). Recall that the \(j\)th autocovariance for an AR(1) process is given by

\[
E(Y_j - \mu)(Y_{j-i} - \mu) = \sigma^2 \phi^j/(1 - \phi^2).
\]  

Hence, [5.2.12] can be written as

\[
\Omega = \sigma^2V,
\]  

where

\[
V = \frac{1}{1 - \phi^2}
\begin{bmatrix}
1 & \phi & \phi^2 & \cdots & \phi^{T-1} \\
\phi & 1 & \phi & \cdots & \phi^{T-2} \\
\phi^2 & \phi & 1 & \cdots & \phi^{T-3} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\phi^{T-1} & \phi^{T-2} & \phi^{T-3} & \cdots & 1
\end{bmatrix}
\]  

Viewing the observed sample \(y\) as a single draw from a \(N(\mu, \Omega)\) distribution, the sample likelihood could be written down immediately from the formula for the multivariate Gaussian density:

\[
f_y(y; \theta) = (2\pi)^{-T/2} |\Omega|^{-1/2} \exp[-(y - \mu)^T \Omega^{-1} (y - \mu)].
\]  

with log likelihood

\[
\ell(\theta) = -(T/2) \log(2\pi) + \frac{1}{2} \log|\Omega| - \frac{1}{2} (y - \mu)^T \Omega^{-1} (y - \mu).
\]  

Evidently, [5.2.17] and [5.2.9] must represent the identical function of \((y_1, y_2, \ldots, y_T)\). To verify that this is indeed the case, define

\[
L = \begin{bmatrix}
\sqrt{1 - \phi^2} & 0 & 0 & \cdots & 0 & 0 \\
-\phi & 1 & 0 & \cdots & 0 & 0 \\
0 & -\phi & 1 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 1 & 0
\end{bmatrix}
\]  

It is straightforward to show that

\[
L^T L = V^{-1},
\]  

\footnote{By direct multiplication, one calculates

\[
LV = \frac{1}{1 - \phi^2}
\begin{bmatrix}
\sqrt{1 - \phi^2} & \phi \sqrt{1 - \phi^2} & \phi \sqrt{1 - \phi^2} & \cdots & \phi^{T-1} \sqrt{1 - \phi^2} \\
0 & (1 - \phi^2) & (1 - \phi^2) & \cdots & \phi^{T-2} (1 - \phi^2) \\
0 & 0 & (1 - \phi^2) & \cdots & \phi^{T-3} (1 - \phi^2) \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & (1 - \phi^2)
\end{bmatrix}
\]  

and premultiplying this by \(L^T\) produces the \((T \times T)\) identity matrix. Thus, \(L^T L = I\), confirming [5.2.19].}

implying from [5.2.14] that

\[
\Omega^{-1} = \sigma^{-2}L^T L.
\]  

Substituting [5.2.20] into [5.2.17] results in

\[
\ell(\theta) = -(T/2) \log(2\pi) + \frac{1}{2} \log|\sigma^{-2}L^T L| - \frac{1}{2} (y - \mu)^T \sigma^{-2}L^T L (y - \mu).
\]  

Define the \((T \times 1)\) vector \(\hat{y}\) to be

\[
\hat{y} = L(y - \mu) = \begin{bmatrix}
\sqrt{1 - \phi^2} & 0 & 0 & \cdots & 0 & y_1 - \mu \\
-\phi & 1 & 0 & \cdots & 0 & y_2 - \mu \\
0 & -\phi & 1 & \cdots & 0 & y_3 - \mu \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & -\phi & 1 & y_T - \mu
\end{bmatrix}
\]  

Substituting \(\mu = c(1 - \phi)\), this becomes

\[
\hat{y} = \begin{bmatrix}
\sqrt{1 - \phi^2} (y_1 - c(1 - \phi)) \\
-\phi c(1 - \phi) y_2 \\
0 & -\phi c(1 - \phi) y_3 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & -\phi c y_{T-1}
\end{bmatrix}
\]  

The last term in [5.2.21] can thus be written

\[
(y - \mu)^T \sigma^{-2}L^T L (y - \mu) = \frac{1}{2(2\pi)} \hat{y}^T \hat{y}
\]  

\[
= \frac{1}{2(2\pi)} [1 - \phi^2] \sum_{i=1}^{T} (y_i - c + \phi y_{i-1})^2
\]  

The middle term in [5.2.21] is similarly

\[
\frac{1}{2} \log|\sigma^{-2}L^T L| = \frac{1}{2} \log|\sigma^{-2}L^T L|
\]

\[
= - \frac{1}{2} \log \sigma^2 + \frac{1}{2} \log|L^T L|
\]

\[
= (T/2) \log \sigma^2 + \frac{1}{2} \log(1 - \phi^2).
\]  

where use has been made of equations [A.4.8], [A.4.9], and [A.4.11] in the Mathematical Review (Appendix A) at the end of the book. Moreover, since \(L\) is lower triangular, its determinant is given by the product of the terms along the principal diagonal: \(|L| = \sqrt{1 - \phi^2}\). Thus, [5.2.24] states that

\[
\frac{1}{2} \log|\sigma^{-2}L^T L| = (T/2) \log \sigma^2 + \frac{1}{2} \log(1 - \phi^2).
\]  

Substituting [5.2.23] and [5.2.25] into [5.2.21] reproduces [5.2.9]. Thus, equations [5.2.17] and [5.2.9] are just two different expressions for the same magnitude, as claimed. Either expression accurately describes the log likelihood function.

5.2. The Likelihood Function for a Gaussian AR(1) Process
5.3. The Likelihood Function for a Gaussian AR(p) Process

This section discusses a Gaussian AR(p) process. The log-likelihood function for a Gaussian AR(p) process is given by

\[\ell(\theta) = -\frac{1}{2} \log(2\pi) - \frac{1}{2} \sum_{t=2}^{T} (x_t - \phi_1 x_{t-1} - \cdots - \phi_p x_{t-p})^2\]

or

\[\ell(\theta) = -\frac{1}{2} \sum_{t=2}^{T} (x_t - \phi_1 x_{t-1} - \cdots - \phi_p x_{t-p})^2\]

where \(\theta = (\phi_1, \phi_2, \ldots, \phi_p, \sigma^2)\) are the parameters to be estimated.

The likelihood function is maximized by finding the values of \(\phi_1, \phi_2, \ldots, \phi_p\) and \(\sigma^2\) that maximize \(\ell(\theta)\). This is typically done using numerical optimization algorithms.

5.3.1. Evaluating the Likelihood Function

Let \(\Sigma_{y} = \sum_{t=1}^{T} E(Y_t | Y_{t-1}, \ldots, Y_1)\) be the conditional covariance matrix of the observations. The maximum likelihood estimate of \(\phi_1, \phi_2, \ldots, \phi_p\) is found by maximizing the log-likelihood function with respect to these parameters.

5.3.2. The Likelihood Function

The likelihood function for a Gaussian AR(p) process is given by

\[\ell(\theta) = -\frac{1}{2} \sum_{t=2}^{T} (x_t - \phi_1 x_{t-1} - \cdots - \phi_p x_{t-p})^2\]

with \(\phi_1, \phi_2, \ldots, \phi_p\) as the parameters to be estimated.

For example, for a first-order autoregression (\(p = 1\)), \(\phi_1 = 0\) is the value of \(\phi_1\) that maximizes the log-likelihood function.

5.3.3. Maximum Likelihood Estimation

The maximum likelihood estimates of \(\phi_1, \phi_2, \ldots, \phi_p\) are found by maximizing the log-likelihood function with respect to these parameters. This is typically done using numerical optimization algorithms.

5.3.4. Properties of the Maximum Likelihood Estimator

The maximum likelihood estimator of \(\phi_1, \phi_2, \ldots, \phi_p\) is consistent and asymptotically normal. The asymptotic covariance matrix of the maximum likelihood estimator is given by

\[\hat{\Sigma}_\phi = \left(\frac{1}{T} \sum_{t=2}^{T} E(Y_t | Y_{t-1}, \ldots, Y_1)\right)^{-1}\]

where \(\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_p\) are the maximum likelihood estimates of \(\phi_1, \phi_2, \ldots, \phi_p\), respectively.
where \( \gamma \), the \( j \)th autocovariance for an AR(\( p \)) process, can be calculated using the methods in Chapter 3. The density of the first \( p \) observations is then that of a \( N(\mu_p, \sigma^2 V_p) \) variable:

\[
\begin{align*}
\gamma_{i} & = (2\pi)^{-\frac{3p}{2}}\sigma^{3p} \exp \left[ -\frac{1}{2\sigma^2}(y_p - \mu_p)^T V_p^{-1} (y_p - \mu_p) \right] \\
& = (2\pi)^{-\frac{3p}{2}}(\sigma^{-3})^{p!} \exp \left[ -\frac{1}{2\sigma^2}(y_p - \mu_p)^T V_p^{-1} (y_p - \mu_p) \right].
\end{align*}
\]

where use has been made of result [A.4.8].

For the remaining observations in the sample, \( (y_{p+1}, y_{p+2}, \ldots, y_t) \), the prediction-error decomposition can be used. Conditional on the first \( t \) – 1 observations, the \( t \)th observation is Gaussian with mean

\[ c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} \]

and variance \( \sigma^2 \). Only the \( p \) most recent observations matter for this distribution. Hence, for \( t > p \),

\[
\begin{align*}
\gamma_{t} & = f_{y_t, y_{t-1}, \ldots, y_{p+1}}(y_t, y_{t-1}, \ldots, y_{p+1}; \theta) \\
& = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(y_t - c - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \cdots - \phi_p y_{t-p})^2}{2\sigma^2} \right].
\end{align*}
\]

The likelihood function for the complete sample is then

\[
\begin{align*}
\gamma & = f_{y_t, y_{t-1}, \ldots, y_{p+1}}(y_t, y_{t-1}, \ldots, y_{p+1}; \theta) \\
& \times \prod_{r=p+1}^{t-1} f_{y_r, y_{r-1}, \ldots, y_{p+1}}(y_r, y_{r-1}, \ldots, y_{p+1}; \theta),
\end{align*}
\]

and the log likelihood is therefore

\[
\begin{align*}
\mathcal{L}(\theta) & = \log f_{y_t, y_{t-1}, \ldots, y_{p+1}}(y_t, y_{t-1}, \ldots, y_{p+1}; \theta) \\
& = -\frac{p}{2} \log(2\pi) - \frac{p}{2} \log(\sigma^2) + \frac{1}{2} \log|V_p^{-1}| \\
& \quad - \frac{1}{2\sigma^2}(y_p - \mu_p)^T V_p^{-1} (y_p - \mu_p) \\
& \quad - \frac{T - p}{2} \log(2\pi) - \frac{T - p}{2} \log(\sigma^2) \\
& \quad - \frac{T}{2} \sum_{i=p+1}^{t-1} \log(2\pi) - \frac{T}{2} \log(\sigma^2) \\
& = -\frac{T}{2} \log(2\pi) - \frac{T}{2} \log(\sigma^2) + \frac{1}{2} \log|V_p^{-1}| \\
& \quad - \frac{1}{2\sigma^2}(y_p - \mu_p)^T V_p^{-1} (y_p - \mu_p) \\
& \quad - \frac{T}{2} \sum_{i=p+1}^{t-1} \log|V_i^{-1}| \\
& \quad - \frac{T}{2} \sum_{i=p+1}^{t-1} \log(2\pi) - \frac{T}{2} \log(\sigma^2). \\
\end{align*}
\]

Evaluation of [5.3.6] requires inverting the \((p \times p)\) matrix \( V_p \). Denote the row \( i \), column \( j \) element of \( V_p^{-1} \) by \( v(i, j) \). Galbraith and Galbraith (1974, equation 16, p. 70) showed that

\[
v(i, j) = \begin{bmatrix} -\sum_{k=0}^{i-1} \phi_k & \sum_{k=p+1}^{i-1} \phi_k \\
\sum_{k=p+1}^{i-1} \phi_k & -\sum_{k=p+1}^{i-1} \phi_k \end{bmatrix} \text{ for } 1 \leq i \leq j \leq p, \tag{5.3.7}
\]

where \( \phi_p = -1 \). Values of \( v(i, j) \) for \( i > j \) can be inferred from the fact that \( V_p^{-1} \) is symmetric \( (v(i, j) = v(j, i)) \). For example, for an AR(1) process, \( V_p^{-1} \) is a scalar whose value is found by taking \( i = j = p = 1 \):

\[
\begin{align*}
V_1^{-1} & = \begin{bmatrix} \phi_1 - \phi_2 & -(\phi_1 + \phi_2) \\
-(\phi_1 + \phi_2) & (1 - \phi_2) \end{bmatrix} \\
& = \begin{bmatrix} \phi_1 - \phi_2 \\
-(\phi_1 + \phi_2) \end{bmatrix}
\end{align*}
\]

Thus \( \sigma^2 V_1 = \sigma^2 \begin{bmatrix} 1 - \phi_2 \\
0 \end{bmatrix} \), which indeed reproduces the formula for the variance of an AR(1) process. For \( p = 2 \), equation [5.3.7] implies

\[
V_2^{-1} = \begin{bmatrix} (1 - \phi_2) & -(\phi_1 + \phi_2) \\
-(\phi_1 + \phi_2) & (1 - \phi_2) \end{bmatrix}
\]

and

\[
\begin{align*}
& \quad (y_2 - \mu_2)^T V_2^{-1} (y_2 - \mu_2) \\
& = [(y_1 - \mu) - (y_2 - \mu)](1 + \phi_2)(1 - \phi_1) \left[ (1 - \phi_2) \right] (y_2 - \mu) \\
& = (1 + \phi_2) \times ((1 - \phi_2)(y_1 - \mu)^2 - 2\phi(y_2 - \mu))(y_2 - \mu). \\
\end{align*}
\]

The exact log likelihood for a Gaussian AR(2) process is thus given by

\[
\mathcal{L}(\theta) = -\frac{T}{2} \log(2\pi) - \frac{T}{2} \log(\sigma^2) + \frac{1}{2} \log((1 + \phi_2)^2 - \phi_1^2) \\
- \frac{1 + \phi_2}{2\sigma^2} \times ((1 - \phi_1)(y_1 - \mu)^2 - 2\phi(y_2 - \mu))(y_2 - \mu) \\
- \frac{T}{2} \sum_{i=p+1}^{t-1} \log(2\pi) - \frac{T}{2} \log(\sigma^2),
\]

where \( \mu = c(1 - \phi_1 - \phi_2) \).

**Conditional Maximum Likelihood Estimates**

Maximization of the exact log likelihood for an AR(\( p \)) process [5.3.6] must be accomplished numerically. In contrast, the log of the likelihood conditional on the first \( p \) observations assumes the simple form

\[
\begin{align*}
\log f_{y_t, y_{t-1}, \ldots, y_{p+1}|y_{p+1}, \ldots, y_{p+1}}(y_p, y_{p+1}, \ldots, y_t; \theta) & = -\frac{T - p}{2} \log(2\pi) - \frac{T - p}{2} \log(\sigma^2) \\
& \quad - \frac{T}{2} \sum_{i=p+1}^{t-1} \log(2\pi) - \frac{T}{2} \log(\sigma^2) \\
& \quad - \frac{T}{2} \sum_{i=p+1}^{t-1} \log(2\pi) - \frac{T}{2} \log(\sigma^2) \\
& \quad - \frac{T}{2} \sum_{i=p+1}^{t-1} \log(2\pi) - \frac{T}{2} \log(\sigma^2) \\
& \quad - \frac{T}{2} \sum_{i=p+1}^{t-1} \log(2\pi) - \frac{T}{2} \log(\sigma^2) \\
& \quad - \frac{T}{2} \sum_{i=p+1}^{t-1} \log(2\pi) - \frac{T}{2} \log(\sigma^2) \\
& \quad - \frac{T}{2} \sum_{i=p+1}^{t-1} \log(2\pi) - \frac{T}{2} \log(\sigma^2) \\
& \quad - \frac{T}{2} \sum_{i=p+1}^{t-1} \log(2\pi) - \frac{T}{2} \log(\sigma^2) \\
& \quad - \frac{T}{2} \sum_{i=p+1}^{t-1} \log(2\pi) - \frac{T}{2} \log(\sigma^2) \\
& \quad - \frac{T}{2} \sum_{i=p+1}^{t-1} \log(2\pi) - \frac{T}{2} \log(\sigma^2) \\
& \quad - \frac{T}{2} \sum_{i=p+1}^{t-1} \log(2\pi) - \frac{T}{2} \log(\sigma^2) \\
\end{align*}
\]

5.3. The Likelihood Function for a Gaussian AR(p) Process 125
The values of \( c, \phi_1, \phi_2, \ldots, \phi_p \) that maximize [5.3.9] are the same as those that minimize

\[
\sum_{i=p+1}^{T} (y_i - c - \phi_1 y_{i-1} - \phi_2 y_{i-2} - \cdots - \phi_p y_{i-p})^2. \tag{5.3.10}
\]

Thus, the conditional maximum likelihood estimates of these parameters can be obtained from an OLS regression of \( y_i \) on a constant and \( p \) of its own lagged values. The conditional maximum likelihood estimate of \( \sigma^2 \) turns out to be the average squared residual from this regression:

\[
\hat{\sigma}^2 = \frac{1}{T - p} \sum_{i=p+1}^{T} (y_i - \hat{c} - \hat{\phi}_1 y_{i-1} - \hat{\phi}_2 y_{i-2} - \cdots - \hat{\phi}_p y_{i-p})^2.
\]

The exact maximum likelihood estimates and the conditional maximum likelihood estimates again have the same large-sample distribution.

### Maximum Likelihood Estimation for Non-Gaussian Time Series

We noted in Chapter 4 that an OLS regression of a variable on a constant and \( p \) of its lags would yield a consistent estimate of the coefficients of the linear projection,

\[ \hat{E}(Y_i | Y_{i-1}, Y_{i-2}, \ldots, Y_{i-p}), \]

provided that the process is ergodic for second moments. This OLS regression also maximizes the Gaussian conditional log likelihood [5.3.9]. Thus, even if the process is non-Gaussian, if we mistakenly form a Gaussian log likelihood function and maximize it, the resulting estimates (\( \hat{c}, \hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_p \)) will provide consistent estimates of the population parameters in [5.3.1].

An estimate that maximizes a misspecified likelihood function (for example, an MLE calculated under the assumption of a Gaussian process when the true data are non-Gaussian) is known as a quasi-maximum likelihood estimate. Sometimes, as turns out to be the case here, quasi-maximum likelihood estimation provides consistent estimates of the population parameters of interest. However, standard errors for the estimated coefficients that are calculated under the Gaussianity assumption need not be correct if the true data are non-Gaussian.2

Alternatively, if the raw data are non-Gaussian, sometimes a simple transformation such as taking logs will produce a Gaussian time series. For a positive random variable \( Y_i \), Box and Cox (1964) proposed the general class of transformations

\[
Y_i^{(\lambda)} = \begin{cases} 
  Y_i^{1 - \lambda} / \lambda & \text{for } \lambda \neq 0 \\
  \log Y_i & \text{for } \lambda = 0.
\end{cases}
\]

One approach is to pick a particular value of \( \lambda \) and maximize the likelihood function for \( Y_i^{(\lambda)} \) under the assumption that \( Y_i^{(\lambda)} \) is a Gaussian ARMA process. The value of \( \lambda \) that is associated with the highest value of the maximized likelihood is taken as the best transformation. However, Nelson and Granger (1979) reported discouraging results from this method in practice.

2These points were first raised by White (1982) and are discussed further in Sections 5.8 and 14.4.

### 5.4. The Likelihood Function for a Gaussian MA(1) Process

#### Conditional Likelihood Function

Calculation of the likelihood function for an autoregression turned out to be much simpler if we conditioned on initial values for the \( Y_i \)'s. Similarly, calculation of the likelihood function for a moving average process is simpler if we condition on initial values for the \( e_i \)'s.

Consider the Gaussian MA(1) process

\[ Y_i = \mu + e_i + \theta e_{i-1} \tag{5.4.1}\]

with \( e_i \) i.i.d. \( N(0, \sigma^2) \). Let \( \theta = (\mu, \sigma^2) \) denote the population parameters to be estimated. If the value of \( e_{i-1} \) were known with certainty, then

\[ Y_i | e_{i-1} \sim N(\mu, \theta e_{i-1}) \]

or

\[
\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(y_i - \mu - \theta e_{i-1})^2}{2\sigma^2} \right). \tag{5.4.2}
\]

Suppose that we knew for certain that \( e_0 = 0 \). Then

\[ (Y_i | e_0 = 0) \sim N(\mu, \sigma^2). \]

Moreover, given observation of \( y_i \), the value of \( e_i \) is then known with certainty as well:

\[ e_i = y_i - \mu, \]

allowing application of [5.4.2] again:

\[
\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(y_i - \mu - \theta e_{i-1})^2}{2\sigma^2} \right).
\]

Since \( e_i \) is known with certainty, \( e_2 \) can be calculated from

\[ e_2 = y_2 - \mu - \theta e_1. \]

Proceeding in this fashion, it is clear that given knowledge that \( e_0 = 0 \), the full sequence \( \{e_1, e_2, \ldots, e_T\} \) can be calculated from \( \{y_1, y_2, \ldots, y_T\} \) by iterating on

\[ e_i = y_i - \mu - \theta e_{i-1} \tag{5.4.3}\]

for \( i = 1, 2, \ldots, T \), starting from \( e_0 = 0 \). The conditional density of the \( n \)th observation can then be calculated from [5.4.2] as

\[
\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{e_n^2}{2\sigma^2} \right), \tag{5.4.4}
\]

where

\[
f_{y_n}(y_n | e_{n-1}; \theta) = f_{y_n}(y_n | e_{n-1}; \theta)
\]

for \( n = 0, 1, \ldots, T \).
The sample likelihood would then be the product of these individual densities:

\[ f_{Y_i|Y_{i-1}, \ldots, Y_1|Y_0 = 0} = f_{Y_i|Y_{i-1}}(Y_i|Y_{i-1}, \ldots, Y_1, Y_0 = 0; \theta) \]

The conditional log likelihood is

\[ L(\theta) = \log f_{Y_i|Y_{i-1}, \ldots, Y_1|Y_0 = 0}(Y_i|Y_{i-1}, \ldots, Y_1, Y_0 = 0; \theta) \]

\[ = -\frac{T}{2} \log(2\pi) - \frac{T}{2} \log(\sigma^2) - \sum_{i=1}^{T} \frac{1}{2} \left( Y_i - \mu \right)^2 \omega_{\theta, i}. \]

For a particular numerical value of \( \theta \), we thus calculate the sequence of \( s_i \) implied by the data from [5.4.3]. The conditional log likelihood [5.4.4] is then a function of the sum of squares of these \( s_i \)'s. Although it is simple to program this iteration by computer, the log likelihood is a fairly complicated nonlinear function of \( \mu \) and \( \theta \), so that an analytical expression for the maximum likelihood estimates of \( \mu \) and \( \theta \) is not readily calculated. Hence, even the conditional maximum likelihood estimates for an MA(1) process must be found by numerical optimization.

Iteration on [5.4.3] from an arbitrary starting value of \( s_0 \) will result in

\[ s_i = (Y_i - \mu) - \theta(Y_{i-1} - \mu) + \theta^2(Y_{i-2} - \mu) + \cdots + (-1)^{i-1} \theta^{i-1}(Y_1 - \mu) + (-1)^i \theta s_0. \]

If \( |\theta| \) is substantially less than unity, the effect of imposing \( s_0 = 0 \) will quickly die out and the conditional likelihood [5.4.4] will give a good approximation to the unconditional likelihood for a reasonably large sample size. By contrast, if \( |\theta| \geq 1 \), the consequences of imposing \( s_0 = 0 \) accumulate over time. The conditional approach is not reasonable in such a case. If numerical optimization of [5.4.3] results in a value of \( \theta \) that exceeds 1 in absolute value, the results must be discarded. The numerical optimization should be attempted again with the reciprocal of \( \theta \) used as a starting value for the numerical search procedure.

**Exact Likelihood Function**

Two convenient algorithms are available for calculating the exact likelihood function for a Gaussian MA(1) process. One approach is to use the Kalman filter discussed in Chapter 13. A second approach uses the triangular factorization of the variance-covariance matrix. The second approach is described here.

As in Section 5.2, the observations on \( y \) can be collected in a \((T \times 1)\) vector \( y = (y_1, y_2, \ldots, y_T)' \) with mean \( \mu = (\mu, \mu, \ldots, \mu)' \) and \((T \times T)\) variance-covariance matrix

\[ \Omega = \Sigma \mu - \mu \Sigma \mu \Sigma \mu. \]

The variance-covariance matrix for \( T \) consecutive draws from an MA(1) process is

\[ \Omega = \sigma_t^2 \begin{bmatrix} (1 + \theta^2) & \theta & 0 & \cdots & 0 \\ \theta & (1 + \theta^2) & \theta & \cdots & 0 \\ 0 & \theta & (1 + \theta^2) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & (1 + \theta^2) \end{bmatrix}. \]

The likelihood function is then

\[ f(y; \theta) = (2\pi)^{-T/2} \left| \Omega \right|^{-1/2} \exp \left\{ -\frac{1}{2} (y - \mu)' \Omega^{-1}(y - \mu) \right\}. \]

A prediction-error decomposition of the likelihood is provided from the triangular factorization of \( \Omega \),

\[ \Omega = \Omega_A, \]

where \( A \) is the lower triangular matrix given in [4.5.18] and \( D \) is the diagonal matrix in [4.5.19]. Substituting [5.4.7] into [5.4.6] gives

\[ f(y; \theta) = (2\pi)^{-T/2} |\Omega_A|^{-1/2} \times \exp \left\{ -\frac{1}{2} (y - \mu)' \Omega_A^{-1}(y - \mu) \right\}. \]

But \( A \) is a lower triangular matrix with 1s along the principal diagonal. Hence, \( |A| = 1 \) and

\[ |\Omega_A| = |A| |D| |A| = |D|. \]

Further defining

\[ \tilde{y} = A^{-1}(y - \mu), \]

the likelihood [5.4.8] can be written

\[ f(y; \theta) = (2\pi)^{-T/2} |D|^{-1/2} \exp \left\{ -\frac{1}{2} \tilde{y}' D^{-1} \tilde{y} \right\}. \]

Notice that [5.4.9] implies

\[ A \tilde{y} = y - \mu. \]

The first row of this system states that \( \tilde{y}_1 = y_1 - \mu \), while the \( t \)th row implies that

\[ \tilde{y}_t = y_t - \mu - \frac{\theta}{1 + \theta^2 + \theta^4 + \cdots + \theta^{2(t-1)}} \tilde{y}_{t-1}. \]

The vector \( \tilde{y} \) can thus be calculated by iterating on [5.4.11] for \( t = 2, 3, \ldots, T \) starting from \( \tilde{y}_1 = y_1 - \mu \). The variable \( \tilde{y}_t \) has the interpretation as the residual from a linear projection of \( y \) onto a constant and \( y_{t-1}, y_{t-2}, \ldots, y_1 \), while the \( t \)th diagonal element of \( D \) gives the MSE of this linear projection:

\[ d_t = E(\tilde{y}_t^2) = \sigma_t^2 + \frac{1 + \theta^2 + \theta^4 + \cdots + \theta^{2(t-1)}}{1 + \theta^2 + \theta^4 + \cdots + \theta^{2T-1}} d_1. \]

Since \( D \) is diagonal, its determinant is the product of the terms along the principal diagonal,

\[ |D| = \prod_{t=1}^{T} d_t. \]

while the inverse of \( D \) is obtained by taking reciprocals of the terms along the principal diagonal. Hence,

\[ \tilde{y}' D^{-1} \tilde{y} = \sum_{t=1}^{T} \frac{\tilde{y}_t^2}{d_t}. \]

Substituting [5.4.13] and [5.4.14] into [5.4.10], the likelihood function is

\[ f(y; \theta) = (2\pi)^{-T/2} \prod_{t=1}^{T} \frac{1}{d_t} \exp \left\{ -\frac{1}{2} \sum_{t=1}^{T} \frac{\tilde{y}_t^2}{d_t} \right\}. \]

The exact log likelihood for a Gaussian MA(1) process is therefore

\[ \log f(y; \theta) = -\frac{T}{2} \log(2\pi) - \frac{T}{2} \log(\sigma_t^2) - \frac{1}{2} \sum_{t=1}^{T} \frac{\tilde{y}_t^2}{d_t}. \]

Given numerical values for \( \mu \), \( \theta \), and \( \sigma_t^2 \), the sequence \( \tilde{y}_t \) is calculated by iterating on [5.4.11] starting with \( \tilde{y}_1 = y_1 - \mu \), while \( d_t \) is given by [5.4.12].

In contrast to the conditional log likelihood function [5.4.5], expression [5.4.16] will be valid regardless of whether \( \theta \) is associated with an invertible MA(1) representation. The value of [5.4.16] at \( \theta = 0, \sigma_t^2 = \sigma^2 \) will be identical to its value at \( \theta = \theta^{-1}, \sigma_t^2 = \sigma^2 \sigma_t^2 \), see Exercise 5.1.

5.4. The Likelihood Function for a Gaussian MA(1) Process 129
5.5. The Likelihood Function for a Gaussian MA(q) Process

Conditional Likelihood Function

For the MA(q) process,
\[ Y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}, \]  
[5.5.1]

a simple approach is to condition on the assumption that the first q values for \( \varepsilon \) were all zero:
\[ \varepsilon_0 = \varepsilon_{-1} = \cdots = \varepsilon_{-q+1} = 0. \]  
[5.5.2]

From these starting values we can iterate on
\[ \varepsilon_t = y_t - \mu - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \]  
[5.5.3]

for \( t = 1, 2, \ldots, T \). Let \( \varepsilon_0 \) denote the \((q \times 1)\) vector \((\varepsilon_0, \varepsilon_{-1}, \ldots, \varepsilon_{-q+1})'\). The conditional log likelihood is then
\[ L(\theta) = \log f_{Y_T, Y_{T-1}, \ldots, Y_{T-q}}(y_T, y_{T-1}, \ldots, y_1; \varepsilon_0 = 0; \theta) = -\frac{T}{2} \log(2\pi) - \frac{T}{2} \log(\sigma^2) - \frac{1}{\sigma^2} \sum_{t=1}^{T} \varepsilon_t^2, \]  
[5.5.4]

where \( \theta = (\mu, \theta_1, \theta_2, \ldots, \theta_q, \sigma^2)' \). Again, expression [5.5.4] is useful only if all values of \( z \) for which
\[ 1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q = 0 \]
lie outside the unit circle.

Exact Likelihood Function

The exact likelihood function is given by
\[ f(y; \theta) = (2\pi)^{-T/2} |\Omega|^{-1/2} \exp[- \frac{1}{2} (y - \mu)' \Omega^{-1} (y - \mu)], \]  
[5.5.5]

where as before \( y = (y_1, y_2, \ldots, y_T)' \) and \( \mu = (\mu, \mu, \ldots, \mu)' \). Here \( \Omega \) represents the variance-covariance matrix of \( T \) consecutive draws from an MA(q) process:
\[ \Omega = \begin{bmatrix} \sigma^2 & \sigma^2 & \cdots & \sigma^2 \\ \sigma^2 & \sigma^2 & \cdots & \sigma^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2 & \sigma^2 & \cdots & \sigma^2 \end{bmatrix} \]  
[5.5.6]

\[ \theta = \begin{bmatrix} \sigma^2 \theta_1 + \theta_{k+1} \theta_1 + \theta_{k+2} \theta_2 + \cdots + \theta_q \theta_{q-k} \\ 0 \end{bmatrix} \quad \text{for } k = 0, 1, \ldots, q \]
\[ \text{for } k > q, \]  
[5.5.7]

where \( \theta_0 = 1 \). Again, the exact likelihood function [5.5.5] can be evaluated using either the Kalman filter of Chapter 13 or the triangular factorization of \( \Omega \),
\[ \Omega = ADA', \]  
[5.5.8]

where \( A \) is the lower triangular matrix given by [4.4.11] and \( D \) is the diagonal matrix given by [4.4.7]. Note that the band structure of \( \Omega \) in [5.5.6] makes \( A \) and \( D \) simple to calculate. After the first \((q + 1)\) rows, all the subsequent entries in the first column of \( \Omega \) are already zero, so no multiple of the first row need be added to make these zero. Hence, \( a_{q+k} = 0 \) for \( i > q + 1 \). Similarly, beyond the first \((q + 2)\) rows of the second column, no multiple of the second row need be added to make these entries zero, meaning that \( a_{q+k} = 0 \) for \( i > q + 2 \). Thus \( A \) is a lower triangular band matrix with \( a_{q+k} = 0 \) for \( i > q + j \):
\[ A = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ a_{21} & 1 & 0 & \cdots & 0 & 0 \\ a_{31} & a_{32} & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{q+1,1} & a_{q+1,2} & a_{q+1,3} & \cdots & 0 & 0 \\ 0 & a_{q+2,2} & a_{q+2,3} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{T,T-1} & 1 \end{bmatrix} \]

A computer can be programmed to calculate these matrices quickly for a given numerical value for \( \theta \).

Substituting [5.5.8] into [5.5.5], the exact likelihood function for a Gaussian MA(q) process can be written as in [5.4.10]:
\[ f(y; \theta) = (2\pi)^{-T/2} |D|^{-1/2} \exp[- \frac{1}{2} y'D^{-1}y], \]

where
\[ A y = y - \mu. \]  
[5.5.9]

The elements of \( y \) can be calculated recursively by working down the rows of [5.5.9]:
\[ y_1 = y_1 - \mu; \]
\[ y_2 = (y_2 - \mu) - a_{21} y_1; \]
\[ y_3 = (y_3 - \mu) - a_{22} y_2 - a_{31} y_1; \]
\[ \vdots \]
\[ y_T = (y_T - \mu) - a_{T,T-1} y_{T-1} - a_{T,T-2} y_{T-2} - \cdots - a_{T,q} y_{q}. \]

The exact log likelihood function can then be calculated as in [5.4.16]:
\[ L(\theta) = \log f(y; \theta) = -\frac{T}{2} \log(2\pi) - \frac{1}{2} \sum_{t=1}^{T} \log(\lambda_t) - \frac{1}{2} \sum_{t=1}^{T} \frac{y_t^2}{\lambda_t}, \]  
[5.5.10]
5.6. The Likelihood Function for a Gaussian ARMA(p, q) Process

Conditional Likelihood Function

A Gaussian ARMA(p, q) process takes the form

\[ Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \epsilon_t \]

where \( \epsilon_t \sim \text{i.i.d. } N(0, \sigma^2) \). The goal is to estimate the vector of population parameters \( \theta = (c, \phi_1, \phi_2, \ldots, \phi_p, \theta_1, \theta_2, \ldots, \theta_q, \sigma^2) \).

The approximation to the likelihood function for an autoregressive conditioned on initial values of the y’s. The approximation to the likelihood function for a moving average process conditioned on initial values of the \( \epsilon \)'s. A common approximation to the likelihood function for an ARMA(p, q) process conditions on both y’s and \( \epsilon \)'s.

Taking initial values for \( y_0 = (y_{-1}, \ldots, y_{-p+1}, \epsilon_0, \epsilon_1, \ldots, \epsilon_{-q}) \) as given, the sequence \( \{\epsilon_t, \epsilon_2, \ldots, \epsilon_T\} \) can be calculated from \( \{y_1, y_2, \ldots, y_T\} \) by iterating on

\[
\epsilon_t = y_t - c - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \cdots - \phi_p y_{t-p} - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \cdots - \theta_q \epsilon_{t-q}
\]

for \( t = 1, 2, \ldots, T \). The conditional log likelihood is then

\[
L(\theta) = \log f(y_T, y_{T-1}, \ldots, y_{y_{-p+1}}, y_T, y_{y_{-p+1}}, \ldots, y_T | \epsilon_0, \epsilon_1, \ldots, \epsilon_{-q}; \theta)
\]

\[
= -\frac{T}{2} \log(2\pi) - \frac{T}{2} \log(\sigma^2) - \frac{T}{2} \sum_{i=1}^{T} \frac{\epsilon_i^2}{2\sigma^2}
\]

One option is to set initial \( y \)'s and \( \epsilon \)'s equal to their expected values. That is, set \( y_0 = c(1 - \phi_1 - \phi_2 - \cdots - \phi_p) \) for \( x = 0, -1, \ldots, -p + 1 \) and set \( \epsilon_0 = 0 \) for \( x = 0, -1, \ldots, -q + 1 \), and then proceed with the iteration in [5.6.2] for \( t = 1, 2, \ldots, T \). Alternatively, Box and Jenkins (1976, p. 211) recommended setting \( \epsilon \)'s to zero but \( y \)'s equal to their actual values. Thus, iteration on [5.6.2] is started at date \( t = p + 1 \) with \( y_1, y_2, \ldots, y_T \) set to the observed values and

\[
\epsilon_0 = \epsilon_1 = \cdots = \epsilon_{q-1} = 0.
\]

Then the conditional likelihood calculated is

\[
\log f(y_T, \ldots, y_{p+1}, y_{p+1}, \epsilon_0 = 0, \ldots, \epsilon_{p+q-1} = 0)
\]

\[
= -\frac{T-p}{2} \log(2\pi) - \frac{T-p}{2} \log(\sigma^2) - \frac{T-p}{2} \sum_{i=p+1}^{T} \epsilon_i^2 / 2\sigma^2
\]

As in the case for the moving average processes, these approximations should be used only if all values of \( \epsilon \) satisfying

\[
1 + \theta_1 \epsilon + \theta_2 \epsilon^2 + \cdots + \theta_q \epsilon^q = 0
\]

lie outside the unit circle.

Alternative Algorithms

The simplest approach to calculating the exact likelihood function for a Gaussian ARMA process is to use the Kalman filter described in Chapter 13. For more

details on exact and approximate maximum likelihood estimation of ARMA models, see Galbraith and Galbraith (1974), Box and Jenkins (1976, Chapter 6), Hannan and Rissanen (1982), and Koreisha and Pukkila (1989).

5.7. Numerical Optimization

Previous sections of this chapter have shown how to calculate the log likelihood function

\[
L(\theta) = \log f(y_T, y_{T-1}, \ldots, y_T | y_T, y_{T-1}, \ldots, y_T; \theta)
\]

for various specifications of the process thought to have generated the observed data \( y_T, y_{T-1}, \ldots, y_T \). Given the observed data, the formulas given could be used to calculate the value of \( L(\theta) \) for any given numerical value of \( \theta \).

This section discusses how to find the value of \( \theta \) that maximizes \( L(\theta) \) given no more knowledge than this ability to calculate the value of \( L(\theta) \) for any particular value of \( \theta \). The general approach is to write a procedure that enables a computer to calculate the numerical value of \( L(\theta) \) for any particular numerical values for \( \theta \) and the observed data \( y_T, y_{T-1}, \ldots, y_T \). We can think of this procedure as a "black box" that enables us to guess some value of \( \theta \) and see what the resulting value of \( L(\theta) \) would be:

\[
\begin{array}{ccc}
\text{Input} & \text{Procedure} & \text{Output} \\
\{y_T, y_{T-1}, \ldots, y_T\} & \text{calculates} & \text{value of} \\
\theta & \text{L(\theta)} & \end{array}
\]

The idea will be to make a series of different guesses for \( \theta \), compare the value of \( L(\theta) \) for each guess, and try to infer from these values for \( L(\theta) \) the value \( \hat{\theta} \) for which \( L(\hat{\theta}) \) is largest. Such methods are described as numerical maximization.

Grid Search

The simplest approach to numerical maximization is known as the grid search method. To illustrate this approach, suppose we have data generated by an AR(1) process, for which the log likelihood was seen to be given by [5.2.9]. To keep the example very simple, it is assumed to be known that the mean of the process is zero (\( c = 0 \)) and that the innovations have unit variance (\( \sigma^2 = 1 \)). Thus the only unknown parameter is the autoregressive coefficient \( \phi \), and [5.2.9] simplifies to

\[
L(\phi) = -\frac{T}{2} \log(2\pi) - \frac{T}{2} \log(1 - \phi^2) - \frac{T}{2} (1 - \phi^2) y_1^2 - \frac{T}{2} \sum_{t=2}^{T} (y_t - \phi y_{t-1})^2.
\]

Suppose that the observed sample consists of the following \( T = 5 \) observations:

\[
y_1 = 0.8, \quad y_2 = 0.2, \quad y_3 = -1.2, \quad y_4 = -0.4, \quad y_5 = 0.0.
\]

If we make an arbitrary guess as to the value of \( \phi \), say, \( \phi = 0.0 \), and plug this guess into expression [5.7.2], we calculate that \( L(\phi) = -5.71 \) at \( \phi = 0.0 \). Trying another guess (\( \phi = 0.1 \)), we calculate \( L(\phi) = -5.71 \) at \( \phi = 0.1 \)—the log likelihood is higher at \( \phi = 0.1 \) than at \( \phi = 0.0 \). Continuing in this fashon, we could calculate the value of \( L(\phi) \) for every value of \( \phi \) between -0.9 and +0.9 in increments of
The results are reported in Figure 5.1. It appears from these calculations that the log likelihood function $\mathcal{L}(\phi)$ is nicely behaved with a unique maximum at some value of $\phi$ between 0.1 and 0.3. We could then focus on this subregion of the parameter space and evaluate $\mathcal{L}(\phi)$ at a finer grid, calculating the value of $\mathcal{L}(\phi)$ for all values of $\phi$ between 0.1 and 0.3 in increments of 0.02. Proceeding in this fashion, it should be possible to get arbitrarily close to the value of $\phi$ that maximizes $\mathcal{L}(\phi)$ by making the grid finer and finer.

Note that this procedure does not find the exact MLE $\hat{\phi}$, but instead approximates it with any accuracy desired. In general, this will be the case with any numerical maximization algorithm. To use these algorithms we therefore have to specify a convergence criterion, or some way of deciding when we are close enough to the true maximum. For example, suppose we want an estimate $\hat{\phi}$ that differs from the true MLE by no more than $\pm 0.0001$. Then we would continue refining the grid until the increments are in steps of 0.0001, and the best estimate among the elements of that grid would be the numerical MLE of $\phi$.

For the simple AR(1) example in Figure 5.1, the log likelihood function is unimodal—there is a unique value $\theta$ for which $\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = 0$. For a general numerical maximization problem, this need not be the case. For example, suppose that we are interested in estimating a scalar parameter $\theta$ for which the log likelihood function is as displayed in Figure 5.2. The value $\theta = -0.6$ is a local maximum, meaning that the likelihood function is higher there than for any other $\theta$ in a neighborhood around $\theta = -0.6$. However, the global maximum occurs around $\theta = 0.2$. The grid search method should work well for a unimodal likelihood as long as $\mathcal{L}(\theta)$ is continuous. When there are multiple local maxima, the grid must be sufficiently fine to reveal all of the local "hills" on the likelihood surface.

### Steepest Ascent

Grid search can be a very good method when there is a single unknown parameter to estimate. However, it quickly becomes intractable when the number of elements of $\theta$ becomes large. An alternative numerical method that often succeeds in maximizing a continuously differentiable function of a large number of parameters is known as steepest ascent.

To understand this approach, let us temporarily disregard the "black box" nature of the investigation and instead examine how we would proceed analytically with a particular maximization problem. Suppose we have an initial estimate of the parameter vector, denoted $\theta^{(0)}$, and wish to come up with a better estimate $\theta^{(1)}$. Imagine that we are constrained to choose $\theta^{(1)}$ so that the squared distance between $\theta^{(0)}$ and $\theta^{(1)}$ is some fixed number $k$:

$$
(\theta^{(1)} - \theta^{(0)})'(\theta^{(1)} - \theta^{(0)}) = k.
$$

The optimal value to choose for $\theta^{(1)}$ would then be the solution to the following constrained maximization problem:

$$
\max_{\theta^{(1)}} \mathcal{L}(\theta^{(1)}) \quad \text{subject to} \quad (\theta^{(1)} - \theta^{(0)})'(\theta^{(1)} - \theta^{(0)}) = k.
$$

To characterize the solution to this problem, form the Lagrangean,

$$
J(\theta^{(1)}) = \mathcal{L}(\theta^{(1)}) + \lambda [k - (\theta^{(1)} - \theta^{(0)})'(\theta^{(1)} - \theta^{(0)})].
$$

where $\lambda$ denotes a Lagrange multiplier. Differentiating [5.7.3] with respect to $\theta^{(1)}$ and setting the result equal to zero yields

$$
\frac{\partial \mathcal{L}(\theta)}{\partial \theta} \bigg|_{\theta = \theta^{(1)}} - (2\lambda)(\theta^{(1)} - \theta^{(0)}) = 0.
$$

Let $g(\theta)$ denote the gradient vector of the log likelihood function:

$$
g(\theta) = \frac{\partial \mathcal{L}(\theta)}{\partial \theta}.
$$

If there are $a$ elements of $\theta$, then $g(\theta)$ is an $(a \times 1)$ vector whose $i$th element represents the derivative of the log likelihood with respect to the $i$th element of $\theta$.

---

**FIGURE 5.2**  Bimodal log likelihood function.

---

**FIGURE 5.1**  Log likelihood for an AR(1) process for various guesses of $\phi$.  

134 \hspace{1em} Chapter 5 \hspace{1em} Maximum Likelihood Estimation
Using this notation, expression [5.7.4] can be written as
\[ \theta^{(1)} - \theta^{(0)} = [1/(2\lambda)] \cdot g(\theta^{(0)}). \] [5.7.5]

Expression [5.7.5] asserts that if we are allowed to change \( \theta \) by only a fixed amount, the biggest increase in the log likelihood function will be achieved if the change in \( \theta \) (the magnitude \( \theta^{(1)} - \theta^{(0)} \) is chosen to be a constant \( 1/(2\lambda) \) times the gradient vector \( g(\theta^{(0)}) \). If we are contemplating a very small step (so that \( k \) is near zero), the value \( g(\theta^{(0)}) \) will approach \( g(\theta^{(0)}) \). In other words, the gradient vector \( g(\theta^{(0)}) \) gives the direction in which the log likelihood function increases most steeply from \( \theta^{(0)} \).

For illustration, suppose that \( a = 2 \) and let the log likelihood be
\[ L(\theta) = -1.5\theta_1^2 - 2\theta_2^2. \] [5.7.6]
We can easily see analytically for this example that the MLE is given by \( \hat{\theta} = (0, 0)' \). Let us nevertheless use this example to illustrate how the method of steepest ascent works. The elements of the gradient vector are
\[ \frac{\partial L(\theta)}{\partial \theta_1} = -3\theta_1, \quad \frac{\partial L(\theta)}{\partial \theta_2} = -4\theta_2. \] [5.7.7]
Suppose that the initial guess is \( \theta^{(0)} = (-1, 1)' \). Then
\[ \frac{\partial L(\theta)}{\partial \theta_1} \bigg|_{\theta^{(0)}} = 3, \quad \frac{\partial L(\theta)}{\partial \theta_2} \bigg|_{\theta^{(0)}} = -4. \]
An increase in \( \theta_1 \) would increase the likelihood, while an increase in \( \theta_2 \) would decrease the likelihood. The gradient vector evaluated at \( \theta^{(0)} \) is
\[ g(\theta^{(0)}) = \begin{bmatrix} 3 \\ -4 \end{bmatrix}, \]
so that the optimal step \( \theta^{(1)} - \theta^{(0)} \) should be proportional to \( (3, -4)' \). For example, with \( k = 1 \) we would choose
\[ \theta_1^{(1)} - \theta_1^{(0)} = \frac{3}{4}, \quad \theta_2^{(1)} - \theta_2^{(0)} = -\frac{4}{4}; \]
that is, the new guesses would be \( \theta_1^{(1)} = -0.4 \) and \( \theta_2^{(1)} = 0.2 \). To increase the likelihood by the greatest amount, we want to increase \( \theta_1 \) and decrease \( \theta_2 \), relative to their values at the initial guess \( \theta^{(0)} \). Since a one-unit change in \( \theta_1 \) has a bigger effect on \( L(\theta) \) than would a one-unit change in \( \theta_2 \), the change in \( \theta_2 \) is larger in absolute value than the change in \( \theta_1 \).

Let us now return to the black box perspective, where the only capability we have is to calculate the value of \( L(\theta) \) for a specified numerical value of \( \theta \). We might start with an arbitrary initial guess for the value of \( \theta \), denoted \( \theta^{(0)} \). Suppose we then calculate the value of the gradient vector at \( \theta^{(0)} \):
\[ g(\theta^{(0)}) = \frac{\partial L(\theta)}{\partial \theta} \bigg|_{\theta^{(0)}}. \] [5.7.8]
This gradient could in principle be calculated analytically, by differentiating the general expression for \( L(\theta) \) with respect to \( \theta \) and writing a computer procedure to calculate each element of \( g(\theta) \) given the data and a numerical value for \( \theta \). For example, expression [5.7.7] could be used to calculate \( g(\theta) \) for any particular value of \( \theta \). Alternatively, if it is too hard to differentiate \( L(\theta) \) analytically, we can always get a numerical approximation to the gradient by seeing how \( L(\theta) \) changes for a small change in each element of \( \theta \). In particular, the \( i \)th element of \( g(\theta^{(0)}) \) might be approximated by
\[ g_i(\theta^{(0)}) \approx \frac{\Delta_i L(\theta^{(0)}; \theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_i^{(0)}, \theta_{i+1}^{(0)}, \ldots)}{\Delta_i \theta_i}, \]
where \( \Delta \) represents some arbitrarily chosen small scalar such as \( \Delta = 10^{-6} \). By numerically calculating the value of \( L(\theta) \) at \( \theta^{(0)} \) and at different values of \( \theta \) corresponding to small changes in each of the individual elements of \( \theta^{(0)} \), an estimate of the full vector \( g(\theta^{(0)}) \) can be uncovered.

Result [5.7.5] suggests that we should change the value of \( \theta \) in the direction of the gradient, choosing
\[ \theta^{(1)} - \theta^{(0)} = s \cdot g(\theta^{(0)}) \]
for some positive scalar \( s \). A suitable choice for \( s \) could be found by an adaptation of the grid search method. For example, we might calculate the value of \( L(\theta^{(0)} + s \cdot g(\theta^{(0)})) \) for \( s = 0, 1, 2, 3, 4, 5 \), and 16 and choose as the new estimate \( \theta^{(1)} \) the value of \( \theta^{(0)} + s \cdot g(\theta^{(0)}) \) for which \( L(\theta) \) is largest. Smaller or larger values of \( s \) could also be explored if the maximum appears to be at one of the extremes. If none of the values of \( s \) improves the likelihood, then a very small value for \( s \) such as the value \( \Delta = 10^{-6} \) used to approximate the derivative should be tried.

We can then repeat the process, taking \( \theta^{(1)} = \theta^{(0)} + s \cdot g(\theta^{(0)}) \) as the starting point, evaluating the gradient at the new location \( g(\theta^{(1)}) \), and generating a new estimate \( \theta^{(2)} \) according to
\[ \theta^{(2)} = \theta^{(1)} + s \cdot g(\theta^{(1)}) \]
for the best choice of \( s \). The process is iterated, calculating
\[ \theta^{(m+1)} = \theta^{(m)} + s \cdot g(\theta^{(m)}) \]
for \( m = 0, 1, 2, \ldots \), until some convergence criterion is satisfied, such as that the gradient vector \( g(\theta^{(m)}) \) is within some specified tolerance of zero, the distance between \( \theta^{(m+1)} \) and \( \theta^{(m)} \) is less than some specified threshold, or the change between \( L(\theta^{(m+1)}) \) and \( L(\theta^{(m)}) \) is smaller than some desired amount.

Figure 5.3 illustrates the method of steepest ascent when \( \theta \) contains \( a = 2 \) elements. The figure displays contour lines for the log likelihood \( L(\theta) \); along a given contour, the log likelihood \( L(\theta) \) is constant. If the iteration is started at the initial guess \( \theta^{(0)} \), the gradient \( g(\theta^{(0)}) \) describes the direction of steepest ascent. Finding the optimal step in that direction produces the new estimate \( \theta^{(1)} \). The gradient at that point \( g(\theta^{(1)}) \) then determines a new search direction on which a new estimate \( \theta^{(2)} \) is based, until the top of the hill is reached.

Figure 5.3 also illustrates a multivariate generalization of the problem with multiple local maxima seen earlier in Figure 5.2. The procedure should converge to a local maximum, which in this case is different from the global maximum \( \theta^{*} \). In Figure 5.3, it appears that if \( \theta^{(0)} \) were used to begin the iteration in place of \( \theta^{(0)} \), the procedure would converge to the true global maximum \( \theta^{*} \). In practice, the only way to ensure that a global maximum is found is to begin the iteration from a number of different starting values for \( \theta^{(0)} \) and to continue the sequence from each starting value until the top of the hill associated with that starting value is discovered.
Let \( \theta^{(0)} \) denote an initial guess as to the value of \( \theta \). One can calculate the derivative of the log likelihood at that initial guess \( g(\theta^{(0)}) \) either analytically, as in [5.7.7], or numerically, as in [5.7.9]. One can also use analytical or numerical methods to calculate the negative of the matrix of second derivatives at the initial guess \( H(\theta^{(0)}) \). Expression [5.7.11] suggests that an improved estimate of \( \theta \) (denoted \( \theta^{(1)} \)) would satisfy

\[
g(\theta^{(0)}) = H(\theta^{(0)})[\theta^{(1)} - \theta^{(0)}]
\]

or

\[
\theta^{(1)} - \theta^{(0)} = [H(\theta^{(0)})]^{-1}g(\theta^{(0)}).
\]  

[5.7.12]

One could next calculate the gradient and Hessian at \( \theta^{(1)} \) and use these to find a new estimate \( \theta^{(2)} \) and continue iterating in this fashion. The \( m \)th step in the iteration updates the estimate of \( \theta \) by using the formula

\[
\theta^{(m+1)} = \theta^{(m)} + [H(\theta^{(m)})]^{-1}g(\theta^{(m)}).
\]  

[5.7.13]

If the log likelihood function happens to be a perfect quadratic function, then [5.7.10] holds exactly and [5.7.12] will generate the exact MLE in a single step:

\[
\theta^{(1)} = \hat{\theta}_{MLE}.
\]

If the quadratic approximation is reasonably good, Newton-Raphson should converge to the local maximum more quickly than the steepest-ascent method. However, if the likelihood function is not concave, Newton-Raphson behaves quite poorly. Thus, steepest ascent is often slower to converge but sometimes proves to be more robust compared with Newton-Raphson.

Since [5.7.10] is usually only an approximation to the true log likelihood function, the iteration on [5.7.13] is often modified as follows. Expression [5.7.13] is taken to suggest the search direction. The value of the log likelihood function at several points in that direction is then calculated, and the best value determines the length of the step. This strategy calls for replacing [5.7.13] by

\[
\theta^{(m+1)} = \theta^{(m)} + s[H(\theta^{(m)})]^{-1}g(\theta^{(m)}),
\]  

[5.7.14]

where \( s \) is a scalar controlling the step length. One calculates \( \theta^{(m+1)} \) and the associated value for the log likelihood \( L(\theta^{(m+1)}) \) for various values of \( s \) in [5.7.14] and chooses as the estimate \( \theta^{(m+1)} \) the value that produces the biggest value for the log likelihood.

Davidon-Fletcher-Powell

If \( \theta \) contains \( a \) unknown parameters, then the symmetric matrix \( H(\theta) \) has \( a(a + 1)/2 \) separable elements. Calculating all these elements can be extremely time-consuming if \( a \) is large. An alternative approach reasons as follows. The matrix of second derivatives \(-H(\theta)\) corresponds to the first derivatives of the gradient vector \( g(\theta) \), which tell us how \( g(\theta) \) changes as \( \theta \) changes. We get some independent information about this by comparing \( g(\theta^{(1)}) - g(\theta^{(0)}) \) with \( \theta^{(1)} - \theta^{(0)} \). This is not enough information by itself to estimate \( H(\theta) \), but it is information that could be used to update an initial guess about the value of \( H(\theta) \). Thus, rather than evaluate \( H(\theta) \) directly each iteration, the idea will be to start with an initial guess about \( H(\theta) \) and update the guess solely on the basis of how much \( g(\theta) \) changes between iterations, given the magnitude of the change in \( \theta \). Such methods are sometimes described as modified Newton-Raphson.

One of the most popular modified Newton-Raphson methods was proposed by Davidon (1959) and Fletcher and Powell (1963). Since it \( H^{-1} \) rather than \( H \)
itself that appears in the updating formula [5.7.14], the Davidon-Fletcher-Powell algorithm updates an estimate of $H^{-1}$ at each step on the basis of the size of the change in $g(\theta)$ relative to the change in $\theta$. Specifically, let $\theta^{(m)}$ denote an estimate of $\theta$ that has been calculated at the $m$th iteration, and let $A^{(m)}$ denote an estimate of $[H(\theta^{(m)})]^{-1}$. The new estimate $\theta^{(m+1)}$ is given by

$$\theta^{(m+1)} = \theta^{(m)} + sA^{(m)}g(\theta^{(m)})$$  \[5.7.15\]

for $s$ the positive scalar that maximizes $\mathcal{L}[\theta^{(m)} + sA^{(m)}g(\theta^{(m)})]$. Once $\theta^{(m+1)}$ and the gradient at $\theta^{(m+1)}$ have been calculated, a new estimate $A^{(m+1)}$ is found from

$$A^{(m+1)} = A^{(m)} - \frac{A^{(m)}(\Delta g^{(m+1)})^2A^{(m)}}{(\Delta g^{(m+1)})^2A^{(m)} + (\Delta g^{(m+1)})(\Delta g^{(m+1)})}$$

$$= -\frac{(\Delta g^{(m+1)})(\Delta g^{(m+1)})}{(\Delta g^{(m+1)})^2A^{(m)}}$$  \[5.7.16\]

where

$$\Delta g^{(m+1)} = g(\theta^{(m+1)}) - g(\theta^{(m)})$$

$$\Delta g^{(m+1)} = g(\theta^{(m+1)}) - g(\theta^{(m)})$$

In what sense should $A^{(m+1)}$ as calculated from [5.7.16] be regarded as an estimate of the inverse of $H(\theta^{(m+1)})$? Consider first the case when $\theta$ is a scalar ($a = 1$). Then [5.7.16] simplifies to

$$A^{(m+1)} = A^{(m)} - \frac{A^{(m)}(\Delta g^{(m+1)})^2A^{(m)}}{(\Delta g^{(m+1)})^2A^{(m)} + (\Delta g^{(m+1)})(\Delta g^{(m+1)})}$$

$$= A^{(m)} - \frac{A^{(m)}(\Delta g^{(m+1)})}{(\Delta g^{(m+1)})^2A^{(m)}}$$

$$= \Delta g^{(m+1)}.$$  \[5.7.17\]

In this case,

$$[A^{(m+1)}]^{-1} = -\frac{\Delta g^{(m+1)}}{\Delta g^{(m+1)}}$$

which is the natural discrete approximation to

$$H(\theta^{(m+1)}) = -\frac{\partial^2 \mathcal{L}}{\partial \theta^2} = -\frac{\partial^2 \mathcal{L}}{\partial \theta^2}.$$

More generally (for $a > 1$), an estimate of the derivative of $g(\cdot)$ should be related to the observed change in $g(\cdot)$ according to

$$g(\theta^{(m+1)}) = g(\theta^{(m)}) + \frac{\partial g}{\partial \theta} \bigg|_{\theta = \theta^{(m+1)}} [\theta^{(m+1)} - \theta^{(m)}].$$

That is,

$$g(\theta^{(m+1)}) = g(\theta^{(m)}) - H(\theta^{(m+1)})[\theta^{(m+1)} - \theta^{(m)}]$$

or

$$\Delta \theta^{(m+1)} = -H(\theta^{(m+1)})^{-1} \Delta g^{(m+1)}.$$  \[5.7.18\]

Hence an estimate $A^{(m+1)}$ of $[H(\theta^{(m+1)})]^{-1}$ should satisfy

$$A^{(m+1)} \Delta g^{(m+1)} = -\Delta \theta^{(m+1)}.$$  \[5.7.19\]

Postmultiplication of [5.7.16] by $\Delta g^{(m+1)}$ confirms that [5.7.17] is indeed satisfied by the Davidon-Fletcher-Powell estimate $A^{(m+1)}$:

$$A^{(m+1)} \Delta g^{(m+1)} = A^{(m)} \Delta g^{(m+1)}$$

$$\Delta g^{(m+1)} - \Delta g^{(m+1)} [\Delta g^{(m+1)}]^2 \Delta g^{(m+1)}$$

$$= A^{(m)} \Delta g^{(m+1)} - A^{(m)} \Delta g^{(m+1)} [\Delta g^{(m+1)}]^2 \Delta g^{(m+1)}$$

$$= -\Delta \theta^{(m+1)}$$

Thus, multiplication [5.7.16] produces an estimate of $[H(\theta^{(m+1)})]^{-1}$ that is consistent with the magnitude of the observed change between $g(\theta^{(m+1)})$ and $g(\theta^{(m)})$ given the size of the change between $\theta^{(m+1)}$ and $\theta^{(m)}$.

The following proposition (proved in Appendix 5A at the end of the chapter) establishes some further useful properties of the updating formula [5.7.16].

**Proposition 5.1.** (Fletcher and Powell (1963)). Consider $\mathcal{L}(\theta)$, where $\mathcal{L}: \mathbb{R}^2 \rightarrow \mathbb{R}$ has continuous first derivatives denoted

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = \frac{\partial \mathcal{L}(\theta)}{\partial \theta}.$$

Suppose that some element of $g(\theta^{(m)})$ is nonzero, and let $A^{(m)}$ be a positive definite symmetric $(a \times a)$ matrix. Then the following hold.

(a) There exists a scalar $s > 0$ such that $\mathcal{L}(\theta^{(m+1)}) > \mathcal{L}(\theta^{(m)})$ for

$$\theta^{(m+1)} = \theta^{(m)} + sA^{(m)}g(\theta^{(m)}).$$  \[5.7.18\]

(b) If $s$ is in [5.7.18] is chosen so as to maximize $\mathcal{L}(\theta^{(m+1)})$, then the first-order conditions for an interior maximum imply that

$$[g(\theta^{(m+1)})]^{-1} [\theta^{(m+1)} - \theta^{(m)}] = 0.$$  \[5.7.19\]

(c) Provided that [5.7.19] holds and that some element of $g(\theta^{(m+1)}) - g(\theta^{(m)})$ is nonzero, then $A^{(m+1)}$ described by [5.7.16] is a positive definite symmetric matrix.

Result (a) establishes that as long as we are not already at an optimum ($g(\theta^{(m)}) \neq 0$), there exists a step in the direction suggested by the algorithm that will increase the likelihood further, provided that $A^{(m)}$ is a positive definite matrix. Result (c) establishes that provided that the iteration is begun with $A^{(0)}$ a positive definite matrix, then the sequence of matrices $(A^{(m)})_{m=1}^{\infty}$ should all be positive definite, meaning that each step of the iteration should increase the likelihood function. A standard procedure is to start the iteration with $A^{(0)} = I_a$, the $(a \times a)$ identity matrix.

If the function $\mathcal{L}(\theta)$ is exactly quadratic, so that

$$\mathcal{L}(\theta) = \mathcal{L}(\theta^{(0)}) + g(\theta^{(0)}) \cdot (\theta - \theta^{(0)}) + \frac{1}{2} (\theta - \theta^{(0)})^T H(\theta) (\theta - \theta^{(0)})$$

with $H$ positive definite, then Fletcher and Powell’s 1964 work showed that the iteration on [5.7.15] and [5.7.16] will converge to the true global maximum in $a$ steps:

$$\theta^{(a)} = \hat{\theta}_{MLE} = \theta^{(0)} + H^{-1} g$$

and the weighting matrix will converge to the inverse of $-1$ times the matrix of second derivatives:

$$A^{(a)} = H^{-1}.$$  \[5.7.20\]
More generally, if \( \mathcal{L}(\theta) \) is well approximated by a quadratic function, then the Davidon-Fletcher-Powell search procedure should approach the global maximum more quickly than the steepest-ascent method,

\[
\theta^{(k)} = \tilde{\theta}_{MLE}
\]

for large \( N \), while \( A^{(k)} \) should converge to the negative of the matrix of second derivatives of the log likelihood function:

\[
A^{(k)} = \left[ \frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta \partial \theta'} \right]_{\theta = \tilde{\theta}_{MLE}}^{-1}.
\]  [5.7.20]

In practice, however, the approximation in [5.7.20] can be somewhat poor, and it is better to evaluate the matrix of second derivatives numerically for purposes of calculating standard errors, as discussed in Section 5.8.

If the function \( \mathcal{L}(\theta) \) is not globally concave or if the starting value \( \theta^{(0)} \) is far from the true maximum, the Davidon-Fletcher-Powell procedure can do very badly. If problems are encountered, it often helps to try a different starting value \( \theta^{(0)} \), to rescale the data or parameters so that the elements of \( \theta \) are in comparable units, or to rescale the initial matrix \( A^{(0)} \)—for example, by setting

\[
A^{(0)} = (1 \times 10^{-3}) I_n.
\]

### Other Numerical Optimization Methods

A variety of other modified Newton-Raphson methods are available, which use alternative techniques for updating an estimate of \( H(\theta^{(m)}) \) or its inverse. Two of the more popular methods are those of Broyden (1965, 1967) and Berndt, Hall, Hall, and Hausman (1974). Surveys of these and a variety of other approaches are provided by Judge, Griffiths, Hill, and Lee (1980, pp. 719-72) and Quandt (1983). Obviously, these same methods can be used to minimize a function \( Q(\theta) \) with respect to \( \theta \). We simply multiply the objective function by \(-1\) and then maximize the function \(-Q(\theta)\).

### 5.8. Statistical Inference with Maximum Likelihood Estimation

The previous section discussed ways to find the maximum likelihood estimate \( \hat{\theta} \) given only the numerical ability to evaluate the log likelihood function \( \mathcal{L}(\theta) \). This section summarizes general approaches that can be used to test a hypothesis about \( \theta \). The section merely summarizes a number of useful results without providing any proofs. We will return to these issues in more depth in Chapter 14, where the statistical foundation behind many of these claims will be developed.

Before detailing these results, however, it is worth calling attention to two of the key assumptions behind the formulas presented in this section. First, it is assumed that the observed data are strictly stationary. Second, it is assumed that neither the estimate \( \hat{\theta} \) nor the true value \( \theta_0 \) falls on the boundary of the allowable parameter space. For example, suppose that the first element of \( \theta \) is a parameter corresponding to the probability of a particular event, which must be between 0 and 1. If the event did not occur in the sample, the maximum likelihood estimate of the probability might be zero. This is an example where the estimate \( \hat{\theta} \) falls on the boundary of the allowable parameter space, in which case the formulas presented in this section will not be valid.

Asymptotic Standard Errors for Maximum Likelihood Estimates

If the sample size \( T \) is sufficiently large, it often turns out that the distribution of the maximum likelihood estimate \( \hat{\theta} \) can be well approximated by the following distribution:

\[
\hat{\theta} \sim N(\theta_0, \ T^{-1} \mathbf{V}^{-1}),
\]  [5.8.1]

where \( \theta_0 \) denotes the true parameter vector. The matrix \( \mathbf{V} \) is known as the information matrix and can be estimated in either of two ways.

The second-derivative estimate of the information matrix is

\[
\hat{\mathbf{V}}_{2D} = - \frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta \partial \theta'} \bigg|_{\theta = \hat{\theta}}^{-1}.
\]  [5.8.2]

Here \( \mathcal{L}(\theta) \) denotes the log likelihood:

\[
\mathcal{L}(\theta) = \sum_{t=1}^{T} \log f(y_{t-1}, y_{t}, \ldots; \theta);
\]

and \( y_t \) denotes the history of observations on \( y \) obtained through date \( t \). The matrix of second derivatives of the log likelihood is often calculated numerically. Substituting [5.8.2] into [5.8.1], the terms involving the sample size \( T \) cancel out so that the variance-covariance matrix of \( \hat{\theta} \) can be approximated by

\[
\mathbf{V}(\hat{\theta}) = \frac{1}{T} \sum_{t=1}^{T} \left( h(\hat{\theta}, \ y_t) - h(\hat{\theta}, \ \bar{y}_T) \right) \left( h(\hat{\theta}, \ y_t) - h(\hat{\theta}, \ \bar{y}_T) \right)^\prime.
\]  [5.8.4]

A second estimate of the information matrix \( \mathbf{V} \) in [5.8.1] is called the outer-product estimate:

\[
\hat{\mathbf{V}}_{OP} = - \left( \frac{\partial f(y_{t-1}, y_{t}, \ldots; \theta)}{\partial \theta} \right) \left( \frac{\partial f(y_{t-1}, y_{t}, \ldots; \theta)}{\partial \theta} \right)^\prime \bigg|_{\theta = \hat{\theta}}^{-1}.
\]

Here \( h(\hat{\theta}, \ y_t) \) denotes the \((a \times 1)\) vector of derivatives of the log of the conditional density of the \( t \)th observation with respect to the \( a \) elements of the parameter vector \( \theta \), with this derivative evaluated at the maximum likelihood estimate \( \hat{\theta} \):

\[
\frac{\partial f(y_{t-1}, y_{t}, \ldots; \theta)}{\partial \theta} \bigg|_{\theta = \hat{\theta}}.
\]

In this case, the variance-covariance matrix of \( \hat{\theta} \) is approximated by

\[
\mathbf{V}(\hat{\theta}) = \frac{1}{T} \sum_{t=1}^{T} \left( \frac{\partial f(y_{t-1}, y_{t}, \ldots; \theta)}{\partial \theta} \right) \left( \frac{\partial f(y_{t-1}, y_{t}, \ldots; \theta)}{\partial \theta} \right)^\prime \bigg|_{\theta = \hat{\theta}}^{-1}.
\]

As an illustration of how such approximations can be used, suppose that the log likelihood is given by expression [5.7.6]. For this case, one can see analytically that

\[
\frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta \partial \theta'} = \begin{bmatrix} -3 & 0 \\ 0 & -4 \end{bmatrix},
\]

and so result [5.8.3] suggests that the variance of the maximum likelihood estimate \( \hat{\theta} \) can be approximated by \( 1 \). The MLE for this example was \( \hat{\theta} = 0 \). Thus an
approximate 95% confidence interval for \( \theta_2 \) is given by

\[
0 \pm 2\sqrt{1} = \pm 1.
\]

Note that unless the off-diagonal elements of \( \hat{F} \) are zero, in general one needs to calculate all the elements of the matrix \( \hat{F} \) and invert this full matrix in order to obtain a standard error for any given parameter.

Which estimate of the information matrix, \( \hat{F}_{2D} \) or \( \hat{F}_{OP} \), is it better to use in practice? Expression [5.8.1] is only an approximation to the true distribution of \( \hat{\theta} \), and \( \hat{F}_{2D} \) and \( \hat{F}_{OP} \) are in turn only approximations to the true value of \( \hat{F} \). The theory that justifies these approximations does not give any clear guidance to which is better to use, and typically, researchers rely on whichever estimate of the information matrix is easiest to calculate. If the two estimates differ a great deal, this may mean that the model is misspecified. White (1982) developed a general test of model specification based on this idea. One option for constructing standard errors when the two estimates differ significantly is to use the "quasi-maximum likelihood" standard errors discussed at the end of this section.

**Likelihood Ratio Test**

Another popular approach to testing hypotheses about parameters that are estimated by maximum likelihood is the likelihood ratio test. Suppose a null hypothesis implies a set of \( m \) different restrictions on the value of the \((a \times 1)\) parameter vector \( \theta \). First, we maximize the likelihood function ignoring these restrictions to obtain the unrestricted maximum likelihood estimate \( \hat{\theta} \). Next, we find an estimate \( \hat{\theta} \) that makes the likelihood as large as possible while still satisfying all the restrictions. In practice, this is usually achieved by defining a new \((a - m) \times 1\) vector \( \lambda \) in terms of which all of the elements of \( \theta \) can be expressed when the restrictions are satisfied. For example, if the restriction is that the last \( m \) elements of \( \theta \) are zero, then \( \lambda \) consists of the first \( a - m \) elements of \( \theta \). Let \( \hat{\theta}(\lambda) \) denote the value of the log likelihood function at the unrestricted estimate, and let \( \hat{\lambda}(\theta) \) denote the value of the log likelihood function at the restricted estimate. Clearly \( \hat{L}(\theta) > \hat{L}(\hat{\theta}(\lambda)) \), and it often proves to be the case that

\[
2(\hat{L}(\hat{\theta}(\lambda)) - \hat{L}(\hat{\theta})) = x^2(m).
\]

For example, suppose that \( a = 2 \) and we are interested in testing the hypothesis that \( \theta_2 = \theta_1 + 1 \). Under this null hypothesis the vector \((\theta_1, \theta_2)'\) can be written as \((\lambda, \lambda + 1)'\), where \( \lambda = \theta_1 \). Suppose that the log likelihood is given by expression [5.7.6]. One can find the restricted MLE by replacing \( \theta_2 \) by \( \theta_1 + 1 \) and maximizing the resulting expression with respect to \( \lambda_1 \):

\[
\hat{L}(\lambda) = -1.5\lambda^2 - 2(\lambda + 1)^2.
\]

The first-order condition for maximization of \( \hat{L}(\lambda) \) is

\[
-3\lambda - 4(\lambda + 1) = 0,
\]

or \( \lambda = -4 \). The restricted MLE is thus \( \hat{\lambda} = (-4, -7)' \), and the maximum value attained for the log likelihood while satisfying the restriction is

\[
\hat{L}(\hat{\lambda}) = (-4)(-7)^2 - (3)(7)^2 = 64 - 147(4 + 3) = 1.71.
\]

The unrestricted MLE is \( \hat{\theta} = (0, 0)' \), at which \( \hat{L}(\hat{\theta}) = 0 \). Hence, [5.8.5] would be

\[
2(\hat{L}(\hat{\theta}(\lambda)) - \hat{L}(\hat{\theta})) = \chi^2(1) = 1.71.
\]

The test here involves a single restriction, so \( m = 1 \). From Table B.2 in Appendix B, the probability that a \( \chi^2(1) \) variable exceeds 3.84 is 0.05. Since 1.71 < 3.84, we accept the null hypothesis that \( \theta_2 = \theta_1 + 1 \) at the 5% significance level.

**Lagrange Multiplier Test**

In order to use the standard errors from [5.8.2] or [5.8.4] to test a hypothesis about \( \theta \), we need only to find the unrestricted MLE \( \hat{\theta} \). In order to use the likelihood ratio test [5.8.5], it is necessary to find both the unrestricted MLE \( \hat{\theta} \) and the restricted MLE \( \hat{\theta} \). The Lagrange multiplier test provides a third principle with which to test a null hypothesis that requires only the restricted MLE \( \hat{\theta} \). This test is useful when it is easier to calculate the restricted estimate \( \hat{\theta} \) than the unrestricted estimate \( \hat{\theta} \).

Let \( \theta \) be an \((a \times 1)\) vector of parameters, and let \( \hat{\theta} \) be an estimate of \( \theta \) that maximizes the log likelihood subject to a set of \( m \) restrictions on \( \theta \). Let \( f(y) \) denote the conditional density of the \( r \)th observation, and let \( h(\theta, y) \) denote the \((a \times 1)\) vector of derivatives of the log of this conditional density evaluated at the restricted estimate \( \hat{\theta} \):

\[
h(\hat{\theta}, y) = \frac{\partial}{\partial \theta} \log f(y|y_{-1}, y_{-2}, \ldots, \theta).
\]

The Lagrange multiplier test of the null hypothesis that the restrictions are true is given by the following statistic:

\[
T^{-1} \left[ \sum_{i=1}^{m} h(\hat{\theta}, y_i) \right] \hat{\theta}^{-1} \left[ \sum_{i=1}^{m} h(\hat{\theta}, y_i) \right] \sim \chi^2(m).
\]

If the null hypothesis is true, then for large \( T \) this should approximately have a \( \chi^2(m) \) distribution. The information matrix \( \hat{F} \) can again be estimated as in [5.8.2] or [5.8.4] with \( \hat{\theta} \) replaced by \( \hat{\theta} \).

**Quasi-Maximum Likelihood Standard Errors**

It is mentioned earlier in this section that if the data were really generated from the assumed density and the sample size is sufficiently large, the second-derivative estimate \( \hat{F}_{2D} \) and the out-of-sample estimate \( \hat{F}_{OP} \) of the information matrix should be reasonably close to each other. However, maximum likelihood estimation may still be a reasonable way to estimate parameters even if the data were not generated by the assumed density. For example, we noted in Section 5.2 that the conditional MLE for a Gaussian AR(1) process is obtained from an OLS regression of \( y_i \) on \( y_{i-1} \). This OLS regression is often a very sensible way to estimate parameters of an AR(1) process even if the true innovations \( \epsilon \) are not i.i.d. Gaussian. Although maximum likelihood may be yielding a reasonable estimate of \( \theta \), when the innovations are not i.i.d. Gaussian, the standard errors proposed in [5.8.2] or [5.8.3] may no longer be valid. An approximate variance-covariance matrix for \( \theta \) that is sometimes valid even if the probability density is misspecified is given by

\[
\hat{L}(\theta - \theta_0)(\theta - \theta_0)' = T^{-1} \hat{F}_{2D} \hat{F}_{OP}^{-1},
\]

This variance-covariance matrix was proposed by White (1982), who described this approach as quasi-maximum likelihood estimation.

5.8. Statistical Inference with Maximum Likelihood Estimation
5.9. Inequality Constraints

A Common Pitfall with Numerical Maximization

Suppose we were to apply one of the methods discussed in Section 5.7 such as steepest ascent to the $AR(1)$ likelihood [5.7.2]. We start with an arbitrary initial guess, say, $\phi = 0.1$. We calculate the gradient at this point, and find that it is positive. The computer is then programmed to try to improve this estimate by evaluating the log likelihood at points described by $\phi(s) = \phi(0) + s \cdot g(\phi(0))$ for various values of $s$, seeing what works best. But if the computer were to try a value for $s$ such that $\phi(s) = \phi(0) + s \cdot g(\phi(0)) = 1.1$, calculation of [5.7.2] would involve finding the log of $1 - 1.1^2 = -0.21$. Attempting to calculate the log of a negative number would typically be a fatal execution error, causing the search procedure to crash.

Often such problems can be avoided by using modified Newton-Raphson procedures, provided that the initial estimate $\theta(0)$ is chosen wisely and provided that the initial search area is kept fairly small. The latter might be accomplished by setting the initial weighting matrix $A(0)$ in [5.7.15] and [5.7.16] equal to a small multiple of the identity matrix, such as $A(0) = (1 \times 10^{-4}) I$. In later iterations, the algorithm should use the shape of the likelihood function in the vicinity of the maximum to keep the search conservative. However, if the true $MLE$ is close to one of the boundaries (for example, if $\phi_{MLE} = 0.998$ in the $AR(1)$ example), it will be virtually impossible to keep a numerical algorithm from exploring what happens when $\phi$ is greater than unity, which would induce a fatal crash.

Solving the Problem by Reparameterizing the Likelihood Function

One simple way to ensure that a numerical search always stays within certain specified boundaries is to reparameterize the likelihood function in terms of an $(n \times 1)$ vector $\lambda$ for which $\theta = g(\lambda)$, where the function $g: \mathbb{R}^n \to \mathbb{R}$ incorporates the desired restrictions. The scheme is then as follows:

\[
\begin{align*}
\text{Input} & \quad \text{Procedure} & \quad \text{Output} \\
\begin{array}{c}
y_1, y_2, \ldots, y_T \\
\text{and } \lambda
\end{array} & \rightarrow \begin{array}{c}
\text{set } \theta = g(\lambda); \\
\text{calculate } L(\theta)
\end{array} & \rightarrow \begin{array}{c}
\text{value of } L(g(\lambda))
\end{array}
\end{align*}
\]

For example, to ensure that $\phi$ is always between $\pm 1$, we could take

\[
\phi = g(\lambda) = \frac{\lambda}{1 + |\lambda|}
\]  

[5.9.1]

The goal is to find the value of $\lambda$ that produces the biggest value for the log likelihood. We start with an initial guess such as $\lambda = 3$. The procedure to evaluate the log likelihood function first calculates

\[
\phi = 3/(1 + 3) = 0.75
\]

and then finds the value for the log likelihood associated with this value of $\phi$ from [5.7.2]. No matter what value for $\lambda$ the computer guesses, the value of $\phi$ in [5.9.1] will always be less than $1$ in absolute value and the likelihood function will be well defined. Once we have found the value of $\hat{\lambda}$ that maximizes the likelihood function, the maximum likelihood estimate of $\phi$ is then given by

\[
\hat{\phi} = \frac{\hat{\lambda}}{1 + |\hat{\lambda}|}
\]

This technique of reparameterizing the likelihood function so that estimates always satisfy any necessary constraints is often very easy to implement. However, one note of caution should be mentioned. If a standard error is calculated from the matrix of second derivatives of the log likelihood as in [5.8.3], this represents the standard error of $\lambda$, not the standard error of $\phi$. To obtain a standard error for $\phi$, the best approach is first to reparameterize the likelihood function in terms of $\lambda$ to find the $MLE$, and then to reparameterize in terms of $\phi$ to calculate the matrix of second derivatives evaluated at $\phi$ to get the final standard error for $\phi$. Alternatively, one can calculate an approximation to the standard error for $\phi$ from the standard error for $\lambda$, based on the formula for a Wald test of a nonlinear hypothesis described in Chapter 14.

Parameterizations for a Variance-Covariance Matrix

Another common restriction one needs to impose is that a variance parameter $\sigma^2$ be positive. An obvious way to achieve this is to parameterize the likelihood in terms of $\lambda$ which represents $\pm 1$ times the standard deviation. The procedure to evaluate the log likelihood then begins by squaring this parameter $\lambda$:

\[
\sigma^2 = \lambda^2
\]

and if the standard deviation $\sigma$ is itself called, it is calculated as

\[
\sigma = \sqrt{\lambda^2}
\]

More generally, let $\Omega$ denote an $(n \times n)$ variance-covariance matrix:

\[
\Omega = \begin{bmatrix}
\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\
\sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn}
\end{bmatrix}
\]

Here one needs to impose the condition that $\Omega$ is positive definite and symmetric. The best approach is to parameterize $\Omega$ in terms of the $n(n + 1)/2$ distinct elements of the Cholesky decomposition of $\Omega$:

\[
\Omega = PP^t,
\]

[5.9.2]

where

\[
P = \begin{bmatrix}
\lambda_{12} & 0 & \cdots & 0 \\
\lambda_{21} & \lambda_{22} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
\lambda_{n1} & \lambda_{n2} & \cdots & \lambda_{nn}
\end{bmatrix}
\]

No matter what values the computer guesses for $\lambda_{12}, \lambda_{21}, \ldots, \lambda_{nn}$, the matrix $\Omega$ calculated from [5.9.2] will be symmetric and positive semidefinite.
Parameterizations for Probabilities

Sometimes some of the unknown parameters are probabilities \( p_1, p_2, \ldots, p_K \), which must satisfy the restrictions

\[
0 \leq p_i \leq 1 \quad \text{for } i = 1, 2, \ldots, K
\]

\[
p_1 + p_2 + \cdots + p_K = 1.
\]

In this case, one approach is to parameterize the probabilities in terms of \( \lambda_1, \lambda_2, \ldots, \lambda_{K-1} \), where

\[
p_i = \lambda_i^2 (1 + \lambda_1^2 + \lambda_2^2 + \cdots + \lambda_{K-1}^2) \quad \text{for } i = 1, 2, \ldots, K - 1
\]

\[
p_K = 1 / (1 + \lambda_1^2 + \lambda_2^2 + \cdots + \lambda_{K-1}^2).
\]

More General Inequality Constraints

For more complicated inequality constraints that do not admit a simple reparameterization, an approach that sometimes works is to put a branching statement in the procedure to evaluate the log likelihood function. The procedure first checks whether the constraint is satisfied. If it is, then the likelihood function is evaluated in the usual way. If it is not, then the procedure returns a large negative number in place of the value of the log likelihood function. Sometimes such an approach will allow an MLE satisfying the specified conditions to be found with simple numerical search procedures.

If these measures prove inadequate, more complicated algorithms are available. Judge, Griffiths, Hill, and Lee (1980, pp. 747–49) described some of the possible approaches.

APPENDIX 5.A. Proofs of Chapter 5 Propositions

- Proof of Proposition 5.1.

(a) By Taylor’s theorem,

\[
\Delta (\theta^{(m+1)}) = \Delta (\theta^{(m)}) + [g(\theta^{(m)})]' (\theta^{(m+1)} - \theta^{(m)}) + R_2(\theta^{(m)}, \theta^{(m+1)}).
\]

Substituting [5.7.18] into [5.5.1],

\[
\Delta (\theta^{(m+1)}) - \Delta (\theta^{(m)}) = [g(\theta^{(m)})]' + \Delta (\theta^{(m+1)}) + R_2(\theta^{(m)}, \theta^{(m+1)}).
\]

Since \( \Delta (\theta^{(m)}) \) is positive definite and since \( g(\theta^{(m)}) \neq 0 \), expression [5.5.2] establishes that

\[
\Delta (\theta^{(m+1)}) - \Delta (\theta^{(m)}) = \sum (\theta^{(m+1)} - \theta^{(m)}) + R_2(\theta^{(m)}, \theta^{(m+1)}),
\]

where \( \Delta (\theta^{(m)}) > 0 \). Moreover, \( x^{-1} R_2(\theta^{(m)}, \theta^{(m+1)}) = 0 \) as \( s \to 0 \). Hence, there exists an \( s \) such that \( \Delta (\theta^{(m+1)}) - \Delta (\theta^{(m)}) > 0 \), as claimed.

(b) Direct differentiation reveals

\[
\frac{\partial \Delta (\theta^{(m+1)})}{\partial \theta} = \frac{\partial \Delta (\theta^{(m)})}{\partial \theta} + \frac{\partial \Delta (\theta^{(m+1)})}{\partial \theta} + \cdots + \frac{\partial \Delta (\theta^{(m+1)})}{\partial \theta},
\]

[5.5.3]

with the last line following from [5.7.18]. The first-order conditions set [5.5.3] equal to zero, which implies

\[
0 = [g(\theta^{(m+1)})]' + [g(\theta^{(m)}')] (\theta^{(m+1)} - \theta^{(m)}),
\]

with the last line again following from [5.7.18]. This establishes the claim in [5.5.19].

(c) Let \( y \) be any \( (n \times 1) \) nonzero vector. The task is to show that \( y' \Delta (\theta^{(m+1)}) y > 0 \). Observe from [5.7.16] that

\[
y' \Delta (\theta^{(m+1)}) y = y' \Delta (\theta^{(m)}) y + y' \Delta (\theta^{(m+1)}) (\Delta (\theta^{(m+1)}) y)
\]

\[
= y' (\Delta (\theta^{(m+1)})(\Delta (\theta^{(m+1)})) \Delta (\theta^{(m+1)})) y
\]

[5.5.4]

Since \( \Delta (\theta^{(m)}) \) is positive definite, there exists a nonsingular matrix \( P \) such that

\[
\Delta (\theta^{(m)}) = P^2.
\]

Define

\[
y^* = P y,
\]

\[
x^* = P \Delta (\theta^{(m+1)}).
\]

Then [5.5.4] can be written as

\[
y' \Delta (\theta^{(m+1)}) y = y' P P^T y - \frac{y' P P^T (\Delta (\theta^{(m+1)}) \Delta (\theta^{(m+1)})) P P^T y}{\Delta (\theta^{(m+1)})}.
\]

[5.5.5]

Recalling equation [4.5.6], the first two terms in the last line of [5.5.5] represent the sum of squared residuals from an OLS regression of \( y \) on \( x^* \). This cannot be negative,

\[
y' \Delta (\theta^{(m+1)}) y = y' P P^T y - \frac{y' P P^T (\Delta (\theta^{(m+1)}) \Delta (\theta^{(m+1))}) P P^T y}{\Delta (\theta^{(m+1)})} \leq 0;
\]

[5.5.6]

it would equal zero only if the OLS regression has a perfect fit, or if \( y^* = \beta x^* \) or \( P y = \beta x^* \). Since \( P \) is nonsingular, expression [5.5.6] would only be zero if \( y = \beta \Delta (\theta^{(m+1)}) \) for some \( \beta \). Consider two cases.

Case 1. There is no \( \beta \) such that \( y = \beta \Delta (\theta^{(m+1)}) \). In this case, the inequality [5.5.6] is strict and [5.5.5] implies

\[
y' \Delta (\theta^{(m+1)}) y > \frac{y' P P^T y}{\Delta (\theta^{(m+1)})}.
\]

Since \( y' \Delta (\theta^{(m+1)}) y \geq 0 \), it follows that \( y' \Delta (\theta^{(m+1)}) y > 0 \), provided that

\[
\Delta (\theta^{(m+1)}) > 0.
\]

[5.5.7]

But, from [5.5.19],

\[
(\Delta (\theta^{(m+1)})) - (\Delta (\theta^{(m+1)})) = [g(\theta^{(m+1)}) - g(\theta^{(m+1}})] - [g(\theta^{(m+1)}) - g(\theta^{(m+1})] = 0
\]

\[
= -g(\theta^{(m+1)}) + g(\theta^{(m+1})]
\]

[5.5.8]

with the last line following from [5.7.18]. But the final term in [5.5.8] must be negative, by virtue of the fact that \( \Delta (\theta^{(m)}) \) is positive definite, \( s > 0 \), and \( g(\theta^{(m)}) = 0 \). Hence, [5.5.7] holds, meaning that \( \Delta (\theta^{(m+1)}) > 0 \) is positive definite for this case.

Case 2. There exists a \( \beta \) such that \( y = \beta \Delta (\theta^{(m+1)}) \). In this case, [5.5.6] is zero, so that

\[
y' \Delta (\theta^{(m+1)}) y = \frac{y' \Delta (\theta^{(m+1)})(\Delta (\theta^{(m+1)})) y}{\Delta (\theta^{(m+1)})}.
\]

\[
= \beta (\Delta (\theta^{(m+1)})) (\Delta (\theta^{(m+1)})) \beta (\Delta (\theta^{(m+1)})
\]

[5.5.9]

\[
= -\beta (\Delta (\theta^{(m+1)})) (\Delta (\theta^{(m+1)})) > 0,
\]

as in [5.5.8].

Appendix 5.A. Proofs of Chapter 5 Propositions

149
Bootstrap Methods in Econometrics

JAMES G. MacKINNON
Department of Economics, Queen's University, Kingston, Ontario, Canada

There are many bootstrap methods that can be used for econometric analysis. In certain circumstances, such as regression models with independent and identically distributed error terms, appropriately chosen bootstrap methods generally work very well. However, there are many other cases, such as regression models with dependent errors, in which bootstrap methods do not always work well. This paper discusses a large number of bootstrap methods that can be useful in econometrics. Applications to hypothesis testing are emphasized, and simulation results are presented for a few illustrative cases.

1 Introduction

Bootstrap methods involve estimating a model many times using simulated data. Quantiles computed from the simulated data are then used to make inferences from the actual data. The term "bootstrap" was coined by Efron (1979), but bootstrap methods did not become popular in econometrics until about 10 years ago. One major reason for their increasing popularity in recent years is the staggering drop in the cost of numerical computation over the past two decades.

Although bootstrapping is quite widely used, it is not always well understood. In practice, bootstrapping is often not as easy to do, and does not work as well, as seems to be widely believed. Although it is common to speak of "the bootstrap," this is a rather misleading term, because there are actually many different bootstrap methods. Some bootstrap methods are very easy to implement, and some bootstrap methods work extraordinarily well in certain cases. But bootstrap methods do not always work well, and choosing among alternative ones is often not easy.

The next section introduces bootstrap methods in the context of hypothesis testing. Section II then discusses methods for bootstrapping regression models. Section IV deals with bootstrap standard errors, and Section V discusses bootstrap confidence intervals. Section VI deals with bootstrap methods for dependent data, and Section VII concludes.

II Hypothesis Testing

Suppose that $h$ is the realized value of a test statistic if we knew the cumulative distribution function (CDF) of $H$ under the null hypothesis, say $F(h)$. If we reject the null hypothesis whenever $h$ is abnormal in some sense, for a test that rejects in the upper tail of the distribution, we might choose to calculate a critical value at level $a$, say $a_n$, as defined by the equation

$$1 - F(a_n) = a. \quad (1)$$

Then we would reject the null whenever $h > a_n$. For example, where $F(h)$ is the $X^2(1)$ distribution and $a = 0.05$, $a_n = 3.84$.

An alternative approach, which is preferable in most circumstances, is to calculate the $P$-value, or marginal significance level,

$$p(h) = 1 - F(h). \quad (2)$$

and reject whenever $p(h) < a$. It is easy to see that these two procedures must yield identical inferences, since they must be greater than $a_n$ whenever $p(h) < a$.

In most cases of interest to econometricians, we do not know $F(h)$. Until recently, the usual approach in such cases has been to replace it by an approximate CDF, say $F_m(h)$, based on asymptotic theory. This approach works well when $F(h)$ is a good approximation to $F_m(h)$, but that is by no means always true.

The bootstrap provides another way to approximate $F_m(h)$, which may provide a better approximation. It can be used even when $h$ is complicated and difficult to analyze theoretically. It is not necessary to have $F_m(h)$ as a known asymptotic distribution.

In order to perform a bootstrap test, we must generate $B$ bootstrap samples, indexed by $j$, that satisfy the null hypothesis. A bootstrap sample is a simulated data set. The procedure for generating the bootstrap samples, which always involves a random number generator, is called a bootstrap data-generating process, or bootstrap DGP. Bootstrap DGP for regression models will be discussed in the next section.

For each bootstrap sample, we compute a bootstrap test statistic, say $h_j$, usually (but not always) by the same procedure used to calculate $h$ from the real sample. The bootstrap $P$-value is then

$$p_j(h) = \frac{1}{B} \sum_{k=1}^{B} I(h_k > h_j). \quad (3)$$

where $I$ denotes the indicator function, which is equal to 1 when its argument is true and 0 otherwise. Equation (3) can also be written as

$$p_j(h) = F_m(h_j). \quad (4)$$

where $F_m(h_j)$ denotes the empirical distribution function, or EDF, of the $h_j$. If we let the number of bootstrap samples, $B$, tend to infinity, then $p_j(h)$ tends to $F_m(h)$, the CDF of the $h_j$.

The bootstrap $P$-value (4) looks just like the true $P$-value (2), but with the EDF of the bootstrap distribution, $F_m(h_j)$, replacing the unknown CDF $F(h)$. From this, it is clear that bootstrap tests generally not be exact. Thus, the probability of rejecting the null at level $a$ will generally not be equal to $a$. Most of the problems with bootstrap tests arise because $F_m(h_j)$ is only an estimate of $F(h)$, but because $F(h)$ may not be a good approximation to $F_m(h)$.

The bootstrap $P$-value (3) is appropriate if we wish to reject the null hypothesis whenever $h$ is sufficiently large and positive. However, for a quantity such as $a$, a statistic, that can take on either sign, it is generally more appropriate to use either

$$p_j(h) = \frac{1}{B} \sum_{k=1}^{B} I(h_k > 0), \quad (5)$$

or, alternatively,

$$p_j(h) = \frac{1}{B} \sum_{k=1}^{B} I(h_k < 0). \quad (6)$$

In Equation (5), we implicitly assume that the distribution of $h$ is symmetric around zero. In Equation (6), however, we make no such assumption. The factors of 2 in (6) is necessary because there are two tails, and $F_m(h)$ could be far out in either tail by chance. Without it, $p_j(h)$ would lie between 0 and 0.5. There is no guarantee that $p_j(h)$ and $p_j(h)$ will be similar. Indeed, if the mean of the $h_j$ is far from zero, they may be quite different, and tests based on $p_j(h)$ may have very different power properties. Unless the sample size is large, tests based on $p_j(h)$ will probably be more reliable, under the null hypothesis, than tests based on $p_j(h)$.

1 Monte Carlo Tests

There is an important special case in which bootstrap tests are exact. For this result to hold, we need two conditions:

1. The test statistic is pivotal, which means that its distribution does not depend on any unknown parameters.
2. The number of bootstrap samples $B$ is such that $o(B+1)$ is an integer, where $o$ is the level of the test.

When these two conditions hold, a bootstrap test is called a Monte Carlo test. It is not difficult to see why Monte Carlo tests are exact. By condition 1, the null distribution, $F(h)$, is the same as the distribution of the test statistic. There are also many other useful results. For example, if $B = 999$ and $a = 0.05$, $p_j(h)$ would be less than 0.05, and we will consequently reject the null, whenever $h$ is one of the 50 largest test statistics. Monte Carlo tests can be applied to many procedures for testing the specification of linear regression models with fixed regressors and normal
and treat \(z_{\alpha} \) like any other test statistic for the purpose of bootstrapping. Note that \(z_{\alpha} \) will be asymptotically normal whenever the test statistics \(T \sim \alpha \) have a jointly asymptotic distribution that is of free of unknown parameters. Whenever we perform two or more tests, it is dangerous to rely on ordinary \(P \) values, because the probability of obtaining a low \(P \) value by chance increases with the number of tests we perform. This can be a serious problem when testing model specification and when estimating models with many parameters the significance of which we wish to test. The overall size of such a procedure can be very much larger than the nominal level of each individual test.

By using the bootstrap, it is remarkable easy to obtain an asymptotically valid \(P \) value for the most extreme test statistic actually observed. By analogy with (7), we can define

\[
p_{\alpha} = \min \{p(z_{1}, \ldots, z_{\alpha}) \mid z_{\alpha} \}
\]

where \(p(z_{1}, \ldots, z_{\alpha}) \) denotes the \(P \) value in most cases computed analytically, for the \(i \)th test statistic \(T_i \). Bootstrapping \(p_{\alpha} \) is just like bootstrapping \(z_{\alpha} \) defined in (7). Westfall and Young (1993) provide an extensive discussion of multiple hypothesis testing based on bootstrap methods.

There is a widespread misconception that bootstrap tests are less powerful than other types of tests. Except for the case where the loss of power is small because of \(B \) is large, this is entirely false. Comparing the powers of tests that are not exactly the same is difficult in general. However, it is evident that the powers of asymptotic and boot tests that are based on the same test statistic are very similar when the tests have been properly size adjusted.

III. Bootstrapping Regression Models

What determines how reliably a bootstrap test performs is how well the bootstrap DGP mimics the features of the true DGP that matter for the distribution of the test statistic. Essentially the same thing can also be said for bootstrap confidence intervals and bootstrap standard errors. In this section, I discuss four different types of bootstrap DGP for regression models with uncorrelated error terms. Models with dependent errors will be discussed in Section VI.

Consider the linear regression model

\[
\hat{\beta} = X \beta + \epsilon, \quad \epsilon \sim N(0, \Sigma)
\]

where there are \(n \) observations. Here \(X \) is a row vector of regressors on \(X \), and \(\beta \) is a column vector of regression coefficients. The regressors may include lagged dependent variables, but \(\epsilon \) is not explosive and does not have a unit root.

There are a great many ways to specify bootstrap DGP for the model (8). Some require very strong assumptions about the error terms \(\epsilon \), whereas others require much weaker ones. In general, making stronger assumptions results in better performance if those assumptions are satisfied, but it leads to asymptotically invalid inferences if they are not. All the methods that will be discussed can also be applied sometimes with minor changes, to non-linear regression models.

1) The Residual Bootstrap

If the error term is \(i \) is independent and identically distributed with common variance \(\sigma^2 \), then we can generally make very accurate inferences by using the residual bootstrap. We do not need to assume that the errors follow the normal distribution or any other known distribution.

The first step in the residual bootstrap is to obtain OLS estimates \(\hat{\beta} \) and residuals \(\hat{\epsilon} \). Unless the quantity to be bootstrapped is invariant to the variance of the error terms (this is true of test statistics for serial correlation, for example), it is advisable to rescale the residuals so that they have the correct distribution.

The simplest type of rescaled residual is

\[
\hat{\epsilon}_{\alpha} = \frac{\hat{\epsilon}_{\alpha}}{\hat{\sigma}_{\alpha}}
\]

The first factor here is the inverse of the square root of the row vector of the factor by which \(i \) times the sum of squared residuals underestimates \(\sigma^2 \). A somewhat more complicated method of rescaling the "hat matrix" \(X (X'X)^{-1} X' \) to rescale each residual by a different factor. It may work a bit better than (9) when some observations have high leverage; details are given in Davidson and MacKinnon (2000b).

The residual bootstrap DGP using rescaled residuals generates a typical observation of the bootstrap sample by the equation

\[
\tilde{X} = X \beta + \tilde{\epsilon}, \quad \tilde{\epsilon} \sim EDP(\hat{\sigma})
\]

The bootstrap errors \(\tilde{\epsilon} \) here are said to be "empirical" from the \(\hat{\sigma} \). That is, they are drawn from the empirical distribution function, or EDF, of the \(\hat{\sigma} \). This function assigns probability \(1/n \) to each of the \(\hat{\epsilon} \). Thus, each of the bootstrap error terms can take on a possible value, namely, the values of the \(\hat{\epsilon} \), each with probability \(1/n \).
When the regressors include lagged dependent variables, the bootstrap DGP (10) is normally implemented recursively, so that $x_t^*$ depends on its own lagged values. Either pre-sample values of $y_t$ or drawings from the unconditional distribution of the $y_t$ may be used to start the recursive process.

(iii) The Parametric Bootstrap

It may seem remarkable that the residual bootstrap should work at all, let alone that it should work well. The reason it often works well (as we shall see in an example below) is that least squares estimates and test statistics are generally not very sensitive to the distribution of the error terms.

Of course, if this distribution is assumed to be normal, we can replace (10) by the parametric bootstrap DGP

$$x_t = X_t \tilde{\beta} + \epsilon_t, \quad \epsilon_t \sim NID(0, \sigma^2).$$

(11)

Here it is assumed that the errors are normally distributed, and so the bootstrap error terms are independent normal random variates with variance $\sigma^2$, the usual OLS estimate of the error variance. Similar methods can be used with any model estimated by maximum likelihood, but their validity generally depends on the strong assumptions inherent in maximum likelihood estimation.

For inferences about regression coefficients, it generally makes very little difference whether we use the usual full bootstrap or the parametric bootstrap with normal errors, whether or not the errors are actually normally distributed. However, for inferences about other aspects of a model, such as possible heteroskedasticity, it can make a large difference.

(iii) Restricted vs. Unrestricted Estimates

As described, the residual and parametric bootstrap procedures use unrestricted estimates of $\beta$. This is inappropriate in the case of specification tests, such as tests for serial correlation or non-stationary hypothesis tests. For example, Davidson and MacKinnon (1987) observe that parametric bootstrap methods tend to test too many and identify too well in the sense that they generally work very well. However, using unrestricted estimates is not appropriate if we are testing a restriction on $\beta$.

Both the methods described so far can easily be modified to impose restrictions on the vector $\beta$. In the first step, we simply need to estimate the model under the null to obtain restricted estimates $\tilde{\beta}$. Then we use these estimates instead of $\beta$ in the bootstrap DGP (11). We can resample from either restricted or unrestricted residuals. In most cases, it seems to make little difference which we use.

There are two reasons to use restricted parameter estimates in the bootstrap DGP when testing restrictions on $\beta$. The first reason is that, if we do not do so, the bootstrap DGP will not satisfy the null hypothesis. If we identify $\tilde{\beta}$ in the $\tilde{\tau}$ such a case, the bootstrap test will be greatly in powers. It is possible to get around this problem by changing the null hypothesis used to compute the $\tilde{t}$, as we shall discuss below to the context of the pairs bootstrap, but it is preferable to avoid the need to do so.

The second reason for using restricted parameter estimates is that imposing the restrictions on the null hypothesis yields more efficient estimates of the nuisance parameters upon which the distribution of the test statistic may depend. This generally makes bootstrap tests more reliable, because the parameters of the bootstrap DGP are estimated more precisely. For a detailed discussion of how the reliability of bootstrap tests depends on the estimates of nuisance parameters, see Davidson and MacKinnon (1999).

(iv) The Wild Bootstrap

The residual bootstrap is not valid if the error terms are not independently and identically distributed, but two other commonly used bootstrap methods are valid in this case. The first of these is the 'wild bootstrap', which was proposed by Wu (1986) for regression models with heteroskedastic errors.

For a model like (8) with independent but heteroskedastic errors, the wild bootstrap DGP is

$$x_t = X_t \tilde{\beta} + \tilde{\epsilon}_t(\tilde{\beta}),$$

(12)

where $\tilde{\epsilon}_t(\tilde{\beta})$ is a transformation of the $t$th residual $\tilde{\epsilon}_t$, and $\tilde{\epsilon}_t$ is a random variable with mean 0 and variance 1. One possible choice for $\tilde{\epsilon}_t(\tilde{\beta})$ is just $\tilde{\epsilon}_t$, but a better choice is

$$\tilde{\epsilon}_t(\tilde{\beta}) = \frac{\tilde{\epsilon}_t}{\sqrt{1 + h_{t-1}^2}} - h_{t-1},$$

(13)

where $h_t$ is the $t$th diagonal of the 'hat matrix' which was defined just after (9). When the $\tilde{\epsilon}_t(\tilde{\beta})$ are defined by (13), they would have a constant variance if the error terms were homoskedastic.

There are various ways to specify the distribution of the $\tilde{\epsilon}_t$. The simplest, but not the most popular, is

$$\tilde{\epsilon}_t = 1 \text{ with probability } \frac{1}{2},$$

$$\tilde{\epsilon}_t = -1 \text{ with probability } \frac{1}{2}.\quad (14)

Thus, each bootstrap error term can take on only two possible values. Davidson and Flaxman (2001) have shown that wild bootstrap tests based on (14) usually perform better than wild bootstrap tests that use one other distributions when the conditional distribution of the error terms is approximately symmetric. When it is sufficiently asymmetric, however, it may be better to use another two-point distribution, which is the one that is most commonly used in practice:

$$\tilde{\epsilon}_t = \frac{(S_t - 1)/2}{\sqrt{(S_t + 1)/2}} \text{ with probability } \frac{(S_t + 1)/2}{2S_t},$$

(15)

$$\text{where } S_t = \frac{(Y_t - Y_t^* - X_t \tilde{\beta})^2}{(X_t X_t)^{-1} X_t \tilde{\beta} \tilde{\beta}' X_t X_t^{-1}}.$$

The wild bootstrap may seem like a rather strange procedure. When a distributions like (14) or (15) is used, each error term can take on only two possible values, which depend on the size of the residuals. Thus, in certain respects, the bootstrap DGP cannot possibly resemble the real one. However, the expectation of the square of $\tilde{\epsilon}_t$ is approximately the variance of $\tilde{\epsilon}_t$. The wild bootstrap error terms will, on average, have about the same variance as the $\tilde{\epsilon}_t$. In many cases, this seems to be enough for the wild bootstrap DGP to mimic the essential features of the true DGP.

As with the residual bootstrap, the null hypothesis can, and should, be imposed whenever we are using the wild bootstrap in test a hypothesis about $\beta$. Although it might seem that the wild bootstrap works only with cross-sectional data or static models, variants of it can be also used with dynamic models, provided the error terms are uncorrelated; see Gonzalo and Kilian (2004).

(v) The Pairs Bootstrap

Another method that can accommodate heteroskedasticity is the 'pairs bootstrap', which was proposed by Freedman (1984). The idea was applied to regressions with instrumental variables by Freedman (1984) and Freedman and Peters (1984). The idea is to assume that the data is divided into pairs of residuals. Then, in the case of the regression model (8), we resample from the matrix (Y X) with typical row $(y_t, X_t)$. Each observation of the bootstrap sample is $(y_t^*$, $X_t^*$), a randomly chosen row from (Y X). This method is called the pairs (or pairs bootstrap) because the dependent variable $y_t^*$ and the independent variables $X_t^*$ are always selected in pairs.

Unlike the residual and wild bootstrap, the pairs bootstrap does not condition on $X$. Instead, each bootstrap sample has a different $X$ matrix. This method implicitly assumes that each observation $(y_t, X_t)$ is an independent random drawing from a multivariate distribution. It does not require that the error terms be homoskedastic, and it even works for dynamic models if we treat lagged dependent variables like any other element of $X$.

In the case of multivariate models, we can combine the pairs and residual bootstraps. We organize the residuals into a matrix and then apply the pairs bootstrap to its rows, adding the bootstrap error terms generated in the appropriate fixed values to yield the bootstrap data. This method preserves the cross-equation correlations of the residuals without imposing any distributional assumptions on the bootstrap error term.

The pairs bootstrap is very easy to implement, and it can be applied to an enormous range of models. However, it suffers from two major disadvantages. The first of these is that the bootstrap DGP does not impose any restrictions on $\beta$. If we are testing such restrictions, as opposed to estimating standard errors or forming confidence intervals, we need to modify the bootstrap test statistic so that it is testing something that is true in the bootstrap DGP. Suppose the actual test statistic takes the form of a statistic for the hypothesis that $\beta = \beta^*$:

$$\tau = \beta - \beta^*.\quad (16)$$

Then $\tilde{\beta}$ is the unrestricted estimate of the parameter $\beta$ that is being tested, and $\beta^*$ is its standard error. Then, for bootstrap testing to be valid, we must use the bootstrap test statistic

$$\tau_t = \tilde{\beta}_t - \beta^*.\quad (17)$$

where $\tilde{\beta}_t$ is the estimate of $\beta$ from the $t$th bootstrap sample, and $\beta^*$ is its standard error, calculated using the bootstrap sample by whatever procedure was employed to calculate $\beta^*$ using the actual sample. As the estimate of $\beta$ from the bootstrap samples should, on average, be equal to $\beta$, at least asymptotically, the null hypothesis tested by $\tau_t$ is 'true' for the pairs bootstrap DGP.

This other deficiency of the pairs bootstrap is thus, compared to the residual bootstrap (when it is valid) and to the wild bootstrap, the pairs bootstrap generally does not yield very accurate results. This is primarily because it does not condition on the actual $X$ matrix of the regression, we will examine in a case in which the pairs bootstrap does not work particularly well.
(iv) A Comparison of Several Methods

To demonstrate how well, or how badly, various bootstrap procedures perform, it is necessary to perform a simulation experiment. As an illustration, consider testing the null hypothesis that \( H_0: \beta_1 = 0.9 \) in the autoregressive model

\[
X_t = \beta_1 X_{t-1} + \epsilon_t, \quad n = N(0, \sigma^2).
\]  

(18)

Standard tests are not exact here, because \( \beta_1 \), the OLS estimate of \( \beta_1 \), is biased. All tests are based on the usual \( t \) statistic

\[
t = \frac{\hat{\beta}_1 - 0.9}{s_{\hat{\beta}_1}}.
\]  

(19)

It may seem odd that the null hypothesis is that \( \beta_1 = 0.9 \) rather than \( \beta_1 = 0 \) or \( \beta_1 = 1 \). The reason for not examining tests for \( \beta_1 = 0 \) or \( \beta_1 = 1 \) is that asymptotic methods work pretty well for that case, and there is not much to be gained by using the bootstrap. The reason for not examining tests for \( \beta_1 = 1 \) is that the asymptotic theory changes drastically when there is a unit root. The values of \( \beta_1 \) and \( \sigma \) seem to have only a small effect on the results; in the experiments, these values were \( \beta_1 = 1 \) and \( \sigma = 1 \).

The experiments deal with five methods of inference. The first uses Student's \( t \) distribution, which is valid only asymptotically in this case. The second is the residual bootstrap using restricted estimates and restricted residuals generated using \( \hat{\beta}_1 \), called the 'RR bootstrap' for short. The third is the residual bootstrap using unrestricted estimates and unrestricted residuals, called the 'UR bootstrap' for short. The fourth is the wild bootstrap using the untransformed distribution (14) and residuals rescaled by (13). Each experiment had 100,000 replications, with \( n = 120 \). There is a reasonably good value of \( \sigma \) that should generally be used in practice, but in a simulation experiment with a large number of replications, the randomness due in \( \sigma \) being small tends to average out across the replications. Thus, when we are studying the properties of a bootstrap test under the null hypothesis, there is generally no reason to use a large value of \( \sigma \). Experiments were performed for each of the following sample sizes: 10, 15, 20, 25, 30, 40, 50, 60, 80, 100, 150, 200, 300, 400, 500, and 1200. Each of these is larger than its predecessor by approximately the square root of 2.

We can see from Figure 1 that inference based on the \( t \) distribution is seriously unreliable. It improves as \( n \) increases, but it is still somewhat

An unsatisfactory feature of the bootstrap is that the standard errors are computed using the empirical standard deviations of the bootstrap estimates. These are not always reliable, especially when the sample size is small. There are several ways to calculate a standard error. The one used in the experiments is based on the bootstrapping distribution on the \( \hat{\beta}_1 \) and \( s_{\hat{\beta}_1} \), which will be described in the next section.

Figure 2 shows the results of this second set of experiments. All the bootstrap methods now reject less frequently than the \( t \) distribution for all sample sizes. However, the RR bootstrap performs quite poorly when \( n \) is small. The wild bootstrap seems to perform best when \( n \) is very large, which is in accord with theory. However, the bootstrap actually underperforms in this case, which is somewhat worrying. The surprisingly good performance of the RR bootstrap, even though it does not provide heteroscedasticity, is presumably because the bias is \( \hat{\beta}_1 \) is much more important than the heteroscedasticity of the error terms.

Although there are many situations in which bootstrap standard errors are useful (we will encounter one in the next section), the error terms may provide no advantage. In the context of ordinary least squares, however, it makes absolute no sense to use bootstrap standard errors. There are two widely used estimators for the covariance matrix of the OLS parameter vector \( \hat{\beta} \) in the model (8) when the error terms are independent. The best known, which is valid when the error terms are homoscedastic, is

\[
\hat{\Sigma}(\hat{\beta}) = (n-1)X'X^{-1}X'X^{-1}.
\]  

(23)

Under heteroscedasticity of unknown form, this estimator is invalid. Instead, we would use a heteroscedastic-consistent covariance matrix estimator, or HCCME, of the form

\[
\hat{\Sigma}(\hat{\beta}) = (n-1)X'X^{-1}X'X^{-1}.
\]  

(24)

where \( \hat{\Sigma} \) is an \( n \times n \) diagonal matrix with diagonal elements equal to the squared residuals or, preferably, some transformation of those that is
designed to offset the tendency of least squares residuals to be too small. The HCCM of (24) divides each of the squared residuals by 1 - \delta, where \delta is a diagonal element of the “no matrix” that was defined just after (9); see Davidson and MacKinnon (2006: Chapter 5).

Whenever the bootstrap DGP, the bootstrap covariance matrix is

\[ V = \frac{1}{B-1} \sum_{b=1}^{B} (Y_{b} - \hat{\beta} Y)^{2} \],

(25)

where the notation should be obvious; compare (22).

For a residual bootstrap DGP like (10), it can be shown that, as \alpha and \beta become large, (25) tends to

\[ \sigma^{2} (X'X)^{-1} \],

(26)

where \sigma^2 is the average variance of the bootstrap error terms. The matrix tends to the same limit as (23). Thus, in this case, the bootstrap covariance matrix (25) is valid if the errors are independent and homoscedastic; not otherwise.

In contrast, for the wild bootstrap, the bootstrap covariance matrix (25) is approximately equal to the matrix

\[ \frac{1}{B} \sum_{b=1}^{B} (X'X)^{-1} (X'X)^{-1} \],

(27)

where \alpha is the vector of bootstrap error terms for the jth bootstrap sample. This looks a lot like the HCCM of (24). The matrix in the middle here is approximately equal to \sigma^2 (X'X)^{-1}. Thus, for \sigma and \alpha reasonably large, we would expect (27) to be very similar to the HCCM of (24). A similar argument can be applied to the paired bootstrap, see Flachaire (2003).

We have seen that, for a linear regression model, there is nothing to be gained by using a bootstrap covariance matrix instead of a conventional one like (24) or (25). However, when convenient analytical results like these are not available, bootstrap covariance matrices and standard errors can be very useful.

© 2006 The Economic Society of Australia
(ii) Percentile t Confidence Intervals
A method that has better properties than the simple bootstrap interval, at least in theory, is the "percentile t" method, also called "bootstrap t" and "Studentized bootstrap", which has been advocated by Hall (1992). A percentile t confidence interval for \( \theta \) at level 1 - \( \alpha \) is
\[
[\hat{\theta} - t_{1-\alpha/2} s_{1}, \hat{\theta} + t_{1-\alpha/2} s_{1}],
\]
where \( s_{1} \) is the standard error of \( \hat{\theta} \) and \( t_{1} \) is the \( \alpha \)-quantile of the bootstrap t statistics
\[
t_{1} = \frac{\hat{\theta} - \bar{\theta}}{s_{t_1}}.
\]
For example, if \( \alpha = 0.05 \) and \( \bar{\theta} = 999 \), \( s_{t_1} \) will be number 975, and \( t_{0.95} \) will be number 2.5, in the sorted list of the \( t_{1} \). The use of \( \bar{\theta} = 999 \) in this example is not an accident. A fairly large value of \( \bar{\theta} \) is needed if the quantities of the distribution of the \( \hat{\theta} \) are to be estimated accurately, and, as with bootstrap tests, it is desirable for \( \bar{\theta} \) to be chosen in such a way that \( \delta(\bar{\theta}) = 1 \) is an integer.

The interval (31) looks very much like an ordinary confidence interval based on inverting a statistic, except that quantities of the bootstrap distribution of the \( \hat{\theta} \) are used instead of quantities of the Student's t distribution. Because of this, the percentile t method implicitly performs a sort of correction. When the median of the \( \hat{\theta} \) is positive (negative), the percentile t interval tends to be shifted to the left (right) relative to an asymptotic interval based on the normal or Student's t distribution.

In theory, percentile t confidence intervals achieve "higher-order accuracy" relative to asymptotic intervals or the simple bootstrap interval (31). This means that the rate at which the error in coverage probability declines as \( n \) increases is faster than it is for asymptotic methods. However, as we will see in the next subsection, percentile t intervals do not always perform well in practice (MacKinnon, 2002).

The percentile t method evidently cannot be used if \( s(\hat{\theta}) \) cannot be calculated. It should not be used if \( s(\hat{\theta}) \) is unreliable or strongly dependent on \( \hat{\theta} \), as its excellent theoretical properties do not seem to apply in practice in such cases. This method seems to be particularly useful when the \( t \) statistic for \( \hat{\theta} \) to equal its true value is not symmetrically distributed around zero, but \( s_{1} \) is a reliable estimator of \( \operatorname{Var}(\hat{\theta}) \).

(iii) Comparing Bootstrap Confidence Intervals
Here we perform a simulation to illustrate the fact that bootstrap confidence intervals do not always work particularly well. Suppose that \( y_{i}, i = 1, \ldots, n \), are drawings from a distribution \( F(x) \). We want to form confidence intervals for some of the quantities of \( F(x) \). If \( \hat{\theta} \) is the true \( \alpha \)-quantile, and \( \theta_{i} \) is the corresponding estimate, then asymptotic theory tells us that
\[
\operatorname{Var}(\hat{\theta}_{i}) = \frac{\hat{\theta} - \bar{\theta}}{s_{t_{1}}^{2}}.
\]

Here \( f(y_{i}) \) is the density of \( y_{i} \), evaluated at \( \hat{\theta}_{i} \). In practice, we replace \( \hat{\theta}_{i} \) by a kernel density estimate \( f_{h}(y_{i}) \) so as to obtain the standard error estimate
\[
\hat{s}_{i}^{2} = \frac{\hat{\theta} - \bar{\theta}}{s_{h}^{2}}.
\]
Thus, the 0.95 asymptotic confidence interval is equal to
\[
[\hat{\theta} - 1.96\hat{s}_{1}, \hat{\theta} + 1.96\hat{s}_{1}].
\]

The simplest bootstrap procedure is just to resample the data, calculate the desired quantile(s) of each bootstrap sample, and then use Equation (22) to estimate the bootstrap standard error. This yields the 0.95 simple bootstrap interval
\[
[\hat{\theta} - 1.96\hat{s}_{1}, \hat{\theta} + 1.96\hat{s}_{1}].
\]
We can also use the percentile t method. This is much more expensive, because it requires kernel estimation for the actual sample and for every bootstrap sample.

In the experiments, \( F(x) \) was \( x^{3} \), which is severely skewed to the right. \( \bar{\theta} \) was 999, and \( n = 0.1, 0.2, \ldots, 0.9 \). The sample size \( n \) varied from 50 to 3,000 by factors of \( 2 \). A very standard method of kernel estimation was employed. It used a Gaussian kernel with bandwidth equal to \( 0.699\bar{n}^{0.25} \) times the sample standard deviation of the \( y \). There were 100,000 replications for each sample size.

Figure 4 shows the coverage frequency of three different confidence intervals for the 0.1 quantile, the 0.5 quantile (the median), and the 0.9 quantile. The coverage frequency is the proportion of the time that the interval includes the true value of the quantile. Ideally, it should be 0.95 here. The simulation results are not in accord with standard bootstrap theory. The asymptotic interval sometimes overcoves and sometimes undercovers, whereas both bootstrap intervals always undercover. The simple bootstrap interval, which is conceptually the easiest to calculate, clearly performs best for both the 0.1 and 0.9 quantiles. The asymptotic interval performs best for the
In contrast, the percentile t interval, which theory seems to recommend, performs least well in almost every case. This is probably because the estimated standard errors, given in Equation (33), are not particularly reliable and are not independent of the quantile estimates.

V Bootstrap DGPs for Dependent Data

All of the bootstrap DGPs that have been discussed so far treat the error terms (or the data, in the case of the bivariate bootstrap) as independent. When that is not the case, these methods are not appropriate. In particular, resampling (whether of residuals or data) breaks up whatever dependence there may be and is therefore unsuitable for use when there is dependence.

Numerous bootstrap DGPs for dependent data have been proposed. The two most popular approaches are the "leave bootstrap" and the "block bootstrap." The former attempts to model the dependence using a parametric model. The latter resamples blocks of consecutive observations instead of individual observations. Each of these methods has a great many variants, and the discussion here is necessarily quite superficial. Recent surveys of bootstrap methods for time-series data include Bilinski (2001), Horvitz (2002b), Politis (2001), and Hall et al. (2001).

(i) The Simple Bootstrap

Suppose that the error terms $e_i$ in a regression model, which for simplicity we may assume to be the linear model (8), follow an unknown stationary process with homoscedastic innovations. The simple bootstrap attempts to approximate this process, generally by using an AR(p) process with $p$ chosen either by some sort of model selection criterion or by sequential testing.

The first step is to estimate the model (8), preferably imposing the null hypothesis if one is to be used, so as to obtain residuals $e_i$. The next step is to estimate the AR(p) model

$$
\hat{e}_i = \sum_{k=1}^p \hat{\rho}_k \hat{e}_{i-k} + \epsilon_i,$$

(36)

for several values of $p$ and choose the best one. This may be performed in a number of ways. A GLS estimation does not ensure that the estimated model is stationary, it may be advisable to use another estimation method, such as full maximum likelihood or the Yule-Walker equations, so as to determine the order of the AR process from the sample.

The second step is to resample from the AR(p) model

$$
x_1 = X_1, x_2, \ldots, x_n = \sum_{i=1}^n \hat{\rho}_k \hat{e}_{i-k} + \epsilon_i,$$

(37)

where $\hat{e}_i$ are the estimated parameters, and $\epsilon_i$ are the innovations. The $\hat{e}_i$ are resampled from the (possibly) recursively estimated residuals. Here $m$ is a somewhat arbitrary number, such as 100, chosen so that the process can be allowed to run for some time before the sample period starts. We set the initial values of $e_1, \ldots, e_n$ to zero and discard the $e_1$ for $r < 1$.

The final step is to generate the bootstrap data by the equation

$$
x_1 = X_1, x_2 = \epsilon_2, \ldots, x_n = \epsilon_n,$$

(38)

where $\epsilon_i$ may be estimated in various ways. If restrictions are being tested, they should always be imposed, but this is not performed when constructing confidence intervals. GLS estimates are typically used, but more efficient estimators can often be obtained by using GLS based on the covariance matrix implied by (37). Obviously, whatever estimator is used must be consistent under the null hypothesis.

The simple bootstrap is somewhat simplistic, because it assumes that the innovations, the $e_i$, are independent and identically distributed. This rules out GARCH models and other forms of conditional heteroscedasticity. Moreover, as we will see below, an AR(p) model with a reasonable value of $p$ does provide a good approximation to every stationary, stochastic process. Nevertheless, the simple bootstrap is quite popular. It has recently been applied in Dickey-Fuller unit root testing by Park (2003) and Chang and Park (2003), and it seems to work quite well in many cases.

(ii) Block Bootstrap Methods

Block bootstrap methods, originally proposed by Künsch (1989), divide the quantities that are being resampled, which might be either recursive residuals or the $X_i$, into blocks of consecu-

itive observations. The blocks, which may be either overlapping or non-overlapping and may be either fixed or variable in length, are then resampled. It appears that the best approach is to use overlapping blocks of fixed length: see Lahiri (1996). This is called the "moving-block bootstrap.

For the moving-block bootstrap, there are $n - b + 1$ blocks. The first contains observations 1 through $b$, the second contains observations 2 through $b + 1$, and the last contains observations $n - b + 1$ through $n$. Each bootstrap sample is then constructed by resampling from these overlapping blocks. Unless $n$ is an integer, one or more of the blocks will have to be truncated to form a sample of length $n$.

The choice of $b$ is critical. In theory, it must be allowed to increase as $n$ increases, and the rate of increase is often proportional to $n^{\alpha}$. Of course, as usual sample sizes are generally fixed, it is not clear what this means in practice. If the blocks are too short, the bootstrap samples cannot possibly mimic the original sample, because the dependence is broken whenever we start a new block. However, if the blocks are too long, the bootstrap samples are not random enough. In many cases, it is only when the sample size is quite large that it is possible to choose $b$ so that the blocks are neither too short nor too long.

The "block-of-blocks" bootstrap (Phillips & Romano, 1992) is the analogue of the paired bootstrap for dynamic models. Consider the dynamic regression model

$$
x_t = c + \gamma x_{t-1} + \delta x_{t-2} + \epsilon_t,$$

(39)

if we define

$$
Z_t = [1, x_{t-1}, x_{t-2}],
$$

(40)

we can construct $n - b + 1$ overlapping blocks as

$$
x_{t_1}, x_{t_2}, \ldots, x_{t_{b-1}}, Z_{t_{b}}, \ldots, Z_{t_{n}},
$$

(41)

There are then resampled in the usual way.

The advantages of the block-of-blocks bootstrap are: it can be used with almost any sort of dynamic model and that it can handle heteroscedasticity as well as serial correlation. However, its finite sample performance is often not very good. Moreover, as we do not impose the null hypothesis, any test statistic must be adjusted so that it is testing a hypothesis that is true for the bootstrap DGP. Ideally, this adjustment should take account of the fact that, because of the overlapping blocks, not all observations appear with equal frequency in the bootstrap samples. See Horvitz (2006).

The theoretical properties of block bootstrap methods are not particularly good. When used for testing and for construction of parametric $t$ confidence intervals, they frequently offer higher-order accuracy than asymptotic methods. However, the rate of improvement is generally quite small; see Hall et al. (1995) and Andrews (2002, 2004). Two other recent theoretical papers that focus on different aspects of block bootstrap methods are Gonçalves and White (2004, 2005).

(3) Example: A Unit Root Test

The asymptotic distributions of many unit root tests do not depend on the process that generates the error terms, but the finite sample distributions do. Consider an augmented Dickey-Fuller test for a time series with $r$th order observation $x_t$ to have a unit root. One popular version of such a test is the $t$ statistic for $b = 0$ in the regression

$$
x_t = \beta_0 + \beta_1 x_{t-1} + \sum_{i=1}^r \beta_i x_{t-i} + \epsilon_t,
$$

(42)

The $p$ lags of $\epsilon_t$ are added to account for serial correlation in the error terms. The value of $p$ can be chosen in a number of different ways, which substantially affect the finite-sample properties of the resulting tests. These include model selection criteria, such as AIC and BIC, and various sequential testing schemes; see, among others, Ng and Perron (2001).

In order to bootstrap this test, we first run the regression under the null that $\beta = 0$ and then generate bootstrap samples that satisfy the null. There are several ways in which to do this, which lead to bootstrap DGPs that can have quite different finite-sample properties.

For the naive bootstrap, we first regress $\Delta x_t$ on a constant and a number of lags of $\Delta x_t$ containing coefficients $\beta_i$ and residuals $\tilde{e}_t$. We then generate data according to the equation

$$
x_t = \gamma x_{t-1} + \phi_{r+1} \Delta x_{t-1} + \tilde{e}_t + \sum_{i=1}^r \beta_i \Delta x_{t-i},
$$

(43)

setting the initial values of $x_0$ to zero. The $\tilde{e}_t$ are resampled from the (rescaled) $\epsilon_t$. Like the value of $p$ in Equation (42), the number of lags, $r$, can be chosen in various ways. In practice, $r$ may or may not equal $p$. The details of how $\Delta x_t$ is chosen may substantially affect the performance of the bootstrap DGP infinite samples.

For the moving-block bootstrap, there are no parameters to estimate, because we are not attempting to estimate the process for the error terms and, under the null hypothesis, $\beta = 0$ and $\beta_1 = 1$. The residuals under the null are just $\tilde{e}_t = \Delta x_t - \sum_{i=1}^r \beta_i \epsilon_{t-i}$, where the second term is included in error terms that they have been zero. We resample the bootstrap errors $\tilde{e}_t$ from overlapping blocks of the $\tilde{e}_t$ and then generate the bootstrap data according to the random walk $x_t = x_{t-1} + \tilde{e}_t$. The easiest way to deal with the initial observations is to start the process at zero and generate $n$ or observations, discarding the first $n$ of them.

If we knew the error terms followed a particular process, we could estimate it and use

\[ \text{Economic Record, September 2006} \]
A semiparametric bootstrap. For example, if they followed an MA(1) process, we could estimate the model

$$\Delta \hat{y} = \alpha \hat{y} + \epsilon + \Delta \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

and generate the bootstrap data according to the equation

$$y^* = y^{*2} + \epsilon^* + \Delta \epsilon^*$$

where the $\epsilon^*$ are resampled from rescaled and demeaned $\epsilon$. For purposes of illustration, I performed a number of experiments in which there were 25 observations and the error actually followed the MA(1) process (43). There were 100 000 replications for each of 30 values of $\alpha$ from -0.95 to 0.95 at intervals of 0.05. The number of lags $p$ in the test regression (42) was chosen by the AIC and forced to be between 4 and 12. This selection procedure was repeated for each bootstrap sample. I used three bootstrap DGPs. The first was a moving-block bootstrap with block length 12. The second was a sieve bootstrap with $\hat{y}$ restricted to lie between 4 and 12 and chosen by the AIC. The third was a semiparametric bootstrap based on (43). Readers may well feel that it is cheating to use the least of these procedures, as practitioners will surely be confident that the data actually come from an MA(1) process.

The results of these experiments are shown in Figure 5. It can be seen that the 'asymptotic' test always overrejects, although the overrejection is only severe for large, negative values of $\alpha$. The results termed 'asymptotic' actually use a feasible critical value, -2.9212, that would be valid if there were no serial correlation and no lags of $\Delta \hat{y}$ in the test regression. It was taken from MacKinnon (1996). Using the genuine asymptotic critical value, -2.8614, would result in slightly higher rejection frequencies.

All three bootstrap methods work remarkably well for $\alpha > 0$, but all three work poorly for $\alpha < -0.9$. Not surprisingly, the semiparametric procedure generally works best, but even it overrejects quite noticeably for large, negative values of $\alpha$. This presumably happens because the estimate of $\alpha$ is biased upwards in this case, so that the bootstrap DGP fails to mimic the true DGP sufficiently well. The sieve and moving-block bootstraps overreject much more severely. In the case of the sieve bootstrap, this reflects the fact that even a fairly high-order AR process does not do a very good job of mimicking an MA(1) process with a large, negative coefficient. Interestingly, although the moving-block bootstrap overrejects severely for large, negative values of $\alpha$, it underrejects quite noticeably for smaller, negative values.

This example illustrates the fact that bootstrap methods may or may not yield accurate inferences, and that different bootstrap methods may perform quite differently. It suggests that bootstrap methods should be used with considerable caution when performing unit root and related tests.

"VII Conclusions"

It is very misleading to talk about 'the bootstrap', because there are actually many different bootstrap methods. Deciding what sort of bootstrap DGP to use is an important issue. The first, and often the hardest, thing that an applied econometrician must do. Conditioned on the choice of bootstrap DGP, there are a number of other substantive decisions to be made.

In the case of hypothesis testing, it is almost always desirable to impose the null hypothesis on the bootstrap DGP, but $\alpha$ may not be feasible to do so. When it is not, we have to change the null hypothesis for the bootstrap samples so that what we are testing is 'true' for the bootstrap data. There is often more than one statistic that could be bootstrapped, and we have to choose among them. For tests based on signed statistics, such as $t$ statistics, we may or may not wish to assume symmetry when calculating $t$ values.

For confidence intervals, the number of options is bewildering. We can use asymptotic intervals constructed using bootstrap standard errors, which may or may not incorporate bias correction. We can use percentile $t$ intervals based on various types of standard errors, which may or may not have symmetry imposed on them. We can also use a number of methods that were not discussed in this paper, including primitive ones like the 'percentile method' and more sophisticated ones like the BC method, see Efron and Tibshirani (1993) and Davison and Hinkley (1997). Whatever bootstrap methods we choose to use, it is always important to make it clear precisely what was performed whenever we report the results of empirical work. Simply saying that something is a 'bootstrap standard error', a 'bootstrap $t$ value', or a 'bootstrap confidence interval' provides the reader with grossly insufficient information. We need to make it clear exactly how the bootstrap data were generated and what procedures were then used to calculate the quantities of interest.

REFERENCES


© 2006 The Economic Society of Australia
brought up in the world of spreadsheets and click-and-press computer packages. Accordingly, in addition to discussing computation in detail in the book itself, the website associated with the book contains MATLAB programs for performing Bayesian analysis in a wide variety of models. In general, the focus of the book is on application rather than theory. Hence, I expect that the applied economist interested in using Bayesian methods will find it more useful than the theoretical econometrician.

I would like to thank the numerous people (some anonymous) who gave me helpful comments at various stages in the writing of this book, including: Luc Bauwens, Jeff Dorfman, David Edgerton, John Geweke, Bill Griffiths, Frank Kleibergen, Tony Lancaster, Jim LeSage, Michel Lubrano, Brendan McCabe, Bill McCausland, Richard Paap, Rodney Strachan, and Arnold Zellner. In addition, I would like to thank Steve Hardman for his expert editorial advice. All I know about Bayesian econometrics comes through my work with a series of exceptional co-authors: Carmen Fernandez, Henk Hoek, Eduardo Ley, Kai Li, Jacek Osiewalski, Dale Poirier, Simon Potter, Mark Steel, Justin Tobias, and Herman van Dijk. Of these, I would like to thank Mark Steel, in particular, for patiently responding to my numerous questions about Bayesian methodology and requests for citations of relevant papers. Finally, I wish to express my sincere gratitude to Dale Poirier, for his constant support throughout my professional life, from teacher and PhD supervisor, to valued co-author and friend.

1.1 BAYESIAN THEORY

Bayesian econometrics is based on a few simple rules of probability. This is one of the chief advantages of the Bayesian approach. All of the things that an econometrician would wish to do, such as estimate the parameters of a model, compare different models or obtain predictions from a model, involve the same rules of probability. Bayesian methods are, thus, universal and can be used any time a researcher is interested in using data to learn about a phenomenon.

To motivate the simplicity of the Bayesian approach, let us consider two random variables, $A$ and $B$.\footnote{This chapter assumes the reader knows the basic rules of probability. Appendix B provides a brief introduction to probability for the reader who does not have such a background or would like a reminder of this material.} The rules of probability imply:

$$p(A, B) = p(A|B)p(B)$$

where $p(A, B)$ is the joint probability of $A$ and $B$ occurring, $p(A|B)$ is the probability of $A$ occurring conditional on $B$ having occurred (i.e., the conditional probability of $A$ given $B$), and $p(B)$ is the marginal probability of $B$. Alternatively, we can reverse the roles of $A$ and $B$ and find an expression for the joint probability of $A$ and $B$:

$$p(A, B) = p(B|A)p(A)$$

Equating these two expressions for $p(A, B)$ and rearranging provides us with Bayes' rule, which lies at the heart of Bayesian econometrics:

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}$$

(1.1)
Econometrics is concerned with using data to learn about something the researcher is interested in. Just what the 'something' is depends upon the context. However, in economics we typically work with models which depend upon parameters. For the reader with some previous training in econometrics, it might be useful to have in mind the regression model. In this model interest often centers on the coefficients in the regression, and the researcher is interested in estimating these coefficients. In this case, the coefficients are the parameters under study. Let \( y \) be a vector or matrix of data and \( \theta \) be a vector or matrix which contains the parameters for a model which seeks to explain \( y \). We are interested in learning about \( \theta \) based on the data, \( y \). Bayesian econometrics uses Bayes’ rule to do so. In other words, the Bayesian would replace \( B \) by \( \theta \) and \( A \) by \( y \) in (1.1) to obtain:

\[
p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{1.2}
\]

Bayesians treat \( p(\theta|y) \) as being of fundamental interest. That is, it directly addresses the question “Given the data, what do we know about \( \theta ? \)”. The treatment of \( \theta \) as a random variable is controversial among some econometricians. The chief competitor to Bayesian econometrics, often called frequentist econometrics, says that \( \theta \) is not a random variable. However, Bayesian econometrics is based on a subjective view of probability, which argues that our uncertainty about anything unknown can be expressed using the rules of probability. In this book, we will not discuss such methodological issues (see Poirier (1995) for more detail). Rather, we will take it as given that econometrics involves learning about something unknown (e.g. coefficients in a regression) given something known (e.g. data) and the conditional probability of the unknown given the known is the best way of summarizing what we have learned.

Having established that \( p(\theta|y) \) is of fundamental interest for the econometrician interested in using data to learn about parameters in a model, let us now return to (1.2). Insofar as we are only interested in learning about \( \theta \), we can ignore the term \( p(y) \), since it does not involve \( \theta \). We can then write:

\[
p(\theta|y) \propto p(y|\theta)p(\theta) \tag{1.3}
\]

The term \( p(\theta|y) \) is referred to as the posterior density, the p.d.f. for the data given the parameters of the model, \( p(y|\theta) \), as the likelihood function and \( p(\theta) \) as the prior density. You often hear this relationship referred to as “posterior is proportional to likelihood times prior”. At this stage, this may seem a little abstract, and the manner in which priors and likelihoods are developed to allow for the calculation of the posterior may be unclear. Things should become clearer to you in the following chapters, where we will develop likelihood functions and priors in specific contexts. Here we provide only a brief general discussion of what these are.

\footnote{Appendix A contains a brief introduction to matrix algebra.}

The prior, \( p(\theta) \), does not depend upon the data. Accordingly, it contains any non-data information available about \( \theta \). In other words, it summarizes what you know about \( \theta \) prior to seeing the data. As an example, suppose \( \theta \) is a parameter which reflects returns to scale in a production process. In many cases, it is reasonable to assume that returns to scale are roughly constant. Thus, before you look at the data, you have prior information about \( \theta \), in that you would expect it to be approximately one. Prior information is a controversial aspect of Bayesian methods. In this book, we will discuss both informative and noninformative priors for various models. In addition, in later chapters, we will discuss empirical Bayes methods. These use data-based information to choose the prior and, hence, violate a basic premise of Bayesian methods. Nevertheless, empirical Bayes methods are becoming increasingly popular for the researcher who is interested in practical, objective, tools that seem to work well in practice.

The likelihood function, \( p(y|\theta) \), is the density of the data conditional on the parameters of the model. It is often referred to as the data generating process. For instance, in the linear regression model (which will be discussed in the next chapter), it is common to assume that the errors have a Normal distribution. This implies that \( p(y|\theta) \) is a Normal density, which depends upon parameters (i.e. the regression coefficients and the error variance).

The posterior, \( p(\theta|y) \), is the density which is of fundamental interest. It summarizes all we know about \( \theta \) after (i.e. posterior to) seeing the data. Equation (1.3) can be thought of as an updating rule, where the data allows us to update our prior views about \( \theta \). The result is the posterior which combines both data and non-data information.

In addition to learning about parameters of a model, an econometrician might be interested in comparing different models. A model is formally defined by a likelihood function and a prior. Suppose we have \( m \) different models, \( M_i \) for \( i = 1, \ldots, m \), which all seek to explain \( y \). \( M_i \) depends upon parameters \( \theta^i \). In cases where many models are being entertained, it is important to be explicit about which model is under consideration. Hence, the posterior for the parameters calculated using \( M_i \) is written as

\[
p(\theta^i|y, M_i) = \frac{p(y|\theta^i, M_i)p(\theta^i|M_i)}{p(y|M_i)} \tag{1.4}
\]

and the notation makes clear that we now have a posterior, likelihood, and prior for each model.

The logic of Bayesian econometrics suggests that we use Bayes’ rule to derive a probability statement about what we do not know (i.e. whether a model is a correct one or not) conditional on what we do know (i.e. the data). This means the posterior model probability can be used to assess the degree of support for

\footnote{Carlin and Louis (2000) is a good reference for the reader interested in developing a deeper understanding of empirical Bayes methods.}
Using (1.1) with \( B = M_i \) and \( A = y \), we obtain
\[
p(M_i | y) = \frac{p(y | M_i) p(M_i)}{p(y)}
\] (1.5)

Of the terms in (1.5), \( p(M_i) \) is referred to as the prior model probability. Since it does not involve the data, it measures how likely we believe \( M_i \) to be the correct one before seeing the data. \( p(y | M_i) \) is called the marginal likelihood, and is calculated using (1.4) and a few simple manipulations. In particular, if we integrate both sides of (1.4) with respect to \( \theta^i \), use the fact that \( \int p(\theta^i | y, M_i) d\theta^i = 1 \) (since probability density functions integrate to one), and rearrange, we obtain:
\[
p(y | M_i) = \int p(y | \theta^i, M_i) p(\theta^i | M_i) d\theta^i
\] (1.6)

Note that the marginal likelihood depends only upon the prior and the likelihood. In subsequent chapters, we discuss how (1.6) can be calculated in practice.

Since the denominator in (1.5) is often hard to calculate directly, it is common to compare two models, \( i \) and \( j \), using the posterior odds ratio, which is simply the ratio of their posterior model probabilities:
\[
PO_{ij} = \frac{p(M_i | y)}{p(M_j | y)} = \frac{p(y | M_i) p(M_i)}{p(y | M_j) p(M_j)}
\] (1.7)

Note that, since \( p(y) \) is common to both models, it cancels out when we take the ratio. As we will discuss in subsequent chapters, there are special techniques in many cases for calculating the posterior odds ratio directly. If we calculate the posterior odds ratio comparing every pair of models, and we assume that our set of models is exhaustive (in that \( p(M_1 | y) + p(M_2 | y) + \cdots + p(M_m | y) = 1 \)), then we can use posterior odds ratios to calculate the posterior model probabilities given in (1.5). For instance, if we have \( m = 2 \) models then we can use the two equations
\[
p(M_1 | y) + p(M_2 | y) = 1
\]
and
\[
PO_{12} = \frac{p(M_1 | y)}{p(M_2 | y)}
\]

to work out
\[
p(M_1 | y) = \frac{PO_{12}}{1 + PO_{12}}
\]
and
\[
p(M_2 | y) = 1 - p(M_1 | y)
\]

Thus, knowledge of the posterior odds ratio allows us to figure out the posterior model probabilities.

To introduce some more jargon, econometricians may be interested in model comparison when equal prior weight is attached to each model. That is, \( p(M_i) = p(M_j) \) or, equivalently, the prior odds ratio which is \( \frac{p(M_i)}{p(M_j)} \) is set to one. In this case, the posterior odds ratio becomes simply the ratio of marginal likelihoods, and is given a special name, the Bayes Factor, defined as:
\[
BF_{ij} = \frac{p(y | M_j)}{p(y | M_i)}
\] (1.8)

Finally, econometricians are often interested in prediction. That is, given the observed data, \( y \), the econometrician may be interested in predicting some future unobserved data \( y^* \). Our Bayesian reasoning says that we should summarize our uncertainty about what we do not know (i.e. \( y^* \)) through a conditional probability statement. That is, prediction should be based on the predictive density \( p(y^* | y) \) (or, if we have many models, we would want to make explicit the dependence of a prediction on a particular model, and write \( p(y^* | y, M_i) \)). Using a few simple rules of probability, we can write \( p(y | y^*) \) in a convenient form. In particular, since a marginal density can be obtained from a joint density through integration (see Appendix B), we can write:
\[
p(y^* | y) = \int p(y^* | y, \theta) d\theta
\]

However, the term inside the integral can be rewritten using another simple rule of probability:
\[
p(y^* | y) = \int p(y^* | y, \theta) d\theta
\] (1.9)

As we shall see in future chapters, the form for the predictive in (1.9) is quite convenient, since it involves the posterior.

On one level, this book could end right here. These few pages have outlined all the basic theoretical concepts required for the Bayesian to learn about parameters, compare models and predict. We stress what an enormous advantage this is. Once you accept that unknown things (i.e. \( \theta \), \( M_i \) and \( y^* \)) are random variables, the rest of Bayesian approach is non-controversial. It simply uses the rules of probability, which are mathematically true, to carry out statistical inference. A benefit of this is that, if you keep these simple rules in mind, it is hard to lose sight of the big picture. When facing a new model (or reading a new chapter in the book), just remember that Bayesian econometrics requires selection of a prior and a likelihood. These can then be used to form the posterior, (1.3), which forms the basis for all inference about unknown parameters in a model. If you have many models and are interested in comparing them, you can use posterior model probabilities (1.5), posterior odds ratios (1.7), or Bayes Factors (1.8). To obtain any of these, we usually have to calculate the marginal likelihood (1.6). Prediction is done through the predictive density, \( p(y^* | y) \), which is usually calculated using (1.9). These few equations can be used to carry out statistical inference in any application you may wish to consider.

The rest of this book can be thought of as simply examples of how (1.5)–(1.9) can be used to carry out Bayesian inference for various models which have been commonly-used by others. Nevertheless, we stress that Bayesian inference can be
done with any model using the techniques outlined above and, when confronting an empirical problem, you should not necessarily feel constrained to work with one of the off-the-shelf models described in this book.

## 1.2 BAYESIAN COMPUTATION

The theoretical and conceptual elegance of the Bayesian approach has made it an attractive one for many decades. However, until recently, Bayesians have been in a distinct minority in the field of econometrics, which has been dominated by the frequentist approach. There are two main reasons for this: prior information and computation. With regards to the former, many researchers object to the use of 'subjective' prior information in the supposedly 'objective' science of economics. There is a long, at times philosophical, debate about the role of prior information in statistical science, and the present book is not the place to attempt to summarize this debate. The interested reader is referred to Poitier (1995), which provides a deeper discussion of this issue and includes an extensive bibliography. Briefly, most Bayesians would argue that the entire model building process can involve an enormous amount of non-data information (e.g. econometricians must decide which models to work with, which variables to include, what criteria to use to compare models or estimate parameters, which empirical results to report, etc.). The Bayesian approach is honest and rigorous about precisely how much non-data information is used. Furthermore, if prior information is available, it should be used on the grounds that more information is preferred to less. As a final line of defense, Bayesians have developed noninformative priors for many classes of model. That is, the Bayesian approach allows for the use of prior information if you wish to use it. However, if you do not wish to use it, you do not have to do so. Regardless of how a researcher feels about prior information, it should in no way be an obstacle to the adoption of Bayesian methods.

Computation is the second, and historically more substantive, reason for the minority status of Bayesian econometrics. That is, Bayesian econometrics has historically been computationally difficult or impossible to do for all but a few specific classes of model. The computing revolution of the last 20 years has overcome this hurdle and has led to a blossoming of Bayesian methods in many fields. However, this has made Bayesian econometrics a field which makes heavy use of the computer, and a great deal of this book is devoted to a discussion of computation. In essence, the ideas of Bayesian econometrics are simple, since they only involve the rules of probability. However, to use Bayesian econometrics in practice often requires a lot of number crunching.

To see why computational issues are so important, let us return to the basic equations which underpin Bayesian econometrics. The equations relating to model comparison and prediction either directly or indirectly involve integrals (i.e. (1.6) and (1.9) involve integrals, and (1.6) is a building block for (1.7) and (1.8)). In some (rare) cases, analytical solutions for these integrals are available. That is, you can sit down with pen and paper and work out the integrals. However, we usually need the computer to evaluate the integrals for us, and many algorithms for doing so have been developed.

The equation defining the posterior does not involve any integrals, but presentation of information about the parameters can often involve substantial computation. This arises since, although $p(\theta|y)$ summarizes all we know about $\theta$ after seeing the data, it is rarely possible to present all the information about $p(\theta|y)$ when writing up a paper. In cases where $p(\theta|y)$ has a simple form or $\theta$ is one-dimensional, it is possible to do so, for instance, by graphing the posterior density. However, in general, econometricians choose to present various numerical summaries of the information contained in the posterior, and these can involve integration. For instance, it is common to present a point estimate, or best guess, of what $\theta$ is. Bayesians typically use decision theory to justify a particular choice of a point estimate. In this book, we will not discuss decision theory. The reader is referred to Poitier (1995) or Berger (1985) for excellent discussions of this topic (see also Exercise 1 below). Suffice it to note here that various intuitively plausible point estimates such as the mean, median, and mode of the posterior can be justified in a decision theoretical framework.

Let us suppose you want to use the mean of the posterior density (or posterior mean) as a point estimate, and suppose $\theta$ is a vector with $k$ elements, $\theta = (\theta_1, \ldots, \theta_k)$. The posterior mean of any element of $\theta$ is calculated as (see Appendix B)

$$E(\theta_i|y) = \int \theta_i p(\theta|y) d\theta$$

(1.10)

Apart from a few simple cases, it is not possible to evaluate this integral analytically, and once again we must turn to the computer.

In addition to a point estimate, it is usually desirable to present a measure of the degree of uncertainty associated with the point estimate. The most common such measure is the posterior standard deviation, which is the square root of the posterior variance. The latter is calculated as

$$\text{var}(\theta_i|y) = E(\theta_i^2|y) - (E(\theta_i|y))^2$$

which requires evaluation of the integral in (1.10), as well as

$$E(\theta_i^2|y) = \int \theta_i^2 p(\theta|y) d\theta$$

Depending on the context, the econometrician may wish to present many other features of the posterior. For instance, interest may center on whether a particular parameter is positive. In this case, the econometrician would calculate

$$p(\theta_i \geq 0|y) = \int_0^\infty p(\theta|y) d\theta$$

and, once again, an integral is involved.
All of these posterior features which the Bayesian may wish to calculate have the form:

\[ E[g(\theta)|y] = \int g(\theta) p(\theta|y) d\theta \]  

where \( g(\theta) \) is a function of interest. For instance, \( g(\theta) = \theta_i \) when calculating the posterior mean of \( \theta_i \) and \( g(\theta) = 1(\theta_i \geq 0) \) when calculating the probability that \( \theta_i \) is positive, where \( 1(A) \) is the indicator function which equals 1 if condition \( A \) holds and equals zero otherwise. Even the predictive density in (1.9) falls in this framework if we set \( g(\theta) = p(y^*|y, \theta) \). Thus, most things a Bayesian would want to calculate can be put in the form (1.11). The chief exceptions which do not have this form are the marginal likelihood and quantiles of the posterior density (e.g. in some cases, one may wish to calculate the posterior median and posterior interquartile range, and these cannot be put in the form of (1.11)). These exceptions will be discussed in the context of particular models in subsequent chapters.

At this point, a word of warning is called for. Throughout this book, we focus on evaluating \( E[g(\theta)|y] \) for various choices of \( g(.) \). Unless otherwise noted, for every model and \( g(.) \) discussed in this book, \( E[g(\theta)|y] \) exists. However, for some models it is possible that \( E[g(\theta)|y] \) does not exist. For instance, for the Cauchy distribution, which is the t distribution with one degree of freedom (see Appendix B, Definition B.26), the mean does not exist. Hence, if we had a model which had a Cauchy posterior distribution, \( E[\theta|y] \) would not exist. When developing methods for Bayesian inference in a new model, it is thus important to prove that \( E[g(\theta)|y] \) does exist. Provided that \( p(\theta|y) \) is a valid probability density function, quantiles will exist. So, if you are unsure that \( E[g(\theta)|y] \) exists, you can always present quantile-based information (e.g. the median and interquartile range).

In rare cases, (1.11) can be worked out analytically. However, in general, we must use the computer to calculate (1.11). There are many methods for doing this, but the predominant approach in modern Bayesian econometrics is posterior simulation. There are a myriad of posterior simulators which are commonly used in Bayesian econometrics, and many of these will be discussed in future chapters in the context of particular models. However, all of these are applications or extensions of laws of large numbers or central limit theorems. In this book, we do not discuss these concepts of asymptotic distribution theory in any detail. The interested reader is referred to Poirier (1995) or Greene (2000). Appendix B provides some simple cases, and these can serve to illustrate the basic ideas of posterior simulation.

A straightforward implication of the law of large numbers given in Appendix B (see Definition B.31 and Theorem B.19) is:

**Theorem 1.1: Monte Carlo integration**

Let \( \theta^{(s)} \) for \( s = 1, \ldots, S \) be a random sample from \( p(\theta|y) \), and define

\[ \tilde{g}_S = \frac{1}{S} \sum_{s=1}^{S} g(\theta^{(s)}) \]  

then \( \tilde{g}_S \) converges to \( E[g(\theta)|y] \) as \( S \) goes to infinity.

In practice, this means that, if we can get the computer to take a random sample from the posterior, (1.12) allows us to approximate \( E[g(\theta)|y] \) by simply averaging the function of interest evaluated at the random sample. To introduce some jargon, this sampling from the posterior is referred to as posterior simulation, and \( \theta^{(s)} \) is referred to as a draw or replication. Theorem 1.1 describes the simplest posterior simulator, and use of this theorem to approximate \( E[g(\theta)|y] \) is referred to as Monte Carlo integration.

Monte Carlo integration can be used to approximate \( E[g(\theta)|y] \), but only if \( S \) were infinite would the approximation error go to zero. The econometrician can, of course, choose any value for \( S \), allowing larger values of \( S \) will increase the computational burden. There are many ways of gauging the approximation error associated with a particular value of \( S \). Some of these will be discussed in subsequent chapters. However, many are based on extensions of the central limit theorem given in Appendix B, Definition B.33 and Theorem B.20. For the case of Monte Carlo integration, this central limit theorem implies:

**Theorem 1.2: A numerical standard error**

Using the setup and definitions of Theorem 1.1,

\[ \sqrt{S} [\tilde{g}_S - E[g(\theta)|y]] \to N(0, \sigma_g^2) \]  

as \( S \) goes to infinity, where \( \sigma_g^2 = \text{var}(g(\theta)|y) \).

Theorem 1.2 can be used to obtain an estimate of the approximation error in a Monte Carlo integration exercise by using properties of the Normal distribution. For instance, using the fact that the standard Normal has 95% of its probability located within 1.96 standard deviations from its mean yields the approximate result that:

\[ \Pr \left[ -1.96 \frac{\sigma_g}{\sqrt{S}} \leq \tilde{g}_S - E[g(\theta)|y] \leq 1.96 \frac{\sigma_g}{\sqrt{S}} \right] = 0.95 \]

By controlling \( S \), the econometrician can ensure that \( \tilde{g}_S - E[g(\theta)|y] \) is sufficiently small with a high degree of probability. In practice, \( \sigma_g \) is unknown, but the Monte Carlo integration procedure allows us to approximate it. The term \( \frac{\sigma_g}{\sqrt{S}} \) is known as the numerical standard error, and the econometrician can simply report it as a measure of approximation error. Theorem 1.2 also implies, for example, that if \( S = 10000 \) then the numerical standard error is 1%, as big as the posterior standard deviation. In many empirical contexts, this may be a nice way of expressing the approximation error implicit in Monte Carlo integration.

Unfortunately, it is not always possible to do Monte Carlo integration. Algorithms exist for taking random draws from many common densities (e.g. the
Normal, the Chi-squared). However, for many models, the posteriors do not have one of these common forms. In such cases, development of posterior simulators is a more challenging task. In subsequent chapters, we describe many types of posterior simulators. However, we introduce Monte Carlo integration here so as to present the basic ideas behind posterior simulation in a simple case.

1.3 BAYESIAN COMPUTER SOFTWARE

There are several computer software packages that are useful for doing Bayesian analysis in certain classes of model. However, Bayesian econometrics still tends to require a bit more computing effort than frequentist econometrics. For the latter, there are many canned packages that allow the user to simply click on an icon in order to carry out a particular econometric procedure. Many would argue that this apparent advantage is actually a disadvantage, in that it encourages the econometrician to simply use whatever set of techniques is available in the computer package. This can lead to the researcher simply presenting whatever estimates, test statistics, and diagnostics that are produced, regardless of whether they are appropriate for the application at hand. Bayesian inference forces the researcher to think in terms of the models (i.e., likelihoods and priors), which are appropriate for the empirical question under consideration. The myriad of possible priors and likelihoods make it difficult to construct a Bayesian computer package that can be used widely. For this reason, many Bayesian econometricians create their own programs in matrix programming languages such as MATLAB, Gauss, or Ox. This is not that difficult to do. It is also well worth the effort, since writing a program is a very good way of forcing yourself to fully understand an econometric procedure. In this book, the empirical illustrations are carried out using MATLAB, which is probably the most commonly-used computer language for Bayesian econometrics and statistics. The website associated with this book contains copies of the programs used in the empirical illustrations, and the reader is encouraged to experiment with these programs as a way of learning Bayesian programming. Furthermore, some of the questions at the end of each chapter require the use of the computer, and provide another route for the reader to develop some basic programming skills.

For readers who do not wish to develop programming skills, there are some Bayesian computer packages that allow for simple analysis of standard classes of models. BUGS, an acronym for Bayesian Inference Using Gibbs Sampling (see Best et al., 1995), handles a fairly wide class of models using a common posterior simulation technique called Gibbs sampling. More directly relevant for econometricians is Bayesian Analysis, Computation and Communication (BACC), which handles a wide range of common models (see McCausland and Stevens, 2001).

The easiest way to use BACC is as a dynamically linked library to another popular language such as MATLAB. In other words, BACC can be treated as a set of MATLAB commands. For instance, instead of programming up a posterior simulator for analysis of the regression model discussed in Chapter 4, BACC allows for Bayesian inference to be done using one simple MATLAB command. Jim LeSage’s Econometrics Toolbox (see LeSage, 1999) also contains many MATLAB functions that can be used for aspects of Bayesian inference. The empirical illustrations in this book which involve posterior simulation use his random number generators. At the time of writing, BUGS, BACC, and the Econometrics Toolbox were available on the web for free for educational purposes. Many other Bayesian software packages exist, although most are more oriented towards the statistician than the econometrician. Appendix C of Carlin and Louis (2000) provides much more information about relevant software.

1.4 SUMMARY

In this chapter, we have covered all the basic issues in Bayesian econometrics at a high level of abstraction. We have stressed that the ability to put all the general theory in one chapter, involving only basic concepts in probability, is an enormous advantage of the Bayesian approach. The basic building blocks of the Bayesian approach are the likelihood function and the prior, the product of these defines the posterior (see (1.3)), which forms the basis for inference about the unknown parameters in a model. Different models can be compared using posterior model probabilities (see (1.5)), which require the calculation of marginal likelihoods (1.6). Prediction is based on the predictive density (1.9). In most cases, it is not possible to work with all these building blocks analytically. Hence, Bayesian computation is an important topic. Posterior simulation is the predominant method of Bayesian computation.

Future chapters go through particular models, and show precisely how these abstract concepts become concrete in practical contexts. The logic of Bayesian econometrics set out in this chapter provides a template for the organization of following chapters. Chapters will usually begin with a likelihood function and a prior. Then a posterior is derived along with computational methods for posterior inference and model comparison. The reader is encouraged to think in terms of this likelihood/prior/posterior/computation organizational structure both when reading this book and when beginning a new empirical project.

1.5 EXERCISES

1.5.1 Theoretical Exercises

Remember that Appendix B describes basic concepts in probability, including definitions of common probability distributions.
The Normal Linear Regression Model with Natural Conjugate Prior and a Single Explanatory Variable

2.1 INTRODUCTION

The regression model is the workhorse of econometrics. A detailed motivation and discussion of the regression model can be found in any standard econometrics text (e.g. Greene (2000), Gujarati (1995), Hill, Griffiths and Judge (1997), or Koop (2000). Briefly, the linear regression model posits a relationship between a dependent variable, $y$, and $k$ explanatory variables, $x_1, \ldots, x_k$, of the form:

$$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

where $\varepsilon$ is the regression error, and $x_1$ is implicitly set to 1 to allow for an intercept.

It is not hard for any economist to think of many examples where a particular variable depends upon others. For instance, an individual's wage depends upon her education, experience, and other characteristics. The level of GDP in a country depends upon the size and quality of its workforce, its capital stock, and many other characteristics. The production costs of a firm depends upon the amount of outputs produced, as well as input prices, etc. The empirical example used in the next chapter involves data on houses in Windsor, Canada. Interest centers on the factors which influence house prices, the dependent variable. The explanatory variables are the lot size of the property, the number of bedrooms, number of bathrooms, and number of storeys in the house. Note that this example (like most in economics) involves many explanatory variables and, hence, we have many parameters. With many parameters, the notation becomes very complicated unless matrix algebra is used. To introduce the basic concepts and motivation for the linear regression model with minimal matrix algebra, we begin with a simple case where there is only one explanatory variable. Subsequently, in Chapter 3, we move to the general case involving many explanatory variables.
2.2 THE LIKELIHOOD FUNCTION

Let \( y_i \) and \( x_i \) denote the observed data on the dependent and explanatory variables, respectively, for individual \( i \) for \( i = 1, \ldots, N \). We use the term ‘individual’ to denote the unit of observation, but we could have data on firms, products, time periods, etc. To simplify the mathematics, we do not allow for an intercept and, hence, the linear regression model becomes:

\[
y_i = \beta x_i + \varepsilon_i
\]

where \( \varepsilon_i \) is an error term. There are many justifications for inclusion of an error term. It can reflect measurement error, or the fact that the linear relationship between \( x \) and \( y \) is only an approximation of the true relationship. More simply, you can imagine the linear regression model as fitting a straight line with slope \( \beta \) through an XY-plot of the data. In all but the most trivial cases, it is not possible to fit a straight line through all \( N \) data points. Hence, it is inevitable that error will result.

Assumptions about \( \varepsilon_i \) and \( x_i \) determine the form of the likelihood function. The standard assumptions (which we will free up in later chapters) are:

1. \( \varepsilon_i \) is Normally distributed with mean 0, variance \( \sigma^2 \), and \( \varepsilon_i \) and \( \varepsilon_j \) are independent of one another for \( i \neq j \). Shorthand notation for this is: \( \varepsilon_i \) is i.i.d. \( N(0, \sigma^2) \), where i.i.d. stands for ‘independent and identically distributed’.
2. The \( x_i \) are either fixed (i.e. not random variables) or, if they are random variables, they are independent of \( \varepsilon_i \) with a probability density function, \( p(x_i | \lambda) \) where \( \lambda \) is a vector of parameters that does not include \( \beta \) and \( \sigma^2 \).

The assumption that the explanatory variables are not random is a standard one in the physical sciences, where experimental methods are common. That is, as part of the experimental setup, the researcher chooses particular values for \( x \) and they are not random. In most economic applications, such an assumption is unreasonable. However, the assumption that the distribution of \( x \) is independent of the error and with a distribution which does not depend upon the parameters of interest is often a reasonable one. In the language of economics, you can think of it as implying that \( x \) is an exogenous variable.

The likelihood function is defined as the joint probability density function for all the data conditional on the unknown parameters (see (1.3)). As shorthand notation, we can stack all our observations of the dependent variable into a vector of length \( N \):

\[
y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}
\]

or, equivalently (and more compactly), \( y = (y_1, y_2, \ldots, y_N)' \). Similarly, for the explanatory variable, we define \( x = (x_1, x_2, \ldots, x_N)' \). The likelihood function then becomes \( p(y, x | \beta, \sigma^2, \lambda) \). The second assumption above implies that we can write the likelihood function as:

\[
p(y, x | \beta, \sigma^2, \lambda) = p(y | x, \beta, \sigma^2) p(x | \lambda)
\]

Insofar as the distribution of \( x \) is not of interest, we can then work with the likelihood function conditional on \( x \), \( p(y | x, \beta, \sigma^2) \). For simplicity of notation, we will not explicitly include \( x \) in our conditioning set for the regression model. It should be remembered that the regression model (whether handled using Bayesian or frequentist methods) implicitly involves working with the conditional distribution of \( y \) given \( x \), and not the joint distribution of these two random vectors.

The assumptions about the errors can be used to work out the precise form of the likelihood function. In particular, using some basic rules of probability and (2.1), we find:

- \( p(y_i | \beta, \sigma^2) \) is Normal (see Appendix B, Theorem B.10).
- \( E(y_i | \beta, \sigma^2) = \beta x_i \) (see Appendix B, Theorem B.2).
- \( var(y_i | \beta, \sigma^2) = \sigma^2 \) (see Appendix B, Theorem B.2).

Using the definition of the Normal density (Appendix B, Definition B.24) we obtain

\[
p(y_i | \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(y_i - \beta x_i)^2}{2\sigma^2} \right)
\]

Finally, since, for \( i \neq j \), \( \varepsilon_i \) and \( \varepsilon_j \) are independent of one another, it follows that \( y_i \) and \( y_j \) are also independent of one another and, thus, \( p(y | \beta, \sigma^2) = \prod_{i=1}^{N} p(y_i | \beta, \sigma^2) \) and, hence, the likelihood function is given by:

\[
p(y | \beta, \sigma^2) = \frac{1}{(2\pi)^{N/2} \sigma^N} \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \beta x_i)^2 \right)
\]

For future derivations, it proves convenient to rewrite the likelihood in a slightly different way. It can be shown that:

\[
\sum_{i=1}^{N} (y_i - \beta x_i)^2 = \nu s^2 + (\beta - \hat{\beta})^2 \sum_{i=1}^{N} x_i^2
\]

where

\[
\nu = N - 1
\]

\[
\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}
\]

\[1\]To prove this, write \( \sum (y_i - \beta x_i)^2 = \sum (y_i - \beta x_i)^2 - (\beta - \beta \hat{x}_i)^2 \) and then expand out the right-hand side:
and
\[ s^2 = \frac{\sum (y_i - \hat{\beta} x_i)^2}{v} \] (2.5)

For the reader with a knowledge of frequentist econometrics, note that \( \hat{\beta} \), \( s^2 \) and \( v \) are the Ordinary Least Squares (OLS) estimator for \( \beta \), standard error and degrees of freedom, respectively. They are also sufficient statistics (see Poirier, 1995, p. 222) for (2.2). Furthermore, for many technical derivations, it is easier to work with the error precision rather than the variance. The error precision is defined as: \( h = \frac{1}{s^2} \).

Using these results, we can write the likelihood function as:
\[
p(y|\beta, h) = \frac{1}{(2\pi)^{\frac{N}{2}}} \left\{ h^\frac{N}{2} \exp \left[ -\frac{h}{2} (\beta - \hat{\beta})^2 \sum_{i=1}^{N} x_i^2 \right] \right\} \left\{ h^\frac{v}{2} \exp \left[ -\frac{hv}{2s^2} \right] \right\}
\] (2.6)

For future reference, note that the first term in curly brackets looks like the kernel of a Normal density for \( \beta \), and the second term looks almost like a Gamma density for \( h \) (see Appendix B, Definitions B.24 and B.22).

### 2.3 THE PRIOR

Priors are meant to reflect any information the researcher has before seeing the data which she wishes to include. Hence, priors can take any form. However, it is common to choose particular classes of priors that are easy to interpret and/or make computation easier. Natural conjugate priors typically have both such advantages. A conjugate prior distribution is one which, when combined with the likelihood, yields a posterior that falls in the same class of distributions. A natural conjugate prior has the additional property that it has the same functional form as the likelihood function. These properties mean that the prior information can be interpreted in the same way as likelihood function information. In other words, the prior can be interpreted as arising from a fictitious data set from the same process that generated the actual data.

In the simple linear regression model, we must elicit a prior for \( \beta \) and \( h \), which we denote by \( p(\beta, h) \). The fact that we are not conditioning on the data means that \( p(\beta, h) \) is a prior density, the posterior density will be denoted by \( p(\beta, h|y) \). It proves convenient to write \( p(\beta, h) = p(\beta|h)p(h) \) and think in terms of a prior for \( \beta|h \) and one for \( h \). The form of the likelihood function in (2.6) suggests that the natural conjugate prior will involve a Normal distribution for \( \beta|h \) and a Gamma distribution for \( h \). This is indeed the case. The name given to a distribution such as this which is a product of a Gamma and a (conditional) Normal is the Normal-Gamma. Appendix B, Definition B.26 provides further details on this distribution. Using notation introduced in Appendix B, if
\[
\beta|h \sim N(\bar{\beta}, h^{-1}v)
\]

and
\[
h \sim G(\bar{s}^{-2}, v)
\]

then the natural conjugate prior for \( \beta \) and \( h \) is denoted by:
\[
\beta, h \sim NG(\bar{\beta}, \bar{V}, \bar{s}^{-2}, v)
\] (2.7)

The researcher would then choose particular values of the so-called prior hyper-parameters \( \bar{\beta}, \bar{V}, \bar{s}^{-2} \) and \( v \) to reflect her prior information. The exact interpretation of these hyperparameters becomes clearer once you have seen their role in the posterior and, hence, we defer a deeper discussion of prior elicitation until the next section.

Throughout this book, we use bars under parameters (e.g. \( \bar{\beta} \)) to denote parameters of a prior density, and bars over parameters (e.g. \( \bar{V} \)) to denote parameters of a posterior density.

### 2.4 THE POSTERIOR

The posterior density summarizes all the information, both prior and data-based, that we have about the unknown parameters, \( \beta \) and \( h \). It is proportional to (see (1.3)) the likelihood (2.2) times the prior density (see (2.7)). For the sake of brevity, we do not provide all the algebraic details here. Poirier (1995, p. 527) or Zellner (1971, pp. 60–61) provide closely related derivations. Messy, but conceptually straightforward, manipulations can be used to show that the posterior density is also of Normal-Gamma form, confirming that the prior of the previous section is indeed a natural conjugate one.

Formally, we have a posterior of the form
\[
\beta, h|y \sim NG(\bar{\beta}, \bar{V}, \bar{s}^{-2}, v)
\] (2.8)

where
\[
\bar{V} = \frac{1}{V^{-1} + \sum x_i^2}
\] (2.9)
\[
\bar{\beta} = \bar{V}(\bar{V}^{-1} \beta + \bar{\beta} \sum x_i^2)
\] (2.10)
\[
\bar{v} = v + N
\] (2.11)

and \( \bar{s}^{-2} \) is defined implicitly through
\[
\bar{v}\bar{s}^2 = v\bar{s}^2 + \nu \bar{v}^2 + \frac{(\bar{\beta} - \beta)^2}{\bar{V} + \left( \frac{1}{\sum x_i^2} \right) ^2}
\] (2.12)

In regression modeling, the coefficient on the explanatory variable, \( \beta \), is usually the primary focus, since it is a measure of the marginal effect of the explanatory
variable on the dependent variable. The posterior mean, \( \hat{\beta} \), is a commonly-used point estimate, and \( \text{var}(\beta|y) \) is a commonly-used metric for the uncertainty associated with the point estimate. Using the basic rules of probability, the posterior mean can be calculated as:

\[
E(\beta|y) = \int \int \beta p(\beta, h|y) dh d\beta = \int \beta p(\beta|y) d\beta
\]

This equation motivates interest in the marginal posterior density, \( p(\beta|y) \). Fortunately, this can be calculated analytically using the properties of the Normal-Gamma distribution (see Appendix B, Theorem B.15). In particular, these imply that, if we integrate out \( h \) (i.e. use the fact that \( p(\beta|y) = \int p(\beta, h|y) dh \)), the marginal posterior distribution for \( \beta \) is a \( t \) distribution. In terms of the notation of Appendix B, Definition B.25:

\[
\beta|y \sim t(\bar{\beta}, \bar{\sigma}^2 V, \bar{v})
\]

and it follows from the definition of the \( t \) distribution that

\[
E(\beta|y) = \bar{\beta}
\]

and

\[
\text{var}(\beta|y) = \frac{\bar{v}\bar{\sigma}^2}{\bar{v} - 2}
\]

The error precision, \( h \), is usually of less interest than \( \beta \), but the properties of the Normal-Gamma imply immediately that:

\[
h|y \sim \mathcal{G}(\bar{v}^{-2}, \bar{v})
\]

and hence that

\[
E(h|y) = \bar{v}^{-2}
\]

and

\[
\text{var}(h|y) = \frac{2\bar{v}^{-2}}{\bar{v}}
\]

Equations (2.9)–(2.18) provide insight into how Bayesian methods combine prior and data information in a very simple model and, hence, it is worth discussing them in some detail. Note, first, that the results the Bayesian econometrician would wish to report all can be written out analytically, and do not involve integration. In Chapter 1, we stressed that Bayesian inference often required posterior simulation. The linear regression model with Normal-Gamma natural conjugate prior is one case where posterior simulation is not required.

The frequentist econometrician would often use \( \hat{\beta} \), the ordinary least squares estimate of \( \beta \). The common Bayesian point estimate, \( \hat{\beta} \), is a weighted average of the OLS estimate and the prior mean, \( \bar{\beta} \). The weights are proportional to \( \sum x_i^2 \) and \( \bar{v}^{-1} \), respectively. The latter of these reflects the confidence in the prior. For instance, if the prior variance you select is high, you are saying you are very uncertain about what likely values of \( \beta \) are. As a result, \( \bar{v}^{-1} \) will be

small and little weight will be attached to \( \hat{\beta} \), your best prior guess at what \( \beta \) is. The sum \( \sum x_i^2 \) plays a similar role with respect to data-based information. Loosely speaking, it reflects the degree of confidence that the data have in its best guess for \( \beta \), the OLS estimate \( \hat{\beta} \). Readers knowledgeable of frequentist econometrics will recognize \( \sum x_i^2 \bar{v}^{-1} \) as being proportional to the variance of \( \hat{\beta} \). Alternative intuition can be obtained by considering the simplest case, where \( x_i = 1 \) for \( i = 1, \ldots, N \). Then \( \sum x_i^2 = N \), and the weight attached to \( \hat{\beta} \) will simply be the sample size, a reasonable measure for the amount of information in the data. Note that, for both prior mean and the OLS estimate, the posterior mean attaches weight proportional to their precisions (i.e. the inverse of their variances). Hence, Bayesian methods combine data and prior information in a sensible way.

In frequentist econometrics, the variance of the OLS estimator for the regression model given in (2.1) is \( s^2(\sum x_i^2)^{-1} \). This variance would be used to obtain frequentist standard errors and carry out various hypothesis tests (e.g. the frequentist \( t \)-statistic for testing \( \beta = 0 \) is \( \frac{\hat{\beta}}{s\sqrt{\sum x_i^2}} \). The Bayesian analogue is the posterior variance of \( \beta \) given in (2.15), which has a very similar form, but incorporates both prior and data information. For instance, (2.9) can be informally interpreted as saying \( \text{"posterior precision is an average of prior precision (}\bar{v}^{-1}\text{) and data precision (}\sum x_i^2\text{)\"} \). Similarly, (2.12) has an intuitive interpretation of \( \text{"posterior sum of squared errors (}\bar{v} s^2\text{) is the sum of prior sum of squared errors (}\bar{v}^2 s^2\text{), OLS sum of squared errors (}\bar{v}^2 s^2\text{), and a term which measures the conflict between prior and data information\"} \).

The other equations above also emphasize the intuition that the Bayesian posterior combines data and prior information. Furthermore, the natural conjugate prior implies that the prior can be interpreted as arising from a fictitious data set (e.g. \( \bar{v} \) and \( N \) play the same role in (2.11) and (2.12) and, hence, \( \bar{v} \) can be interpreted as a prior sample size).

For the reader trained in frequentist econometrics, it is useful to draw out the similarities and differences between what a Bayesian would do and what a frequentist would do. The latter might calculate \( \hat{\beta} \) and its variance, \( s^2(\sum x_i^2)^{-1} \), and estimate \( \sigma^2 \) by \( s^2 \). The former might calculate the posterior mean and variance of \( \beta \) (i.e. \( \bar{\beta} \) and \( \frac{\bar{v} s^2}{\bar{v}^2} \)) and estimate \( h = \sigma^2 \) by its posterior mean, \( \bar{v}^{-2} \). These are very similar strategies, except for two important differences. First, the Bayesian formulae all combine prior and data information. Secondly, the Bayesian interprets \( \beta \) as a random variable, whereas the frequentist interprets \( \hat{\beta} \) as a random variable.

The fact that the natural conjugate prior implies prior information enters in the same manner as data information helps with prior elicitation. For instance, when choosing particular values for \( \bar{\beta}, \bar{v}, \bar{v}^{-2} \) and \( v \) it helps to know that \( \beta \) is equivalent to the OLS estimate from an imaginary data set of \( y \) observations with an imaginary \( \sum x_i^2 \) equal to \( \bar{v}^{-1} \) and an imaginary \( s^2 \) given by \( \bar{v}^2 \). However, econometrics is a public science where empirical results are presented to a wide
variety of readers. In many cases, most readers may be able to agree on what a sensible prior might be (e.g. economic theory often specifies what reasonable parameter values might be). However, in cases where different researchers can approach a problem with very different priors, a Bayesian analysis with only a single prior can be criticized. There are two main Bayesian strategies for surmounting such a criticism. First, a prior sensitivity analysis can be carried out. This means that empirical results can be presented using various priors. If empirical results are basically the same for various sensible priors, then the reader is reassured that researchers with different beliefs can, after looking at the data, come to agreement. If results are sensitive to choice of prior, then the data is not enough to force agreement on researchers with different prior views. The Bayesian approach allows for the scientifically honest finding of such a state of affairs. There is a substantive literature which finds bounds on, for example, the posterior mean of a parameter. We do not discuss this so-called extreme bounds analysis literature in any detail. A typical result in this literature is of the form: “for any possible choice of \( \mathcal{V} \), \( \beta \) must lie between specified upper and lower bounds”. Poirier (1995, pp. 532–536) provides an introduction to this literature, and further references (see also Exercise 6 in Chapter 3).

A second strategy for prior elicitation in cases where wide disagreement about prior choice could arise is to use a noninformative prior. The Bayesian literature on noninformative priors is too voluminous to survey here. Poirier (1995, pp. 318–331) and Zellner (1971, pp. 41–53) provide detailed discussion about this issue (see also Chapter 12, Section 12.3). Suffice it to note here that, in many cases, it is desirable for data information to be predominant over prior information. In the context of the natural conjugate prior above, it is clear how one can do this. Given the ‘fictitious prior sample’ interpretation of the natural conjugate prior, it can be seen that setting \( \mathcal{V} \) small relative to \( N \) and \( \mathcal{V} \) to a large value will ensure that prior information plays little role in the posterior formula (see (2.9)–(2.12)). We refer to such a prior as a relatively noninformative prior.

Taking the argument in the previous paragraph to the limit suggests that we can create a purely noninformative prior by setting \( \mathcal{V} = 0 \) and \( \mathcal{V}^{-1} = 0 \) (i.e. \( \mathcal{V} \rightarrow \infty \)). Such choices are indeed commonly made, and they imply \( \beta, h, |y \sim N(\bar{\beta}, \mathcal{V}, \mathcal{V}^{-2}, \mathcal{V}) \), where

\[
\bar{\mathcal{V}} = \frac{1}{\sum x_i^2}
\]

\[
\bar{\beta} = \bar{\beta}
\]

\[
\bar{\nu} = N
\]

and

\[
\bar{\nu}^2 = \nu^2
\]

With this noninformative prior, all of these formulae involve only data information and, in fact, are equal to ordinary least squares results.

**Linear Regression Model with a Single Variable**

In one sense, this noninformative prior has very attractive properties and, given the close relationship with OLS results, provides a bridge between the Bayesian and frequentist approaches. However, it has one undesirable property: this prior ‘density’ is not, in fact, a valid density, in that it does not integrate to one. Such priors are referred to as improper. The Bayesian literature has many examples of problems caused by the use of improper priors. We will see below problems which occur in a model comparison exercise when improper priors are used.

To see the impropriety of this noninformative prior, note that the posterior results (2.19)–(2.22) can be justified by as combining the likelihood function with the following ‘prior density’:

\[
p(\beta, h) = \frac{1}{h}
\]

where \( h \) is defined over the interval \((0, \infty)\). If you try integrating this ‘prior density’ over \((0, \infty)\), you will find that the result is \(\infty\), not one as would occur for a valid p.d.f. Bayesians often write this prior as:

\[
p(\beta, h) \propto \frac{1}{h}
\]

but it should be stressed that this notation is not formally correct, since \( p(\beta, h) \) is not a valid density function.

It is worth digressing and noting that noninformative priors tend to be improper in most models. To see why this is, consider a continuous scalar parameter \( \theta \), which is defined on an interval \([a, b] \). A researcher who wishes to be noninformative about \( \theta \) would allocate equal prior weight to each equally sized sub-interval (e.g. each interval of width \(0.01\) should be equally likely). This implies that a Uniform prior over the interval \([a, b]\) is a sensible noninformative prior for \( \theta \). However, in most models we do not know \( a \) and \( b \), so they should properly be set to \(-\infty\) and \(\infty\), respectively. Unfortunately, any Uniform density which yields non-zero probability to a finite bounded interval will integrate to infinity over \((-\infty, \infty)\). Formally, we should not even really speak of the Uniform density in this case, since it is only defined for finite values of \(a\) and \(b\). Thus, any Uniform ‘noninformative’ prior will be improper.

**2.5 Model Comparison**

Suppose we have two simple regression models, \( M_1 \) and \( M_2 \), which purport to explain \( y \). These models differ in their explanatory variables. We distinguish the two models by adding subscripts to the variables and parameters. That is, \( M_j \) for \( j = 1, 2 \) is based on the simple linear regression model:

\[
y_i = \beta_j x_{ji} + \epsilon_{ji}
\]

for \( i = 1, \ldots, N \). Assumptions about \( \epsilon_{ji} \) and \( x_{ji} \) are the same as those about \( \epsilon_i \) and \( x_i \) in the previous section (i.e. \( \epsilon_{ji} \) is i.i.d. \( N(0, \sigma^2) \), and \( x_{ji} \) is either not random or exogenous for \( j = 1, 2 \)).
For the two models, we write the Normal-Gamma natural conjugate priors as:

$$\beta_j, \h_j | M_j \sim NG(\bar{\beta}_j, \bar{\gamma}_j, \bar{s}_j^{-2}, \gamma_j)$$  \hspace{1cm} (2.25)

which implies posteriors of the form:

$$\beta_j, \h_j | y, M_j \sim NG(\bar{\beta}_j, \bar{\gamma}_j, \bar{s}_j^{-2}, \bar{\nu}_j)$$  \hspace{1cm} (2.26)

where

$$\bar{\nu}_j = \frac{1}{\bar{\nu}_j^{-1} + \sum x_{j}^2}$$  \hspace{1cm} (2.27)

$$\bar{\beta}_j = \bar{\nu}_j (\bar{\nu}_j^{-1} \bar{\beta}_j + \bar{\beta}_j \sum x_{j}^2)$$  \hspace{1cm} (2.28)

$$\bar{\nu}_j = \nu_j + N$$  \hspace{1cm} (2.29)

and \(\bar{s}_j^{-2}\) is defined implicitly through

$$\bar{\nu}_j \bar{s}_j^2 = \nu_j s_j^2 + v_j s_j^2 + \frac{(\bar{\beta}_j - \beta_j)^2}{\bar{\nu}_j + \sum x_{j}^2}$$  \hspace{1cm} (2.30)

\(\bar{\beta}_j, s_j^2 \) and \(\nu_j \) are OLS quantities analogous to those defined in (2.3)–(2.5). In other words, everything is as in (2.7)–(2.12), except that we have added \(\bar{j}\) subscripts to distinguish between the two models.

Equations (2.26)–(2.30) can be used to carry out posterior inference in either of the two models. However, our purpose here is to discuss model comparison. As described in Chapter 1, a chief tool of Bayesian model comparison is the posterior odds ratio:

$$PO_{12} = \frac{p(M_1|y)}{p(M_2|y)}$$

The prior model probabilities, \(p(M_i)\) for \(i = 1, 2\), must be selected before seeing the data. The noninformative choice, \(p(M_1) = p(M_2) = \frac{1}{2}\), is commonly made. The marginal likelihood, \(p(y|M_j)\), is calculated as:

$$p(y|M_j) = \int \int p(y|\beta_j, h_j) p(\beta_j, h_j) d\beta_j dh_j$$  \hspace{1cm} (2.31)

Unlike with many models, in the Normal linear regression model with natural conjugate prior, the integrals in (2.31) can be calculated analytically. Potier (1995, pp. 542–543) or Zellner (1971, pp. 72–75) provide details of this calculation, which allows us to write:

$$p(y|M_j) = c_j \left( \frac{\bar{\nu}_j}{\bar{\gamma}_j} \right)^{\frac{\nu_j}{2}} \left( \bar{\nu}_j s_j^2 \right)^{-\frac{\nu_j}{2}}$$  \hspace{1cm} (2.32)

A linear regression model with a single variable

$$c_j = \frac{\Gamma \left( \frac{\nu_j}{2} \right) \left( \nu_j s_j^2 \right)^{\frac{\nu_j}{2}}}{\Gamma \left( \frac{\nu_j}{2} \right) \nu_j s_j^2}$$  \hspace{1cm} (2.33)

and \(\Gamma()\) is the Gamma function. The posterior odds ratio comparing \(M_1\) to \(M_2\) becomes:

$$PO_{12} = \frac{\left( \frac{\nu_1}{\nu_1} \right)^{\frac{\nu_1}{2}} \left( \nu_1 s_1^2 \right)^{-\frac{\nu_1}{2}} p(M_1)}{\left( \frac{\nu_2}{\nu_2} \right)^{\frac{\nu_2}{2}} \left( \nu_2 s_2^2 \right)^{-\frac{\nu_2}{2}} p(M_2)}$$  \hspace{1cm} (2.34)

The posterior odds ratio can be used to calculate the posterior model probabilities, \(p(M_j|y)\), using the relationships:

$$p(M_1|y) = \frac{PO_{12}}{1 + PO_{12}}$$

and

$$p(M_2|y) = \frac{1}{1 + PO_{12}}$$

A discussion of (2.34) offers insight into the factors which enter a Bayesian comparison of models. First, the greater is the prior odds ratio, \(\frac{p(M_1)}{p(M_2)}\), the higher the support for \(M_1\). Note, secondly, that \(\bar{\nu}_j s_j^2\) contains the term \(\nu_j s_j^2\) which is the sum of squared errors (see (2.3) and (2.5)). The sum of squared errors is a common measure of the model fit, with lower values indicating a better model fit. Hence, the posterior odds ratio rewards models which fit the data better. Thirdly, other things being equal, the posterior odds ratio will indicate support for the model where there is the greatest coherency between prior and data information (i.e. \(\bar{\beta}_j - \beta_j\)) enters \(\bar{\nu}_j s_j^2\). Finally, \(\left[ \frac{\nu_1}{\nu_2} \right]^{\frac{\nu_1}{2}}\) is the ratio of prior to posterior variances. This term can be interpreted as saying, all else being equal, the model with more prior information (i.e. smaller prior variance) relative to posterior information receives most support.

As we shall see in the next chapter, posterior odds ratios also contain a reward for parsimony in that, all else being equal, posterior odds favor the model with fewer parameters. The two models compared here have the same number of parameters (i.e. \(\beta_j\) and \(h_j\)) and, hence, this reward for parsimony is not evident. However, in general, this is an important feature of posterior odds ratios.

\(^2\)See Potier (1995, p. 98) for a definition of the Gamma function. All that you need to know here is that the Gamma function is calculated by the type of software used for Bayesian analysis (e.g. MATLAB or Gauss).
Under the noninformative variant of the natural conjugate prior (i.e., $\nu_j = 0$, $V_j^{-1} = 0$), the marginal likelihood is not defined and, hence, the posterior odds ratio is undefined. This is one problem with the use of noninformative priors for model comparison (we will see another problem in the next chapter). However, in the present context, a common solution to this problem is to set $\nu_1 = \nu_2$ equal to an arbitrarily small number and do the same with $V_1^{-1}$ and $V_2^{-1}$. Also, set $s_1^2 = s_2^2$. Under these assumptions, the posterior odds ratio is defined and simplifies and becomes arbitrarily close to:

$$P_{O12} = \left( \frac{1}{\sum s_1^2} \right)^{\frac{h}{2}} \left( \frac{1}{\sum s_2^2} \right)^{-\frac{h}{2}} \frac{p(M_1)}{p(M_2)}$$

(2.35)

In this case, the posterior odds ratio reflects only the prior odds ratio, the relative goodness of fit of the two models, and the ratio of terms involving $\sum s_j^2$, which reflect the precision of the posterior for $M_j$. However, as we shall see in the next chapter, this solution to the problem which arises from the use of the noninformative prior will not work when the number of parameters is different in the two models being compared.

In this section, we have shown how a Bayesian would compare two models. If you have many models, you can compare any or all pairs of them or calculate posterior model probabilities for each model (see the discussion after (1.7) in Chapter 1).

### 2.6 PREDICTION

Now let us drop the $j$ subscript and return to the single model with likelihood and prior defined by (2.6) and (2.7). Equations (2.8)–(2.12) describe Bayesian methods for learning about the parameters $\beta$ and $h$, based on a data set with $N$ observations. Suppose interest centers on predicting an unobserved data point generated from the same model. Formally, assume we have the equation:

$$y^* = \beta x^* + \varepsilon^*$$

(2.36)

where $y^*$ is not observed. Other than this, all the assumptions of this model are the same as for the simple regression model discussed previously (i.e., $\varepsilon^*$ is independent of $\varepsilon_i$ for $i = 1, \ldots, N$ and is $N(0, h^{-1})$, and the $\beta$ in (2.36) is the same as the $\beta$ in (2.1)). It is also necessary to assume $x^*$ is observed. To understand why the latter assumption is necessary, consider an application where the dependent variable is a worker’s salary, and the explanatory variable is some characteristic of the worker (e.g., years of education). If interest focuses on predicting the wage of a new worker, we would have to know her years of education in order to form a meaningful prediction.

As described in Chapter 1, Bayesian prediction is based on calculating:

$$p(y^*|y) = \int p(y^*|y, \beta, h) p(\beta, h|y) d\beta dh$$

(2.37)

The fact that $\varepsilon^*$ is independent of $\varepsilon_i$ implies that $y$ and $y^*$ are independent of one another and, hence, $p(y^*|y, \beta, h) = p(y^*|\beta, h)$. The terms inside the integral in (2.37) are thus the posterior, $p(\beta, h|y)$, and $p(y^*|\beta, h)$. Using a similar reasoning to that used for deriving the likelihood function, we find that

$$p(y^*|\beta, h) = \frac{h}{(2\pi)^{\frac{N}{2}}} \exp \left[ -\frac{h}{2}(y^* - \beta x^*)^2 \right]$$

(2.38)

Multiplying (2.38) by the posterior given in (2.8) and integrating as described in (2.37) yields (Zellner, 1971, pp. 72–75):

$$p(y^*|y) \propto (\bar{v} + (y^* - \bar{\beta} x^*)^2 h^{-2})(1 + V x^*)^{-1}$$

(2.39)

It can be verified (see Appendix B, Definition B.25) that this is a univariate $t$-density with mean $\bar{\beta} x^*$, variance $\frac{h^2}{V x^2}$, and degrees of freedom $\bar{v}$. In other words,

$$y^* | y \sim t(\bar{\beta} x^*, \bar{v}^2 (1 + V x^2), \bar{v})$$

(2.40)

These results can be used to provide point predictions and measures of uncertainty associated with the point prediction (e.g., the predictive standard deviation).

Our discussion of prediction is a logical place to introduce an important Bayesian concept: model averaging. In the previous section, we have shown how to calculate posterior model probabilities, $p(M_j|y)$, for $j = 1, 2$. These can be used to select one of the two models to work with. However, it is not always desirable to simply choose the one model with highest posterior model probability and throw away the other (or others). Bayesian model averaging involves keeping all models, but presenting results averaged over all models. In terms of the rules of probability, it is simple to derive:

$$p(y^*|y) = p(y^*|y, M_1) p(M_1|y) + p(y^*|y, M_2) p(M_2|y)$$

(2.41)

In words, insofar as interest centers on $p(y^*|y)$, you should not simply choose one model and work with, e.g., $p(y^*|y, M_1)$, but rather average results over the two models with weights given by the posterior model probabilities. Using the properties of the expected value operator (see Appendix B, Definition B.8), it follows immediately that:

$$E(y^*|y) = E(y^*|y, M_1) p(M_1|y) + E(y^*|y, M_2) p(M_2|y)$$
which can be used to calculate point predictions averaged over the two models. If
\( g(\cdot) \) is any function of interest (see (1.11)), then the preceding result generalizes to
\[
E[g(y^*)|y] = E[g(y^*)|y, M_1]p(M_1|y) + E[g(y^*)|y, M_2]p(M_2|y)
\] (2.42)
which can be used to calculate other functions of the predictive such as the predictive variance.

These results can be generalized to the case of many models and to the case
where the function of interest involves parameters instead of \( y^* \). Bayesian model
averaging is discussed in much greater detail in Chapter 11.

2.7 EMPIRICAL ILLUSTRATION

The regression model outlined in this chapter is probably too simple to be used
for any serious empirical work. For one thing, to simplify the algebra, we have
not included an intercept in the model. Furthermore, virtually any serious application
will involve several explanatory variables. Hence, to illustrate the basic
corcepts discussed in this chapter, we will work with a data set artificially generated by the computer. That is, we set \( N = 50 \). We begin by generating values of
the explanatory variable, \( x_i \), which are i.i.d. draws from the \( N(0, 1) \) distribution
for \( i = 1, \ldots , 50 \). We then generate values for the errors, \( \epsilon_i \), which
are \( N(0, h^{-1}) \). Finally, we use the explanatory variables and errors to generate the dependent variable \( y_i = \beta x_i + \epsilon_i \). We set \( \beta = 2 \) and \( h = 1 \). We use two
priors, the noninformative one given in (2.23) and the informative natural conjugate prior given in (2.7) with \( \beta = 1.5, \sigma = 0.25, \gamma = 10 \) and \( \xi^{-2} = 1 \). The choices of data generating process and prior hyperparameter values are purely illustrative.

Tables 2.1 and 2.2 present prior and posterior properties of the model parameters, \( \beta \) and \( h \), respectively, using (2.7)–(2.22). Figure 2.1 plots posteriors for \( \beta \) under the informative and noninformative priors as well as the informative prior itself (the noninformative prior for \( \beta \) is simply a flat line). From (2.13) it follows that the plotted p.d.f.s are all t-densities. Posterior properties based on the noninformative prior reflect only likelihood function information and are equivalent to frequentist OLS quantities (see (2.19)–(2.22)). For this reason, the

| Table 2.1 Prior and Posterior Properties of \( \beta \) |
|---|---|---|
| | Informativ | Using Noninformative Prior | Using Informative Prior |
| Mean | 1.50 | 2.06 | 1.96 |
| St. Deviation | 0.56 | 0.24 | 0.22 |

Figure 2.1 Marginal Prior and Posteriors for \( \beta \)

The marginal posterior for \( \beta \) under the noninformative prior is labeled 'Likelihood' in Figure 2.1.

The tables and figure show clearly how Bayesian inference involves combining
prior and data information to form a posterior. For instance, in Figure 2.1, it can be seen that the posterior based on the informative prior looks to be
an average of the prior density and the likelihood function. Tables 2.1 and 2.2
show that the posterior means of both parameters, \( E(\beta|y) \) and \( E(h|y) \), using the informative prior lie between the relevant prior mean and the likelihood-based
quantity (i.e. the posterior mean using the noninformative prior). The prior we have selected contains less information than the data. This can be seen either
in the figure (i.e. the prior p.d.f. is more dispersed than the likelihood) or in
the tables (i.e. the prior standard deviations are larger than the likelihood-based
quantities).
Remember that, since the data set has been artificially created, we know that the true parameter values are \( \beta = 2 \) and \( h = 1 \). You would, of course, never expect a point estimate like a posterior mean or an OLS quantity to be precisely equal to the true value. However, the posterior means are quite close to their true values relative to their posterior standard deviations. Note also that the posterior standard deviations using the informative prior are slightly smaller than those using the noninformative prior. This reflects the intuitive notion that, in general, more information allows for more precise estimation. That is, it is intuitively sensible that a posterior which combines both prior and data information will be less dispersed than one which uses a noninformative prior and is based only on data information. In terms of the formulae, this intuitive notion is captured through (2.9) being smaller than (2.19) if \( V > 0 \). Note, however, that this intuition is not guaranteed to hold in every case since, if prior and data information are greatly different from one another, then (2.12) can become much bigger than (2.22). Since both \( V \) and \( \sigma^2 \) enter the formula for the posterior standard deviation of \( \beta \), it is possible (although unusual) for the posterior standard deviation using an informative prior to be larger than that using a noninformative prior.

To illustrate model comparison, let us suppose we are interested in comparing the model we have been discussing to another linear regression model which contains only an intercept (i.e. in this second model \( x_i = 1 \) for \( i = 1, \ldots, 50 \)). For both models, we use the same informative prior described above (i.e. both priors are \( N(0.75, 0.25, 1, 10) \)). Assuming a prior odds ratio equal to one, (2.34) can be used to calculate the posterior odds ratio comparing these two models. Of course, we know our first model is the correct one and, hence, we would expect the posterior odds ratio to indicate this. This does turn out to be the case, since we find a posterior odds ratio of 3749.7. In words, we are finding overwhelming support for our correct first model. It is almost 4000 times more likely to be true than the second model. In terms of posterior model probabilities, the posterior odds ratio implies that \( p(M_1 | y) = 0.9997 \) and \( p(M_2 | y) = 0.0003 \). If we were to do Bayesian model averaging using these two models, we would attach 99.97% weight to results from the first model and only 0.03% weight to results from the second (see (2.41)).

Predictive inference can be carried out using (2.40). We illustrate how this is done by selecting the point \( x^* = 0.5 \). Using the informative prior, it turns out that

\[
y^* | y \sim \mathcal{N}(0.98, 0.97, 60)
\]

Using the noninformative prior, it turns out that

\[
y^* | y \sim \mathcal{N}(1.03, 0.95, 50)
\]

Either of these probability statements can be used to present point predictions, predictive standard deviations, or any other predictive function of interest you may wish to calculate.

2.8 SUMMARY

In this chapter, we have gone through a complete Bayesian analysis (i.e. likelihood, prior, posterior, model comparison and prediction), for the Normal linear regression model with a single explanatory variable and a so-called natural conjugate prior. For the parameters of this model, \( \beta \) and \( h \), this prior has a Normal-Gamma distribution. The natural conjugate nature of the prior means that the posterior also has a Normal-Gamma distribution. For this prior, posterior and predictive inference and model comparison can be done analytically and no posterior simulation is required. Other themes introduced in this chapter include the concept of a noninformative prior and Bayesian model averaging.

2.9 EXERCISES

2.9.1 Theoretical Exercises

1. Prove the result in (2.8). Hint: This is a standard derivation proved in many other textbooks such as Poirier (1995, p. 527) or Zellner (1971, pp. 60–61), and you may wish to examine some of these references if you are having trouble.

2. For this question, assume the likelihood function is as described in Section 2.2 with known error precision, \( h = 1 \), and \( x_i = 1 \) for \( i = 1, \ldots, N \).

(a) Assume a Uniform prior for \( \beta \) such that \( \beta \sim U(\alpha, \gamma) \). Derive the posterior \( p(\beta | y) \).

(b) What happens to \( p(\beta | y) \) as \( \alpha \to -\infty \) and \( \gamma \to \infty \)?

(c) Use the change-of-variable theorem (Appendix B, Theorem B.21) to derive the prior for a one-to-one function of the regression coefficient, \( g(\beta) \), assuming that \( \beta \) has the Uniform prior given in (a). Sketch the implied prior for several choices of \( g \) (e.g. \( g(\beta) = \log(\beta) \), \( g(\beta) = \exp(\beta) \), \( g(\beta) = \exp(g) \), etc.).

(d) Consider what happens to the priors in part (c) as \( \alpha \to -\infty \) and \( \gamma \to \infty \).

(e) Given your answers to part (d), discuss whether a prior which is 'noninformative' when the model is parameterized in one way is also 'noninformative' when the model is parameterized in a different way.

2.9.2 Computer-Based Exercises

Remember that some data sets and MATLAB programs are available on the website associated with this book.

3. Generating artificial data sets. This is an important skill since they can be used to understand the properties of models and investigate the performance of a particular computer algorithm. Since you have chosen the values for
Chapter 7
Simulation by MCMC Methods

The basis of an MCMC algorithm is the construction of a transition kernel (see Section 6.3), \( p(x, y) \), that has an invariant density equal to the target density. Given such a kernel, the process can be started at \( x_0 \) to yield a draw \( x_1 \) from \( p(x_0, x_1) \), \( x_2 \) from \( p(x_1, x_2) \), \ldots, and \( x_G \) from \( p(x_{G-1}, x_G) \), where \( G \) is the desired number of simulations. After a transient period, the distribution of the \( x_G \) is approximately equal to the target distribution. The question is how to find a kernel that has the target as its invariant distribution. It is remarkable that there is a general principle for finding such kernels, the Metropolis–Hastings (MH) algorithm. We first discuss a special case—the Gibbs algorithm or Gibbs sampler—and then explain a more general version of the MH algorithm.

It is important to distinguish between the number of simulated values \( G \) and the number of observations \( n \) in the sample of data that is being analyzed. The former may be made very large—the only restriction comes from computer time and capacity, but the number of observations is fixed at the time the data are collected. Larger values of \( G \) lead to more accurate approximations. MCMC algorithms provide an approximation to the exact posterior distribution of a parameter; that is, they approximate the posterior distribution of the parameters, taking the number of observations to be fixed at \( n \). In contrast, frequentist procedures that invoke such criteria as consistency are concerned with the effects of letting \( n \) become large.

A brief comment on notation: when discussing simulation techniques in this chapter, we follow the literature in denoting random variables by such symbols as \( x \), which usually denotes the current value of the chain, and \( y \), which usually denotes the next value, and the target distribution by \( f(\cdot) \). For applications to Bayesian inference, the random variables of interest are parameters \( \theta \) and the target is the posterior distribution \( \pi(\theta|y) \), where \( y \) represents the data. We utilize the latter notation in Part III.

7.1 Gibbs Algorithm

7.1.1 Basic Algorithm

The Gibbs algorithm is a special case of the MH algorithm that can be used when it is possible to sample from each conditional distribution. For example, suppose we wish to sample from a nonstandard joint distribution \( f(x_1, x_2) \), where the variables appear in two blocks, both of which may be vectors. Further suppose that the two conditional distributions \( f(x_1|x_2) \) and \( f(x_2|x_1) \) are distributions for which simulation algorithms are known. Then consider the following algorithm.

**Algorithm 7.1: Gibbs algorithm with two blocks**

1. Choose a starting value \( x_2^{(0)} \).
2. At the first iteration, draw

\[
\begin{align*}
x_1^{(1)} & \text{ from } f(x_1|x_2^{(0)}), \\
x_2^{(1)} & \text{ from } f(x_2|x_1^{(1)}).
\end{align*}
\]

3. At the \( g \)th iteration, draw

\[
\begin{align*}
x_1^{(g)} & \text{ from } f(x_1|x_2^{(g-1)}), \\
x_2^{(g)} & \text{ from } f(x_2|x_1^{(g)}),
\end{align*}
\]

until the desired number of iterations is obtained. (The roles of \( x_1 \) and \( x_2 \) may be interchanged.)

Because the starting value is not drawn from the invariant distribution, some portion of the initial sample must be discarded; this portion is the transient or burn-in sample. The burn-in sample size \( B \) is usually set at several hundred to several thousand, and checks can be made to see whether the choice matters; in most cases, there are no theorems that indicate what \( B \) should be. For \( g > B \), the distribution of the draws is approximately the target distribution. We denote by \( G \) the sample size after discarding the first \( B \) observations. Convergence diagnostics are discussed in Section 7.3.

We now show that the invariant distribution of the Gibbs kernel is the target distribution. To simplify the notation, let \( x = (x_1, x_2) \) be the values of the random variables at the beginning of one iteration of the algorithm and \( y = (y_1, y_2) \) be the values at the end of the iteration. The Gibbs kernel is

\[
p(x, y) = f(y_1|x_2)f(y_2|x_1),
\]
from which we can compute
\[
\int p(x, y)f(x)\,dx = \int f(y_1|x_2)f(y_2|y_1)f(x_1, x_2)\,dx_1\,dx_2
\]
\[
= f(y_2|y_1)\int f(y_1|x_2)f(x_2)\,dx_2
\]
\[
= f(y_2|y_1)f(y_1)
\]
\[
= f(y).
\]
which proves that \( f(\cdot) \) is the invariant distribution for the Gibbs kernel \( p(x, y) \).

Proof that the invariant distribution of the Gibbs kernel is the target distribution is a necessary, but not sufficient condition for the kernel to converge to the target. Such conditions are very technical and difficult to verify for particular cases, but some general results are available. For example, Tierney (1994, p. 1712) states that most Gibbs samplers satisfy the conditions of the following theorem.

**Theorem 7.1** Suppose \( P \) is \( \pi \)-irreducible and has \( \pi \) as its invariant distribution. If \( P(\cdot, \cdot) \) is absolutely continuous with respect to \( \pi \) for all \( x \), then \( P \) is Harris recurrent.

Extending Gibbs sampling to \( d \) blocks of variables is possible when all of the conditional densities \( f(x_i|x_{-i}) \) are distributions from which random draws can be generated, where \( x_{-i} \) are all the variables in the joint distribution other than \( x_i \). The algorithm proceeds as follows (the ordering of the \( x_i \) is arbitrary).

**Algorithm 7.2: Gibbs algorithm with \( d \) blocks**

1. Choose \( x_2^{(0)}, \ldots, x_d^{(0)} \).
2. Draw
   \[
   x_1^{(1)} \text{ from } f(x_1|x_2^{(0)}, \ldots, x_d^{(0)})
   \]
   \[
   x_2^{(1)} \text{ from } f(x_2|x_1^{(1)}, x_3^{(0)}, \ldots, x_d^{(0)})
   \]
   \[
   \vdots
   \]
   \[
   x_d^{(1)} \text{ from } f(x_d|x_1^{(1)}, \ldots, x_{d-1}^{(1)}).
   \]
3. At the \( g \)th iteration, draw
   \[
   x_1^{(g)} \text{ from } f(x_1|x_2^{(g-1)}, \ldots, x_d^{(g-1)})
   \]
   \[
   x_2^{(g)} \text{ from } f(x_2|x_1^{(g)}, x_3^{(g-1)}, \ldots, x_d^{(g-1)})
   \]
   \[
   \vdots
   \]
   \[
   x_d^{(g)} \text{ from } f(x_d|x_1^{(g)}, \ldots, x_{d-1}^{(g)}).
   \]

Since many applications of Gibbs sampling are presented in Part III, we offer only two examples here.

Let \( y_i \sim N(\mu, h^{-1}) \), \( i = 1, \ldots, n \), be independently distributed, where the distribution has been parameterized in terms of the precision. We assume the conditionally conjugate priors \( \mu \sim N(\mu_0, h_0^{-1}) \) and \( h \sim G(\alpha_0/2, \delta_0/2) \). Verify that this model is a special case of the normal linear regression model. From this specification, we have

\[
\pi(\mu, h|y) \propto h^{n/2} \exp \left[ -\frac{h}{2} \sum (y_i - \mu)^2 \right] \exp \left[ -\frac{h_0}{2} (\mu - \mu_0)^2 \right] 
\]
\[
\times h^{\alpha_0/2-1} \exp \left[ -\frac{\delta_0 h}{2} \right].
\]

From here, it is easy to derive the conditional posterior distribution of \( h \),

\[
\pi(h|\mu, y) \propto h^{(\alpha_0+n)/2-1} \exp \left[ -h \frac{\delta_0 + \sum (y_i - \mu)^2}{2} \right].
\]  
(7.1)

Equation (7.1) is recognized as the density function of \( G(\alpha_0 + n)/2, (\delta_0 + \sum (y_i - \mu)^2)/2 \), which is available for sampling in all statistical packages.

We complete the square in \( \mu \) to obtain

\[
\pi(\mu|h, y) \propto \exp \left[ -\frac{h_0 + hn}{2} \left( \mu - \frac{h_0 \mu_0 + hn \bar{y}}{h_0 + hn} \right)^2 \right],
\]  
(7.2)

which should be recognized as \( N((h_0 \mu_0 + hn \bar{y})/(h_0 + hn), (h_0 + hn)^{-1}) \) and is available for sampling in all statistical packages. In algorithmic form,

**Algorithm 7.3: Mean and precision for normal model**

1. Choose a starting value for \( \mu = \mu^{(0)} \).
2. Sample \( h^{(1)} \) from \( G(\alpha_1/2, \delta_1/2) \), where \( \alpha_1 = \alpha_0 + n \) and \( \delta_1 = \delta_0 + \sum (y_i - \mu^{(0)})^2 \).
3. At the \( g \)th iteration, draw
   \[
   \mu^{(g)} \text{ from } N((h_0 \mu_0 + h^{(g-1)}n \bar{y})/(h_0 + h^{(g-1)}n), (h_0 + h^{(g-1)}n)^{-1})
   \]
   \[
   h^{(g)} \text{ from } G(\alpha_1/2, \delta_0 + \sum (y_i - \mu^{(g)})^2)/2).
   \]

If desired, the sampling can begin with \( h^{(0)} \) and the algorithm modified accordingly.

As a second example, we consider a more general version of the Poisson model with changing parameters that is described in Exercise 3.8. We assume that

\[
p(y_i) = \begin{cases} 
\frac{e^{-\theta} \theta^{y_i}}{y_i!}, & \text{for } i = 1, \ldots, k, \\
\frac{e^{-\theta} \theta^{y_i}}{y_i!}, & \text{for } i = k + 1, \ldots, n,
\end{cases}
\]  
(7.3)
where \( y_j = 0, 1, \ldots \), and the switch point \( k \) is unknown. The specification is completed by assigning the conditionally conjugate priors,

\[
\theta_1 \sim G(\alpha_{10}, \beta_{10}), \quad \theta_2 \sim G(\alpha_{20}, \beta_{20}), \quad \pi(k = j) = 1/n, \quad j = 1, \ldots, n. \tag{7.4}
\]

We have assigned gamma distributions to \( \theta_1 \) and \( \theta_2 \) because they are positive and the discrete uniform distribution to \( k \) over the values \( 1, \ldots, n \), which includes the possibility that no change occurs; that is, \( k = n \). The details of the algorithm are taken up in Exercise 7.3, and references may be found in Section 7.4.

Although the Gibbs sampler usually works well in practice, there are some situations in which it does not. If there is a high correlation between one or more of the random variables in different blocks, the algorithm may not “mix well.” This means that the sampler fails to traverse the full support of the sample space, generating iterations from only a limited portion. For a large enough number of iterations, it will traverse the space, but it may fail for the number of iterations generated in practice.

As an example, consider the problem of sampling \( X = (X_1, X_2) \) from the bivariate normal distribution \( N_2(0, \Sigma) \), where

\[
\Sigma = \begin{pmatrix}
1 & \rho \\
\rho & 1
\end{pmatrix}.
\]

We emphasize that most statistical packages allow efficient and independent sampling from the bivariate normal distribution, and the method we examine is not used in practice. We employ the Gibbs sampler with \( X_1 \) and \( X_2 \) as our two blocks. You should verify that \( f(X_1|x_2) \sim N(\rho x_2, 1 - \rho^2) \) and \( f(X_2|x_1) \sim N(\rho x_1, 1 - \rho^2) \). This algorithm performs badly if \( \rho \approx 1 \), which implies that the conditional variance of both variables \( (1 - \rho^2) \) is close to zero. Accordingly, in each iteration, the sampler generates values that are very close to the value of the previous iteration, which implies that the initial value \( x_2^{(0)} \) or \( x_2^{(0)} \) can play a large role in the generated sample. Since the marginal distribution is known to be \( X_1 \sim N(0, 1) \), we can compare the results of a Gibbs sampler, where \( G = 5,000 \) and \( N = 500 \), to the true distribution. With \( \rho = 0.999 \) we find that \( \hat{x}_1 = 1.084 \) when \( x_2^{(0)} = 1 \) and \( \hat{x}_1 = -0.623 \) when \( x_2^{(0)} = -2 \). Note that the mean is greatly affected by the starting value of the algorithm. In contrast, when \( \rho = 0.5 \), the starting value of \(-2\) has little effect, yielding a mean of 0.0587. We discuss in Section 7.3 some methods for detecting poor mixing.

### 7.1 Gibbs Algorithm

#### 7.1.2 Calculation of Marginal Likelihood

We next consider the problem of computing the marginal likelihood when working with a nonstandard distribution. In Section 3.2.4, we point out that the marginal likelihood is the inverse of the normalizing constant of the posterior distribution. The normalizing constant is unknown when working with nonstandard distributions, but it is not needed for the implementation of the Gibbs and MH algorithms. It is, however, needed for computing Bayes factors. Several methods have been proposed, and we describe Chib’s widely used approach to estimating the marginal likelihood when a sample is generated from a Gibbs algorithm.

The Chib method begins with the identity

\[
\pi(\theta^*|y) = \frac{f(y|\theta^*) \pi(\theta^*)}{f(y)},
\]

where \( \theta^* \) is a particular value of \( \theta \) and \( f(y) \) is the marginal likelihood. For numerical accuracy, \( \theta^* \) is usually chosen to be the mean of the sample values. The identity can be written as

\[
f(y) = \frac{f(y|\theta^*) \pi(\theta^*)}{\pi(\theta^*|y)},
\]

and the Chib method computes the right-hand side from the output of a Gibbs sampler. The terms in the numerator of the right-hand side are readily computed; they are the likelihood function and prior distribution, respectively, evaluated at \( \theta^* \). The main problem is to compute \( \pi(\theta^*|y) \), for which the normalizing constant is not known.

Consider the simple case where the Gibbs algorithm is run in two blocks, denoted by \( \theta_1 \) and \( \theta_2 \). We may write

\[
\pi(\theta_1^*, \theta_2^*|y) = \pi(\theta_1^*|\theta_2^*, y)\pi(\theta_2^*|y).
\]

The first term on the right can be evaluated immediately because the conditional distributions are known when running the Gibbs sampler. To compute the second, Chib employs the identity

\[
\pi(\theta_2^*|y) = \int \pi(\theta_1, \theta_2^*|y) d\theta_1
\]

\[
= \int \pi(\theta_2^*|\theta_1, y)\pi(\theta_1|y) d\theta_1,
\]

which can be approximated by

\[
\hat{\pi}(\theta_2^*|y) = \frac{1}{G} \sum \pi(\theta_2^*|\theta_1^{(g)}, y),
\]

where the values of \( \theta_1^{(g)} \) are taken from the Gibbs output.
When there are three or more blocks, the computation requires additional simulations. Taking the three-block case as an example, we start with the identity

\[ f(y) = \frac{f(y|\theta_1^*, \theta_2^*, \theta_3^*) \pi(\theta_1^*, \theta_2^*, \theta_3^*)}{\pi(\theta_1^*, \theta_2^*, \theta_3^*|y)}. \]

The numerator is readily available, and we write the denominator as

\[ \pi(\theta_1^*, \theta_2^*, \theta_3^*|y) = \pi(\theta_1^*|y) \pi(\theta_2^*|\theta_1^*, y) \pi(\theta_3^*|\theta_1^*, \theta_2^*, y). \]

The Gibbs output can be used to approximate the first term as

\[ \hat{\pi}(\theta_1^*|y) = \frac{1}{G} \sum \pi(\theta_1^*|\theta_2^{(g)}, \theta_3^{(g)}, y). \]

For the second term, we use

\[ \pi(\theta_2^*|\theta_1^*, y) = \int \pi(\theta_2^*|\theta_1^*, \theta_3, y) \pi(\theta_3|\theta_1^*, y) d\theta_3. \]

Then

\[ \hat{\pi}(\theta_2^*|\theta_1^*, y) = \frac{1}{G} \sum \pi(\theta_2^*|\theta_1^*, \theta_3^{(g)}, y), \]

where the \( \theta_3^{(g)} \) are generated from a “reduced run,” in which \( \theta_2^{(g)} \) and \( \theta_3^{(g)} \) are sampled from \( \pi(\theta_2|\theta_1^*, \theta_3, y) \) and \( \pi(\theta_3|\theta_1^*, \theta_2, y) \), respectively, and \( \theta_1 \) is fixed at \( \theta_1^* \). Computations for the reduced run can use the same code as the original run, but \( \theta_1 \) is held constant at \( \theta_1^* \). Finally, the value of \( \pi(\theta_3^*|\theta_1^*, \theta_2^*, y) \) is available directly from the conditional distribution.

Since many of the models discussed in Part III use only the Gibbs algorithm, readers interested in applications may proceed to Section 7.3 and then to Part III, returning to Section 7.2 when the MH algorithm is encountered.

### 7.2 Metropolis–Hastings Algorithm

#### 7.2.1 Basic Algorithm

The MH algorithm is more general than the Gibbs sampler because it does not require that the full set of conditional distributions be available for sampling. Although it can be used in blocks, we first explain it in the one-block case. To generate a sample from \( f(X) \), where \( X \) may be a scalar or vector random variable, the first step is to find a kernel \( p(X, Y) \) that has \( f(\cdot) \) as its invariant distribution. Since the Gibbs sampler is of no use when one or more of the conditionals are not available for sampling, a different approach to finding a kernel is necessary. To that end, we introduce the idea of a **reversible** kernel, defined as a kernel \( q(\cdot, \cdot) \) for which

\[ f(x)q(x, y) = f(y)q(y, x). \]

If \( q \) is reversible,

\[
P(y \in A) = \int_A \int \mathbb{R}^d f(x)q(x, y) dx \, dy
= \int_A \int \mathbb{R}^d f(y)q(y, x) dx \, dy
= \int_A f(y) dy.
\]

This shows that \( f(\cdot) \) is the invariant distribution for the kernel \( q(\cdot, \cdot) \) because the probability that \( y \) is contained in \( A \) is computed from \( f(\cdot) \).

The fact that a reversible kernel has this property can help in finding a kernel that has the desired target distribution. Chib and Greenberg (1995b) show how this can be done when starting with a nonreversible proposal density. We now follow their derivation of the algorithm. The trick is to make an irreversible kernel reversible. If a kernel is not reversible, for some pair \((x, y)\) we have

\[ f(x)q(x, y) > f(y)q(y, x). \]

The MH algorithm deals with this situation by multiplying both sides by a function \( \alpha(\cdot, \cdot) \) that turns the irreversible kernel \( q(\cdot, \cdot) \) into the reversible kernel \( p(x, y) = \alpha(x, y)q(x, y) \):

\[ f(x)\alpha(x, y)q(x, y) = f(y)\alpha(y, x)q(y, x). \]

The expression \( \alpha(x, y)q(x, y) \) is interpreted as follows: if the present state of the process is \( x \), generate a value \( y \) from the kernel \( q(x, y) \) and make the move to \( y \) with probability \( \alpha(x, y) \). If the move to \( y \) is rejected, the process remains at \( x \). We call \( q(x, y) \) the proposal density because of its analogous role in the AR algorithm, but the MH algorithm is very different from the latter. In the AR case, the algorithm continues to generate values until a candidate is accepted. In contrast, in the MH algorithm, the procedure returns the current state of the process as the next state when a candidate is rejected and continues to the next iteration; this implies that values may be repeated in a simulation run. Note that this transition kernel combines a continuous kernel \( q(x, y) \) and a probability mass function \( \alpha(x, y) \).

How to define \( \alpha(x, y) \) is the next question. Suppose that

\[ f(x)q(x, y) > f(y)q(y, x). \]

Roughly speaking, this means that the kernel goes from \( x \) to \( y \) with greater probability than it goes from \( y \) to \( x \). Accordingly, if the process is at \( y \) and the kernel
proposes a move to \( x \), that move should be made with high probability. This can be done by setting \( \alpha(y, x) = 1 \). But then, \( \alpha(x, y) \) is determined because, from (7.5),

\[
f(x)q(x, y)\alpha(x, y) = f(y)q(y, x)
\]

implies

\[
\alpha(x, y) = \begin{cases} 
\min \left\{ \frac{f(y)q(y, x)}{f(x)q(x, y)}, 1 \right\}, & \text{if } f(x)q(x, y) \neq 0, \\
0, & \text{otherwise.}
\end{cases} \tag{7.6}
\]

The condition that \( f(x)q(x, y) \neq 0 \) is usually satisfied in practice because the starting value is always chosen in the support of the distribution and the kernel usually generates values in the support of the distribution. In some cases, however, it is convenient to generate values outside the support. As an example, draws from an untruncated proposal distribution are sometimes used to generate values from a truncated distribution because it is difficult to specify an appropriate truncated proposal density.

It is important to recognize that, in computing \( \alpha(\cdot, \cdot) \), an unknown constant in the target distribution is not needed, because it cancels out via the fraction \( f(y)/f(x) \).

To summarize in algorithmic form:

\textbf{Algorithm 7.4: MH algorithm}

1. Given \( x \), generate \( Y \) from \( q(x, y) \).
2. Generate \( U \) from \( U(0, 1) \). If
   \[
   U \leq \alpha(x, Y) = \min \left\{ \frac{f(Y)q(Y, x)}{f(x)q(x, Y)}, 1 \right\},
   \]
   return \( Y \). Otherwise, return \( x \) and go to 1.

Although we have shown that the MH kernel has the desired target distribution, this is only a necessary condition for convergence to the target. For Metropolis kernels, Tierney (1994, p. 1713) provides the following theorem.

\textbf{Theorem 7.2} Suppose \( P \) is a \( \pi \)-irreducible Metropolis kernel. Then \( P \) is Harris recurrent.

The next implementation issue is how to choose the proposal density \( q(\cdot, \cdot) \). In many cases, there are several possible choices, and which is best is often a matter of judgment. Several considerations enter into this decision. On the one hand, the kernel should generate proposals that have a reasonably good probability of acceptance; if not, the same value will be returned often, and the algorithm will mix poorly. On the other hand, there may be a high acceptance rate if the kernel generates only proposals that are close to the current point, but the sampling may then be confined to a small part of the support, again leading to poor mixing.

Two straightforward, but not necessarily good, kernels are the random-walk kernel and the independence kernel. For the former, the proposal \( y \) is generated from the current value \( x \) by the addition of a random variable or vector \( u, y = x + u \), where the distribution of \( u \) is specified. If that distribution is symmetric around zero, that is, \( h(u) = h(-u) \), the kernel has the property that \( q(x, y) = q(y, x) \), which implies that \( \alpha(x, y) = f(y)/f(x) \). Accordingly, with a random-walk kernel, a move from \( x \) to \( y \) is made for certain if \( f(y) > f(x) \). A move from a higher density point to a lower density point is not ruled out, but the probability of such a move \( f(x)/f(y) \) is less than one.

The independence kernel has the property \( q(x, y) = q(y) \); that is, the proposal density is independent of the current state of the chain. For this type of kernel

\[
\alpha(x, y) = \frac{f(y)/q(y)}{f(x)/q(x)},
\]

and our comments about the probability of a move are similar to those about the random-walk chain if \( f(\cdot) \) is replaced by \( f(\cdot)/q(\cdot) \).

As a simple example of an independence chain, we generate data from a Beta(3, 4) distribution with \( U(0, 1) \) as the proposal density. In algorithm form, we have the following.

\textbf{Algorithm 7.5: MH for Beta(3, 4) with \( U(0, 1) \) proposal}

1. Set \( x^{(0)} \) equal to a number between zero and one.
2. At the \( g \)th iteration, generate \( U_1 \) and \( U_2 \) from \( U(0, 1) \).
3. If
   \[
   U_1 \leq \alpha(x^{(g-1)}, U_2) = \frac{U_2^2(1 - U_2)^2}{(x^{(g-1)})^3(1 - x^{(g-1)})^3},
   \]
   set \( x^{(g)} = U_2 \). Otherwise set \( x^{(g)} = x^{(g-1)} \).
4. Go to 2 and continue until the desired number of iterations is obtained.

Figure 7.1 displays results for 5,000 iterations after discarding the first 500; it indicates a good fit between the generated values, plotted as a histogram, and the true distribution, plotted as a solid line. The acceptance probability is 0.57, meaning that 57% of the proposals were accepted. The mean of the sample values is 0.4296, compared to the theoretical mean of 3/7 = 0.4286.

We recommend a “tailored” kernel: construct a kernel that approximates the target distribution. This may be done by choosing a fat-tailed distribution, such as the multivariate \( t \) with small \( v \), whose mean and scale matrix are chosen to coincide with the mode and negative inverse of the second-derivative matrix at the mode, respectively. An example of a tailored kernel may be found in Section 9.2. If there is just one parameter block, the tailored kernel is an independence kernel. If there
is more than one block, the tailored kernel for the block being updated may depend on the current values of parameters in the other blocks.

Implementation of the MH algorithm is often facilitated by blocking, as in the Gibbs sampler. Blocking allows sampling for a few variables at a time, which may make it easier to find suitable proposal densities. Thus, if the target density \( f(X_1, X_2) \) has two blocks, we can use the MH algorithm to generate from each block.

**Algorithm 7.6: MH algorithm with two blocks**

1. Let the state be \((x_1, x_2)\) at the \(g\)th iteration and \((y_1, y_2)\) at the \(g + 1\)st iteration. Draw

   \[ Z_1 \text{ from } q_1(x_1, Z_1|x_2) \text{ and } U_1 \text{ from } U(0, 1). \]

2. If

   \[ U_1 \leq \alpha(x_1, Z_1|x_2) = \frac{f(Z_1, x_2)q_1(Z_1, x_1|x_2)}{f(x_1, x_2)q_1(x_1, Z_1|x_2)}, \]

   return \( y_1 = Z_1 \). Otherwise return \( y_1 = x_1 \).

3. Draw

   \[ Z_2 \text{ from } q_2(x_2, Z_2|y_1) \text{ and } U_2 \text{ from } U(0, 1). \]

4. If

   \[ U_2 \leq \alpha(x_2, Z_1|y_1) = \frac{f(y_1, Z_2)q_2(Z_2, x_2|y_1)}{f(y_1, x_2)q_2(x_2, Z_2|y_1)}, \]

   return \( y_2 = Z_2 \). Otherwise return \( y_2 = x_2 \).

In this algorithm, the kernel \( q_1(x_1, Y_1|x_2) \) is analogous to \( q(x, Y) \); it generates a value \( Y_1 \) conditional on the current value \( x_1 \) in the same block and the current value \( x_2 \) in the other block. If "tailored" proposal densities are used, new densities are specified for \( q_1(x_1, Z_1|x_2) \) and \( q_2(x_2, Z_2|y_1) \) for each value of \( x_2 \) and \( y_1 \), respectively. This algorithm can be extended to an arbitrary number of blocks.

Having introduced blocks of parameters, we can show that the Gibbs sampler is a special case of the MH algorithm. Consider \( \alpha(\cdot, \cdot) \) when the kernel for moving from the current value \( x_1 \) to the proposal value \( Y_1 \) is the conditional distribution \( f(x_1|x_2) \), which is assumed to be available for sampling. Then

\[ \alpha(x_1, Y_1|x_2) = \frac{f(Y_1, x_2)f(x_1|x_2)}{f(x_1, x_2)f(Y_1|x_2)}, \]

but, since \( f(Y_1|x_2) = f(Y_1, x_2)/f(x_2) \) and \( f(x_1|x_2) = f(x_1, x_2)/f(x_2) \), it follows that \( \alpha(x_1, Y_1|x_2) = 1 \), showing that the Gibbs algorithm is an MH algorithm where the proposal is always accepted.

When implementing the MH algorithm to blocks of parameters, Gibbs sampling may be employed in any blocks for which the conditional distributions are available for sampling. In the remaining blocks, the MH algorithm may be employed in the usual way, that is, by finding suitable proposal densities and accepting with probability \( \alpha(x, y) \). At each iteration, the algorithm works through the blocks, either moving to a new value or retaining the current value of the variables in the block. Examples of such algorithms appear in Part III.

### 7.2.2 Calculation of Marginal Likelihood

Chib and Jeliazkov (2001) have developed a modification of the Chib method for computing the marginal likelihood from MH output. We explain it for the one-block case; extensions may be found in their article. The method again starts with the identity

\[ f(y) = \frac{\pi(y|\theta^*)\pi(\theta^*)}{\pi(\theta^*|y)}, \]
where computing the denominator on the right-hand side is the biggest problem. From (7.5),

$$\alpha(\theta, \theta^*_{-}y)q(\theta, \theta^*_{-}y)\pi(\theta|y) = \alpha(\theta^*, \theta|y)q(\theta^*, \theta|y)\pi(\theta^*|y),$$

from which we find

$$\pi(\theta^*|y) = \frac{\int \alpha(\theta, \theta^*_{-}y)q(\theta, \theta^*_{-}y)\pi(\theta|y) d\theta}{\int \alpha(\theta^*, \theta|y)q(\theta^*, \theta|y) d\theta}.$$

The numerator of this expression can be estimated directly from the MH output $\theta^{(t)}$ as

$$\frac{1}{G} \sum_g \alpha(\theta^{(t)}, \theta^*_{-}y)q(\theta^{(t)}, \theta^*|y),$$

and the denominator can be estimated by drawing $\theta^{(j)}$ from $q(\theta^*, \theta|y)$, $j = 1, \ldots, J$; that is, generate $J$ values from the proposal density, all of which are conditioned on $\theta^*$, and then compute

$$\frac{1}{J} \sum_J \alpha(\theta^*, \theta^{(j)}|y).$$

It is important for this computation that $\alpha(\theta^{(t)}, \theta^*|y)$ be computed from (7.6) as the minimum of 1 and the fraction in the curly braces.

**7.3 Numerical Standard Errors and Convergence**

The n.s.e. is defined in Section 5.6 as $\sqrt{\text{Var}(Z)/G}$. For independent samples, the variance can be estimated as

$$\hat{\text{Var}}(Z) = \frac{1}{G-1} \sum(Z^{(t)} - \bar{Z})^2.$$

When the sample is not independent, as in the case of MCMC output, this expression is not valid. For nonindependent samples, the expectation of the sum of squared deviations from the mean includes covariance terms that must be taken into account, which the aforementioned expression fails to do. Several methods have been proposed to do this; we present the “batch means” method and refer to others in Section 7.4.

The logic behind the batch means method applied to MCMC output is that the autocorrelations of high-enough order tend to zero. By examining the empirical autocorrelations, we can determine a lag at which the autocorrelation is small enough that it can be neglected; a cutoff value of 0.05 is often adopted. Let $j_0$ be a lag length such that the absolute value of the autocorrelation at lag $j_0$ is less than 0.05. Then divide the sample of $G$ observations into $G/j_0 = N$ groups or batches, adjusting $j_0$ or $G$ if necessary to get an integer value for $N$. Compute the mean of the simulated output in batch $b$, $b = 1, \ldots, N$, and then estimate the variance by

$$\hat{\text{Var}}(\bar{Z}_b) = \frac{1}{N-1} \sum_{b=1}^{N} (\bar{Z}_b - \bar{Z})^2,$$

where $\bar{Z}_b$ is the mean of batch $b$ and $\bar{Z}$ is the overall mean. The n.s.e. is then computed as $\sqrt{\hat{\text{Var}}(\bar{Z}_b)/N}$.

Since positive autocorrelation leads to a larger variance than would be obtained with independent samples, several measures have been proposed to compare the variances as an indication of the cost of being forced to work with correlated samples. One is the relative numerical efficiency (r.n.e.), defined as

$$\text{r.n.e.} = \sqrt{\frac{\sum_b (Z^{(t)} - \bar{Z})^2}{G}}/\frac{\hat{\text{Var}}(\bar{Z}_b)}{N}.$$ 

Small values indicate substantial autocorrelation in the chain. The inefficiency factor or autocorrelation time is defined as the reciprocal of the r.n.e. Large values of the inefficiency factor indicate the cost of working with correlated samples.

Knowing when an MCMC chain has converged is, at this writing, as much an art as a science. Although there are a few indicators of nonconvergence, there are no widely applicable theorems to compute the length of the burn-in period. Several indicators have been proposed, many of which are concerned with attempting to measure the amount of mixing that is taking place as the sampler runs. The justification is that a well-mixing sampler is likely to provide a good approximation to the target distribution. Although this justification may not work in all cases, in practice it seems adequate. A measure of the degree of mixing is the autocorrelation function, which is the estimated autocorrelation of the series at various lags. If the algorithm is mixing well, we would expect that the autocorrelation would drop off to zero within a few lags; in contrast, a poorly mixing algorithm may have substantial correlation at very large lags.

As an example, consider the bivariate normal sampling problem of Section 7.1. Figure 7.3 presents autocorrelations. As we saw previously, there is extremely poor mixing when $\rho = 0.999$. The autocorrelations for these cases show this clearly: they are over 0.900 even at lag 20. In contrast, the absolute values of the autocorrelations for the $\rho = 0.50$ case are less than 0.05 at lag 3.

**7.4 Further Reading and References**

Chapter 7 Robert and Casella (2004) presents the material covered in this chapter at a reasonable level. An advanced book, used as the source for much of the
research in this area, is Nummelin (1984). It is heavy going, and the pertinent material is scattered throughout the book. An excellent book that covers much of the same material is Meyn and Tweedie (1993). It takes up many applications of the theory and is somewhat difficult, but well worth the effort for those who wish to go more deeply into these topics. Two summary articles for MCMC methods in econometrics are Chib (2001) and Geweke and Keane (2001); the former's approach is closer to the one taken in this book.

Section 7.1 The Poisson model with an unknown switch point is analyzed from the Bayesian viewpoint in Carlin et al. (1992a); they consider prior distributions more general than those we have specified and take up more complex versions of the model, including multiple switch points and the hidden Markov switch point model. An alternative approach to computations for the latter is presented in Chib (1998).

If there are at least two blocks of parameters, say \((\theta_1, \theta_2)\), and we are interested in estimating the marginal density function of one of the blocks, say \(\pi(\theta_2 | y)\), it is sometimes preferable to compute it by averaging \(\pi(\theta_2 | \theta_1, y)\), which is known when a Gibbs algorithm is employed, over the realizations of \(\theta_1\) than to kernel-smooth the realizations of \(\theta_2\). A good discussion of this idea, called Rao–Blackwellization, can be found in Lancaster (2004, sec. 4.4.5–4.4.6).

Chib (1995) shows how to compute the marginal likelihood from the Gibbs output.


Section 7.3 The numerical standard error is proposed in Geweke (1989), and the inefficiency factor is used by Chib (2001). Details about estimating the autocovariance function may be found in Geyer (1992), Ripley (1987, chap. 6), and Priestley (1981, chaps. 5–7). The latter is particularly strong on window estimation and spectral methods. Ripley discusses the batch means method.

Robert and Casella (2004, chap. 12) is a clear and thorough discussion of convergence issues, and it contains many further references. Programs to compute convergence diagnostics for MCMC simulations are contained in coda at its Web site, http://www-fis.iarc.fr/coda/.

7.5 Exercises

7.1 Verify (7.1) and (7.2).

7.2 Download the data in "Birth weight I," from the StatLabs Web site, www.stat.berkeley.edu/users/statlabs/labs.html. Specify a prior distribution for the average birth weight and the variance of birth weights, and then implement Algorithm 7.3 to generate a sample from the joint distribution. Compute the n.s.e. for your output. (It would be good practice to program this algorithm yourself; otherwise the calculations can be done in BACC or one of the other packages discussed in Appendix B.)
Forecasts Based on Conditional Expectation

Suppose we are interested in forecasting the value of a variable \( Y_{t+1} \) based on a set of variables \( X_t \) observed at date \( t \). For example, we might want to forecast \( Y_{t+1} \) based on its most recent values. In this case, \( X_t \) would consist of a constant plus \( Y_t, Y_{t-1}, \ldots, \) and \( Y_{t-m} \).

Let \( Y_{t+1}^* \) denote a forecast of \( Y_{t+1} \) based on \( X_t \). To evaluate the usefulness of this forecast, we need to specify a loss function, or a summary of how concerned we are if our forecast is off by a particular amount. Very convenient results are obtained from assuming a quadratic loss function. A quadratic loss function means choosing the forecast \( Y_{t+1}^* \) so as to minimize

\[
E(Y_{t+1} - Y_{t+1}^*)^2. \tag{4.1.1}
\]

Expression [4.1.1] is known as the mean squared error associated with the forecast \( Y_{t+1}^* \), denoted

\[
MSE(Y_{t+1}^*) = E(Y_{t+1} - Y_{t+1}^*)^2.
\]

The forecast with the smallest mean squared error turns out to be the expectation of \( Y_{t+1} \) conditional on \( X_t \):

\[
Y_{t+1}^* = E(Y_{t+1}|X_t). \tag{4.1.2}
\]

To verify this claim, consider basing \( Y_{t+1}^* \) on any function \( g(X_t) \) other than the conditional expectation,

\[
Y_{t+1}^* = g(X_t). \tag{4.1.3}
\]

For this candidate forecasting rule, the MSE would be

\[
E(Y_{t+1} - g(X_t))^2 = E(Y_{t+1} - E(Y_{t+1}|X_t) + E(Y_{t+1}|X_t) - g(X_t))^2
\]

\[
= E(Y_{t+1} - E(Y_{t+1}|X_t))^2
\]

\[
+ 2E([Y_{t+1} - E(Y_{t+1}|X_t)] [E(Y_{t+1}|X_t) - g(X_t)])
\]

\[
+ E([E(Y_{t+1}|X_t) - g(X_t)])^2. \tag{4.1.4}
\]

Write the middle term on the right side of [4.1.4] as

\[
2E[\eta_{t+1}] = C, \tag{4.1.5}
\]

where

\[
\eta_{t+1} = [Y_{t+1} - E(Y_{t+1}|X_t)] [E(Y_{t+1}|X_t) - g(X_t)].
\]

Consider first the expectation of \( \eta_{t+1} \) conditional on \( X_t \). Conditional on \( X_t \), the terms \( E(Y_{t+1}|X_t) \) and \( g(X_t) \) are known constants and can be factored out of this expectation:

\[
E[\eta_{t+1}|X_t] = [E(Y_{t+1}|X_t) - g(X_t)] \times E([Y_{t+1} - E(Y_{t+1}|X_t)]|X_t)
\]

\[
= [E(Y_{t+1}|X_t) - g(X_t)] \times 0
\]

\[
= 0.
\]

By a straightforward application of the law of iterated expectations, equation [A.5.10], it follows that

\[
E[\eta_{t+1}] = E_X[E[\eta_{t+1}|X_t]] = 0.
\]

Substituting this back into [4.1.4] gives

\[
E(Y_{t+1} - g(X_t))^2 = E(Y_{t+1} - E(Y_{t+1}|X_t))^2 + E([E(Y_{t+1}|X_t) - g(X_t)])^2. \tag{4.1.6}
\]

The second term on the right side of [4.1.6] cannot be made smaller than zero, and the first term does not depend on \( g(X_t) \). The function \( g(X_t) \) that makes the mean squared error [4.1.6] as small as possible is the function that sets the second term in [4.1.6] to zero:

\[
E(Y_{t+1}|X_t) = g(X_t). \tag{4.1.7}
\]

Thus the forecast \( g(X_t) \) that minimizes the mean squared error is the conditional expectation \( E(Y_{t+1}|X_t) \), as claimed.

The MSE of this optimal forecast is

\[
E(Y_{t+1}^* - g(X_t))^2 = E(Y_{t+1}^* - E(Y_{t+1}|X_t))^2. \tag{4.1.8}
\]

The conditional expectation \( E(Y_{t+1}|X_t) \) represents the conditional population moment of the random variable \( Y_{t+1} \) and is not a function of the random variable \( Y_{t+1} \) itself. For example, if \( Y_{t+1}|X_t = N(\alpha X_t, \sigma^2) \), then \( E(Y_{t+1}|X_t) = \alpha X_t \), which does not depend on \( Y_{t+1} \).
Forecasts Based on Linear Projection

We now restrict the class of forecasts considered by requiring the forecast \( Y_{t+1} \) to be a linear function of \( X_t \):

\[
Y_{t+1} = \alpha' X_t. \tag{4.1.9}
\]

Suppose we were to find a value for \( \alpha \) such that the forecast error \( (Y_{t+1} - \alpha' X_t) \) is uncorrelated with \( X_t \):

\[
E[(Y_{t+1} - \alpha' X_t)X_t] = 0. \tag{4.1.10}
\]

If \( \tag{4.1.10} \) holds, then the forecast \( \alpha' X_t \) is called the linear projection of \( Y_{t+1} \) on \( X_t \).

The linear projection turns out to produce the smallest mean squared error among the class of all linear forecasting rules. The proof of this claim closely parallels the demonstration of the optimality of the conditional expectation among the set of all possible forecasts. Let \( g' X \), denote any arbitrary linear forecasting rule. Note that its MSE is

\[
E[(Y_{t+1} - g' X_t)^2] = E(Y_{t+1} - \alpha' X_t + \alpha' X_t - g' X_t)^2
= E(Y_{t+1} - \alpha' X_t)^2 + 2E[(Y_{t+1} - \alpha' X_t)[\alpha' X_t - g' X_t]]
+ E(\alpha' X_t - g' X_t)^2. \tag{4.1.11}
\]

As in the case of \( \tag{4.1.4} \), the middle term on the right side of \( \tag{4.1.11} \) is zero:

\[
E[(Y_{t+1} - \alpha' X_t)[\alpha' X_t - g' X_t]] = (E(Y_{t+1} - \alpha' X_t)[\alpha' X_t - g' X_t]) = 0 \tag{4.1.12}
\]

by virtue of \( \tag{4.1.11} \). Thus \( \tag{4.1.11} \) simplifies to

\[
E(Y_{t+1} - g' X_t)^2 = E(Y_{t+1} - \alpha' X_t)^2 + E(\alpha' X_t - g' X_t)^2. \tag{4.1.12}
\]

The optimal linear forecast \( g' X \), is the value that sets the second term in \( \tag{4.1.12} \) equal to zero:

\[
g' X_t = \alpha' X_t,
\]

where \( \alpha' X \) satisfies \( \tag{4.1.10} \).

For \( \alpha' X \), satisfying \( \tag{4.1.10} \), we will use the notation

\[
\hat{P}(Y_{t+1} | X_t) = \alpha' X_t, \tag{4.1.13}
\]

or

\[
\hat{Y}_{t+1} = \alpha' X_t,
\]

to indicate the linear projection of \( Y_{t+1} \) on \( X_t \). Notice that

\[
MSE[\hat{P}(Y_{t+1} | X_t)] \leq MSE[E(Y_{t+1} | X_t)],
\]

since the conditional expectation offers the best possible forecast.

For most applications a constant term will be included in the projection. We will use the symbol \( \hat{E} \) to indicate a linear projection on a vector of random variables \( X_t \) along with a constant term:

\[
\hat{E}(Y_{t+1} | X_t) = \hat{P}(Y_{t+1} | X_t, 1). \tag{4.1.14}
\]

Properties of Linear Projection

It is straightforward to use \( \tag{4.1.10} \) to calculate the projection coefficient \( \alpha \) in terms of the moments of \( Y_{t+1} \) and \( X_t \):

\[
E(Y_{t+1} X_t) = \alpha' E(X_t X_t'). \tag{4.1.15}
\]


greek letters \( \alpha' = E(Y_{t+1} X_t') [E(X_t X_t')]^{-1} \).

assuming that \( E(X_t X_t') \) is nonsingular. When \( E(X_t X_t') \) is singular, the coefficient vector \( \alpha \) is not uniquely determined by \( \tag{4.1.10} \), though the product of this vector with the explanatory variables, \( \alpha' X_t \), is uniquely determined by \( \tag{4.1.10} \).

The MSE associated with a linear projection is given by

\[
E[(Y_{t+1} - \alpha' X_t)^2] = E(Y_{t+1})^2 - 2E(\alpha' X_t Y_{t+1}) + E(\alpha' X_t \alpha' X_t'). \tag{4.1.16}
\]

Substituting \( \tag{4.1.13} \) into \( \tag{4.1.14} \) produces

\[
E(Y_{t+1} - \alpha' X_t)^2 = E(Y_{t+1})^2 - 2E(Y_{t+1} X_t') [E(X_t X_t')]^{-1} E(X_t Y_{t+1})
+ E(Y_{t+1} X_t') [E(X_t X_t')]^{-1} E(X_t Y_{t+1})
\times [E(X_t X_t')]^{-1} E(X_t' Y_{t+1})
\times [E(X_t X_t')]^{-1} E(X_t Y_{t+1}). \tag{4.1.17}
\]

Notice that if \( X_t \) includes a constant term, then the projection of \( (a Y_{t+1} + b) \) on \( X_t \) (where \( a \) and \( b \) are deterministic constants) is equal to

\[
\hat{P}(a Y_{t+1} + b | X_t) = a' \hat{P}(Y_{t+1} | X_t) + b.
\]

To see this, observe that \( a' \hat{P}(Y_{t+1} | X_t) + b \) is a linear function of \( X_t \). Moreover, the forecast error,

\[
[a Y_{t+1} + b] - [a' \hat{P}(Y_{t+1} | X_t) + b] = a [Y_{t+1} - \hat{P}(Y_{t+1} | X_t)],
\]

is uncorrelated with \( X_t \), as required of a linear projection.

Linear Projection and Ordinary Least Squares Regression

Linear projection is closely related to ordinary least squares regression. This subsection discusses the relationship between the two concepts.

A linear regression model relates an observation on \( y_{t+1} \) to \( x_t \):

\[
y_{t+1} = \beta' x_t + \epsilon_t. \tag{4.1.18}
\]

Given a sample of \( T \) observations on \( y \) and \( x \), the sample sum of squared residuals is defined as

\[
\sum_{t=1}^{T} (y_{t+1} - \beta' x_t)^2. \tag{4.1.19}
\]

The value of \( \beta \) that minimizes \( \tag{4.1.17} \), denoted \( b \), is the ordinary least squares (OLS) estimate of \( \beta \). The formula for \( b \) turns out to be

\[
b = \left( \sum x_t x_t' \right)^{-1} \left( \sum x_t y_{t+1} \right). \tag{4.1.20}
\]

1 If \( E(X_t X_t') \) is singular, there exists a nonzero vector \( e \) such that \( E(X_t X_t') e = E(e X_t X_t') = 0 \), so that some linear combination \( \alpha' X_t \) is equal to zero for all realizations. For example, if \( X_t \) consists of two random variables, the second variable must be a rescaled version of the first, \( X_2 = c X_1 \). One could simply drop the redundant variable from such a system and calculate the linear projection of \( Y_{t+1} \) on \( X_t' \), where \( X_t' \) is a vector consisting of the nonredundant elements of \( X_t \). This linear projection \( \alpha' X_t' \) can be uniquely calculated from \( \tag{4.1.3} \) with \( X_t \) in \( \tag{4.1.3} \) replaced by \( X_t' \). Any linear combination of the original variables \( \alpha' X_t \), satisfying \( \tag{4.1.10} \) represents this same random variable; that is, \( \alpha' X_t = \alpha' X_t' \) for all values of \( \alpha \) consistent with \( \tag{4.1.10} \).
which equivalently can be written

\[ b = \left[ \frac{1}{T} \sum_{t=1}^{T} x_t' \right]^{-1} \left[ \frac{1}{T} \sum_{t=1}^{T} x_t y_{t+1} \right]. \]  

[4.1.19]

Comparing the OLS coefficient estimate \( b \) in equation 4.1.19 with the linear projection coefficient \( \alpha \) in equation 4.1.13, we see that \( b \) is constructed from the sample moments \( (1/T) \sum_{t=1}^{T} x_t' x_t \) and \((1/T) \sum_{t=1}^{T} x_t y_{t+1} \) while \( \alpha \) is constructed from population moments \( E(x' x) \) and \( E(x' y_{t+1}) \). Thus OLS regression is a summary of the particular sample observations \( (x_1, x_2, \ldots, x_T) \) and \( (y_2, y_3, \ldots, y_{T+1}) \) whereas linear projection is a summary of the population characteristics of the stochastic process \( \{x_t, y_{t+1}\}_{t=1}^{\infty} \).

Although linear projection describes population moments and ordinary least squares describes sample moments, there is a formal mathematical sense in which the two operations are the same. Appendix 4.A to this chapter discusses this parallel and shows how the formulas for an OLS regression can be viewed as a special case of the formulas for a linear projection.

Notice that if the stochastic process \( \{x_t, y_{t+1}\} \) is covariance-stationary and ergodic for second moments, then the sample moments will converge to the population moments as the sample size \( T \) goes to infinity:

\[ (1/T) \sum_{t=1}^{T} x_t x_t' \to E(x' x) \]

\[ (1/T) \sum_{t=1}^{T} x_t y_{t+1} \to E(x' y_{t+1}). \]

implying

\[ b \to \alpha. \]  

[4.1.20]

Thus OLS regression of \( y_{t+1} \) on \( x_t \) yields a consistent estimate of the linear projection coefficient \( \alpha \). Note that this result requires only that the process be ergodic for second moments. By contrast, structural econometric analysis requires much stronger assumptions about the relation between \( x \) and \( y \). The difference arises because structural analysis seeks the effect of \( x \) on \( y \). In structural analysis, changes in \( x \) are associated with a particular structural event such as a change in Federal Reserve policy, and the objective is to evaluate the consequences for \( y \). Where that is the objective, it is very important to consider the nature of the correlation between \( x \) and \( y \) before relying on OLS estimates. In the case of linear projection, however, the only concern is forecasting, for which it does not matter whether it is \( x \) that causes \( y \) or \( y \) that causes \( x \). Their observed historical comovements (as summarized by \( E(x' y_{t+1}) \)) are all that is needed for calculating a forecast. Result 4.1.20 shows that ordinary least squares regression provides a sound basis for forecasting under very mild assumptions.

One possible violation of these assumptions should nevertheless be noted. Result 4.1.20 was derived by assuming a covariance-stationary, ergodic process. However, the moments of the data may have changed over time in fundamental ways, or the future environment may be different from that in the past. Where this is the case, ordinary least squares may be undesirable, and better forecasts can emerge from careful structural analysis.

### Forecasts Based on an Infinite Number of Observations

#### Forecasting Based on Lagged \( \epsilon \)'s

Consider a process with an MA(\( \infty \)) representation

\[ (y_t - \mu) = \psi(L) \epsilon_t, \]  

[4.2.1]

with \( \epsilon_t \) white noise and

\[ \psi(L) = \sum_{j=0}^{\infty} \psi_j L^j \]

\[ \psi_0 = 1 \]

\[ \sum_{j=0}^{\infty} |\psi_j| < \infty. \]  

[4.2.2]

Suppose that we have an infinite number of observations on \( y \) through date \( t \), \( \{\epsilon_t, \epsilon_{t-1}, \epsilon_{t-2}, \ldots\} \), and further know the values of \( \mu \) and \( \{\psi_0, \psi_1, \ldots\} \). Say we want to forecast the value of \( y_{t+1} \) that is, the value that \( y \) will take on \( s \) periods from now. Note that 4.2.1 implies

\[ y_{t+1} = \mu + \psi_1 \epsilon_t + \psi_2 \epsilon_{t-1} + \cdots + \psi_{t+1} \epsilon_{t+1} + \psi_{t+2} \epsilon_{t+2} + \cdots \]  

[4.2.3]

The optimal linear forecast takes the form

\[ \hat{y}_{t+1} = \mu + \psi_1 \epsilon_t + \psi_2 \epsilon_{t-1} + \psi_3 \epsilon_{t-2} + \cdots \]  

[4.2.4]

#### 4.2. Forecasts Based on an Infinite Number of Observations
That is, the unknown future $e'$s are set to their expected value of zero. The error associated with this forecast is

$$Y_{t+1} - \hat{E}[Y_{t+1} | \epsilon_t, \epsilon_{t-1}, \ldots] = e_{t+1} + \psi_1 e_{t+2} + \psi_2 e_{t+3} + \cdots.$$  \[4.2.5\]

In order for [4.2.4] to be the optimal linear forecast, condition [4.1.10] requires the forecast error to have mean zero and to be uncorrelated with $e_t, e_{t-1}, \ldots$. It is readily confirmed that the error in [4.2.5] has these properties, so [4.2.4] must indeed be the linear projection, as claimed. The mean squared error associated with this forecast is

$$E(Y_{t+1} - \hat{E}[Y_{t+1} | \epsilon_t, \epsilon_{t-1}, \ldots])^2 = (1 + \psi_1^2 + \psi_2^2 + \cdots + \psi_q^2) \sigma^2. \quad [4.2.6]$$

For example, for an $MA(q)$ process,

$$\phi(L) = 1 - \theta_1 L - \theta_2 L^2 - \cdots - \theta_q L^q,$$

the optimal linear forecast is

$$\hat{E}[Y_{t+1} | \epsilon_t, \epsilon_{t-1}, \ldots] = \begin{cases} \mu + \theta_1 \epsilon_t + \theta_2 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} & \text{for } s = 1, 2, \ldots, q \\ \mu & \text{for } s = q + 1, q + 2, \ldots. \end{cases} \quad [4.2.7]$$

The MSE is

$$\sigma^2 (1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2) \sigma^2 \quad \text{for } s = 1, 2, 3, \ldots, q,$$

$$\sigma^2 (1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2) \sigma^2 \quad \text{for } s = q + 1, q + 2, \ldots. \quad [4.2.8]$$

The MSE increases with the forecast horizon $s$ up until $s = q$. If we try to forecast an $MA(q)$ farther than $q$ periods into the future, the forecast is simply the unconditional mean of the series $E(Y_t | \mu)$, and the MSE is the unconditional variance of the series $\text{Var}(Y_t) = (1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2) \sigma^2$.

These properties also characterize the $MA(\infty)$ case as the forecast horizon $s$ goes to infinity. It is straightforward to establish from [4.2.2] that as $s \to \infty$, the forecast in [4.2.4] converges in mean square to $\mu$, the unconditional mean. The MSE [4.2.6] likewise converges to $\sigma^2 \sum_{j=0}^{\infty} \theta_j^2$, which is the unconditional variance of the $MA(\infty)$ process [4.2.1].

A compact lag operator expression for the forecast in [4.2.4] is sometimes used. Consider taking the polynomial $\phi(L)$ and dividing by $L$: \[4.2.10\]

$$\phi(L) L^{-1} = L^{-1} + \psi_1 L^{-2} + \psi_2 L^{-3} + \cdots + \psi_q L^{-1} + \psi_0 L^{0}$$

$$+ \psi_{q+1} L^{1} + \psi_{q+2} L^{2} + \cdots.$$ 

The annihilation operator $^+\phi(L)$ (indicated by $[\phi(L)]^+$) replaces negative powers of $L$ by zero; for example,

$$[\phi(L) L^{-1}]^+ = \psi_1 L^{-1} + \psi_2 L^{2} + \cdots. \quad [4.2.8]$$

Comparing [4.2.8] with [4.2.4], the optimal forecast could be written in lag operator notation as

$$\hat{E}[Y_{t+1} | \epsilon_t, \epsilon_{t-1}, \ldots] = \mu + \left[\frac{\phi(L)}{L} \right]^+ \epsilon_t. \quad [4.2.9]$$

\[3\]This discussion of forecasting based on the annihilation operator is similar to that in Sargent (1987).

Forecasting Based on Lagged $Y$'s

The previous forecasts were based on the assumption that $e_t$ is observed directly. In the usual forecasting situation, we actually have observations on lagged $Y$'s, not lagged $e$'s. Suppose that the process [4.2.1] has an $AR(\infty)$ representation given by

$$\eta(L)(Y_t - \mu) = e_t, \quad [4.2.10]$$

where $\eta(L) = \sum_{j=0}^{\infty} \eta_j L^j$, $\eta_0 = 1$, and $\sum_{j=0}^{\infty} |\eta_j| < \infty$. Suppose further that the $AR$ polynomial $\eta(L)$ and the MA polynomial $\phi(L)$ are related by

$$\eta(L) = (\phi(L))^{-1}. \quad [4.2.11]$$

A covariance-stationary $AR(p)$ model of the form

$$(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p)(Y_t - \mu) = e_t, \quad [4.2.12]$$

or, more compactly,

$$\phi(L)(Y_t - \mu) = e_t,$$

clearly satisfies these requirements, with $\eta(L) = \phi(L)$ and $\psi(L) = (\phi(L))^{-1}$. An $MA(q)$ process

$$Y_t - \mu = (1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q)e_t, \quad [4.2.13]$$

or

$$Y_t - \mu = \theta(L)e_t,$$

is also of this form, with $\phi(L) = \theta(L)$ and $\eta(L) = (\theta(L))^{-1}$, provided that [4.2.13] is based on the invertible representation. With a noninvertible $MA(q)$, the roots must first be flipped as described in Section 3.7 before applying the formulas given in this section. An $ARMA(p, q)$ process also satisfies [4.2.10] and [4.2.11] with $\phi(L) = \theta(L)\phi(L)$, provided that the autoregressive operator $\theta(L)$ satisfies the stationarity condition (roots of $\theta(z) = 0$ lie outside the unit circle) and that the moving average operator $\theta(L)$ satisfies the invertibility condition (roots of $\theta(z) = 0$ lie outside the unit circle).

Where the restrictions associated with [4.2.10] and [4.2.11] are satisfied, observations on $\{Y_t, Y_{t-1}, \ldots\}$ will be sufficient to construct $\{e_t, e_{t-1}, \ldots\}$. For example, for an $AR(1)$ process [4.2.10] would be

$$(1 - \phi L)(Y_t - \mu) = e_t. \quad [4.2.14]$$

Thus, given $\phi$ and $\mu$ and observation of $Y_t$ and $Y_{t-1}$, the value of $e_t$ can be constructed from

$$e_t = (Y_t - \mu) - \phi(Y_{t-1} - \mu).$$

For an $MA(1)$ process written in invertible form, [4.2.10] would be

$$(1 + \theta L)(Y_t - \mu) = e_t.$$

Given an infinite number of observations on $Y$, we could construct $e$ from

$$e_t = (Y_t - \mu) - \theta(Y_{t-1} - \mu) + \theta(Y_{t-2} - \mu)$$

$$(1 - \theta L)(Y_t - \mu) = e_t. \quad [4.2.15]$$

Under these conditions, [4.2.10] can be substituted into [4.2.9] to obtain the forecast of $Y_{t+s}$ as a function of lagged $Y$'s:

$$\hat{E}[Y_{t+s} | Y_t, Y_{t-1}, \ldots] = \mu + \left[\frac{\phi(L)}{L} \right]^+ \eta(L)(Y_t - \mu).$$

4.2. Forecasts Based on an Infinite Number of Observations
or, using \([4.2.11]\),
\[
E[Y_{i+1}|Y_i, Y_{i-1}, \ldots] = \mu + \left[ \frac{\phi(L)}{L} \right] \frac{1}{\psi(L)} (Y_i - \mu) \tag{4.2.16}
\]

Equation \([4.2.16]\) is known as the Wiener-Kolmogorov prediction formula. Several examples of using this forecasting rule follow.

**Forecasting an AR(1) Process**

For the covariance-stationary AR(1) process \([4.2.14]\), we have
\[
\phi(L) = 1/(1 - \phi L) = 1 + \phi L + \phi^2 L^2 + \phi^3 L^3 + \cdots \tag{4.2.17}
\]
and
\[
\left[ \frac{\psi(L)}{L} \right] = \phi^0 + \phi^1 L + \phi^2 L^2 + \cdots = \phi^0 (1 - \phi L) \tag{4.2.18}
\]

Substituting \([4.2.18]\) into \([4.2.16]\) yields the optimal linear \(s\)-period-ahead forecast for a stationary AR(1) process:
\[
E[Y_{i+s}|Y_i, Y_{i-1}, \ldots] = \mu + \frac{\phi^s}{1 - \phi} \frac{1}{\psi(L)} (Y_i - \mu) \tag{4.2.19}
\]

The forecast decays geometrically from \((Y_i - \mu)\) toward \(\mu\) as the forecast horizon \(s\) increases. From \([4.2.17]\), the moving average weight \(\psi_i\) is given by \(\phi^i\), so from \([4.2.6]\), the mean squared \(s\)-period-ahead forecast error is
\[
[1 + \phi^2 + \phi^3 + \cdots + \phi^{2s-1}] \sigma^2.
\]

Notice that this grows with \(s\) and asymptotically approaches \(\sigma^2/(1 - \phi^2)\), the unconditional variance of \(Y\).

**Forecasting an AR(p) Process**

Next consider forecasting the stationary AR(p) process \([4.2.12]\). The Wiener-Kolmogorov formula in \([4.2.16]\) essentially expresses the value of \((Y_{i+s} - \mu)\) in terms of initial values \((Y_i - \mu), (Y_{i-1} - \mu), \ldots\) and subsequent values of \(\{\epsilon_{i+1}, \epsilon_{i+2}, \ldots, \epsilon_{i+s}\}\) and then drops the terms involving future \(\epsilon\)'s. An expression of this form was provided by equation \([1.2.26]\), which described the value of a variable subject to a pth-order difference equation in terms of initial conditions and subsequent shocks:
\[
Y_{i+s} - \mu = f_{i+s}^0 (Y_i - \mu) + f_{i+s}^1 (Y_{i-1} - \mu) + \cdots + f_{i+s}^{p-1} (Y_{i+p-1} - \mu)
\]
\[
+ \epsilon_{i+s} + \phi_1 \epsilon_{i+s-1} + \phi_2 \epsilon_{i+s-2} + \cdots + \phi_{p-1} \epsilon_{i+1},
\]

where
\[
\psi_i = f_{i+1}^0.
\]

Recall that \(f_{i+s}^0\) denotes the \((1, 1)\) element of \(F\), \(f_{i+s}^1\) denotes the \((1, 2)\) element of \(F\), and so on, where \(F\) is the following \((p \times p)\) matrix:
\[
F = \begin{bmatrix}
\phi_1 & \phi_2 & \cdots & \phi_{p-1} & \phi_p \\
1 & 0 & 0 & \cdots & 0 \\
0 & 1 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 1
\end{bmatrix}
\]

The optimal \(s\)-period-ahead forecast is thus
\[
\hat{Y}_{i+s} = \mu + f_{i+s}^0 (Y_i - \mu) + f_{i+s}^1 (Y_{i-1} - \mu) + \cdots + f_{i+s}^{p-1} (Y_{i+p-1} - \mu).
\]

Notice that for any forecast horizon \(s\) the optimal forecast is a constant plus a linear function of \((Y_n, Y_{n-1}, \ldots, Y_{n+p-1})\). The associated forecast error is
\[
Y_{i+s} - \hat{Y}_{i+s} = \epsilon_{i+s} + \phi_1 \epsilon_{i+s-1} + \phi_2 \epsilon_{i+s-2} + \cdots + \phi_{p-1} \epsilon_{i+1}.
\]

The easiest way to calculate the forecast in \([4.2.22]\) is through a simple recursion. This recursion can be deduced independently from a principle known as the law of iterated projections, which will be proved formally in Section 4.5. Suppose that at date \(t\) we wanted to make a one-period-ahead forecast of \(Y_{t+1}\). The optimal forecast is clearly
\[
\hat{Y}_{t+1} - \mu = \phi_1 (Y_t - \mu) + \phi_2 (Y_{t-1} - \mu) + \cdots + \phi_p (Y_{t-p+1} - \mu).
\]

Consider next a two-period-ahead forecast. Suppose that at date \(t + 1\) we were to make a one-period-ahead forecast of \(Y_{t+2}\). Replacing \(t\) with \(t + 1\) in \([4.2.24]\) gives the optimal forecast as
\[
\hat{Y}_{t+2} - \mu = \phi_1 (Y_{t+1} - \mu) + \phi_2 (Y_t - \mu) + \cdots + \phi_p (Y_{t-p+2} - \mu).
\]

The law of iterated projections asserts that if this date \(t + 1\) forecast of \(Y_{t+2}\) is projected on date \(t\) information, the result is the date \(t\) forecast of \(Y_{t+2}\). At date \(t\) the values \(Y_n, Y_{n-1}, \ldots, Y_{n+p-2}\) in \([4.2.25]\) are known. Thus,
\[
\hat{Y}_{t+p} - \mu = \phi_1 (Y_{t+p-1} - \mu) + \phi_2 (Y_{t+p-2} - \mu) + \cdots + \phi_p (Y_{t+p-2} - \mu).
\]

Substituting \([4.2.24]\) into \([4.2.26]\) then yields the two-period-ahead forecast for an AR(p) process:
\[
\hat{Y}_{t+2} - \mu = \phi_1 (Y_{t+1} - \mu) + \phi_2 (Y_t - \mu) + \cdots + \phi_p (Y_{t+p-1} - \mu)
\]
\[
+ \phi_1 (Y_{t+p-2} - \mu) + \phi_2 (Y_{t+p-3} - \mu) + \cdots + \phi_p (Y_{t+p-2} - \mu)
\]
\[
= (\phi_1 + \phi_2) (Y_{t+1} - \mu) + (\phi_2 + \phi_3) (Y_t - \mu) + \cdots + (\phi_{p-1} + \phi_p) (Y_{t+p-2} - \mu) + \phi_p (Y_{t+p-1} - \mu).
\]

The two-period-ahead forecasts of an AR(p) process can be obtained by iterating
\[
\hat{Y}_{t+s} - \mu = \phi_1 (Y_{t+s-1} - \mu) + \phi_2 (Y_{t+s-2} - \mu) + \cdots + \phi_p (Y_{t+s-p} - \mu).
\]

4.2. Forecasts Based on an Infinite Number of Observations

81
for \( j = 1, 2, \ldots, s \) where
\[
\hat{Y}_t = Y_t \quad \text{for } t = t.
\]

**Forecasting an MA(1) Process**

Next consider an invertible MA(1) representation,
\[
Y_t - \mu = (1 + \theta L)e_t, \quad [4.2.28]
\]
with \(|\theta| < 1\). Replacing \(\phi(L)\) in the Wiener-Kolmogorov formula [4.2.16] with \((1 + \theta L)\) gives
\[
\hat{Y}_{t+1} = \mu + \frac{1 + \theta L}{L^*} \frac{1}{1 + \theta L} (Y_t - \mu). \quad [4.2.29]
\]

To forecast an MA(1) process one period into the future \((s = 1)\),
\[
\left[ \frac{1 + \theta L}{L^*} \right] = \theta, \quad [4.2.30]
\]
and so
\[
\hat{Y}_{t+1} = \mu + \frac{\theta}{1 + \theta L} (Y_t - \mu)
\]
\[
= \mu + \theta(Y_t - \mu) - \theta^2(Y_{t-1} - \mu) + \theta^3(Y_{t-2} - \mu) - \cdots. \quad [4.2.31]
\]

It is sometimes useful to write [4.2.28] as
\[
e_t = \frac{1}{1 + \theta L} (Y_t - \mu)
\]
and view \(e_t\) as the outcome of an infinite recursion,
\[
\hat{e}_s = (Y_s - \mu) - \theta \hat{e}_{s-1}. \quad [4.2.32]
\]

The one-period-ahead forecast [4.2.30] could then be written as
\[
\hat{Y}_{t+1} = \mu + \theta e_t. \quad [4.2.33]
\]

Equation [4.2.31] is in fact an exact characterization of \(e_n\), deduced from simple rearrangement of [4.2.28]. The “hat” notation \((\hat{e}_s)\) is introduced at this point in anticipation of the approximations to \(e_t\) that will be introduced in the following section and substituted into [4.2.31] and [4.2.32].

To forecast an MA(1) process for \( s = 2, 3, \ldots \) periods into the future,
\[
\left[ \frac{1 + \theta L}{L^*} \right] = 0 \quad \text{for } s = 2, 3, \ldots;
\]\nand so, from [4.2.29],
\[
\hat{Y}_{t+s} = \mu \quad \text{for } s = 2, 3, \ldots. \quad [4.2.33]
\]

**Forecasting an MA(q) Process**

For an invertible MA(q) process,
\[
(Y_t - \mu) = (1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q)e_t,
\]
the forecast [4.2.16] becomes
\[
\hat{Y}_{t+s} = \mu + \frac{1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q}{L^*} \frac{1}{1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q} (Y_t - \mu). \quad [4.2.34]
\]

Now
\[
\frac{1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q}{L^*} = \begin{cases}
\theta_1 + \theta_1 \theta_2 L^2 + \cdots + \theta_q L^{q-1} & \text{for } s = 1, 2, \ldots, q \\
0 & \text{for } s = q + 1, q + 2, \ldots.
\end{cases}
\]

Thus, for horizons of \( s = 1, 2, \ldots, q \), the forecast is given by
\[
\hat{Y}_{t+s} = \mu + (\theta_1 + \theta_2 L^2 + \cdots + \theta_q L^{q-1}) \hat{e}_s, \quad [4.2.35]
\]
where \(\hat{e}_s\) can be characterized by the recursion
\[
\hat{e}_s = (Y_s - \mu) - \theta_1 \hat{e}_{s-1} - \theta_2 \hat{e}_{s-2} - \cdots - \theta_q \hat{e}_{s-q}. \quad [4.2.36]
\]

A forecast farther than \( q \) periods into the future is simply the unconditional mean \(\mu\).

**Forecasting an ARMA(1, 1) Process**

For an ARMA(1, 1) process
\[
(1 - \phi L)(Y_t - \mu) = (1 + \theta L)e_t,
\]
that is stationary \((|\phi| < 1)\) and invertible \((|\theta| < 1)\),
\[
\hat{Y}_{t+s} = \mu + \frac{1}{1 - \phi L} \frac{1 - \phi L^s}{1 + \theta L} (Y_t - \mu). \quad [4.2.37]
\]

Here
\[
\left[ \frac{1}{1 - \phi L} \right] = \begin{cases}
(1 - \phi L + \phi^2 L^2 + \cdots) & \text{for } L^s = (1 - \phi L + \phi^2 L^2 + \cdots) \\
(1 - \phi L + \phi^2 L^2 + \cdots) & \text{for } L^s = \phi L^s + \phi^2 L^2 + \cdots.
\end{cases}
\]

Substituting [4.2.38] into [4.2.37] gives
\[
\hat{Y}_{t+s} = \mu + \frac{\phi + \phi \phi^{-1}}{1 - \phi L} (Y_t - \mu) \quad [4.2.39]
\]

4.2. Forecasts Based on an Infinite Number of Observations
Note that for $s = 2, 3, \ldots$, the forecast [4.2.39] obeys the recursion
\[(\hat{Y}_{s+1} - \mu) = \phi(\hat{Y}_{s+2} - \mu).\]
Thus, beyond one period, the forecast decays geometrically at the rate $\phi$ toward the unconditional mean $\mu$. The one-period-ahead forecast $(s = 1)$ is given by
\[\hat{Y}_{t+1} = \mu + \frac{\phi + \theta}{1 + \theta L} (Y_t - \mu).\] [4.2.40]
This can equivalently be written
\[(\hat{Y}_{t+1} - \mu) = \phi(1 + \phi L) + \theta(1 - \phi L) (Y_t - \mu) = \phi(Y_t - \mu) + \theta \hat{\epsilon}_t.\] [4.2.41]
where
\[\hat{\epsilon}_t = \frac{(1 - \phi L)}{(1 + \theta L)} (Y_t - \mu)\]
or
\[\hat{\epsilon}_t = (Y_t - \mu) - \phi(Y_{t-1} - \mu) - \theta \hat{\epsilon}_{t-1} = Y_t - \hat{Y}_{t-1}.\] [4.2.42]

### 4.3. Forecasts Based on a Finite Number of Observations

The formulas in the preceding section assumed that we had an infinite number of past observations on $Y_t$, $(Y_t, Y_{t-1}, \ldots)$, and knew with certainty population parameters such as $\mu$, $\phi$, and $\theta$. This section continues to assume that population parameters are known with certainty, but develops forecasts based on a finite number of observations $(Y_n, Y_{n-1}, \ldots, Y_{n-m+1})$.

For forecasting an $AR(p)$ process, an optimal $s$-period-ahead linear forecast based on an infinite number of observations $(Y_n, Y_{n-1}, \ldots)$ in fact makes use of only the $p$ most recent values $(Y_n, Y_{n-1}, \ldots, Y_{n-p})$. For an MA or $ARMA$ process, however, we would in principle require all of the historical values of $Y$ in order to implement the formulas of the preceding section.

### Approximations to Optimal Forecasts

One approach to forecasting based on a finite number of observations is to act as if presample $\epsilon$'s were all equal to zero. The idea is thus to use the approximation
\[\hat{E}(Y_{t+s}) = \hat{E}(Y_{t+s}|Y_{t}, Y_{t-1}, \ldots) \ni \hat{E}(Y_{t+s}|Y_{t}, Y_{t-1}, \ldots, Y_{t-m+1}, \epsilon_{t-m} = 0, \epsilon_{t-m-1} = 0, \ldots).\] [4.3.1]

For example, consider forecasting an $MA(q)$ process. The recursion [4.2.36] can be started by setting
\[\hat{\epsilon}_{t-m} = \hat{\epsilon}_{t-m-1} = \cdots = \hat{\epsilon}_{t-m-q+1} = 0\] [4.3.2]
and then iterating on [4.2.36] to generate $\hat{\epsilon}_{t-m+1}, \hat{\epsilon}_{t-m+2}, \ldots, \hat{\epsilon}_t$. These calculations produce
\[\hat{\epsilon}_{t-m+1} = (Y_{t-m+1} - \mu),\]
\[\hat{\epsilon}_{t-m+2} = (Y_{t-m+2} - \mu) - \theta \hat{\epsilon}_{t-m+1},\]
\[\hat{\epsilon}_{t-m+3} = (Y_{t-m+3} - \mu) - \theta \hat{\epsilon}_{t-m+2} - \theta^2 \hat{\epsilon}_{t-m+1},\]
and so on. The resulting values for $(\hat{\epsilon}_t, \hat{\epsilon}_{t-1}, \ldots, \hat{\epsilon}_{t-q+1})$ are then substituted directly into [4.2.35] to produce the forecast [4.3.1]. For example, for $s = q = 1$, the forecast would be
\[\hat{Y}_{t+1} = \mu + \theta(Y_t - \mu) - \theta^2(Y_{t-1} - \mu) + \theta^3(Y_{t-2} - \mu) - \cdots + (1 - \theta + \theta^2 - \cdots - \theta^{m-1}) \theta^m (Y_{t-m+1} - \mu),\] [4.3.3]
which is to be used as an approximation to the $AR(\infty)$ forecast,
\[\mu + \theta(Y_t - \mu) - \theta^2(Y_{t-1} - \mu) + \theta^3(Y_{t-2} - \mu) - \cdots,\] [4.3.4]

For $m$ large and $|\theta|$ small, this clearly gives an excellent approximation. For $|\theta|$ closer to unity, the approximation may be poorer. Note that if the moving average operator is noninvertible, the forecast [4.3.1] is inappropriate and should not be used.

#### 4.3. Forecasts Based on a Finite Number of Observations

84 Chapter 4 | Forecasting
Exact Finite-Sample Forecasts

An alternative approach is to calculate the exact projection of \( Y_{t+1} \) on its \( m \) most recent values. Let

\[
X_m = \begin{bmatrix}
1 \\
y_t \\
y_{t-1} \\
\vdots \\
y_{t-m+1}
\end{bmatrix}
\]

We thus seek a linear forecast of the form

\[
\alpha^{(m)} X_m = \alpha_0^{(m)} + \alpha_1^{(m)} Y_t + \cdots + \alpha_{m-1}^{(m)} Y_{t-m+1}.
\]

The coefficient relating \( Y_{t+1} \) to \( Y_t \) in a projection of \( Y_{t+1} \) on the \( m \) most recent values of \( Y \) is denoted \( \alpha^{(m)} \) in [4.3.5]. This will in general be different from the coefficient relating \( Y_{t+1} \) to \( Y_t \) in a projection of \( Y_{t+1} \) on the \( m+1 \) most recent values of \( Y \); the latter coefficient would be denoted \( \alpha^{(m+1)} \).

If \( Y \) is covariance-stationary, then \( E(Y_t Y_{t-h}) = \gamma_h + \mu^2 \). Setting \( X_t = (1, Y_t, Y_{t-1}, \ldots, Y_{t-m+1})' \) in [4.1.13] implies

\[
\alpha^{(m)} = \begin{bmatrix}
\alpha_0^{(m)} \\
\alpha_1^{(m)} \\
\vdots \\
\alpha_m^{(m)}
\end{bmatrix}
= \begin{bmatrix}
1 \\
\gamma_0 + \mu^2 \\
\gamma_1 + \mu^2 \\
\gamma_2 + \mu^2 \\
\vdots \\
\gamma_{m-1} + \mu^2 \\
\gamma_m + \mu^2
\end{bmatrix}^{-1}
\times
\begin{bmatrix}
\mu \\
\gamma_0 \\
\gamma_1 \\
\gamma_2 \\
\vdots \\
\gamma_{m-1} \\
\gamma_m
\end{bmatrix}.
\]

When a constant term is included in \( X_t \), it is more convenient to express variables in deviations from the mean. Then we could calculate the projection of \( (Y_{t+1} - \mu) \) on \( X_t = [(Y_t - \mu), (Y_{t-1} - \mu), \ldots, (Y_{t-m+1} - \mu)]' \):

\[
\hat{Y}_{t+1} = \mu = \alpha_1^{(m)} (Y_t - \mu) + \alpha_2^{(m)} (Y_{t-1} - \mu) + \cdots + \alpha_m^{(m)} (Y_{t-m+1} - \mu).
\]

For this definition of \( X_t \), the coefficients can be calculated directly from [4.1.13] to be

\[
\begin{bmatrix}
\alpha_1^{(m)} \\
\alpha_2^{(m)} \\
\vdots \\
\alpha_m^{(m)}
\end{bmatrix} = \begin{bmatrix}
\gamma_0 \\
\gamma_1 \\
\gamma_2 \\
\vdots \\
\gamma_{m-1} \\
\gamma_m
\end{bmatrix}^{-1}
\begin{bmatrix}
\gamma_0 \\
\gamma_1 \\
\gamma_2 \\
\vdots \\
\gamma_{m-1} \\
\gamma_m
\end{bmatrix}.
\]

To generate an \( h \)-period-ahead forecast \( \hat{Y}_{t+h} \), we would use

\[
\hat{Y}_{t+h} = \mu + \alpha_1^{(m)} (Y_t - \mu) + \alpha_2^{(m)} (Y_{t-1} - \mu) + \cdots + \alpha_m^{(m)} (Y_{t-m+1} - \mu),
\]

where

\[
\begin{bmatrix}
\alpha_0^{(m)} \\
\alpha_1^{(m)} \\
\alpha_2^{(m)} \\
\vdots \\
\alpha_m^{(m)}
\end{bmatrix} = \begin{bmatrix}
\gamma_0 \\
\gamma_1 \\
\gamma_2 \\
\vdots \\
\gamma_{m-1} \\
\gamma_m
\end{bmatrix}^{-1}
\begin{bmatrix}
\gamma_0 \\
\gamma_1 \\
\gamma_2 \\
\vdots \\
\gamma_{m-1} \\
\gamma_m
\end{bmatrix}.
\]

Using expressions such as [4.3.8] requires inverting an \((m \times m)\) matrix. Several algorithms can be used to evaluate [4.3.8] using relatively simple calculations. One approach is based on the Kalman filter discussed in Chapter 13, which can generate exact finite-sample forecasts for a broad class of processes including any ARMA specification. A second approach is based on triangular factorization of the matrix in [4.3.8]. This second approach is developed in the next two sections. This approach will prove helpful for the immediate question of calculating finite-sample forecasts and is also a useful device for establishing a number of later results.

### 4.4. The Triangular Factorization of a Positive Definite Symmetric Matrix

Any positive definite symmetric \((n \times n)\) matrix \( \Omega \) has a unique representation of the form

\[
\Omega = ADA',
\]

where \( A \) is a lower triangular matrix with 1 along the principal diagonal,

\[
A = \begin{bmatrix}
1 & 0 & 0 & \cdots & 0 \\
ad_{21} & 1 & 0 & \cdots & 0 \\
ad_{31} & ad_{32} & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
ad_{m1} & ad_{m2} & ad_{m3} & \cdots & 1
\end{bmatrix}
\]

and \( D \) is a diagonal matrix,

\[
D = \begin{bmatrix}
da_{11} & 0 & 0 & \cdots & 0 \\
0 & d_{22} & 0 & \cdots & 0 \\
0 & 0 & d_{33} & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & d_{mm}
\end{bmatrix}
\]

where \( d_{ii} > 0 \) for all \( i \). This is known as the triangular factorization of \( \Omega \).
and similarly, \[4.7.23\) could be multiplied by \((1 - \pi L)\):
\[
(1 - \pi L)(1 - \rho L)W_t = (1 - \pi L)v_t. \tag{4.7.25}
\]

Adding \[4.7.24\) to \[4.7.25\) produces
\[
(1 - \rho L)(1 - \pi L)(X_t + W_t) = (1 - \rho L)u_t + (1 - \pi L)v_t. \tag{4.7.26}
\]

From \[4.7.21\), the right side of \[4.7.26\) has an MA(1) representation. Thus, we could write
\[
(1 - \phi_1 L - \phi_2 L^2)Y_t = (1 + \theta L)e_t,
\]

where
\[
(1 - \phi_1 L - \phi_2 L^2) = (1 - \rho L)(1 - \pi L)
\]

and
\[
(1 + \theta L)e_t = (1 - \rho L)u_t + (1 - \pi L)v_t.
\]

In other words,
\[
AR(1) + AR(1) = ARMA(2, 1). \tag{4.7.27}
\]

In general, adding an \(AR(p_1)\) process
\[
\pi(L)X_t = u_t,
\]

to an \(AR(p_2)\) process with which it is uncorrelated at all leads and lags,
\[
\rho(L)W_t = v_t,
\]

produces an \(ARMA(p_1 + p_2, \max(p_1, p_2))\) process,
\[
\phi(L)Y_t = \theta(L)e_t,
\]

where
\[
\phi(L) = \pi(L)\rho(L)
\]

and
\[
\theta(L)e_t = \rho(L)u_t + \pi(L)v_t.
\]

### 4.8. Wold's Decomposition and the Box-Jenkins Modeling Philosophy

**Wold's Decomposition**

All of the covariance-stationary processes considered in Chapter 3 can be written in the form
\[
Y_t = \mu + \sum_{j=0}^{\infty} \psi_t e_{t-j} \tag{4.8.1}
\]

where \(e_t\) is the white noise error one would make in forecasting \(Y_t\), as a linear function of lagged \(Y_t\) and where \(\sum_{j=0}^{\infty} \psi_j^2 < \infty\) with \(\psi_0 = 1\). One might think that we were able to write all these processes in the form of \[4.8.1\] because the discussion was restricted to a convenient class of models. However, the following result establishes that the representation \[4.8.1\] is in fact fundamental for any covariance-stationary time series.

**Proposition 4.1:** (Wold’s decomposition). Any zero-mean covariance-stationary process \(Y_t\) can be represented in the form
\[
Y_t = \sum_{j=0}^{\infty} \psi_t e_{t-j} + \kappa_t. \tag{4.8.2}
\]

where \(\psi_0 = 1\) and \(\sum_{j=0}^{\infty} \psi_j^2 < \infty\). The term \(e_t\) is white noise and represents the error made in forecasting \(Y_t\) on the basis of a linear function of lagged \(Y_t\):
\[
e_t = Y_t - \hat{Y}_t = \sum_{j=1}^{\infty} \psi_j e_{t-j-1}, \tag{4.8.3}
\]

The value of \(\kappa_t\) is uncorrelated with \(e_{t-j}\) for any \(j\), though \(\kappa_t\) can be predicted arbitrarily well from a linear function of past values of \(Y_t\):
\[
\kappa_t = \hat{\kappa}_t = \sum_{j=1}^{\infty} \psi_j Y_{t-j}. \tag{4.8.4}
\]

The term \(\kappa_t\) is called the linearly deterministic component of \(Y_t\) while \(\Sigma_{j=0}^{\infty} \psi_j e_{t-j}\) is called the linearly indeterministic component. If \(\kappa_t = 0\), then the process is called purely linearly indeterministic.

This proposition was first proved by Wold (1938). The proposition relies on stable second moments of \(Y_t\) but makes no use of higher moments. It thus describes only optimal linear forecasts of \(Y_t\).

Finding the Wold representation in principle requires fitting an infinite number of parameters \((\psi_0, \psi_1, \ldots)\) to the data. With a finite number of observations on \((Y_1, Y_2, \ldots, Y_T)\), this will never be possible. As a practical matter, we therefore need to make some additional assumptions about the nature of \((\psi_1, \psi_2, \ldots)\). A typical assumption in Chapter 3 was that \(\phi(L)\) can be expressed as the ratio of two finite-order polynomials:
\[
\frac{\sum_{j=0}^{\infty} \psi_j L^j}{\phi(L)} = \frac{\theta(L)}{\theta(L)} = \frac{1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q}{1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p}. \tag{4.8.4}
\]

Another approach, based on the presumed “smoothness” of the population spectrum, will be explored in Chapter 6.

**The Box-Jenkins Modeling Philosophy**

Many forecasters are persuaded of the benefits of parsimony, or using as few parameters as possible. Box and Jenkins (1976) have been influential advocates of this view. They noted that in practice, analysts end up replacing the true operators \(\theta(L)\) and \(\phi(L)\) with estimates \(\hat{\theta}(L)\) and \(\hat{\phi}(L)\) based on the data. The more parameters to estimate, the more room there is to go wrong.

Although complicated models can track the data very well over the historical period for which parameters are estimated, they often perform poorly when used for out-of-sample forecasting. For example, the 1960s saw the development of a number of large macroeconomic models purporting to describe the economy using hundreds of macroeconomic variables and equations. Part of the disillusionment with such efforts was the discovery that univariate ARMA models with small values of \(p\) or \(q\) often produced better forecasts than the big models (see for example Nelson, 1972). As we shall see in later chapters, large size alone was hardly the only liability of these large-scale macroeconomic models. Even so, the claim that simpler models provide more robust forecasts has a great many believers across disciplines.

---

6See Sargent (1987, pp. 286–90) for a nice sketch of the intuition behind this result.

7For more recent pessimistic evidence about current large-scale models, see Ashley (1988).
The approach to forecasting advocated by Box and Jenkins can be broken down into four steps:

1. Transform the data, if necessary, so that the assumption of covariance-stationarity is a reasonable one.
2. Make an initial guess of small values for \( p \) and \( q \) for an ARMA\((p, q)\) model that might describe the transformed series.
3. Estimate the parameters \( \phi(L) \) and \( \theta(L) \).
4. Perform diagnostic analysis to confirm that the model is indeed consistent with the observed features of the data.

The first step, selecting a suitable transformation of the data, is discussed in Chapter 15. For now we merely remark that for economic series that grow over time, many researchers use the change in the natural logarithm of the raw data. For example, if \( X_t \) is the level of real GNP in year \( t \), then

\[
Y_t = \log X_t - \log X_{t-1} \quad [4.8.5]
\]

might be the variable that an ARMA model purports to describe.

The third and fourth steps, estimation and diagnostic testing, will be discussed in Chapters 5 and 14. Analysis of seasonal dynamics can also be an important part of step 2 of the procedure; this is briefly discussed in Section 6.4. The remainder of this section is devoted to an exposition of the second step in the Box-Jenkins procedure on nonseasonal data, namely, selecting candidate values for \( p \) and \( q \).

**Sample Autocorrelations**

An important part of this selection procedure is to form an estimate \( \hat{\rho}_j \) of the population autocorrelation \( \rho_j \). Recall that \( \rho_j \) was defined as

\[
\rho_j = \frac{\gamma_j}{\gamma_0}
\]

where

\[
\gamma_j = E(Y_t - \mu)(Y_{t-j} - \mu).
\]

A natural estimate of the population autocorrelation \( \rho_j \) is provided by the corresponding sample moments:

\[
\hat{\rho}_j = \hat{\gamma}_j/\hat{\gamma}_0,
\]

where

\[
\hat{\gamma}_j = \frac{1}{T} \sum_{t=j+1}^{T} (y_t - \bar{y})(y_{t-j} - \bar{y}) \quad \text{for } j = 0, 1, 2, \ldots, T - 1 \quad [4.8.6]
\]

\[
\bar{y} = \frac{1}{T} \sum_{t=1}^{T} y_t \quad [4.8.7]
\]

Note that even though only \( T - j \) observations are used to construct \( \hat{\gamma}_j \), the denominator in [4.8.6] is \( T \) rather than \( T - j \). Thus, for large \( j \), expression [4.8.6] shrinks the estimates toward zero, as indeed the population autocovariances go to zero as \( j \to \infty \), assuming covariance-stationarity. Also, the full sample of observations is used to construct \( \bar{y} \).

Box and Jenkins refer to this step as 'identification' of the appropriate model. We avoid Box and Jenkins's terminology, because 'identification' has a quite different meaning for econometricians.

Recall that if the data really follow an MA\((q)\) process, then \( \rho_j \) will be zero for \( j > q \). By contrast, if the data follow an AR\((p)\) process, then \( \rho_j \) will gradually decay toward zero as a mixture of exponentials or damped sinusoids. One guide for distinguishing between MA and AR representations, then, would be the decay properties of \( \rho_j \). Often, we are interested in a quick assessment of whether \( \rho_j = 0 \) for \( j = q + 1, q + 2, \ldots \). If the data were really generated by a Gaussian MA\((q)\) process, then the variance of the estimate \( \hat{\rho}_j \) could be approximated by

\[
\text{Var}(\hat{\rho}_j) \approx \frac{1}{T} \left( 1 + 2 \sum_{j=1}^{q} \rho_j^2 \right) \quad \text{for } j = q + 1, q + 2, \ldots \quad [4.8.8]
\]

Thus, in particular, if we suspect that the data were generated by Gaussian white noise, then \( \hat{\rho}_j \) for any \( j \neq 0 \) should lie between \( \pm 2/\sqrt{T} \) about 95% of the time.

In general, if there is autocorrelation in the process that generated the original data \( \{Y_t\} \), then the estimate \( \hat{\rho}_j \) will be correlated with \( \hat{\rho}_i \) for \( i \neq j \). Thus patterns in the estimated \( \hat{\rho}_j \) may represent sampling error rather than patterns in the true \( \rho_j \).

**Partial Autocorrelation**

Another useful measure is the partial autocorrelation. The \( m \)th population partial autocorrelation (denoted \( \alpha_m^{(m)} \)) is defined as the last coefficient in a linear projection of \( Y_t \) on its \( m \) most recent values (equation [4.3.7]):

\[
\hat{Y}_{t+m} - \mu = \alpha_1^{(m)}(Y_t - \mu) + \alpha_2^{(m)}(Y_{t-1} - \mu) + \cdots + \alpha_m^{(m)}(Y_{t-m+1} - \mu).
\]

We saw in equation [4.3.8] that the vector \( \alpha_m^{(m)} \) can be calculated from

\[
\begin{bmatrix}
\alpha_1^{(m)} \\
\vdots \\
\alpha_m^{(m)}
\end{bmatrix} = 
\begin{bmatrix}
\gamma_0 & \gamma_1 & \cdots & \gamma_{m-1} \\
\gamma_1 & \gamma_0 & \cdots & \gamma_{m-2} \\
\vdots & \vdots & \ddots & \vdots \\
\gamma_{m-1} & \gamma_{m-2} & \cdots & \gamma_0
\end{bmatrix}^{-1} 
\begin{bmatrix}
\gamma_1 \\
\vdots \\
\gamma_m
\end{bmatrix}
\]

Recall that if the data were really generated by an AR\((p)\) process, only the \( p \) most recent values of \( Y_t \) would be useful for forecasting. In this case, the projection coefficients on \( Y_t \)'s more than \( p \) periods in the past are zero to equal to:

\[
\alpha_m^{(m)} = 0 \quad \text{for } m = p + 1, p + 2, \ldots
\]

By contrast, if the data really were generated by an MA\((q)\) process with \( q \geq 1 \), then the partial autocorrelation \( \alpha_m^{(m)} \) asymptotically approaches zero instead of cutting off abruptly.

A natural estimate of the \( m \)th partial autocorrelation is the last coefficient in an OLS regression of \( y \) on a constant and its \( m \) most recent values:

\[
y_{t+m} = \hat{\epsilon} + \hat{\alpha}_1^{(m)}y_t + \hat{\alpha}_2^{(m)}y_{t-1} + \cdots + \hat{\alpha}_m^{(m)}y_{t-m+1} + \hat{\eta}_m,
\]

where \( \hat{\epsilon} \) denotes the OLS regression residual. If the data were really generated by an AR\((p)\) process, then the sample estimate \( \hat{\alpha}_m^{(m)} \) would have a variance around the true value \( 0 \) that could be approximated by

\[
\text{Var}(\hat{\alpha}_m^{(m)}) = 1/T \quad \text{for } m = p + 1, p + 2, \ldots
\]

*See Box and Jenkins (1976, p. 35).

*Again, see Box and Jenkins (1976, p. 35).

*Box and Jenkins (1976, p. 65).

4.8. Wold's Decomposition and the Box-Jenkins Modeling Philosophy
Moreover, if the data were really generated by an AR(p) process, then $\hat{\phi}_j^{(i)}$ and $\hat{\phi}_j^{(j)}$ would be asymptotically independent for $i, j > p$.

**Example 4.1**

We illustrate the Box-Jenkins approach with seasonally adjusted quarterly data on U.S. real GNP from 1947 through 1988. The raw data ($x_t$) were converted to log changes ($y_t$) as in [4.8.5]. Panel (a) of Figure 4.2 plots the sample autocorrelations of $y$ ($\hat{\phi}_j$ for $j = 0, 1, \ldots, 20$), while panel (b) displays the sample partial autocorrelations ($\hat{\psi}_m^{(m)}$ for $m = 0, 1, \ldots, 20$). Ninety-five percent confidence bands ($\pm 2/\sqrt{T}$) are plotted on both panels; for panel (a), these are appropriate under the null hypothesis that the data are really white noise, whereas for panel (b), these are appropriate if the data are really generated by an AR(p) process for $p$ less than $m$.

The first two autocorrelations appear nonzero, suggesting that $q = 2$ would be needed to describe these data as coming from a moving average process. On the other hand, the pattern of autocorrelations appears consistent with the simple geometric decay of an AR(1) process,

$$\rho_j = \phi^j$$

with $\phi = 0.4$. The partial autocorrelation could also be viewed as dying out after one lag, also consistent with the AR(1) hypothesis. Thus, one’s initial guess for a parsimonious model might be that GNP growth follows an AR(1) process, with MA(2) as another possibility to be considered.

**APPENDIX 4.A. Parallel Between OLS Regression and Linear Projection**

This appendix discusses the parallel between ordinary least squares regression and linear projection. This parallel is developed by introducing an artificial random variable specifically constructed so as to have population moments identical to the sample moments of a particular sample. Say that in some particular sample on which we intend to perform OLS we have observed $T$ particular values for the explanatory vector, denoted $x_1, x_2, \ldots, x_T$. Consider an artificial discrete-valued random variable $\xi$ that can take on only one of these particular $T$ values, each with probability $(1/T)$:

$$P(\xi = x_1) = 1/T$$

$$P(\xi = x_2) = 1/T$$

$$\vdots$$

$$P(\xi = x_T) = 1/T.$$

Thus $\xi$ is an artificially constructed random variable whose population probability distribution is given by the empirical distribution function of $x_t$. The population mean of the random variable $\xi$ is

$$E(\xi) = \frac{1}{T} \sum_{t=1}^{T} x_t; P(\xi = x_t) = \frac{1}{T} \sum_{t=1}^{T} x_t.$$

Thus, the population mean of $\xi$ equals the observed sample mean of the true random variable $X_t$. The population second moment of $\xi$ is

$$E(\xi^2) = \frac{1}{T} \sum_{t=1}^{T} x_t^2, \quad [4.4.1]$$

which is the sample second moment of $(x_1, x_2, \ldots, x_T)$.

We can similarly construct a second artificial variable $\omega$ that can take on one of the discrete values $(1, 2, \ldots, T)$. Suppose that the joint distribution of $\omega$ and $\xi$ is given by

$$P(\xi = x_t, \omega = y_{t+1}) = 1/T \quad \text{for } t = 1, 2, \ldots, T.$$

Then

$$E(\omega) = \frac{1}{T} \sum_{t=1}^{T} x_t y_{t+1}. \quad [4.4.2]$$

The coefficient for a linear projection of $\omega$ on $\xi$ is the value of $\alpha$ that minimizes

$$E(\omega - \alpha \xi)^2 = \frac{1}{T} \sum_{t=1}^{T} (y_{t+1} - \alpha x_t)^2. \quad [4.4.3]$$

This is algebraically the same problem as choosing $\beta$ so as to minimize [4.1.17]. Thus, ordinary least squares regression (choosing $\beta$ so as to minimize [4.1.17]) can be viewed as a special case of linear projection (choosing $\alpha$ so as to minimize [4.4.3]). The value of $\alpha$
Optimal forecast combinations under general loss functions and forecast error distributions

Graham Elliott, Allan Timmermann*

Department of Economics, University of California-San Diego, La Jolla, CA 92037-0508, USA
Accepted 23 October 2003

Abstract
Existing results on the properties and performance of forecast combinations have been derived in the context of mean squared error loss. Under this loss function empirical studies have generally found that equally-weighted combined forecasts lead to better performance than estimates of optimal forecast combination weights which in turn outperform the best individual predictions. We show that this and other results can be overturned when asymmetries in the loss function and the forecast error distribution are skew. We characterize the optimal combination weights for the most commonly used alternatives to mean squared error loss and demonstrate how the degree of asymmetry in the loss function and skewness of the underlying forecast error distribution can significantly change the optimal combination weights. We also propose estimation methods and investigate their small sample properties in simulations and in an inflation forecasting exercise.

© 2003 Elsevier B.V. All rights reserved.
JEL classification: C13; C53

Key-reords: Forecast combination; Loss function; Asymmetric loss

1. Introduction

Economic decision makers are often presented with numerous competing forecasts generated by different economic models or forecasting services with access to different information sets. A decision maker who needs a forecast could choose a single forecast from a method that has historically performed well according to a particular loss function, or some combination of the available forecasts could be considered. If a combination of forecasts proves to be superior to its individual component, s better

model should be available by pooling the information sets used to form the various forecasts. In practice all the basic information sets are often not available, so this is not a feasible approach. Combining forecasts becomes an attractive strategy in this situation and combinations have indeed proven empirically to lead to substantial improvements over the forecasts produced by the best single model.1

The majority of theoretical results on combination weights and empirical findings on forecast combination performance assume a symmetric, quadratic loss function. However, in economics and finance forecasting performance is increasingly evaluated under more general loss functions that account for asymmetries; see, e.g., Christoffersen and Diebold (1997), Diebold (2001), Granger and Newbold (1986), Granger and Pesaran (2000), Varian (1974), West et al. (1993) and Zellner (1986). For example, financial forecasting performance is often measured through the Sharpe ratio which divides mean returns over the standard deviation of returns on a portfolio constructed on the basis of forecasting information.2 Similarly, it is common to have a value at risk (VaR) objective to account for the disproportionately high costs associated with large losses. This leads to a loss function that puts particular emphasis on draws from the left tail of the forecast error distribution. Central Banks may also have asymmetric preferences which will affect their optimal policies, cf. Peel and Nabay (1998).

This paper examines forecast combinations under general loss functions and forecast error distributions. We establish in general situations the factors that are important in determining the optimal weights in population; numerical examples are then used to demonstrate the extent to which the degree of asymmetry in the loss function interacts with the skew and kurtosis of the underlying forecast errors from the individual models in determining optimal combination weights. It is shown that it is the combination of asymmetry in both loss and data that is required for weights to differ from the mean square error (MSE) optimal weights. Under elliptical symmetry, our paper shows that the linear combination weights (other than the constant term) are identical for (almost) all loss functions. This invariance result follows since for any set of forecast combination weights, the constant can be adjusted to choose the bias that optimally trades off the bias and variance of the forecast error. The weights on the forecasts are left free to minimize the variance and can thus be found as the solution to a standard quadratic optimization problem.

This result has useful practical implications. Since the forecast combination weights are the same as those under MSE loss, the estimation problem reduces to a simple two-stage procedure: least squares estimation of the optimal combination weights followed by estimation of the constant term which controls the optimal bias. This simplifies the estimation problem greatly since typical estimation procedures require

1 See, e.g., the extensive survey by Cheung (1999). More recently Chen et al. (1999), Marcellino (2002), Newbold and Harvey (2001) and Stock and Watson (1999) have found evidence that combined forecasts often outperform forecasts from the best individual model. Reasons offered for the empirical success of forecast combinations include model misspecification, changes in the underlying parameters, different modeling and estimation strategies and the use of heterogeneous information and shrinkage estimation effects. See Hendry and Clements (2002) for a discussion of these points.

2 Hence even if the mapping from forecasts to portfolio weights is linear, there will be a highly nonlinear mapping from the forecasts to the performance metric.
search methods which can be difficult if the dimensions of the parameter vector is large. An often quoted ‘folk theorem’ in the forecast combination literature is that using average as opposed to optimal weights often works better in practice. The reason is that it can be difficult to precisely estimate the optimal forecast combination weights. Equal-weighted forecast combinations may be biased but they also reduce the forecast error variance by not relying on estimated combination weights that depend on second moments of forecast errors. This finding depends, however, on the trade-off between forecast error bias and variance (as well as higher order moments in the case of asymmetric forecast errors) which in turn depends on the shape of the loss function. Consequently this ‘folk theorem’ can be overturned under asymmetric loss functions and asymmetrically distributed forecast errors.

We finally address estimation of optimal combination weights in a much more general context than the existing literature. For general differentiable loss functions such as linear we propose to estimate the forecast combination weights using M estimation. For the lin–lin loss function estimation can be cast in the context of a quantile regression problem. For asymptotic quadratic loss, we propose an iterated weighted least squares estimation approach. The performance of these and other estimators is examined in Monte Carlo simulations.

The plan of the paper is as follows. Section 2 reviews results for the standard case with mean squared error loss. Section 3 provides theoretical results for the general case with arbitrary loss function and forecast error distribution. Section 4 considers three commonly used asymmetric loss functions and Section 5 investigates forecast combinations in the context of Gaussian mixture distributions. Section 6 discusses estimation of the optimal combination weights, while Section 7 presents Monte Carlo simulation results. Section 8 provides an empirical application to inflation forecasting and Section 9 concludes.

2. Forecast combination under mean squared error loss

Suppose a decision maker is interested in forecasting some variable, yt, on the basis of an m-vector of forecasts of this variable ŷ. Each element of ŷ is determined exactly and is adapted to an expanding sequence of information sets, Ωt, which constitutes a standard filtration. Hence ŷt is adapted to Ωt, whereas yt is not. Ωt comprises Ωt−1 in addition to other variables used to predict yt. We will assume that yt and ŷt have joint distribution P(yt, ŷt) with finite first and second moments

\[
E(\begin{pmatrix} \hat{y}_t \\ y_t \end{pmatrix}) = \begin{pmatrix} \mu_y \\ \mu \end{pmatrix}
\]

and

\[
\text{Var}(\begin{pmatrix} \hat{y}_t \\ y_t \end{pmatrix}) = \begin{pmatrix} \sigma^2_y & \sigma_{y\hat{y}} \\ \sigma_{\hat{y}y} & \Sigma_{12} \end{pmatrix}.
\]  

To keep the notation simple, we suppress the dependence of the joint distribution and moments on Ωt−1, simply using P(·) and E(·) in place of P(·|Ωt−1), E(·|Ωt−1) and Var(·|Ωt−1).

Motivated by the seminal paper of Bates and Granger (1969), the forecast combination literature has studied the class of linear forecast combinations, \( \omega^\top \hat{y}_t \), where \( \omega \) is an m-vector of combination weights and \( \omega^\top \) is a scalar constant. This gives rise to a forecast error, \( \varepsilon_t \), from the combination,

\[
\varepsilon_t = y_t - \omega^\top \hat{y}_t.
\]

Under the assumed moment structure \( \varepsilon_t \), has first and second moments

\[
\begin{align*}
\mu_\varepsilon & = \mu_y - \omega^\top \mu, \\
\sigma^2_\varepsilon & = \sigma^2_y + \omega^\top \Sigma_{22} \omega - 2 \omega^\top \Sigma_{21}.
\end{align*}
\]

Assuming the loss is quadratic and symmetric in the forecast error, \( L(\varepsilon_t) = \varepsilon_t^2 \), the objective is to minimize the following expression:

\[
E((\varepsilon_t - \omega^\top \hat{y}_t)^2) = (\mu_y - \omega^\top \mu)^2 + \sigma^2_y + \omega^\top \Sigma_{22} \omega - 2 \omega^\top \Sigma_{21}.
\]

The optimal population values of the constant and the vector of combination weights, \( \omega^* \) and \( \omega_{00} \), are thus

\[
\begin{align*}
\omega^* & = \mu_y - \omega^* \mu, \\
\omega_{00} & = \Sigma_{22}^{-1} \sigma_{12}.
\end{align*}
\]

The sample analog of these weights is of course the usual least squares estimator for the outcomes regressed on a constant and the vector of forecasts (Granger and Ramanathan, 1984). This solution takes an interesting form: the vector of combination weights depends only on the variance–covariance matrix of the outcome and the predictions. Given the optimal weights, \( \omega_{00} \), the optimal constant depends on the biases in the forecasts but not directly on the variance–covariance matrix, although it will of course depend indirectly through \( \omega_{00} \). The inclusion of a constant ensures that the combined forecast is unbiased since \( \omega^* \) can be chosen such that \( \mu_y = 0 \), which is optimal under MSE loss. For any variance–covariance matrix, and hence \( \omega_{00} \), \( \omega^* \) adjusts to ensure that the combined forecast has the optimal bias.

Bayesian solutions to the forecast combination problem under quadratic loss have also been proposed. Palm and Zellner (1992) develop a general Bayesian framework for combinations where individual forecasts can be biased and the covariance matrix of the forecast errors may be unknown.

\[\text{We shall generally assume that the covariance matrix } \Sigma_{22} \text{ is invertible, but see further below for a discussion of this point.}\]
3. Forecast combination under general loss functions and forecast error error distributions

While the MSE loss function may be a useful approximation in some cases, recent interest has focused on extending existing results to allow for more general asymmetric loss functions. Tay and Wallis (2000) provide several references to this literature. Christoffersen and Diebold (1997) develop the theory for optimal bias adjustment in the case of commonly used asymmetric loss functions when the moments of the underlying error distribution vary over time, such as under volatility clustering, but maintain the assumption that (conditionally) the forecast errors follow a Gaussian distribution.

None of these results have been extended to the forecast combination literature. To fill out this gap, this section develops results for optimal forecast combinations under general loss functions and forecast error error distributions. We then explore the significance of the factors determining the optimal combination weights for the most commonly used asymmetric loss functions such as lin-lin, linx and asymmetric quadratic loss.

3.1. The loss function

We will be concerned with loss functions that only depend on the forecast error and thus take the form $L(\epsilon)$. We assume $L(\cdot)$ has the following properties (Granger, 1999):

1. $L(0) = 0$;
2. $\min_{\epsilon} L(\epsilon) = 0$, so $L(\epsilon) \geq 0$;
3. $L(\epsilon)$ is monotonic nondecreasing as $\epsilon$ moves away from zero:
   
4. $L(\epsilon_1) \geq L(\epsilon_2)$ if $\epsilon_1 > \epsilon_2 \geq 0$ or if $\epsilon_1 < \epsilon_2 \leq 0$.

In addition to these properties, the loss function may also be symmetric ($L(-\epsilon) = L(\epsilon)$), homogeneous ($L(\lambda \epsilon) = \lambda L(\epsilon)$ for some positive function $h(\cdot)$) and differentiable up to some order. Under quadratic loss forecasting and decision problems can be separated and only the conditional mean of the predicted random variable matter due to certainty equivalence. This does not hold for more general loss functions. Thus, if the objective of the analysis is to derive a decision maker’s optimal actions, in general the forecast density matters rather than simply the point forecast. However, there are cases where certainty equivalence can still be established for risk-averse or risk-seeking decision makers, cf. Whittle’s (1983) risk-sensitive optimal control method. Alternatively, one can think of situations where the forecast is the decision, as in the case of information services that provide a single forecast for multiple users (e.g., financial analysts or international organizations such as the IMF or OECD).

The decision maker’s problem is to find the optimal forecast combination weights and a constant that minimize expected loss:

$$\min_{\alpha, b} \int L(\epsilon) \, dF(\epsilon), \quad (7)$$

where $F(\epsilon)$ is the cumulative density of $\epsilon$.

Optimal forecast combination weights generally depend on the shape of both the loss function and the forecast error distribution. This can be seen from the following result:

**Proposition 1.** Suppose that (1) the expected loss is finite, (2) the loss function is analytic except for possibly at a finite number of points (occurring with probability zero) at which it is continuous, and (3) all moments of the forecast error distribution up to the highest non-zero derivative of the loss function with respect to the forecast combination weights exist.

Then the optimal combination weights solve the expression

$$\min_{\alpha, b} \left\{ L(\mu_\alpha) + \frac{1}{2} \int p^2 \, d\bar{F}(\epsilon) \right\} = \min_{\alpha, b} \left\{ L(\mu_\alpha) + \frac{1}{2} \int p^2 \, d\bar{F}(\epsilon) \right\} = \min_{\alpha, b} \left\{ L(\mu_\alpha) + \frac{1}{2} \int p^2 \, d\bar{F}(\epsilon) \right\}$$

where $\int \alpha = \beta^2 \, L(\epsilon)/\int p^2 \, d\bar{F}(\epsilon)$. The result follows from noting that, at the points where $L(\cdot)$ is analytic and thus can be represented by a power series, cf. Rudin (1964, p. 158), it lends itself to a Taylor-series expansion around $\mu_\alpha$:

$$L(\epsilon) = L(\mu_\alpha) + \frac{1}{2} \int p^2 \, d\bar{F}(\epsilon) = L(\mu_\alpha) + \frac{1}{2} \int p^2 \, d\bar{F}(\epsilon)$$

Taking expectations, the finite number of points where $L(\cdot)$ is not analytic can be ignored since they are assumed to occur with probability zero. Thus:

$$E[L(\epsilon)] = \int L(\epsilon) \, d\bar{F}(\epsilon)$$

$$= \int L(\mu_\alpha) + \frac{1}{2} \int p^2 \, d\bar{F}(\epsilon)$$

$$= \int L(\mu_\alpha) + \frac{1}{2} \int p^2 \, d\bar{F}(\epsilon)$$

$$= \int L(\mu_\alpha) + \frac{1}{2} \int p^2 \, d\bar{F}(\epsilon)$$

The second assumption does not rule out loss functions such as lin-lin which are non-differentiable at a single point. The third assumption requires that all moments exist for which the corresponding derivative of the loss function with respect to the forecast error is non-zero. This is a strong requirement and rules out many potential combinations of loss functions and forecast error distributions.6

---

6 For example, combining a t-distribution with three degrees of freedom with a loss function whose fourth derivative is non-zero will result in a non-existing expected loss. Another example comes from the linear loss function, whose higher order derivatives are all non-zero. All moments of the forecast error distribution must therefore exist to ensure that the expected loss is well defined.
When the loss function is not linear or quadratic, higher order moments of the forecast error distribution such as the skew generally matter. It is not surprising that asymmetry in the loss function interacts with asymmetry in the underlying distributions of the outcomes and forecasts that are being combined. Suppose that the decision maker dislikes positive forecast errors more than negative ones. Now consider moving mass of the distribution to the right, so the underlying distribution is positively skewed. The loss function prefers lower probability of positive forecast errors when choosing the combination weights. The extent of the asymmetry in the loss function drives the size of this effect. Hence the combination weights are clearly a function of the parameters of the loss function. There is also a large effect on the optimal constant. When positive forecast errors are to be avoided, the constant is chosen to ensure that the forecast error distribution is centered to the left of zero.

3.2. Elliptically symmetric data

In special cases only the constant ($\alpha'$) depends on the shape of the loss function, while the forecast combination weights are unaltered from the case with MSE loss:

Proposition 2. Suppose that the expected loss can be written as $\mathbb{E}[L(\mu,\epsilon)] = q(\mu, \sigma^2)$ and let $\alpha'$ be the optimal value for $\mu$. Then if the partial derivative $\frac{\partial q(\mu^*, \sigma^2)}{\partial \mu}$ is nonzero,

(i) $\alpha' = \Sigma_{21}^{-1} \alpha_2$;

(ii) $\alpha'$ is the solution to $\frac{\partial q(\mu^*, \sigma^2)}{\partial \mu} = 0$.

The result follows since $\alpha'$ only appears in $\mu_1$, so the optimal value for this parameter solves

$$\frac{\partial q(\mu_1, \sigma^2)}{\partial \sigma^2} = \frac{\partial q(\mu_1, \sigma^2)}{\partial \mu} \frac{\partial \mu}{\partial \sigma^2} = \frac{\partial q(\mu_1, \sigma^2)}{\partial \mu} = 0.$$ 

This is the result in part (ii) of the proposition. This choice for $\alpha'$ sets $\mu_1$ at its optimal point ($\mu^*$). The optimal value for $\alpha$ therefore solves

$$\frac{\partial q(\mu^*, \sigma^2)}{\partial \sigma^2} = \frac{\partial q(\mu^*, \sigma^2)}{\partial \mu} \frac{\partial \mu}{\partial \sigma^2} = 0.$$ 

Furthermore, $\frac{\partial q(\mu^*, \sigma^2)}{\partial \sigma^2} = 2(\Sigma_{21}^{-1} - \sigma_{12})$ so the solution is as stated so long as $\frac{\partial q(\mu^*, \sigma^2)}{\partial \sigma^2} \neq 0$. The second order condition is satisfied as long as the loss function satisfies conditions (1)--(3) above.

Proposition 2 shows that the optimal forecast combination weights are identical to the MSE weights for a wide set of loss functions provided that the marginal distribution of the forecast errors depends only on the first two moments of the forecast errors. The general class of distributions for which this is true is for $\mathcal{P}(X_1, X_2)$ elliptically symmetric. The multivariate normal is a special case of this family, as is the multivariate $t$-distribution. We shall for simplicity refer to distributions satisfying the assumptions of Proposition 2 as elliptically symmetric.

Proposition 2 has two important implications. First, estimation of forecast combination weights under asymmetric loss simplifies to a two-stage procedure (see Section 6.2 below for details). The second implication is that departures from elliptical symmetry are needed to drive the forecast combination weights away from their solution under MSE loss. If, under elliptical symmetry, a forecast adds information (in the sense that it has a non-zero combination weight) under MSE loss, it will also add information under arbitrarily asymmetric loss functions. The converse also holds—if the forecast is not useful under MSE loss then it will not be useful for any other loss function.

Corollary 1. Suppose that the conditions of Proposition 2 hold. Then a forecast gets assigned a non-zero weight in the combined forecast under a general loss function if and only if it’s weight under MSE loss is non-zero.

This result stands in stark contrast to the results we obtain under general forecast error distributions. For sufficient deviation from elliptical symmetry (particularly when there is skew in the forecast errors) the weights on the forecasts will depend on the loss function. However, under elliptically symmetric forecast errors, decision makers agree on the value of prediction signals irrespective of their loss function. They could well disagree about the need to include an intercept as this depends on the shape of the loss function.

When elliptical symmetry is abandoned, a particular forecast or subset of forecasts may contain information under MSE loss but not under asymmetric loss. The converse may also hold: a forecast could be assigned a non-zero weight under asymmetric loss but not under MSE loss. We state this result in Corollary 2 which is the natural converse of Corollary 1 which was established under elliptical symmetry:

Corollary 2. Suppose that the conditions of Proposition 2 do not hold. Then

(i) a forecast may contain information under MSE loss but not under alternative (asymmetric) loss functions;

(ii) a forecast may contain information under asymmetric loss but not under MSE loss.

---

1. This implication is closely related to the result in finance that the distribution of a portfolio is determined by the mean and variance in these situations, see Chamberlain (1993). Under symmetry restrictions on both the loss function and the forecast error distribution Cochrane (1996) obtains a related result.
2. X is elliptically symmetric, $E(Y|\Sigma, x)$, if its density takes the form $f(x) = (2\pi)^{-1/2} |\Sigma|^{-1/2} f(x - \mu)/\Sigma (x - \mu)$, where $\mu$ is an arbitrary constant and $f(.)$ is any function that represents a density.
3. Notice also the relationship between Corollary 1 and the result in Diebold et al. (1998) that decision makers with different loss functions cannot generally agree on the ranking of misspecified predictive density models. Provided that the forecast errors are elliptically symmetric, our result says that decision makers can agree about what is a useful forecast even if the forecasts are individually misspecified.
The above results allow for arbitrary correlations between the forecast errors, as long as the covariance matrix $\Sigma$ is invertible. This condition is unimportant since linearly dependent forecasts would always be dropped from the combination. Serial correlation, on the other hand, is more difficult to handle in our setup. One possibility is to include extra regressors in the forecast combination and use these to absorb any serial correlation that is still present in the combined forecast. An iterative procedure would (a) estimate the combined forecast under the assumed loss function; (b) test for serial correlation in the errors from the combined forecast; (c) add the lagged combined or individual forecasts as extra 'forecasts' if there is evidence of serial correlation in the (combined) forecast error; repeat step (a) and continue the iteration until there is no serial correlation. An alternative strategy, proposed by Diebold (1988), is to use an estimator that allows the forecast errors from the combination model to follow a finite-order ARMA process.

3.3. Forecast encompassing tests

An alternative to the strategy of forecast combination is to simply use a single forecast. This situation arises when the forecast produced by one model, $B$, does not add anything to the forecast from another model, $A$. In this situation, model $A$ is said to forecast encompass model $B$. Under MSE loss this is tested by least squares estimation of the equation:

$$y_{t+1} = \alpha + \beta_0 y_{t+1}^A + \beta_1 y_{t+1}^B + \epsilon_{t+1}. \quad (8)$$

If the joint hypothesis that $\beta_0 = 1$, $\beta_1 = 0$ cannot be rejected, model $A$ forecast encompasses model $B$.

An implication of the previous two propositions and corollaries is the following

**Proposition 3.** Suppose that the assumptions of Proposition 2 hold. Then if model $A$ forecast encompasses model $B$ under MSE loss, model $A$ also encompasses model $B$ for any other loss function.

Conversely, if these assumptions do not hold, it is possible that model $A$ forecast encompasses model $B$ under MSE loss but not under an alternative loss function or that model $A$ does not forecast encompass model $B$ under MSE loss, but does so under the alternative loss function.

When forecast errors are not generated by an elliptically symmetric distribution, it follows that it is not possible to set up a universal encompassing test since the outcome will depend on the specific loss function. Encompassing tests will be specific to the loss function.

Min and Zellner (1993) determine the factors that may lead a single forecast to dominate the strategy of combining multiple forecasts. When the individual forecasts are either unbiased or have a bias of the same magnitude, some degree of forecast combination is often optimal. In contrast, when one forecast is more heavily biased than the others and a constant is not included in the forecast combination, it becomes more likely that individual forecasts will dominate forecast combinations.

4. Results for specific loss functions

To demonstrate the general results from the previous section, this section considers three commonly entertained asymmetric loss functions, namely linex, linlin and asymmetric quadratic loss.

4.1. Linex loss

The linex loss function was introduced by Varian (1974) in the context of losses associated with real estate assessments. Zellner (1986) gave a comprehensive treatment of Bayesian predictions for this loss function. Linex loss takes the form

$$L(e_t) = \exp(\alpha e_t) - \alpha e_t - 1. \quad (9)$$

The parameter $\alpha$ controls the extent of asymmetry. If $\alpha > 0$, there are large losses from positive forecast errors and the losses are larger the larger the value of $\alpha$. If $\alpha < 0$, large losses result from negative forecast errors and the losses are larger the smaller $\alpha$. The expected loss is

$$E[L(e_t)] = M_\alpha(a) - a \mu + 1, \quad (10)$$

where $M_\alpha(a)$ is the moment generating function of the forecast errors and thus depends on their marginal distribution. Eq. (10) shows that the expected linex loss only exists when the forecast error distribution has an infinite number of moments.\(^{10}\) Differentiating the expected loss with respect to $\alpha$ and $\mu$ yields the first order conditions

$$\frac{\partial E[L(\epsilon_t)]}{\partial \alpha} = \frac{\partial M_\alpha(a)}{\partial \alpha} + a = 0,$n

$$\frac{\partial E[L(\epsilon_t)]}{\partial \mu} = \frac{\partial M_\alpha(a)}{\partial \mu} + \mu = 0. \quad (11)$$

These equations fully characterize the solution to the optimal weights as a function of the parameters of the underlying forecast error distribution and of the loss function ($\alpha$).

In common with most other asymmetric loss functions, (11) is generally a highly nonlinear function in the forecast combination weights and simple closed form solutions are typically not available. An exception arises when the forecast errors are normally distributed. In this case $M_\alpha(a) = \exp(\mu \epsilon + (\alpha^2/2) \epsilon^2)$, and the optimal constant is given by

$$\omega_\alpha = \mu + \frac{\alpha}{2} \mu^2.$$

\(^{10}\) This follows from Proposition 1 since all derivatives of the linex loss function are nonzero. The primitive condition in Proposition 2 requires that

$$\frac{\partial M_\alpha(a)}{\partial \alpha} = a \exp \left\{ \frac{\mu^2 + \alpha^2}{2} \right\} \neq 0.$$

This is always satisfied, except in the trivial case of perfect forecasts ($\epsilon_t = 0$).
Under normality of the forecast errors, the optimal population bias from the combined forecasts is $\mu_0 = -(\alpha/2)\sigma_0^2$. This is the optimal bias derived in Christoffersen and Diebold (1997) for the univariate linear forecasting problem. The expression has an intuitive interpretation: when $\alpha > 0$, large losses follow from positive forecast errors and it is optimal to choose a positive constant so that the forecast errors have a negative mean. The forecast error distribution is hence biased towards negative losses to avoid the high penalties associated with positive losses. The reverse is true when the decision maker tries to avoid negative losses (i.e. $\alpha < 0$). In both cases, the larger $\alpha$ is the larger the penalty so the larger the bias. The larger the variance of the forecast errors the larger the chance of large forecast errors and hence the more the combination gets biased to reduce the chance of large losses.

### 4.2. Lin–lin loss

A popular way of capturing asymmetries in the loss function is to let losses take one form for forecast errors over and above a certain threshold, $\kappa$, and another form for forecast errors below this value:

\[
L(e_i) = \begin{cases} 
L_1(e_i) & \text{if } e_i > \kappa, \\
L_2(e_i) & \text{if } e_i \leq \kappa.
\end{cases}
\]

$L_1(\kappa) = L_2(\kappa)$ ensures continuity. Setting $\kappa = 0$ and assuming that $L(\cdot)$ is piecewise linear, we get the lin–lin loss function:

\[
L(e_i) = \begin{cases} 
(1 - \theta)e_i^2 & \text{if } e_i > 0, \\
\theta |e_i| & \text{if } e_i \leq 0,
\end{cases}
\]

where $0 < \theta < 1$. Using an indicator function, the lin–lin loss function can be rewritten as

\[
L(e_i) = (-\theta + 1_{e_i > 0})e_i.
\]

Minimizing this expression over the constant and combination weights is equivalent to solving the problem

\[
\min_{\theta, \mu_0} \left\{ -\theta + 1_{e_i > 0} \right\} e_i.
\]

The expected loss to be minimized is thus

\[
E[L(e_i)] = \int_{-\infty}^{\infty} e_i \ dF(e_i) - \theta \mu_0,
\]

where $F(e_i)$ is the cumulative density function of the forecast error. Expressing $e_i$ as $e_i = \mu_0 + \sigma_0 x_i$, so $z = x$ is the centered and standardized forecast error with density $f_z(\cdot)$ and cumulative density $F_z(\cdot)$, the expected loss is

\[
E[L(e_i)] = \mu_0 \left( 1 - \theta - F_z \left( -\frac{\mu_0}{\sigma_0} \right) \right) + \sigma_0 \int_{-\infty}^{\infty} z \ dF_z(z).
\]

Derivatives of $L(\cdot)$ with respect to $\sigma^2$ and $\omega$ are difficult to interpret as in general the marginal distribution $F_z(\cdot)$ will itself depend on $(\sigma^2, \omega)$. An exception to this occurs when $\mu_0$ is Gaussian and, using Leibniz’s rule of integration, the first order condition for the constant simplifies to

\[
1 - \theta - \Phi \left( -\frac{\mu_0}{\sigma_0} \right) = 0,
\]

where $\Phi(\cdot)$ is the cumulative density of a standard normal variate.

Recalling that $\mu_0 = \mu_0 - \sigma_0 \mu - \sigma_0 \omega \Phi^{-1}(\theta)$ and using the symmetry of $\Phi(\cdot)$ we obtain a closed-form solution for the scalar intercept term:\n
\[
\omega_0 = \mu_0 - \sigma_0 \mu - \sigma_0 \Phi^{-1}(\theta).
\]

The direction of the bias can be evaluated using $\mu_0 = \sigma_0 \Phi^{-1}(\theta)$. For $\theta > 0$, the loss function is symmetric and the forecast error is unbiased. For $\theta < 1/2$, higher weight is placed on positive errors. To lower the probability of large losses associated with positive forecast errors the optimal combination biases the mean forecast error to be negative ($\Phi^{-1}(\theta) < 0$) irrespective of the covariance matrix of the underlying forecasts. The reverse is true when $\theta > 1/2$. As with linx loss, the size of the bias increases as the forecast error variance rises, to better avoid high loss outcomes.

### 4.3. Asymmetric squared loss

The asymmetric squared loss function takes the form

\[
L(e_i) = \begin{cases} 
(1 - \theta)e_i^2 & \text{if } e_i > 0, \\
\theta e_i^2 & \text{if } e_i \leq 0,
\end{cases}
\]

where $0 < \theta < 1$. Minimizing this expression over $\sigma^2, \omega$ is equivalent to solving

\[
\min_{\sigma^2, \omega} \left\{ (1 - \theta)\sigma^2 + \theta \omega^2 \right\}.
\]

Again we can write $e_i = \mu_0 + \sigma_0 x_i$, and change variables in the expression for the expected loss:

\[
E[L(e_i)] = \theta(\sigma_0^2 + \omega_0^2) + (1 - \theta) \int_{-\infty}^{\infty} z^2 \ dF_z(z) + 2\mu_0 \theta \int_{-\infty}^{\infty} z \ dF_z(z) - \theta \mu_0.
\]

\[\text{by symmetry of } dF_z(\cdot) \text{ equal zero.}\]
Taking derivatives with respect to \( t \) yields the first order condition

\[
0 = \alpha_1 - \mu_1 + \beta_1 - 2 \gamma_1 \int_{-\infty}^{\infty} \left[ \phi_x(t) - \phi_x(\mu_1) \right] dt
\]

This means that the solution to the first order condition for \( \theta_1 \) is equal to the solution to the second order condition for \( \theta_1 \). The solution to the second order condition for \( \theta_1 \) is identical to the solution to the second order condition for \( \theta_2 \).

5. Mixture distributions

Section 5 shows that distributions from elliptical statistics may be used to obtain a closed-form solution for \( \theta_1 \) in the case of \( \mu = 0 \) and \( \Sigma \). This section analyzes the effect of \( \mu \) on the distribution of \( \theta_1 \). The joint distribution of \( \theta_1 \) is a mixture of normals, which is approximately a normal distribution when \( \mu = 0 \) and \( \Sigma \) is diagonal. This section considers models that are close to approximating a normal distribution, and applies a second order condition to the first order condition.

[1] A. T. Bhar, D. R. Nachmias, and G. L. Jones (1992) show that high volatility states often have small values for the first two moments of the distribution. For example, when \( \mu = 0 \) and \( \Sigma \) is diagonal, the distribution is normal.

[2] For example, a normal distribution with mean zero and standard deviation one is \( N(0, \Sigma) \), while a normal distribution with mean one and standard deviation one is \( N(1, \Sigma) \).
6. Estimation of forecast combination weights

In practice, forecast combination weights must be estimated from past data. This involves assumptions as to the stability over time of the joint distribution of the predicted variable and prediction signals. The empirical evidence of employing estimation methods based largely on least squares is surveyed in Clemens (1989) and further discussed by Diebold and Lopez (1996).

6.1. Moment estimators

In general we are interested in choosing the constant and the combination weights \( \omega^f, \omega^o \) that minimize the expected loss. The realized loss in period \( t \), which we denote by \( Q_i(\omega^f_i, \omega^o) \), can be written as

\[
Q_i(\omega^f_i, \omega^o) = L(\omega^f_i, \omega^o | y_{t+1}, \theta),
\]

where \( \theta \) are the parameters of the loss function. \( \omega^f \) and \( \omega^o \) can be obtained as an M-estimator based on the sample analog of \( \mathbb{E} (Q_i) \) using \( n \) observations \( \{y_{t+1}, \hat{y}_t\}_{t=1}^n \):

\[
\hat{Q}_i(\omega^f_i, \omega^o) = n^{-1} \sum_{t=1}^n Q_i(\omega^f_i, \omega^o).
\]

An estimator that is available when we can interchange the expectation operator and the derivative of \( Q_i \) with respect to \( \omega^f \) and \( \omega^o \),

\[
\partial Q_i/\partial (\omega^f, \omega^o) = \mathbb{Q}_i, \quad \text{selects } \omega^f, \omega^o \text{ to satisfy the first order condition}
\]

\[
\mathbb{E} (\mathbb{Q}_i(\omega^f, \omega^o)) = 0.
\]

The corresponding sample analog is the average over the \( n \) observations. One problem with this method is that it can give rise to multiple solutions even when the M-statistic is regular.

A second method is related to instrumental variables estimation of a nonlinear regression and can therefore be handled in a GMM framework. The GMM approach estimates \( \omega \) by minimizing the quadratic form

\[
V = \left( \sum_{i=1}^n Q_i(\omega^f_i, \omega^o) \right)^T A^{-1} \left( \sum_{i=1}^n Q_i(\omega^f_i, \omega^o) \right),
\]

where \( A \) is some positive definite matrix. The optimal weighting matrix is standard from GMM problems. An advantage of the M-estimator is that there are a wealth of results available for consistency and asymptotic normality of the estimated weights. A similar result is true for the GMM estimator.

[15] Other examples of estimators of this form include maximum likelihood estimators, where \( Q \) is replaced by minus one times the log-likelihood.

[16] This assumption obviously imposes a restriction on the distribution of the forecast error.
6.1.1. Linex loss
These estimation methods can be demonstrated for the linex loss function. For this case, $Q_\ell(\omega', \omega)$ takes the form
$$Q_\ell(\omega', \omega) = \exp(a(y_i - \omega' y_i)) - a(y_i - \omega' y_i) - 1.$$ 
GMM estimation is based on the derivative, $Q'$:
$$Q'_\ell = a(1 - \exp(a(y_i - \omega' y_i))) \left( 1 \over y_i \right).$$ (23)

The solution to this is equivalent to the nonlinear IV estimator obtained from the regression
$$1 \approx \exp(a(y_i - \omega' y_i)) + \gamma_i$$
using $(y_i, y_i')'$ as instruments.

6.1.2. Lin-lin loss
Recall that the lin-lin loss can be written
$$L(\epsilon) = \{(1 - \theta)\epsilon \mid \epsilon > 0 \} + \theta\mid \epsilon \mid_{\epsilon \leq 0}.$$ 
This is precisely the loss function associated with the quantile regression problem, cf. Koenker and Bassett (1978) and Buchinsky (1992).
Assuming independently and identically distributed forecast errors, the function to be minimized becomes
$$Q_\ell(\omega', \omega) = \frac{1}{T} \left( 1 - \theta \right) \sum_{t=1}^{T} \left| y_t - \omega' y_t \right| + \theta \sum_{t=1}^{T} \left| y_t - \omega' y_t \right|.$$
Buchinsky (1992) shows that this can be rewritten to fit in a method of moments framework. Define the sign function $s(\lambda) = 1_{\lambda > 0} - 1_{\lambda < 0}$. Then the objective function takes the form
$$Q_\ell(\omega) = \frac{1}{T} \sum_{t=1}^{T} \left( 1 - \theta + \frac{1}{2} s(y_t - \omega' y_t) \right) (y_t - \omega' y_t).$$ (24)

Differentiating with respect to $(\omega', \omega)$, the first order conditions are
$$Q'_\ell = \left( \frac{1}{T} \sum_{t=1}^{T} \left( 1 - \theta + \frac{1}{2} s(y_t - \omega' y_t) \right) y_t \right) \omega' = 0,$$ (25)
which again fits into the GMM Framework. Buchinsky (1992) suggests using the simplex method to solve for the parameters.

6.1.3. Asymmetric squared loss
The asymmetric squared loss function can be written as
$$L(\epsilon) = \{(1 - \theta)\epsilon \mid \epsilon > 0 \} + \theta\mid \epsilon \mid_{\epsilon \leq 0}.$$ 
The loss function,
$$L'(\epsilon) = |1 - \theta - 1_{\epsilon > 0}|^2 \epsilon$$ (26)
is in the same homogenous class as $L(\epsilon)$ and the solution $(\omega', \omega'')$ minimizing the two is the same. $L'(\epsilon)$ is the loss function associated with the "expectile" regression problem examined in Newey and Powell (1987).

We choose the constant and forecast combination weights $(\omega', \omega'')$ to minimize
$$Q_\ell(\omega', \omega) = \frac{1}{T} \sum_{t=1}^{T} \left| 1 - \theta - 1_{\gamma_t \epsilon > 0}\epsilon \right| (y_t - \omega' y_t)^2.$$ (27)

This function is continuously differentiable and iterated weighted least squares can therefore be used to estimate the forecast combination weights. The estimator at iteration $n$ becomes
$$\hat{\omega}_n = \left( \frac{1}{T} \sum_{t=1}^{T} k_{n-1} / y'_t \right)^{-1} \left( \frac{1}{T} \sum_{t=1}^{T} k_{n-1} y_t y_t' \right),$$ (28)
where $y_t = (1, y_t')'$ and the scalar weights $k_{n-1}$ are constructed according to
$$k_{n-1} = |1 - \theta - 1_{\gamma_n \epsilon > 0}|\gamma_n.$$(29)

6.2. Estimation under elliptical symmetry

When the forecast errors are elliptically symmetric, a simple two stage procedure is suggested by our theoretical results in Section 3:

(i) Estimate $\omega$ by OLS, regressing $y_t$ on a constant and $\bar{y}$.
(ii) Use the constructed variable $y_t = \omega' \bar{y}$ to estimate $\omega'$ based on the relevant loss function (as above).

In this case the estimation problem is much simpler since the optimization procedure only has to search over a one dimensional parameter space regardless of the number of forecasts we are combining.

The methods require $P(\gamma_{i1}, \gamma_{i1}')$ to be elliptically symmetric for optimality but may work well for minor deviations from this distribution. This suggests that it is important to have a test for elliptical symmetry. Zhu and Neubaus (2000) propose nonparametric Monte Carlo tests that could be useful. Another strategy is to test for elliptical symmetry on the marginals. This is a useful procedure since, under joint elliptical symmetry, the marginals are also symmetric and departures from symmetry of the marginals imply that the joint distribution is not elliptically symmetric. Thus simple tests for skewness in the marginals will have power to detect deviations from elliptical symmetry.
Table 1. Average loss for lin-lik loss function

<table>
<thead>
<tr>
<th>( \rho )</th>
<th>Loss</th>
<th>M-est</th>
<th>OLS</th>
<th>OLS(_c)</th>
<th>AVE</th>
<th>AVE(_c)</th>
<th>Forl</th>
<th>Forlc</th>
<th>For2</th>
<th>For2c</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skewed mixture</td>
<td>(-1.25)</td>
<td>0.28</td>
<td>1.08</td>
<td>1.36</td>
<td>1.08</td>
<td>1.50</td>
<td>1.05</td>
<td>1.35</td>
<td>1.09</td>
<td>1.46</td>
</tr>
<tr>
<td></td>
<td>(-1.00)</td>
<td>0.27</td>
<td>1.06</td>
<td>1.23</td>
<td>1.07</td>
<td>1.18</td>
<td>1.04</td>
<td>1.22</td>
<td>1.07</td>
<td>1.31</td>
</tr>
<tr>
<td></td>
<td>(-0.75)</td>
<td>0.26</td>
<td>1.05</td>
<td>1.14</td>
<td>1.05</td>
<td>1.10</td>
<td>1.03</td>
<td>1.13</td>
<td>1.06</td>
<td>1.21</td>
</tr>
<tr>
<td></td>
<td>(-0.50)</td>
<td>0.32</td>
<td>1.04</td>
<td>1.05</td>
<td>1.04</td>
<td>1.01</td>
<td>1.02</td>
<td>1.04</td>
<td>1.03</td>
<td>1.11</td>
</tr>
<tr>
<td></td>
<td>(-0.25)</td>
<td>0.52</td>
<td>1.04</td>
<td>1.05</td>
<td>1.04</td>
<td>1.01</td>
<td>1.02</td>
<td>1.04</td>
<td>1.03</td>
<td>1.11</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>0.26</td>
<td>1.05</td>
<td>1.14</td>
<td>1.06</td>
<td>1.10</td>
<td>1.03</td>
<td>1.14</td>
<td>1.06</td>
<td>1.21</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>0.27</td>
<td>1.06</td>
<td>1.23</td>
<td>1.07</td>
<td>1.19</td>
<td>1.04</td>
<td>1.23</td>
<td>1.08</td>
<td>1.22</td>
</tr>
<tr>
<td></td>
<td>1.25</td>
<td>0.38</td>
<td>1.07</td>
<td>1.36</td>
<td>1.08</td>
<td>1.31</td>
<td>1.06</td>
<td>1.36</td>
<td>1.09</td>
<td>1.47</td>
</tr>
<tr>
<td>Kurtosis mixture</td>
<td>(-1.25)</td>
<td>0.13</td>
<td>1.17</td>
<td>1.21</td>
<td>1.11</td>
<td>1.12</td>
<td>1.05</td>
<td>1.15</td>
<td>1.07</td>
<td>1.33</td>
</tr>
<tr>
<td></td>
<td>(-1.00)</td>
<td>0.12</td>
<td>1.13</td>
<td>1.16</td>
<td>1.10</td>
<td>1.10</td>
<td>1.04</td>
<td>1.10</td>
<td>1.05</td>
<td>1.23</td>
</tr>
<tr>
<td></td>
<td>(-0.75)</td>
<td>0.11</td>
<td>1.10</td>
<td>1.12</td>
<td>1.09</td>
<td>1.09</td>
<td>1.05</td>
<td>1.10</td>
<td>1.05</td>
<td>1.20</td>
</tr>
<tr>
<td></td>
<td>(-0.50)</td>
<td>0.16</td>
<td>1.08</td>
<td>1.06</td>
<td>1.04</td>
<td>1.02</td>
<td>1.02</td>
<td>1.03</td>
<td>1.04</td>
<td>1.11</td>
</tr>
<tr>
<td></td>
<td>(-0.25)</td>
<td>0.16</td>
<td>1.08</td>
<td>1.06</td>
<td>1.04</td>
<td>1.02</td>
<td>1.02</td>
<td>1.03</td>
<td>1.04</td>
<td>1.11</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>0.11</td>
<td>1.11</td>
<td>1.12</td>
<td>1.09</td>
<td>1.03</td>
<td>1.06</td>
<td>1.05</td>
<td>1.20</td>
<td>1.18</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>0.13</td>
<td>1.13</td>
<td>1.16</td>
<td>1.10</td>
<td>1.08</td>
<td>1.04</td>
<td>1.10</td>
<td>1.05</td>
<td>1.23</td>
</tr>
<tr>
<td></td>
<td>1.25</td>
<td>0.13</td>
<td>1.17</td>
<td>1.21</td>
<td>1.11</td>
<td>1.12</td>
<td>1.05</td>
<td>1.15</td>
<td>1.07</td>
<td>1.33</td>
</tr>
</tbody>
</table>

Note: The first column reports the value of the parameter, \( \rho \), controlling asymmetry. The second column gives the expected loss averaged over 5,000 Monte Carlo replications based on estimates of the forecast combination weights and the constant from 100 observations. These losses are scaled by the population loss. OLS refers to the estimation of the combination weights by OLS. AVE refers to the equal-weighted forecast combination. Forl refers to forecast l alone. Names ending with c refer to the corresponding estimation technique for the weights followed by optimal estimators of the constant.

7. Monte Carlo results

This section reports on the performance of the various estimation procedures applied to the lin-lin and asymmetric quadratic loss functions under the two mixture models in Section 5. In each case we combine two forecasts and assume that a sample of 100 observations is available to estimate the combination weights. All results are based on 5,000 Monte Carlo simulations.

7.1. Expected losses

Tables 1–3 examine the expected loss under lin-lin and asymmetric loss. We report the ratio of the expected loss under the estimated weights relative to the loss under the optimal population weights:

\[
\frac{E[L_{\text{lin}}(\hat{\theta})]}{E[L_{\text{lin}}(\hat{\theta}_{\text{opt}})]}
\]

Results are shown for each of the estimation methods considered in Section 6. The first column reports the value of the asymmetry parameter of the loss function. The second column gives the expected loss using the population weights. The third column evaluates the expected loss when the forecast combination weights and the constant are based on the preferred (among estimators that consistently estimate the weights) estimation method. The fourth and fifth columns give results from OLS estimation and OLS estimation of the weights followed by estimation of the constant term (which is the procedure proposed above if we knew the data were elliptically symmetric), respectively. We also report losses for the average (equal-weighted) and individual forecasts (unit or zero weight on each forecast) as well as losses for each of these where an optimally estimated constant is employed to obtain the correct bias (these are subscripts c in the tables). Separate results are reported for the skewed and kurtotic mixture models.

The first forecast is quite good on its own, the second quite a bit less so. The fact that combination methods do better than simply using the first forecast on its own highlights the value of the forecast combination approach. Consistent with theory, even when the first forecast is obviously better than the second, the latter can still be very useful in constructing a better combined forecast.

The results are similar qualitatively—and to an extent quantitatively as well—across loss functions. Under both the skewed and kurtotic mixture models the 'optimal'

---

17 The expectation is taken with respect to the true distribution of the forecast errors.
Table 3
Average loss for asymmetric quadratic loss function.

<table>
<thead>
<tr>
<th>θ</th>
<th>Loss</th>
<th>WLS</th>
<th>OLS</th>
<th>OLSc</th>
<th>AVE</th>
<th>AVEc</th>
<th>Fwr1</th>
<th>Fwr1c</th>
<th>Fwr2</th>
<th>Fwr2c</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.14</td>
<td>0.89</td>
<td>1.08</td>
<td>1.21</td>
<td>1.05</td>
<td>1.09</td>
<td>1.09</td>
<td>1.99</td>
<td>1.15</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.26</td>
<td>1.05</td>
<td>1.34</td>
<td>1.06</td>
<td>1.30</td>
<td>1.03</td>
<td>1.34</td>
<td>1.97</td>
<td>1.42</td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>0.26</td>
<td>1.12</td>
<td>1.05</td>
<td>1.11</td>
<td>1.02</td>
<td>1.14</td>
<td>1.05</td>
<td>1.24</td>
<td>1.31</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>0.26</td>
<td>1.04</td>
<td>1.06</td>
<td>1.04</td>
<td>1.03</td>
<td>1.02</td>
<td>1.06</td>
<td>1.05</td>
<td>1.12</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>0.26</td>
<td>1.04</td>
<td>1.04</td>
<td>1.03</td>
<td>1.00</td>
<td>1.03</td>
<td>1.04</td>
<td>1.10</td>
<td>1.11</td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>0.26</td>
<td>1.04</td>
<td>1.04</td>
<td>1.03</td>
<td>1.00</td>
<td>1.06</td>
<td>1.05</td>
<td>1.12</td>
<td>1.11</td>
<td></td>
</tr>
<tr>
<td>0.7</td>
<td>0.26</td>
<td>1.04</td>
<td>1.06</td>
<td>1.05</td>
<td>1.11</td>
<td>1.08</td>
<td>1.14</td>
<td>1.05</td>
<td>1.23</td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>0.26</td>
<td>1.05</td>
<td>1.34</td>
<td>1.06</td>
<td>1.29</td>
<td>1.09</td>
<td>1.33</td>
<td>1.06</td>
<td>1.41</td>
<td></td>
</tr>
<tr>
<td>0.9</td>
<td>0.15</td>
<td>1.07</td>
<td>1.33</td>
<td>1.07</td>
<td>1.82</td>
<td>1.05</td>
<td>1.87</td>
<td>1.08</td>
<td>1.58</td>
<td></td>
</tr>
</tbody>
</table>

Kurtosis mixture

<table>
<thead>
<tr>
<th>θ</th>
<th>Loss</th>
<th>WLS</th>
<th>OLS</th>
<th>OLSc</th>
<th>AVE</th>
<th>AVEc</th>
<th>Fwr1</th>
<th>Fwr1c</th>
<th>Fwr2</th>
<th>Fwr2c</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.07</td>
<td>1.16</td>
<td>1.09</td>
<td>1.55</td>
<td>1.04</td>
<td>1.58</td>
<td>1.06</td>
<td>1.76</td>
<td>1.18</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.09</td>
<td>1.17</td>
<td>1.08</td>
<td>1.22</td>
<td>1.03</td>
<td>1.24</td>
<td>1.04</td>
<td>1.38</td>
<td>1.17</td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>0.10</td>
<td>1.09</td>
<td>1.08</td>
<td>1.05</td>
<td>1.03</td>
<td>1.11</td>
<td>1.04</td>
<td>1.24</td>
<td>1.16</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>0.10</td>
<td>1.08</td>
<td>1.06</td>
<td>1.03</td>
<td>1.02</td>
<td>1.05</td>
<td>1.04</td>
<td>1.17</td>
<td>1.16</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>0.10</td>
<td>1.08</td>
<td>1.06</td>
<td>1.03</td>
<td>1.02</td>
<td>1.05</td>
<td>1.04</td>
<td>1.17</td>
<td>1.16</td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>0.10</td>
<td>1.08</td>
<td>1.08</td>
<td>1.03</td>
<td>1.02</td>
<td>1.05</td>
<td>1.04</td>
<td>1.17</td>
<td>1.16</td>
<td></td>
</tr>
<tr>
<td>0.7</td>
<td>0.10</td>
<td>1.09</td>
<td>1.16</td>
<td>1.09</td>
<td>1.09</td>
<td>1.11</td>
<td>1.04</td>
<td>1.24</td>
<td>1.16</td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>0.09</td>
<td>1.11</td>
<td>1.29</td>
<td>1.06</td>
<td>1.23</td>
<td>1.03</td>
<td>1.24</td>
<td>1.04</td>
<td>1.38</td>
<td></td>
</tr>
<tr>
<td>0.9</td>
<td>0.07</td>
<td>1.10</td>
<td>1.65</td>
<td>1.07</td>
<td>1.55</td>
<td>1.04</td>
<td>1.56</td>
<td>1.06</td>
<td>1.76</td>
<td></td>
</tr>
</tbody>
</table>

Note: Columns as in Table 2.

Table 4

Similar results arise under averaging which performs poorly on its own but does much better after the bias adjustment. When the intercept is corrected (using the relevant estimator for each loss function) the losses perform very well in the Monte Carlo design examined here. This, however, strongly depends on the underlying distribution of the data. To demonstrate this point Fig. 2a graphs the average loss for the quantile regression, OLS and AVEc against the asymmetry parameter, θ. For this case the AVEc method does very well against the other methods. The ‘folk theorem’ on the superiority of using even weights reflects that when the weights are ‘close’ to even, imposing this constraint does not result in a large bias and reduces the variance due to estimation error. However, more diverse weights overturn this result, as can be seen in Fig. 2b. In this case we have designed the Monte Carlo so that the weight on the first forecast ranges from 0.4 to 0.6 (close to one-half) but the weight on the second forecast ranges from -0.5 to 0. In this case the constant adjusted average weighting method does extremely poorly. More generally it is easy to find cases where this method does well and cases where it does poorly. Note that the OLSc method still works very well for this Monte Carlo design.

We also examine Diebold and Mariano (1995) tests that compare various estimators, reporting results for the lin-lin case only as results for the other loss functions are similar. Table 4 shows rejection frequencies for a size of 5% using the

Fig. 2. Average relative loss: (a) Model 1, (b) Model 2. Note: Model 1 refers to the skewed mixture. Model 2 is as the skewed mixture but with a different variance-covariance matrix. The modelled series for state 1 are: (1,3) = -0.65; (2,1) = 0.5; and (3,3) = 1/3.
Table 4

<table>
<thead>
<tr>
<th>$\delta$</th>
<th>OLS</th>
<th>OLS2</th>
<th>AVE</th>
<th>AVE2</th>
<th>Fore1</th>
<th>Fore2</th>
<th>ForeA</th>
<th>ForeB</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>1.00</td>
<td>1.03</td>
<td>1.00</td>
<td>0.22</td>
<td>1.00</td>
<td>0.27</td>
<td>1.00</td>
<td>0.39</td>
</tr>
<tr>
<td>0.2</td>
<td>1.00</td>
<td>0.08</td>
<td>1.00</td>
<td>0.21</td>
<td>1.00</td>
<td>0.29</td>
<td>1.00</td>
<td>0.43</td>
</tr>
<tr>
<td>0.3</td>
<td>0.43</td>
<td>0.01</td>
<td>0.82</td>
<td>0.10</td>
<td>0.86</td>
<td>0.18</td>
<td>0.92</td>
<td>0.38</td>
</tr>
<tr>
<td>0.4</td>
<td>0.13</td>
<td>0.00</td>
<td>0.26</td>
<td>0.05</td>
<td>0.57</td>
<td>0.13</td>
<td>0.57</td>
<td>0.33</td>
</tr>
<tr>
<td>0.5</td>
<td>0.00</td>
<td>0.00</td>
<td>0.56</td>
<td>0.03</td>
<td>0.16</td>
<td>0.11</td>
<td>0.38</td>
<td>0.31</td>
</tr>
<tr>
<td>0.6</td>
<td>0.12</td>
<td>0.00</td>
<td>0.26</td>
<td>0.05</td>
<td>0.38</td>
<td>0.13</td>
<td>0.58</td>
<td>0.32</td>
</tr>
<tr>
<td>0.7</td>
<td>0.81</td>
<td>0.02</td>
<td>0.83</td>
<td>0.09</td>
<td>0.87</td>
<td>0.18</td>
<td>0.92</td>
<td>0.46</td>
</tr>
<tr>
<td>0.8</td>
<td>1.00</td>
<td>0.07</td>
<td>1.00</td>
<td>0.17</td>
<td>1.00</td>
<td>0.24</td>
<td>1.00</td>
<td>0.42</td>
</tr>
<tr>
<td>0.9</td>
<td>1.00</td>
<td>0.12</td>
<td>1.00</td>
<td>0.20</td>
<td>1.00</td>
<td>0.26</td>
<td>1.00</td>
<td>0.37</td>
</tr>
</tbody>
</table>

Kurtosis mixture:

<table>
<thead>
<tr>
<th>$\delta$</th>
<th>OLS</th>
<th>OLS2</th>
<th>AVE</th>
<th>AVE2</th>
<th>Fore1</th>
<th>Fore2</th>
<th>ForeA</th>
<th>ForeB</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>1.00</td>
<td>0.01</td>
<td>1.00</td>
<td>0.05</td>
<td>1.00</td>
<td>0.07</td>
<td>1.00</td>
<td>0.20</td>
</tr>
<tr>
<td>0.2</td>
<td>1.00</td>
<td>0.01</td>
<td>1.00</td>
<td>0.05</td>
<td>0.99</td>
<td>0.06</td>
<td>1.00</td>
<td>0.29</td>
</tr>
<tr>
<td>0.3</td>
<td>0.81</td>
<td>0.00</td>
<td>0.83</td>
<td>0.05</td>
<td>0.84</td>
<td>0.08</td>
<td>0.95</td>
<td>0.34</td>
</tr>
<tr>
<td>0.4</td>
<td>0.19</td>
<td>0.00</td>
<td>0.29</td>
<td>0.06</td>
<td>0.32</td>
<td>0.09</td>
<td>0.64</td>
<td>0.29</td>
</tr>
<tr>
<td>0.5</td>
<td>0.01</td>
<td>0.00</td>
<td>0.99</td>
<td>0.06</td>
<td>0.12</td>
<td>0.09</td>
<td>0.47</td>
<td>0.40</td>
</tr>
<tr>
<td>0.6</td>
<td>0.83</td>
<td>0.01</td>
<td>0.23</td>
<td>0.05</td>
<td>0.33</td>
<td>0.08</td>
<td>0.62</td>
<td>0.38</td>
</tr>
<tr>
<td>0.7</td>
<td>0.84</td>
<td>0.01</td>
<td>0.83</td>
<td>0.05</td>
<td>0.84</td>
<td>0.07</td>
<td>0.92</td>
<td>0.35</td>
</tr>
<tr>
<td>0.8</td>
<td>0.01</td>
<td>0.01</td>
<td>0.99</td>
<td>0.05</td>
<td>1.00</td>
<td>0.07</td>
<td>1.00</td>
<td>0.28</td>
</tr>
<tr>
<td>0.9</td>
<td>1.00</td>
<td>0.01</td>
<td>1.00</td>
<td>0.06</td>
<td>1.00</td>
<td>0.08</td>
<td>1.00</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Note: The first column reports the value of the parameter, $\delta$, controlling skewness. Other entries give the percentage of rejections, of one-sided Dickey-Fuller tests (5% significance level) of the optimal combination vs. the corresponding entry. The tests are based on estimates of the forecast combination weights and the constant from 100 observations and 5,000 Monte Carlo replications. OLS refers to the estimation of the combination weights by OLS; AVE refers to equal-weighted forecast combination. Fore refers to forecasts $t$ steps. Names ending with a refer to the corresponding estimation technique for the weights followed by optimal estimation of the constant.

estimator as the base model and testing the null that the alternate methods give similar results against the one-sided alternative that the other methods are worse. A result of 0.05 indicates that other methods perform equally well, a larger frequency of rejection suggests that the base estimator does better and a number lower than 0.05 suggests that it does better for less than the expected number of cases under the null. In most cases the OLS2 method produces results around zero suggesting either that we would prefer the OLS2 method if we used this test or (far more likely) that the test is not sufficiently powerful to distinguish between the forecasts from the two methods. The exception is when there is extreme loss asymmetry. The AVE2 method tends also to do well, especially under asymmetry. But for the most asymmetric cases it is found to be worse by this metric. Part of the reason for this is that the AVE2 forecasts are less variable than the other forecasts and thus contribute less to the standard error. Hence when this method is better it is more likely to be statistically significantly so. These results, as we expect from the theory in Section 3, are more pronounced for the skewed mixture than for the elliptically symmetric kurtotic model.

7.2. Biases in the estimators

Table 5 shows both population values and the bias arising from various estimation methods under lin-lin loss. For the other loss functions results were again similar. The first column shows the value assumed for the asymmetry parameter in the loss function ($\delta$). The next three columns show the population values of the optimal weights. Then follows the bias in the estimates from the preferred estimation method and the bias arising from ordinary least squares estimation. The final columns report the bias in the constant term when the forecast combination weights are estimated by OLS or by averaging followed by optimal estimation of the constant.

For the lin-lin loss function, the quantile regression method only introduces very minor biases in the estimated forecast combination weights and the constant term, irrespective of whether the data generating process is the skewed or kurtotic mixture model and even under very large degrees of asymmetry in the loss
function. The OLS estimator, optimal when \( \theta = 0.5 \), is quite heavily biased even for small departures from symmetric loss. The biases in the constant are mitigated to a large extent when estimating the forecast combination weights by OLS followed by quantile regression of the constant. Nevertheless, the estimates have a bias of around 10% for the most asymmetric loss functions and skewed data.

These simulations show that, on the whole, the simple two-step procedure that estimates the combination weights by OLS followed by quantile regression tends to work very well, provided that the degree of asymmetry in the loss function is not very large. This procedure is of course very simple to use in practice and eliminates most of the loss from using the simple OLS estimator to estimate both the combination weights and the intercept.

These results also carry over to data that simultaneously display skew and kurtosis. Table 6 shows results for a mixture model that sets \( \rho_1 = \rho_2 = 0.5 \) has the same covariance matrices as in the kurtosis mixture model, zero means in state 1 and means of \(-0.25, 0.25\) in state 2. This model matches more closely the moments of the empirical data in the next section. We report results for the asymmetric quadratic loss function but the results are very similar for linear and pin–pin loss. OLS continues to generate substantial biases in the parameter estimates under asymmetric loss and this leads to a large increase in the expected loss. In contrast the weighted least squares estimator or the bias–adjusted least squares method produce only small biases and relatively modest values of the expected loss.

### 8. Empirical application to inflation forecasting

To demonstrate the theoretical ideas developed so far, we consider an application that combines predictions of changes in the consumer price index (CPI) from a simple autoregressive time-series model with survey predictions. The source of the latter is the Livingston Survey data base maintained by the Federal Reserve Bank of Philadelphia. This provides a time series of predictions of the consumer price index six months ahead in time over the period 1946Q1 to 2001Q4. All forecasts are thus one-step ahead for a forecast horizon of 6 months. The forecasts are likely to be based on diverse information sets that comprise a much larger set of public and private information than is typically considered in econometric models.

We combine the aggregate survey predictions\(^{22}\) with predictions from a simple autoregressive forecasting model. Schwarz’s information criterion supported a simple first-order autoregressive (AR(1)) model which we use to predict the inflation rate—defined as the log-difference in the CPI, \( \Delta CPI \) —one-step, or six-months, ahead. One-step-ahead forecasts from the autoregressive model are computed using recursively updated parameter estimates.\(^{23}\)

The estimated correlation between the forecast errors from the autoregressive model and the Livingston survey data is 0.63 over the full sample. While high, this estimate only reflects the linear correlation between the two sets of forecast errors. Inspection of the density of the forecast errors from the AR(1) model and the Livingston survey data suggested that the two forecast error distributions have more probability mass in the center and in the tails than a Gaussian distribution, thus indicating leptokurtosis.

---

\(^{22}\) It can be difficult to interpret the absolute size of the bias in these simulations as it depends on the scaling of the problem which is arbitrary. We use the terms “small” and “large” bias in a relative sense, i.e., relative to the biases observed in the other simulations.

\(^{23}\) This reduces to a simple LAD estimation problem.

\(^{24}\) This survey data has previously been used in numerous studies. For a comprehensive list of these, see the web site maintained by the Federal Reserve Bank of Philadelphia.

\(^{25}\) These are computed as the arithmetic mean of the individual forecasters’ predictions.
The density generated by the autoregressive model also has more probability mass in the right shoulder and tail than the density associated with the Livingston data which has a thicker left shoulder.

The two sets of forecasts thus contain very different information. For small values of \( \theta \) the Livingston forecasts were found to strongly outperform the AR(1) forecasts. As \( \theta \) gets larger, this difference declines. When \( \theta \) is 0.7 or larger the AR(1) forecasts outperform the Livingston forecasts in terms of the average loss over the sample.

To explore the importance the loss function plays in determining the optimal combination weights, we varied the asymmetry parameter, \( \theta \), between 0.1 and 0.9. This moves us from strong aversion against large positive forecast errors via a symmetric loss function towards strong aversion against large negative forecast errors. The optimal combination weights under lin-lin loss are shown in the upper window of Fig. 3. Under standard, symmetric loss (\( \theta = 0.5 \)), the optimal weight on the AR(1) forecast is close to zero. However as the loss function becomes increasingly asymmetric (in either direction), it becomes optimal to put a non-zero weight on the time-series forecast. For small values of \( \theta \) a positive weight is put on both the AR(1) forecasts and the density forecasts. For large values of \( \theta \), the optimal weight on the AR(1) forecasts becomes negative while the weight on the Livingston data increases. Strong aversion against positive forecast errors thus means assigning positive combination weights to both the autoregressive and Livingston predictions. Under strong aversion against negative forecast errors the combination weight on the autoregressive model is strongly negative while the weight on the Livingston prediction is strongly positive. The lower window of Fig. 3 repeats this exercise for the asymmetric quadratic loss function. The actual estimates for the weights differ somewhat but the story is basically the same as under lin-lin loss.

One way to understand the negative weight on the forecasts from the AR(1) model when \( \theta \) is large is by considering scenarios under which large negative forecast errors occur. The tail correlation between the autoregressive and the Livingston forecast errors is greater than one because of the correlations in the center of the forecast error distribution. To avoid large negative forecast errors (which occur under positive combination weights if the two forecast errors are simultaneously negative or under negative weights if the forecast errors are simultaneously positive), the optimal weights have opposite signs. Combinations with weights of opposite signs are similar to setting up a spread in a financial trading strategy, a practice commonly used when return correlations are higher the more volatile markets are. Notice, however, that while large positive forecast errors are more common for the AR(1) model than in the Livingston data, the positive weight on the Livingston forecast is almost twice as large as the negative weight on the AR(1) forecast.

Since the forecast combination weights are very sensitive to the parameters of the loss function, we expect that the distribution of the forecast errors (and indeed the individual forecasts as well as the change in inflation) are far from symmetric. In Table 7 we provide descriptive statistics for the first four moments of the various forecasts and also test for skew and kurtosis for each of the underlying variables as well as the forecast errors for three values of \( \theta \), namely \( \theta = 0.1, 0.5, 0.9 \). These tests merely serve the purpose of characterizing the distribution of the various forecasts. As expected we strongly reject the null of no skew for the individual forecasts and the actual time series while this null is only rejected for the forecast error distribution.
Table 8
Average losses from combined and equal weighted forecasts

<table>
<thead>
<tr>
<th>WLS estimates</th>
<th>Losses</th>
<th>OLSc</th>
<th>AVEc</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \theta )</td>
<td>( w_1 )</td>
<td>( w_2 )</td>
<td>Constant</td>
</tr>
<tr>
<td>0.1</td>
<td>0.594</td>
<td>0.597</td>
<td>0.006</td>
</tr>
<tr>
<td>0.5</td>
<td>-0.001</td>
<td>0.824</td>
<td>0.604</td>
</tr>
<tr>
<td>0.9</td>
<td>-0.428</td>
<td>0.926</td>
<td>0.002</td>
</tr>
</tbody>
</table>

\( \theta \) | \( \omega^{A1} \) | \( \omega^{A2} \) | \( \omega^{B1} \) | \( \omega^{B2} \) | WLS | Equal weights |
| 0.1 | 0.53 | 0.56 | 0.00 | -0.27 | 0.87 | 0.027 | 0.033 |
| 0.5 | 0.51 | -0.02 | 0.30 | -0.06 | 0.61 | 0.052 | 0.045 |
| 0.9 | 0.23 | 0.53 | 0.48 | -0.10 | 0.60 | 0.026 | 0.057 |

*Note:* WLS estimates are estimates of the parameters using the WLS method described for the symmetric quadratic loss function with asymmetry parameter \( \theta \). The loss reported is average out of sample one step ahead loss. In the OLSc case OLS weights are used for the forecast combination weights but the constant is estimated using WLS. AVEc gives results for even weights on the forecasts and a constant estimated using WLS. The second panel is based on combinations of the autoregressive (AR) and disaggregated academic (A1, bank (B1) and industry (I) forecasts.

when \( \theta = 0.1 \) and a strong negative skew emerges. Notice also the shift from negative to positive skew as \( \theta \) increases from 0.1 to 0.9.24

Finally, an implication of the Monte Carlo results in Section 7 is that simply getting the constant correct (imposing the correct bias) is a major part of the attempt to obtain combined forecasts that perform relatively well. In the numerical work we confirmed that simply using the OLS weights on the actual forecasts with the ‘optimal’ method for estimating the constant entailed only minor additional losses. The top panel of Table 8 examines this possibility with the inflation rate data using out-of-sample forecast errors under asymmetric quadratic loss. Except when \( \theta = 0.9 \), the average loss from using OLS forecast combination weights and WLS to estimate the constant is very similar to the loss based on the iterative WLS estimates. When \( \theta = 0.9 \) the WLS estimation method does lead to better results than both the OLS and the averaging methods followed by adjustment.

8.1. Disaggregated survey predictions

The Livingston survey data lists predictions according to the forecasters’ broad affiliation. The most frequent affiliations are nonfinancial business (‘industry’), academic institution (‘academic’) and commercial and investment banking (‘banking’). Forecasters from the same area of business are perhaps more likely to use the same information sets and/or models, so we now consider how the weights on these disaggregated forecasts depend on the loss function. We compute averages of the optimal combination weights on the AR(1) predictions and the three survey predictions as a function of \( \theta \). All predictions are out-of-sample and the superscripts \( AR, A,B,I \) refer to the autoregressive, academic, banking and industry predictions. The bottom panel of Table 8 shows the outcome. When the loss function assigns a very small weight to large negative forecast errors (\( \theta = 0.1 \)), the academic forecasts get a zero weight. Otherwise this weight is large and positive. The weight on the ‘banking’ inflation forecasts is always negative and tends to be small. Most weight is put on the forecasts produced by professionals in nonfinancial business. These consistently receive a weight above 0.5 irrespective of the shape of the loss function. The weight on the autoregressive forecast does not appear to be much affected by using disaggregate rather than aggregate survey data. It continues to be large and positive when \( \theta = 0.1 \) and large and negative when \( \theta = 0.9 \).

We finally investigated whether the common practice of selecting the optimal weights based on the MSE loss (\( \theta = 0.5 \)) of the forecasts produced by independent forecasters. Our results—reported in the bottom panel of Table 8—confirm the frequent finding that equal-weights outperformed estimated optimal weights under MSE loss (\( \theta = 0.5 \)). However, they also show very clearly that this result is overturned under asymmetric loss where use of estimated optimal weights leads to far smaller average losses out-of-sample.

In an analysis of forecasts of GNP/GDP growth from the Blue Chip Economic Indicators, Laster et al. (1999) find evidence that independent forecasters are more likely to produce ‘extreme’ forecasts that differ most from the consensus or median forecast. They propose an economic model where such behaviour is rational. Forecasters attract two types of clients, namely intensive users (who demand consistently precise forecasts) and occasional users that select last period’s star forecaster, i.e. the forecaster whose prediction was closest to the outcome. Producers of forecasts have an incentive to attract new occasional use clients. They can only do this by being more accurate in their prediction than other forecasters. This gives an incentive to generate extreme forecasts. The incentive is stronger, the more the forecaster relies on occasional use clients. Differences in opinions across forecasters may thus be driven by forecasters with different ratios of occasional and intensive forecast users. This story is consistent with our framework where different forecasts are driven by different information sets or by different loss functions. In fact, if the forecaster producers act as agents for forecast users, asymmetric loss could well be caused by differences across forecasters in the ratios of occasional and intensive use clients.

9. Conclusion

Several conclusions emerge from our analysis. If the forecast error distribution is elliptically symmetric then the forecast combination weights are identical for a wide range of loss functions, including the popular mean squared error loss function. This implies that if a particular forecast is informative under MSE loss, then this forecast will be informative for a very wide range of loss functions. If there is little departure from the mean of the error distribution, then the combination weights are close to the optimal combination weights.
from elliptical symmetry, then the optimal forecast combination weights are near the 
standard weights derived under MSE loss. This has practical implications since such 
weights are often much easier to compute. 
Under departures from elliptically symmetric distributions, the optimal combination 
weights under asymmetric loss can be very different than under MSE loss. For example, 
if negative forecast errors are associated with much lower losses than positive ones, 
then it will be optimal to select combination weights that give rise to a much larger 
upward bias than under MSE loss.
Irrespective of the distribution of the forecast errors, the constant term in the forecast 
combination is affected by the loss function. Our simulations suggest that even with 
heavily skewed forecast errors, the loss from using MSE combination weights (which 
are suboptimal) and optimally adjusting the constant results in fairly small increases 
in expected loss. Since this strategy is simple and has good small sample properties, 
for many applications this approach will perhaps work well.

Our empirical application showed that the optimal weights in a combination of infla-
tion survey forecasts and forecasts from a simple autoregressive model strongly depend 
on the degree of asymmetry in the loss function. In the absence of loss asymmetry, 
the autoregressive forecast does not add much information. However, under asym-
metric loss (in either direction), both sets of forecasts appear to contain information 
and have non-zero weights in the combined forecast. The application confirmed the 
frequent finding that equal-weights outperform estimated optimal weights under MSE 
loss. However, it also showed very clearly that this result can be overturned under 
asymmetric loss where use of estimated optimal weights led to much smaller average 
losses out-of-sample.
There are many directions for generalizing our results. While we do not study 
time-variability in higher order conditional moments, it is possible to extend our frame-
work to include such effects. Elliott and Timmermann (2002) consider a framework 
with time-varying mixes of Gaussian distributions that generate time-varying higher 
order moments. First order conditions for the optimal combination weights can be 
derived under a variety of loss functions. In general these depend on the full set of pa-
rameters that determine the unconditional density. In this framework the optimal intercept 
and combination weights become time-varying. Many of the ideas in this paper can 
also be applied to multi-step-ahead forecasting. Some GARCH models imply that the 
one-step-ahead forecast errors are elliptically symmetric, while multi-step-ahead fore-
cast errors are not. This raises the possibility of letting the estimation method depend 
on the forecast horizon.
Another possibility is to adopt the Bayesian methods to forecast combination that 
have been developed by Palm and Zellner (1992), including shrinkage methods. There 
are good theoretical reasons to consider such methods. Non-Bayesian estimators need 
not be admissible in cases where Bayesian methods may have this property. Sim-
ilarly, Bayesian model averaging can be shown to provide better predictive ability 
than individual models along a logarithmic scoring rule. These points would suggest 
potentially important gains from using Bayesian methods. In their extensive analysis 
of a wide range of prediction models for international growth rates Min and Zellner 
(1993) did not find much evidence that combination methods improved MSE 
performance. However they did find that the best Bayesian combination methods dom-
inated their non-Bayesian counterparts. Similarly, recent work by Raftery et al. (1997) 
finds promising gains from Bayesian model averaging techniques in cases with many 
predictors. In future work we plan to extend these methods to cover asymmetric loss 
functions such as the ones considered in this paper.

Acknowledgements
We wish to thank three anonymous referees, an associate editor and the editor, 
Arnold Zellner, for making many constructive suggestions. Seminar participants at Boc-
coni and Emory University and the 2001 NBER conference on macroeconomics and 
forecasting, particularly Frank Diebold, Carlo Favero, Massimiliano Marcellino and 
Paul Ruud, also provided helpful comments. Carlos Capizzano provided excellent re-
search assistance. We are grateful for financial support from the NSF under grant SES 
0111238.

References
451–468.
Journal of Economic Theory 29, 185–201.
Spanish Economic Review 1, 91–121.
806–817.
Forecasting 5, 559–581.
Statistics 6, 105–111.
Master’s Thesis, University of California, San Diego.
In: Clements, M.P., Hendry, D.F. (Eds.), A Companion to Economic Forecasting. Blackwell Publishers, 
20, 199–207.
1, 161–173.


Comparing Predictive Accuracy

Francis X. Diebold
Department of Economics, University of Pennsylvania, Philadelphia, PA 19104-6297, and National Bureau of Economic Research, Cambridge, MA 02138

Roberto S. Mariano
Department of Economics, University of Pennsylvania, Philadelphia, PA 19104-6297

We propose and evaluate explicit tests of the null hypothesis of no difference in the accuracy of two competing forecasts. In contrast to previously developed tests, a wide variety of accuracy measures can be used (in particular, the loss function need not be quadratic and need not even be symmetric), and forecast errors can be non-Gaussian, nonzero mean, serially correlated, and contemporaneously correlated. Asymptotic and exact finite-sample tests are proposed, evaluated, and illustrated.

KEY WORDS: Economic loss function; Exchange rates; Forecast evaluation; Forecasting; Nonparametric tests; Sign test.

Prediction is of fundamental importance in all of the sciences, including economics. Forecast accuracy is of obvious importance to users of forecasts because forecasts are used to guide decisions. Forecast accuracy is also of obvious importance to producers of forecasts, whose reputations (and fortunes) rise and fall with forecast accuracy. Comparisons of forecast accuracy are also of importance to economists more generally who are interested in discriminating among competing economic hypotheses (models). Predictive performance and model adequacy are intrinsically linked—predictive failure implies model inadequacy.

Given the obvious desirability of a formal statistical procedure for forecast-accuracy comparisons, one is struck by the casual manner in which such comparisons are typically carried out. The literature contains literally thousands of forecast-accuracy comparisons; almost without exception, point estimates of forecast accuracy are examined, with no attempt to assess their sampling uncertainty. On reflection, the reason for the casual approach is clear: Correlation of forecast errors across space and time, as well as several additional complications, makes formal comparison of forecast accuracy difficult. Dhrymes et al. (1972) and Howrey, Klein, and McCarthy (1974), for example, offered pessimistic assessments of the possibilities for formal testing.

In this article we propose widely applicable tests of the null hypothesis of no difference in the accuracy of two competing forecasts. Our approach is similar in spirit to that of Vuong (1989) in the sense that we propose methods for measuring and assessing the significance of divergences between models and data. Our approach, however, is based directly on predictive performance, and we entertain a wide class of accuracy measures that users can tailor to particular decision-making situations. This is important because, as is well known, realistic economic loss functions frequently do not conform to stylized textbook favorites like mean squared prediction error (MSPE). [For example, Leitch and Tanner (1991) and Chinn and Meese (1991) stressed direction of change, Cumby and Modest (1987) stressed market and country timing, McCulloch and Rossi (1990), and West, Edison, and Cho (1993) stressed utility-based criteria, and Clements and Hendry (1993) proposed a new accuracy measure, the generalized forecast-error second moment.] Moreover, we allow for forecast errors that are potentially non-Gaussian, nonzero mean, serially correlated, and contemporaneously correlated.

We proceed by detailing our test procedures in Section 1. Then, in Section 2, we review the small extant literature to provide necessary background for the finite-sample evaluation of our tests in Section 3. In Section 4 we provide an illustrative application, and in Section 5 we offer conclusions and directions for future research.

1. TESTING EQUALITY OF FORECAST ACCURACY

Consider two forecasts, \( \{ \hat{y}_n \}_{n=1}^T \) and \( \{ \tilde{y}_n \}_{n=1}^T \), of the time series \( \{ y_n \}_{n=1}^T \). Let the associated forecast errors be \( \{ e_n \}_{n=1}^T \) and \( \{ \tilde{e}_n \}_{n=1}^T \). We wish to assess the expected loss associated with each of the forecasts (or its negative, accuracy). Of great importance, and almost always ignored, is the fact that the economic loss associated with a forecast may be poorly assessed by the usual statistical metrics. That is, forecasts are used to guide decisions, and the loss associated with a forecast error of a particular sign and size is induced directly by the nature of the decision problem at hand. When one considers the variety of decisions undertaken by economic agents guided by forecasts (e.g., risk-hedging decisions, inventory-stocking decisions, policy decisions, advertising-expenditure decisions, public-utility rate-setting decisions, etc.), it is clear that the loss associated with a particular forecast error is in general an asymmetric function of the error and, even if symmetric, certainly need not conform to stylized textbook examples like MSPE.
Thus, we allow the time-\(t\) loss associated with a forecast (say \(\tilde{y}_n\)) to be an arbitrary function of the realization and prediction, \(g(y_n, \tilde{y}_n)\). In many applications, the loss function will be a direct function of the forecast error; that is, \(g(y_n, \tilde{y}_n) = g(e_n)\). To economize on notation, we write \(g(e_n)\) from this point on, recognizing that certain loss functions (like direction-of-change) do not collapse to \(g(e_n)\) form, in which case the full \(g(y_n, \tilde{y}_n)\) form would be used. The null hypothesis of equal forecast accuracy for two forecasts is \(E[g(e_n)] = E[g(e_p)]\), or \(E[d] = 0\), where \(d_i = [g(e_n) - g(e_p)]\) is the loss differential. Thus, the “equal accuracy” null hypothesis is equivalent to the null hypothesis that the population mean of the loss-differential series is 0.

1.1 An Asymptotic Test

Consider a sample path \(\{d_i\}_{i=1}^T\) of a loss-differential series. If the loss-differential series is covariance stationary and short memory, then standard results may be used to deduce the asymptotic distribution of the sample mean loss differential. We have

\[
\sqrt{T}(d - \mu) \xrightarrow{d} N(0, 2\pi f_d(0)),
\]

where

\[
d = \frac{1}{T} \sum_{i=1}^T (g(e_n) - g(e_p))
\]

is the sample mean loss differential,

\[
f_d(0) = \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} \gamma_d(\tau)
\]

is the spectral density of the loss differential at frequency 0, \(\gamma_d(\tau) = E[(d_i - \mu)(d_{i-\tau} - \mu)]\) is the autocovariance of the loss differential at displacement \(\tau\), and \(\mu\) is the population mean loss differential. The formula for \(f_d(0)\) shows that the correction for serial correlation can be substantial, even if the loss differential is only weakly serially correlated, due to cumulation of the autocovariance terms.

Because in large samples the sample mean loss differential \(\bar{d}\) is approximately normally distributed with mean \(\mu\) and variance \(2\pi f_d(0)/T\), the obvious large-sample \(N(0, 1)\) statistic for testing the null hypothesis of equal forecast accuracy is

\[
S_1 = \frac{\bar{d}}{\sqrt{2\pi f_d(0)/T}},
\]

where \(\hat{f}_d(0)\) is a consistent estimate of \(f_d(0)\).

Following standard practice, we obtain a consistent estimate of \(2\pi \hat{f}_d(0)\) by taking a weighted sum of the available sample autocovariances,

\[
2\pi \hat{f}_d(0) = \sum_{\tau=T-1}^{T-1} \left( \frac{\tau}{S(T)} \right) \hat{\gamma}_d(\tau),
\]

where

\[
\hat{\gamma}_d(\tau) = \frac{1}{T} \sum_{i=\tau+1}^{T} (d_i - \bar{d})(d_{i-\tau} - \bar{d}),
\]

1(\(\tau/S(T)\)) is the lag window, and \(S(T)\) is the truncation lag.

To motivate a choice of lag window and truncation lag that we have often found useful in practice, recall the familiar result that optimal \(k\)-step-ahead forecast errors are at most \((k-1)\)-dependent. In practical applications, of course, \((k-1)\)-dependence may be violated for a variety of reasons. Nevertheless, it seems reasonable to take \((k-1)\)-dependence as a reasonable benchmark for a \(k\)-step-ahead forecast error (and the assumption may be readily assessed empirically). This suggests the attractiveness of the uniform, or rectangular, lag window, defined by

\[
1 \left( \frac{\tau}{S(T)} \right) = \begin{cases} 1 & \text{for } \left| \frac{\tau}{S(T)} \right| \leq 1 \\ 0 & \text{otherwise.} \end{cases}
\]

\((k-1)\)-dependence implies that only \((k-1)\) sample autocovariances need be used in the estimation of \(f_d(0)\) because all the others are 0, so \(S(T) = (k-1)\). This is legitimate (i.e., the estimator is consistent) under \((k-1)\)-dependence so long as a uniform window is used because the uniform window assigns unit weight to all included autocovariances.

Because the Dirichlet spectral window associated with the rectangular lag window dips below 0 at certain locations, the resulting estimator of the spectral density function is not guaranteed to be positive semidefinite. The large positive weight near the origin associated with the Dirichlet kernel, however, makes it unlikely to obtain a negative estimate of \(f_d(0)\). In applications, in the rare event that a negative estimate arises, we treat it as 0 and automatically reject the null hypothesis of equal forecast accuracy. If it is viewed as particularly important to impose nonnegativity of the estimated spectral density, it may be enforced by using a Bartlett lag window, with corresponding nonnegative Fejer spectral window, as in the work of Newey and West (1987), at the cost of having to increase the truncation lag "appropriately" with sample size. Other lag windows and truncation lag selection procedures are of course possible as well. Andrews (1991), for example, suggested using a quadratic spectral lag window, together with a "plug-in" automatic bandwidth selection procedure.

1.2 Exact Finite-Sample Tests

Sometimes only a few forecast-error observations are available in practice. One approach in such situations is to bootstrap our asymptotic test statistic, as done by Mark (1995). Ashley’s (1994) work is also very much in that spirit. Little is known about the first-order asymptotic validity of the bootstrap in this situation, however, let alone higher-order asymptotics or actual finite-sample performance. Therefore, it is useful to have available exact finite-sample tests of predictive accuracy, to complement the asymptotic test presented previously. Two powerful such tests are based on the observed loss differentials (the sign test) or their ranks (Wilcoxon’s signed-rank test). [These tests are standard, so our discussion is terse. See, for example, Lehmann (1975) for details.]

1.2.1 The Sign Test. The null hypothesis is a zero-median loss differential: \(\text{med}(g(e_n) - g(e_p)) = 0\). Note that the null of a zero-median loss differential is not the same
as the null of zero difference between median losses; that is, 

$$
\text{med}(g(e_a) - g(e_b)) \neq \text{med}(g(e_a)) - \text{med}(g(e_b)).
$$

For that reason, the null differs slightly in spirit from that associated with our earlier discussed asymptotic test statistic $S_1$, but it nevertheless has an intuitive and meaningful interpretation—namely, that $P(g(e_a) > g(e_b)) = P(g(e_a) < g(e_b))$.

If, however, the loss differential is symmetrically distributed, then the null hypothesis of a zero-median loss differential corresponds precisely to the earlier null because in that case the median and mean are equal. Symmetry of the loss differential will obtain, for example, if the distributions of $g(e_a)$ and $g(e_b)$ are the same up to a location shift. Symmetry is ultimately an empirical matter and may be assessed using standard procedures. We have found roughly symmetric loss-differential series to be quite common in practice.

The construction and intuition of a test statistic are straightforward. Assuming that the loss-differential series is iid (and we shall relax that assumption shortly), the number of positive loss-differential observations in a sample of size $T$ has the binomial distribution with parameters $T$ and $\frac{1}{2}$ under the null hypothesis. The test statistic is therefore simply

$$
S_2 = \sum_{i=1}^{T} I_i(d_i),
$$

where

$$
I_i(d_i) = 1 \quad \text{if } d_i > 0
$$

$$
0 \quad \text{otherwise.}
$$

Significance may be assessed using a table of the cumulative binomial distribution. In large samples, the studentized version of the sign-test statistic is standard normal:

$$
S_{2a} = \frac{S_2 - \frac{.5T}{2}}{\sqrt{\frac{T(T+1)}{12}}} \sim N(0, 1).
$$

1.2.2 Wilcoxon's Signed-Rank Test. A related distribution-free procedure that requires symmetry of the loss differential (but can be more powerful than the sign test in that case) is Wilcoxon's signed-rank test. We again assume for the moment that the loss-differential series is iid. The test statistic is

$$
S_1 = \sum_{i=1}^{T} I_i(d_i) \text{rank}(|d_i|),
$$

the sum of the ranks of the absolute values of the positive observations. The exact finite-sample critical values of the test statistic are invariant to the distribution of the loss differential—it need be only zero-mean and symmetric—and have been tabulated. Moreover, its studentized version is asymptotically standard normal,

$$
S_{1a} = \frac{S_1 - \frac{T(T+1)}{4}}{\sqrt{\frac{T(T+1)(2T+1)}{24}}} \sim N(0, 1).
$$

1.3 Discussion

Here we highlight some of the virtues and limitations of our tests. First, as we have stressed repeatedly, our tests are valid for a very wide class of loss functions. In particular, the loss function need not be quadratic and need not even be symmetric or continuous.

Second, a variety of realistic features of forecast errors are readily accommodated. The forecast errors can be nonzero-mean, non-Gaussian, and contemporaneously correlated. Allowance for contemporaneous correlation, in particular, is important because the forecasts being compared are forecasts of the same economic time series and because the information sets of forecasters are largely overlapping so that forecast errors tend to be strongly contemporaneously correlated.

Moreover, the asymptotic test statistic $S_1$ can of course handle a serially correlated loss differential. This is potentially important because, as discussed earlier, even optimal forecast errors are serially correlated in general. Serial correlation presents more of a problem for the exact finite-sample test statistics $S_1$ and $S_2$ and their asymptotic counterparts $S_{2a}$ and $S_{1a}$ because the elements of the set of all possible rearrangements of the sample loss differential series are not equally likely when the data are serially correlated, which violates the assumptions on which such randomization tests are based. Nevertheless, serial correlation may be handled via Bonferroni bounds, as suggested in a different context by Campbell and Ghyssels (1995). Under the assumption that the forecast errors and hence the loss differential are (k - 1)-dependent, each of the following k sets of loss differentials will be free of serial correlation: $\{d_{t+1,1}, d_{t+1,2}, \ldots, d_{t+1,k}\}, \{d_{t+2,1}, d_{t+2,2}, \ldots, d_{t+2,k}\}, \ldots, \{d_{t+k,1}, d_{t+k,2}, \ldots, d_{t+k,k}\}$. Thus, a test with size bounded by $\alpha$ can be obtained by performing $k$ tests, each of size $\alpha/k$, on each of the $k$ loss-differential sequences and rejecting the null hypothesis if the null is rejected for any of the $k$ samples. Finally, it is interesting to note that, in multistep forecast comparisons, forecast-error serial correlation may be a "common feature," in the terminology of Engle and Kozicki (1993), because it is induced largely by the fact that the forecast horizon is longer than the interval at which the data are sampled and may therefore not be present in loss differentials even if present in the forecast errors themselves. This possibility can of course be checked empirically.

2. EXTANT TESTS

In this section we provide a brief description of three existing tests of forecast accuracy that have appeared in the literature and will be used in our subsequent Monte Carlo comparison.

2.1 The Simple $F$ Test: A Naive Benchmark

If (1) loss is quadratic and (2) the forecast errors are (a) zero mean, (b) Gaussian, (c) serially uncorrelated, or (d) contemporaneously uncorrelated, then the null hypothesis of equal forecast accuracy corresponds to equal forecast error variances [by (1) and (2a)], and by (2b)–(2d), the ratio of sample variances has the usual $F$ distribution under the null hypothesis. More precisely, the test statistic

$$
F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}.
$$
is distributed as $F(T, T)$, where the forecast error series have been stacked into the $(T \times 1)$ vectors $\epsilon_t$ and $\delta_t$.

Test statistic $F$ is of little use in practice, however, because the conditions required to obtain its distribution are too restrictive. Assumption (2d) in particular is unpalatable for reasons discussed earlier. Its violation produces correlation between the numerator and denominator of $F$, which will not then have the $F$ distribution.

### 2.2 The Morgan–Granger–Newbold Test

The contemporaneous correlation problem led Granger and Newbold (1977) to apply an orthogonalizing transformation due to Morgan (1939–1940) that enables relaxation of Assumption (2d). Let $x_t = (e_{t1} + e_{p1})$ and $z_t = (e_{t2} - e_{p2})$, and let $x = (e_t + e_p)$ and $z = (e_t - e_p)$. Then, under the maintained Assumptions (1) and (2a)–(2c), the null hypothesis of equal forecast accuracy is equivalent to zero correlation between $x$ and $z$ (i.e., $\rho_{xz} = 0$) and the test statistic

$$MGN = \frac{\hat{\rho}_{xz}}{\sqrt{\frac{1}{T-1}}}$$

is distributed as Student's $t$ with $T - 1$ df, where

$$\hat{\rho}_{xz} = \frac{x'z}{\sqrt{(x'x)(z'z)}}$$

(e.g., see Hogg and Craig 1978, pp. 300–303).

Let us now consider relaxing the Assumptions (1) and (2a)–(2c) underlying the Morgan–Granger–Newbold (MGN) test. It is clear that the entire framework depends crucially on the assumption of quadratic loss (1), which cannot be relaxed. The remaining assumptions, however, can be weakened in varying degrees; we shall consider them in turn.

First, it is not difficult to relax the unbiasedness Assumption (2a), while maintaining Assumptions (1), (2b), and (2c). Second, the normality Assumption (2b) may be relaxed, while maintaining (1), (2a), and (2c), at the cost of substantial tedium involved with accounting for the higher-order moments that then enter the distribution of the sample correlation coefficient (e.g., see Kendall and Stuart 1979, chap. 26). Finally, the no-serial-correlation Assumption (2c) may be relaxed in addition to the no-contemporaneous-correlation Assumption (2d) while maintaining (1), (2a), and (2b), as discussed in Subsection 2.3.

### 2.3 The Meese–Rogoff Test

Under Assumptions (1), (2a), and (2b), Meese and Rogoff (1988) showed that

$$\sqrt{T} \hat{\gamma}_{zt} \overset{d}{\to} N(0, \Sigma),$$

where $\hat{\gamma}_{zt} = x'z / T$, $\Sigma = \sum_{r=-}\gamma_{zt}(r)\gamma_{zt}(r) + \gamma_{zt}(r)\gamma_{zt}(r)$, $\gamma_{zt}(r) = \text{cov}(x_t, z_{t-r})$, $\gamma_{zt}(r) = \text{cov}(z_t, x_{t-r})$, and $\gamma_{zt}(r) = \text{cov}(z_t, z_{t-r})$. This is a well-known result (e.g., Priestley 1981, pp. 692–693) for the distribution of the sample cross-covariance function, $\text{cov}(\hat{\gamma}_{zt}(s), \hat{\gamma}_{zt}(u))$, specialized to a displacement of 0.

A consistent estimator of $\Sigma$ is

$$\hat{\Sigma} = \sum_{r=-}^{s(T)} \left[ 1 - \frac{\tau}{T} \right] \left[ \hat{\gamma}_{zt}(r)\hat{\gamma}_{zt}(r) + \hat{\gamma}_{zt}(r)\hat{\gamma}_{zt}(r) \right],$$

where

$$\hat{\gamma}_{zt}(r) = \frac{1}{T} \sum_{t=r+1}^{T} x_t z_{t-r}, \quad \tau \geq 0$$

and the truncation lag $s(T)$ grows with the sample size but at a slower rate. Alternatively, following Diebold and Rudebusch (1991), one may use the closely related covariance matrix estimator,

$$\hat{\Sigma}^* = \sum_{r=-}^{s(T)} \left[ \hat{\gamma}_{zt}(r)\hat{\gamma}_{zt}(r) + \hat{\gamma}_{zt}(r)\hat{\gamma}_{zt}(r) \right].$$

Either way, the test statistic is

$$MR = \frac{\hat{\rho}_{zt}}{\sqrt{\hat{\Sigma}} / T},$$

Under the null hypothesis and the maintained Assumptions (1), (2a), and (2b), $MR$ (Meese–Rogoff) is asymptotically distributed as standard normal. It is easy to show that, if the null hypothesis and Assumptions (1), (2a), (2b), and (2c) are satisfied, then all terms in $\Sigma$ are 0 except $\gamma_{zt}(0)$ and $\gamma_{zt}(0)$ so that $MR$ coincides asymptotically with $\Sigma$. It is interesting to note also that reformulation of the test in terms of correlation rather than covariance would have enabled Meese and Rogoff to dispense with the normality assumption because the sample autocorrelations are asymptotically normal even for non-Gaussian time series (e.g., Brockwell and Davis 1992, pp. 221–222).

### 2.4 Additional Extensions

In Subsection 2.3, we considered relaxation of Assumptions (2a)–(2c), one at a time, while consistently maintaining Assumption (1) and consistently relaxing Assumption (2d). Simultaneous relaxation of multiple assumptions is possible within the MGN orthogonalizing transformation framework but much more tedious. The distribution theory required for joint relaxation of (2b) and (2c), for example, is complicated by the presence of fourth-order cumulants in the distribution of the the sample autocovariances, as shown, for example, by Hannan (1970, p. 209) and Mizrahi (1991). More importantly, however, any procedure based on the MGN orthogonalizing transformation is inextricably wed to the assumption of quadratic loss.
3. MONTE CARLO ANALYSIS

3.1 Experimental Design

We evaluate the finite-sample size of test statistics $F$, MGN, MR, $S_1$, $S_2$, $S_2a$, $S_3$, and $S_3a$ under the null hypothesis and various of the maintained assumptions. The design includes a variety of specifications of forecast-error contemporaneous correlation, forecast-error serial correlation, and forecast-error distributions. To maintain applicability of all test statistics for comparison purposes, we use quadratic loss; that is, the null hypothesis is an equality of MSPE's. We emphasize again, however, that an important advantage of test statistics $S_1$, $S_2$, $S_2a$, $S_3$, and $S_3a$ in substantive economic applications—and one not shared by the others—is their direct applicability to analyses with nonquadratic loss functions.

Consider first the case of Gaussian forecast errors. We draw realizations of the bivariate forecast-error process, $\{e_t, e_t\}_{t=1}^T$, with varying degrees of contemporaneous and serial correlation in the generated forecast errors. This is achieved in two steps. First, we build in the desired degree of contemporaneous correlation by drawing a $(2 \times 1)$ forecast error innovation vector $u_i$ from a bivariate standard normal distribution, $u_i \sim N(0, I_2)$, and then premultiplying by the Choleski factor of the desired contemporaneous innovation correlation matrix. Let the desired correlation matrix be

$$R = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \quad \rho \in [0, 1).$$

Then the Choleski factor is

$$P = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1 - \rho^2} \end{bmatrix}.$$

Thus, the transformed $(2 \times 1)$ vector $v_i = Pu_i \sim N(0_2, R)$. This operation is repeated $T$ times, yielding $\{v_{1t}, v_{2t}\}_{t=1}^T$.

Second, (moving average) MA(1) serial correlation (with parameter $\theta$) is introduced by taking

$$\begin{bmatrix} e_{1t} \\ e_{2t} \end{bmatrix} = \begin{bmatrix} \frac{1 - \theta}{\sqrt{1 - \theta^2}} & 0 \\ 0 & \frac{1 + \theta}{\sqrt{1 + \theta^2}} \end{bmatrix} \begin{bmatrix} v_{1t} \\ v_{2t} \end{bmatrix}, \quad t = 1, \ldots, T.$$

We use $\nu_0 = 0$. Multiplication by $(1 + \theta^2)^{-1/2}$ is done to keep the unconditional variance normalized to 1.

We consider sample sizes of $T = 8, 16, 32, 64, 128, 256$, and 512, contemporaneous correlation parameters of $\rho = 0, .5, .9$, and MA parameters of $\theta = .5, .9$. Simple calculations reveal that $\rho$ is not only the correlation between $v_1$ and $v_2$, but also the correlation between the forecast errors $e_1$ and $e_2$, so that varying the correlation of $v_1$ and $v_2$ through $[0, .9]$ effectively varies the correlation of the observed forecast errors through the same range.

We also consider non-Gaussian forecast errors. The design is the same as for the Gaussian case described previously but driven by fat-tailed variates $(u_{1t}', u_{2t}')$ [rather than $(u_{1t}, u_{2t})$], which are independent standardized $t$ random variables with 6 df. The variance of a $t(6)$ random variable is $3/2$. Thus, standardization amounts to dividing the $t(6)$ random variable by $\sqrt{3/2}$.

Throughout, we perform tests at the $\alpha = .1$ level. When using the exact sign and signed-rank tests, restriction of nominal size to precisely 10% is impossible (without introducing randomization), so we use the obtainable exact size closest to 10%, as specified in the tables. We perform at least 5,000 Monte Carlo replications. The truncation lag is set at 1, reflecting the fact that the experiment is designed to mimic the comparison of two-step-ahead forecast errors, with associated MA(1) structure.

3.2 Results

Results appear in Tables 1–6, which show the empirical size of the various test statistics in cases of Gaussian and non-Gaussian forecast errors as the degree of contemporaneous correlation, the degree of serial correlation, and sample size are varied.

Let us first discuss the case of Gaussian forecast errors. The results may be summarized as follows:

1. $F$ is correctly sized in the absence of both contemporaneous and serial correlation but is missized in the presence of either contemporaneous or serial correlation. Serial correlation pushes empirical size above nominal size, but contemporaneous correlation pushes empirical size drastically below nominal size. In combination, and particularly for large $\rho$ and $\theta$, contemporaneous correlation dominates and $F$ is undersized.

2. MGN is designed to remain unaffected by contemporaneous correlation and therefore remains correctly sized so long as $\theta = 0$. Serial correlation, however, pushes empirical size above nominal size.

3. As expected, MR is robust to contemporaneous and serial correlation in large samples, but it is oversized in small samples in the presence of serial correlation. The asymptotic distribution obtains rather quickly, however, resulting in approximately correct size for $T > 64$.

4. The behavior of $S_1$ is similar to that of MR. $S_1$ is robust to contemporaneous and serial correlation in large samples, but it is oversized in small samples, with nominal and empirical size converging a bit more slowly than for MR.

5. The Bonferroni bounds associated with $S_2$ and $S_3$ work well, with nominal and empirical size in close agreement throughout. Moreover, the asymptotics on which $S_{2a}$ and $S_{3a}$ depend obtain quickly.

Now consider the case of non-Gaussian forecast errors. The striking and readily apparent result is that $F$, MGN, and MR are drastically missized in large as well as small samples. $S_1$, $S_{2a}$, and $S_{3a}$, on the other hand, maintain approximately correct size for all but the very small sample sizes. In those cases, $S_1$ and $S_3$ continue to perform well. The results are well summarized by Figure 1, p. 261, which charts the dependence of $F$, MGN, MR, and $S_1$ on $T$ for the non-Gaussian case with $\rho = \theta = .5$. 
### Table 1. Empirical Size Under Quadratic Loss, Test Statistic $F$

<table>
<thead>
<tr>
<th>$T$</th>
<th>$\rho$</th>
<th>$\theta = .0$</th>
<th>$\theta = .5$</th>
<th>$\theta = .9$</th>
<th>$\theta = .0$</th>
<th>$\theta = .5$</th>
<th>$\theta = .9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>.0</td>
<td>9.85</td>
<td>12.14</td>
<td>14.10</td>
<td>14.28</td>
<td>15.76</td>
<td>17.21</td>
</tr>
<tr>
<td>8</td>
<td>.5</td>
<td>7.02</td>
<td>9.49</td>
<td>11.42</td>
<td>9.61</td>
<td>11.64</td>
<td>13.02</td>
</tr>
<tr>
<td>8</td>
<td>.9</td>
<td>.58</td>
<td>1.26</td>
<td>1.86</td>
<td>.57</td>
<td>1.13</td>
<td>1.79</td>
</tr>
<tr>
<td>16</td>
<td>.9</td>
<td>.47</td>
<td>.99</td>
<td>1.55</td>
<td>.34</td>
<td>.70</td>
<td>1.13</td>
</tr>
<tr>
<td>32</td>
<td>.0</td>
<td>9.88</td>
<td>12.68</td>
<td>14.34</td>
<td>18.06</td>
<td>19.55</td>
<td>20.35</td>
</tr>
<tr>
<td>32</td>
<td>.5</td>
<td>6.98</td>
<td>9.50</td>
<td>11.22</td>
<td>21.30</td>
<td>21.00</td>
<td>21.37</td>
</tr>
<tr>
<td>32</td>
<td>.9</td>
<td>.23</td>
<td>.55</td>
<td>1.00</td>
<td>.01</td>
<td>.07</td>
<td>.23</td>
</tr>
<tr>
<td>64</td>
<td>.0</td>
<td>9.71</td>
<td>13.05</td>
<td>14.62</td>
<td>29.84</td>
<td>29.72</td>
<td>29.96</td>
</tr>
<tr>
<td>64</td>
<td>.5</td>
<td>6.48</td>
<td>9.25</td>
<td>10.62</td>
<td>23.48</td>
<td>23.93</td>
<td>24.15</td>
</tr>
<tr>
<td>64</td>
<td>.9</td>
<td>.16</td>
<td>.47</td>
<td>.79</td>
<td>.02</td>
<td>.12</td>
<td>.29</td>
</tr>
<tr>
<td>128</td>
<td>.0</td>
<td>10.30</td>
<td>13.41</td>
<td>14.99</td>
<td>30.34</td>
<td>30.95</td>
<td>31.26</td>
</tr>
<tr>
<td>128</td>
<td>.5</td>
<td>7.01</td>
<td>10.13</td>
<td>11.64</td>
<td>24.89</td>
<td>25.01</td>
<td>25.16</td>
</tr>
<tr>
<td>128</td>
<td>.9</td>
<td>.16</td>
<td>.50</td>
<td>.74</td>
<td>.11</td>
<td>.44</td>
<td>.79</td>
</tr>
<tr>
<td>256</td>
<td>.0</td>
<td>10.01</td>
<td>13.05</td>
<td>14.65</td>
<td>31.07</td>
<td>31.12</td>
<td>31.24</td>
</tr>
<tr>
<td>256</td>
<td>.5</td>
<td>7.37</td>
<td>10.31</td>
<td>11.78</td>
<td>25.48</td>
<td>25.45</td>
<td>25.70</td>
</tr>
<tr>
<td>256</td>
<td>.9</td>
<td>.19</td>
<td>.51</td>
<td>.80</td>
<td>.51</td>
<td>1.13</td>
<td>1.44</td>
</tr>
<tr>
<td>512</td>
<td>.0</td>
<td>10.22</td>
<td>13.51</td>
<td>15.25</td>
<td>31.45</td>
<td>32.38</td>
<td>32.60</td>
</tr>
<tr>
<td>512</td>
<td>.5</td>
<td>7.53</td>
<td>10.16</td>
<td>11.49</td>
<td>26.35</td>
<td>26.92</td>
<td>16.95</td>
</tr>
<tr>
<td>512</td>
<td>.9</td>
<td>.18</td>
<td>.50</td>
<td>.85</td>
<td>.81</td>
<td>1.58</td>
<td>2.06</td>
</tr>
</tbody>
</table>

**Note:** $T$ is sample size, $\rho$ is the contemporaneous correlation between the innovations underlying the forecast errors, and $\theta$ is the coefficient of the MA(1) forecast error. All tests are at the 10% level. 10,000 Monte Carlo replications are performed.

### Table 2. Empirical Size Under Quadratic Loss, Test Statistic $MGN$

<table>
<thead>
<tr>
<th>$T$</th>
<th>$\rho$</th>
<th>$\theta = .0$</th>
<th>$\theta = .5$</th>
<th>$\theta = .9$</th>
<th>$\theta = .0$</th>
<th>$\theta = .5$</th>
<th>$\theta = .9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>.0</td>
<td>10.19</td>
<td>14.14</td>
<td>17.94</td>
<td>18.10</td>
<td>21.89</td>
<td>25.65</td>
</tr>
<tr>
<td>8</td>
<td>.5</td>
<td>9.96</td>
<td>14.66</td>
<td>18.61</td>
<td>16.00</td>
<td>20.51</td>
<td>24.19</td>
</tr>
<tr>
<td>8</td>
<td>.9</td>
<td>9.75</td>
<td>14.53</td>
<td>18.67</td>
<td>11.76</td>
<td>16.31</td>
<td>20.00</td>
</tr>
<tr>
<td>16</td>
<td>.0</td>
<td>10.07</td>
<td>14.34</td>
<td>17.54</td>
<td>20.33</td>
<td>24.54</td>
<td>27.08</td>
</tr>
<tr>
<td>16</td>
<td>.5</td>
<td>9.56</td>
<td>14.37</td>
<td>17.95</td>
<td>37.15</td>
<td>36.18</td>
<td>25.66</td>
</tr>
<tr>
<td>16</td>
<td>.9</td>
<td>10.02</td>
<td>14.70</td>
<td>18.20</td>
<td>12.01</td>
<td>16.76</td>
<td>19.81</td>
</tr>
<tr>
<td>32</td>
<td>.0</td>
<td>9.89</td>
<td>15.04</td>
<td>18.00</td>
<td>22.94</td>
<td>26.32</td>
<td>28.72</td>
</tr>
<tr>
<td>32</td>
<td>.5</td>
<td>10.08</td>
<td>15.11</td>
<td>17.85</td>
<td>20.23</td>
<td>23.76</td>
<td>26.20</td>
</tr>
<tr>
<td>32</td>
<td>.9</td>
<td>9.59</td>
<td>15.32</td>
<td>18.25</td>
<td>12.75</td>
<td>17.78</td>
<td>20.54</td>
</tr>
<tr>
<td>64</td>
<td>.0</td>
<td>10.09</td>
<td>15.37</td>
<td>17.99</td>
<td>24.56</td>
<td>28.15</td>
<td>30.00</td>
</tr>
<tr>
<td>64</td>
<td>.5</td>
<td>9.95</td>
<td>15.18</td>
<td>15.15</td>
<td>21.10</td>
<td>25.18</td>
<td>27.28</td>
</tr>
<tr>
<td>64</td>
<td>.9</td>
<td>10.26</td>
<td>15.67</td>
<td>18.49</td>
<td>12.98</td>
<td>18.09</td>
<td>20.53</td>
</tr>
<tr>
<td>128</td>
<td>.0</td>
<td>9.96</td>
<td>15.09</td>
<td>17.59</td>
<td>26.47</td>
<td>29.50</td>
<td>30.94</td>
</tr>
<tr>
<td>128</td>
<td>.5</td>
<td>10.23</td>
<td>15.07</td>
<td>17.48</td>
<td>23.62</td>
<td>26.82</td>
<td>28.51</td>
</tr>
<tr>
<td>128</td>
<td>.9</td>
<td>10.11</td>
<td>15.05</td>
<td>18.05</td>
<td>14.34</td>
<td>18.89</td>
<td>21.56</td>
</tr>
<tr>
<td>256</td>
<td>.0</td>
<td>10.28</td>
<td>15.62</td>
<td>18.37</td>
<td>27.39</td>
<td>30.74</td>
<td>32.46</td>
</tr>
<tr>
<td>256</td>
<td>.5</td>
<td>10.60</td>
<td>16.02</td>
<td>18.44</td>
<td>23.81</td>
<td>28.38</td>
<td>30.31</td>
</tr>
<tr>
<td>256</td>
<td>.9</td>
<td>10.11</td>
<td>15.48</td>
<td>17.91</td>
<td>14.15</td>
<td>19.43</td>
<td>22.03</td>
</tr>
<tr>
<td>512</td>
<td>.0</td>
<td>10.12</td>
<td>15.34</td>
<td>17.68</td>
<td>27.64</td>
<td>30.55</td>
<td>32.14</td>
</tr>
<tr>
<td>512</td>
<td>.5</td>
<td>10.05</td>
<td>14.96</td>
<td>17.66</td>
<td>24.10</td>
<td>27.40</td>
<td>29.28</td>
</tr>
<tr>
<td>512</td>
<td>.9</td>
<td>9.90</td>
<td>15.09</td>
<td>17.53</td>
<td>14.78</td>
<td>19.16</td>
<td>21.49</td>
</tr>
</tbody>
</table>

**Note:** $T$ is sample size, $\rho$ is the contemporaneous correlation between the innovations underlying the forecast errors, and $\theta$ is the coefficient of the MA(1) forecast error. All tests are at the 10% level. 10,000 Monte Carlo replications are performed.
### Table 3. Empirical Size Under Quadratic Loss, Test Statistic MR

<table>
<thead>
<tr>
<th>$T$</th>
<th>$\rho$</th>
<th>$\theta = .0$</th>
<th>$\theta = .5$</th>
<th>$\theta = .9$</th>
<th>$\theta = .0$</th>
<th>$\theta = .5$</th>
<th>$\theta = .9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>.0</td>
<td>9.67</td>
<td>19.33</td>
<td>22.45</td>
<td>16.16</td>
<td>25.26</td>
<td>27.62</td>
</tr>
<tr>
<td>8</td>
<td>.5</td>
<td>9.50</td>
<td>19.00</td>
<td>22.07</td>
<td>14.81</td>
<td>24.50</td>
<td>26.99</td>
</tr>
<tr>
<td>8</td>
<td>.9</td>
<td>9.66</td>
<td>19.51</td>
<td>22.85</td>
<td>11.23</td>
<td>21.28</td>
<td>24.14</td>
</tr>
<tr>
<td>16</td>
<td>.0</td>
<td>9.62</td>
<td>13.92</td>
<td>14.72</td>
<td>19.94</td>
<td>22.56</td>
<td>23.06</td>
</tr>
<tr>
<td>16</td>
<td>.5</td>
<td>10.02</td>
<td>13.88</td>
<td>14.66</td>
<td>17.70</td>
<td>21.04</td>
<td>21.28</td>
</tr>
<tr>
<td>16</td>
<td>.9</td>
<td>10.04</td>
<td>13.82</td>
<td>14.94</td>
<td>11.76</td>
<td>15.68</td>
<td>16.70</td>
</tr>
<tr>
<td>32</td>
<td>.0</td>
<td>9.96</td>
<td>10.98</td>
<td>11.12</td>
<td>22.78</td>
<td>22.86</td>
<td>21.72</td>
</tr>
<tr>
<td>32</td>
<td>.5</td>
<td>9.88</td>
<td>11.46</td>
<td>11.66</td>
<td>19.78</td>
<td>20.32</td>
<td>20.14</td>
</tr>
<tr>
<td>32</td>
<td>.9</td>
<td>9.86</td>
<td>11.62</td>
<td>11.96</td>
<td>12.42</td>
<td>13.54</td>
<td>13.46</td>
</tr>
<tr>
<td>64</td>
<td>.0</td>
<td>10.32</td>
<td>11.02</td>
<td>11.04</td>
<td>24.50</td>
<td>19.48</td>
<td>18.84</td>
</tr>
<tr>
<td>64</td>
<td>.5</td>
<td>9.84</td>
<td>10.56</td>
<td>10.64</td>
<td>21.44</td>
<td>19.48</td>
<td>18.84</td>
</tr>
<tr>
<td>64</td>
<td>.9</td>
<td>9.58</td>
<td>10.58</td>
<td>10.34</td>
<td>13.38</td>
<td>13.38</td>
<td>13.20</td>
</tr>
<tr>
<td>128</td>
<td>.0</td>
<td>9.78</td>
<td>10.54</td>
<td>10.44</td>
<td>25.66</td>
<td>22.90</td>
<td>21.54</td>
</tr>
<tr>
<td>128</td>
<td>.5</td>
<td>10.02</td>
<td>11.04</td>
<td>11.18</td>
<td>22.76</td>
<td>20.26</td>
<td>19.44</td>
</tr>
<tr>
<td>128</td>
<td>.9</td>
<td>10.76</td>
<td>11.28</td>
<td>11.38</td>
<td>13.44</td>
<td>13.52</td>
<td>12.92</td>
</tr>
<tr>
<td>256</td>
<td>.0</td>
<td>10.04</td>
<td>9.90</td>
<td>9.58</td>
<td>27.16</td>
<td>23.74</td>
<td>22.70</td>
</tr>
<tr>
<td>256</td>
<td>.5</td>
<td>10.32</td>
<td>9.92</td>
<td>9.82</td>
<td>24.00</td>
<td>20.50</td>
<td>19.18</td>
</tr>
<tr>
<td>256</td>
<td>.9</td>
<td>9.92</td>
<td>10.16</td>
<td>10.34</td>
<td>13.38</td>
<td>12.70</td>
<td>12.24</td>
</tr>
<tr>
<td>512</td>
<td>.0</td>
<td>9.94</td>
<td>10.48</td>
<td>10.56</td>
<td>26.92</td>
<td>23.40</td>
<td>21.78</td>
</tr>
<tr>
<td>512</td>
<td>.5</td>
<td>9.52</td>
<td>10.56</td>
<td>10.48</td>
<td>23.56</td>
<td>20.52</td>
<td>19.36</td>
</tr>
<tr>
<td>512</td>
<td>.9</td>
<td>9.80</td>
<td>9.82</td>
<td>9.88</td>
<td>13.96</td>
<td>12.98</td>
<td>12.74</td>
</tr>
</tbody>
</table>

**NOTE:** $T$ is sample size, $\rho$ is the contemporaneous correlation between the innovations underlying the forecast errors, and $\theta$ is the coefficient of the MA(1) forecast error. All tests are at the 10% level. At least 5,000 Monte Carlo replications are performed.

### Table 4. Empirical Size Under Quadratic Loss, Test Statistic $S_1$

<table>
<thead>
<tr>
<th>$T$</th>
<th>$\rho$</th>
<th>$\theta = .0$</th>
<th>$\theta = .5$</th>
<th>$\theta = .9$</th>
<th>$\theta = .0$</th>
<th>$\theta = .5$</th>
<th>$\theta = .9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>.0</td>
<td>31.39</td>
<td>31.10</td>
<td>31.03</td>
<td>31.62</td>
<td>29.51</td>
<td>29.07</td>
</tr>
<tr>
<td>8</td>
<td>.5</td>
<td>31.37</td>
<td>30.39</td>
<td>29.93</td>
<td>31.21</td>
<td>29.71</td>
<td>29.36</td>
</tr>
<tr>
<td>8</td>
<td>.9</td>
<td>31.08</td>
<td>30.19</td>
<td>30.18</td>
<td>31.18</td>
<td>30.12</td>
<td>29.75</td>
</tr>
<tr>
<td>16</td>
<td>.0</td>
<td>20.39</td>
<td>19.11</td>
<td>18.94</td>
<td>19.26</td>
<td>18.50</td>
<td>18.32</td>
</tr>
<tr>
<td>16</td>
<td>.5</td>
<td>20.43</td>
<td>19.52</td>
<td>18.86</td>
<td>19.57</td>
<td>17.67</td>
<td>17.63</td>
</tr>
<tr>
<td>16</td>
<td>.9</td>
<td>20.90</td>
<td>19.55</td>
<td>19.59</td>
<td>20.15</td>
<td>18.38</td>
<td>18.16</td>
</tr>
<tr>
<td>32</td>
<td>.0</td>
<td>12.42</td>
<td>12.28</td>
<td>12.18</td>
<td>11.30</td>
<td>11.64</td>
<td>11.56</td>
</tr>
<tr>
<td>32</td>
<td>.5</td>
<td>13.32</td>
<td>13.22</td>
<td>12.94</td>
<td>11.54</td>
<td>10.66</td>
<td>10.64</td>
</tr>
<tr>
<td>32</td>
<td>.9</td>
<td>12.60</td>
<td>13.38</td>
<td>13.22</td>
<td>11.16</td>
<td>11.22</td>
<td>11.50</td>
</tr>
<tr>
<td>64</td>
<td>.0</td>
<td>12.47</td>
<td>12.11</td>
<td>11.94</td>
<td>12.44</td>
<td>11.62</td>
<td>11.36</td>
</tr>
<tr>
<td>64</td>
<td>.5</td>
<td>12.76</td>
<td>12.49</td>
<td>12.35</td>
<td>12.10</td>
<td>12.26</td>
<td>12.10</td>
</tr>
<tr>
<td>64</td>
<td>.9</td>
<td>12.21</td>
<td>12.23</td>
<td>12.03</td>
<td>13.00</td>
<td>12.36</td>
<td>12.16</td>
</tr>
<tr>
<td>128</td>
<td>.0</td>
<td>11.72</td>
<td>11.94</td>
<td>12.04</td>
<td>11.48</td>
<td>10.72</td>
<td>10.28</td>
</tr>
<tr>
<td>128</td>
<td>.5</td>
<td>11.44</td>
<td>11.72</td>
<td>11.60</td>
<td>10.84</td>
<td>10.96</td>
<td>10.96</td>
</tr>
<tr>
<td>128</td>
<td>.9</td>
<td>11.76</td>
<td>11.26</td>
<td>11.34</td>
<td>11.50</td>
<td>10.66</td>
<td>10.86</td>
</tr>
<tr>
<td>256</td>
<td>.0</td>
<td>11.11</td>
<td>10.65</td>
<td>10.66</td>
<td>12.06</td>
<td>11.67</td>
<td>11.79</td>
</tr>
<tr>
<td>256</td>
<td>.5</td>
<td>10.90</td>
<td>10.39</td>
<td>10.48</td>
<td>12.16</td>
<td>11.46</td>
<td>11.60</td>
</tr>
<tr>
<td>256</td>
<td>.9</td>
<td>10.89</td>
<td>10.79</td>
<td>10.75</td>
<td>11.51</td>
<td>11.59</td>
<td>11.16</td>
</tr>
<tr>
<td>512</td>
<td>.0</td>
<td>11.15</td>
<td>10.67</td>
<td>10.63</td>
<td>10.06</td>
<td>9.46</td>
<td>9.62</td>
</tr>
<tr>
<td>512</td>
<td>.5</td>
<td>10.90</td>
<td>10.39</td>
<td>10.48</td>
<td>9.94</td>
<td>9.66</td>
<td>9.76</td>
</tr>
<tr>
<td>512</td>
<td>.9</td>
<td>10.31</td>
<td>10.09</td>
<td>10.05</td>
<td>10.12</td>
<td>10.12</td>
<td>10.06</td>
</tr>
</tbody>
</table>

**NOTE:** $T$ is sample size, $\rho$ is the contemporaneous correlation between the innovations underlying the forecast errors, and $\theta$ is the coefficient of the MA(1) forecast error. All tests are at the 10% level. At least 5,000 Monte Carlo replications are performed.
### Table 5. Empirical Size Under Quadratic Loss, Test Statistics $S_2$ and $S_{2e}$

<table>
<thead>
<tr>
<th>$T$</th>
<th>$\rho$</th>
<th>$\theta = .0$</th>
<th>$\theta = .5$</th>
<th>$\theta = .9$</th>
<th>$\theta = .0$</th>
<th>$\theta = .5$</th>
<th>$\theta = .9$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$S_2$, nominal size = 25%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>.0</td>
<td>22.24</td>
<td>22.48</td>
<td>22.38</td>
<td>23.94</td>
<td>23.46</td>
<td>23.34</td>
</tr>
<tr>
<td>8</td>
<td>.5</td>
<td>22.14</td>
<td>23.46</td>
<td>22.16</td>
<td>23.08</td>
<td>24.80</td>
<td>23.06</td>
</tr>
<tr>
<td>8</td>
<td>.9</td>
<td>22.24</td>
<td>23.02</td>
<td>22.66</td>
<td>22.92</td>
<td>23.26</td>
<td>22.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S_2$, nominal size = 14.08%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>.9</td>
<td>13.08</td>
<td>13.84</td>
<td>13.28</td>
<td>12.86</td>
<td>13.06</td>
<td>13.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S_2$, nominal size = 15.36%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>.5</td>
<td>14.36</td>
<td>14.06</td>
<td>13.94</td>
<td>15.08</td>
<td>14.36</td>
<td>15.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S_{2e}$, nominal size = 10%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>.0</td>
<td>9.72</td>
<td>9.92</td>
<td>9.42</td>
<td>9.68</td>
<td>10.36</td>
<td>10.44</td>
</tr>
<tr>
<td>64</td>
<td>.5</td>
<td>9.66</td>
<td>10.34</td>
<td>9.68</td>
<td>9.52</td>
<td>10.06</td>
<td>10.00</td>
</tr>
<tr>
<td>64</td>
<td>.9</td>
<td>10.84</td>
<td>9.46</td>
<td>10.34</td>
<td>9.40</td>
<td>8.98</td>
<td>10.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S_{2e}$, nominal size = 10%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>128</td>
<td>.0</td>
<td>11.62</td>
<td>11.62</td>
<td>11.84</td>
<td>12.22</td>
<td>12.20</td>
<td>11.42</td>
</tr>
<tr>
<td>128</td>
<td>.5</td>
<td>11.66</td>
<td>11.62</td>
<td>11.90</td>
<td>12.06</td>
<td>11.94</td>
<td>11.44</td>
</tr>
<tr>
<td>128</td>
<td>.9</td>
<td>11.22</td>
<td>11.72</td>
<td>11.28</td>
<td>12.06</td>
<td>10.76</td>
<td>11.40</td>
</tr>
</tbody>
</table>

NOTE: $T$ is sample size, $\rho$ is the contemporaneous correlation between the innovations underlying the forecast errors, and $\theta$ is the coefficient of the MA(1) forecast error. At least 5,000 Monte Carlo replications are performed.

### Table 6. Empirical Size Under Quadratic Loss, Test Statistics $S_3$ and $S_{3e}$

<table>
<thead>
<tr>
<th>$T$</th>
<th>$\rho$</th>
<th>$\theta = .0$</th>
<th>$\theta = .5$</th>
<th>$\theta = .9$</th>
<th>$\theta = .0$</th>
<th>$\theta = .5$</th>
<th>$\theta = .9$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$S_3$, nominal size = 25%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>.0</td>
<td>22.50</td>
<td>22.92</td>
<td>22.90</td>
<td>23.26</td>
<td>23.34</td>
<td>21.96</td>
</tr>
<tr>
<td>8</td>
<td>.5</td>
<td>22.98</td>
<td>22.26</td>
<td>23.06</td>
<td>23.42</td>
<td>23.86</td>
<td>22.88</td>
</tr>
<tr>
<td>8</td>
<td>.9</td>
<td>23.16</td>
<td>22.36</td>
<td>24.24</td>
<td>24.26</td>
<td>23.32</td>
<td>23.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S_3$, nominal size = 10.92%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>.0</td>
<td>10.62</td>
<td>10.06</td>
<td>10.40</td>
<td>10.16</td>
<td>10.42</td>
<td>9.84</td>
</tr>
<tr>
<td>16</td>
<td>.5</td>
<td>10.38</td>
<td>10.92</td>
<td>10.32</td>
<td>10.54</td>
<td>10.94</td>
<td>10.34</td>
</tr>
<tr>
<td>16</td>
<td>.9</td>
<td>10.64</td>
<td>10.18</td>
<td>9.62</td>
<td>10.58</td>
<td>10.96</td>
<td>10.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S_3$, nominal size = 10.12%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>.0</td>
<td>10.72</td>
<td>10.28</td>
<td>9.30</td>
<td>9.90</td>
<td>10.00</td>
<td>9.98</td>
</tr>
<tr>
<td>32</td>
<td>.5</td>
<td>10.56</td>
<td>10.00</td>
<td>10.02</td>
<td>10.40</td>
<td>10.64</td>
<td>10.30</td>
</tr>
<tr>
<td>32</td>
<td>.9</td>
<td>10.92</td>
<td>10.44</td>
<td>10.30</td>
<td>10.46</td>
<td>9.98</td>
<td>10.70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S_{3e}$, nominal size = 10%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>.0</td>
<td>9.38</td>
<td>9.54</td>
<td>9.16</td>
<td>9.64</td>
<td>9.24</td>
<td>8.84</td>
</tr>
<tr>
<td>64</td>
<td>.5</td>
<td>9.80</td>
<td>10.02</td>
<td>9.66</td>
<td>9.58</td>
<td>8.82</td>
<td>8.78</td>
</tr>
<tr>
<td>64</td>
<td>.9</td>
<td>9.90</td>
<td>9.24</td>
<td>9.68</td>
<td>9.92</td>
<td>9.78</td>
<td>10.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S_{3e}$, nominal size = 10%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>128</td>
<td>.0</td>
<td>9.94</td>
<td>9.70</td>
<td>9.12</td>
<td>9.82</td>
<td>9.04</td>
<td>8.46</td>
</tr>
<tr>
<td>128</td>
<td>.5</td>
<td>9.52</td>
<td>10.00</td>
<td>9.32</td>
<td>10.08</td>
<td>9.24</td>
<td>9.20</td>
</tr>
<tr>
<td>128</td>
<td>.9</td>
<td>9.46</td>
<td>9.64</td>
<td>9.42</td>
<td>9.28</td>
<td>9.22</td>
<td>9.26</td>
</tr>
</tbody>
</table>

NOTE: $T$ is sample size, $\rho$ is the contemporaneous correlation between the innovations underlying the forecast errors, and $\theta$ is the coefficient of the MA(1) forecast error. At least 5,000 Monte Carlo replications are performed.
4. AN EMPIRICAL EXAMPLE

We shall illustrate the practical use of the tests with an application to exchange-rate forecasting. The series to be forecast, measured monthly, is the three-month change in the nominal dollar/Dutch guilder end-of-month spot exchange rate (in U.S. cents, noon, New York interbank), from 1977.01 to 1991.12. We assess two forecasts, the "no change" (0) forecast associated with a random-walk model and the forecast implicit in the three-month forward rate (the difference between the three-month forward rate and the spot rate).

The actual and predicted changes are shown in Figure 2. The random-walk forecast, of course, is just constant at 0, whereas the forward market forecast moves over time. The movements in both forecasts, however, are dwarfed by the realized movements in exchange rates.

We shall assess the forecasts' accuracy under absolute error loss. In terms of point estimates, the random-walk forecast is more accurate. The mean absolute error of the random-walk forecast is 1.42, as opposed to 1.53 for the forward market forecast; as one hears so often, "The random walk wins." The loss-differential series is shown in Figure 3, in which no obvious nonstationarities are visually apparent. Approximate stationarity is also supported by the sample autocorrelation function of the loss differential, shown in Figure 4, which decays quickly.

Because the forecasts are three-step-ahead, our earlier arguments suggest the need to allow for at least two-dependent forecast errors, which may translate into a two-dependent loss differential. This intuition is confirmed by the sample autocorrelation function of the loss differential, in which sizable and significant sample autocorrelations appear at lags 1 and 2 and nowhere else. The Box–Pierce $\chi^2$ test of jointly zero autocorrelations at lags 1 through 15 is 51.12, which is highly significant relative to its asymptotic null distribution of $\chi^2_{13}$. Conversely, the Box–Pierce $\chi^2$ test of jointly zero autocorrelations at lags 3 through 15 is 12.79, which is insignificant relative to its null distribution of $\chi^2_{13}$.

We now proceed to test the null of equal expected loss. $F$, MGN, and MR are inapplicable because one or more of their
maintained assumptions are explicitly violated. We therefore focus on our test statistic $S_1$, setting the truncation lag at two in light of the preceding discussion. We obtain $S_1 = -1.3$, implying a $p$ value of .19. Thus, for the sample at hand, we do not reject at conventional levels the hypothesis of equal expected absolute error—the forward rate is not a statistically significantly worse predictor of the future spot rate than is the current spot rate.

5. CONCLUSIONS AND DIRECTIONS FOR FUTURE RESEARCH

We have proposed several tests of the null hypothesis of equal forecast accuracy. We allow the forecast errors to be non-Gaussian, nonzero mean, serially correlated, and contemporaneously correlated. Perhaps most importantly, our tests are applicable under a very wide variety of loss structures.

We hasten to add that comparison of forecast accuracy is but one of many diagnostics that should be examined when comparing models. Moreover, the superiority of a particular model in terms of forecast accuracy does not necessarily imply that forecasts from other models contain no additional information. That, of course, is the well-known message of the forecast combination and encompassing literature; see, for example, Clemen (1989), Chong and Hendry (1986), and Fair and Shiller (1990).

Several extensions of the results presented here appear to be promising directions for future research. Some are obvious, such as generalization to comparison of more than two forecasts or, perhaps most generally, multiple forecasts for each of multiple variables. Others are less obvious and more interesting. We shall list just a few:

1. Our framework may be broadened to examine not only whether forecast loss differentials have nonzero mean but also whether other variables may explain loss differentials. For example, one could regress the loss differential not only on a constant but also on a “stage of the business cycle” indicator to assess the extent to which relative predictive performance differs over the cycle.
2. The ability to formally compare predictive accuracy afforded by our tests may prove useful as a model specification diagnostic, as well as a means to test both nested and nonnested hypotheses under nonstandard conditions, in the tradition of Ashley, Granger, and Schmalensee (1980) and Mariano and Brown (1983).
3. Explicit account may be taken of the effects of uncertainty associated with estimated model parameters on the behavior of the test statistics, as shown by West (1994).

Let us provide some examples of the ideas sketched in 2. First, consider the development of a test of exclusion restrictions in time series regression that is valid regardless of whether the data are stationary or cointegrated. The desirability of such a test is apparent from works like those of Stock and Watson (1989), Christiano and Eichenbaum (1990), Rudebusch (1993), and Toda and Phillips (1993), in which it is simultaneously apparent that (a) it is difficult to determine reliably the integration status of macroeconomic time series and (b) the conclusions of macroeconomic studies are often critically dependent on the integration status of the relevant time series. One may proceed by noting that tests of exclusion restrictions amount to comparisons of restricted and unrestricted sums of squares. This suggests estimating the restricted and unrestricted models using part of the available data and then using our test of equality of the mean squared errors of the respective one-step-ahead forecasts.

As a second example, it would appear that our test is applicable in nonstandard testing situations, such as when a nuisance parameter is not identified under the null. This occurs, for example, when testing for the appropriate number of states in Hamilton’s (1989) Markov-switching model. In spite of the fact that standard tests are inapplicable, certainly the null and alternative models may be estimated and their out-of-sample forecasting performance compared rigorously, as shown by Engel (1994).

In closing, we note that this article is part of a larger research program aimed at doing model selection, estimation, prediction, and evaluation using the relevant loss function, whatever that loss function may be. This article has addressed evaluation. Granger (1969) and Christoffersen and Diebold (1994) addressed prediction. These results, together with those of Weiss and Andersen (1984) and Weiss (1991, 1994) on estimation under the relevant loss function will make feasible recursive, real-time, prediction-based model selection under the relevant loss function.

ACKNOWLEDGMENTS

We thank the editor, associate editor, and two referees for constructive comments. Seminar participants at Chicago, Cornell, the Federal Reserve Board, London School of Economics, Maryland, the Model Comparison Seminar, Oxford, Pennsylvania, Pittsburgh, and Santa Cruz provided helpful input, as did Rob Engle, Jim Hamilton, Hashem Pesaran, Ingmar Prucha, Peter Robinson, and Ken West, but all errors are ours alone. Portions of this article were written while the first author visited the Financial Markets Group at the London School of Economics, whose hospitality is gratefully acknowledged. Financial support from the National Science Foundation, the Sloan Foundation, and the University of Pennsylvania Research Foundation is gratefully acknowledged. Ralph Bradley, José A. Lopez, and Gretchen Weinbach provided research assistance.

[Received March 1994. Revised December 1994.]

REFERENCES

Diebold and Mariano: Comparing Predictive Accuracy

The Past, Present, and Future of Macroeconomic Forecasting

Francis X. Diebold

The reports of the death of large-scale macroeconomic forecasting models are not exaggerated. But many observers interpret the failure of the early models as indicative of a bleak future for macroeconomic forecasting more generally. Such is not the case. Although the large-scale macroeconomic forecasting models didn’t live up to their original promise, they nevertheless left a useful legacy of lasting contributions from which macroeconomic forecasting will continue to benefit: they spurred the development of powerful identification and estimation theory, computational and simulation techniques, comprehensive machine-readable macroeconomic databases, and much else. Moreover, past failures do not necessarily imply a bleak future: we learn from our mistakes. Just as macroeconomics has benefited from rethinking since the 1970s, so too will macroeconomic forecasting.

Understanding the future of macroeconomic forecasting requires understanding the interplay between measurement and theory, and the corresponding evolution of the nonstructural and structural approaches to forecasting. Nonstructural macroeconomic forecasting methods attempt to exploit the reduced-form correlations in observed macroeconomic time series, with little reliance on economic theory. Structural models, in contrast, view and interpret economic data through the lens of a particular economic theory.

Structural econometric forecasting, because it is based on explicit theory, rises and falls with theory, typically with a lag. Structural Keynesian macroeconomic forecasting, based on postulated systems of decision rules, enjoyed a golden age in the 1950s and 1960s following the advances in Keynesian theory in the 1930s and 1940s, and the two then declined together in the 1970s and 1980s. The evolution of nonstructural economic forecasting, in contrast, is less bound to fashions in economic theory; its origins long predate structural Keynesian macroeconomic forecasting and progress continues at a rapid pace.

One is naturally led to a number of important questions. What of the impressive advances in economic theory of the 1980s and 1990s? Should we not expect them to be followed by a new wave of structural macroeconomic forecasting, or has nonstructural forecasting permanently replaced structural forecasting? Is it necessary to choose between the structural and nonstructural approaches, or might the two be complements rather than substitutes? If a new structural forecasting is likely to emerge, in what ways will it resemble its ancestors? Our answers will take us on a whirlwind tour of the past, present and future of both structural and nonstructural forecasting. We’ll begin by tracing the rise and fall of the structural Keynesian system-of-equations paradigm, and then we’ll step back to assess the long-running and ongoing progress in the nonstructural tradition. Finally, we’ll assess the rise of modern dynamic stochastic general equilibrium macroeconomic theory, its relationship to nonstructural methods, and its implications for a new structural macroeconomic forecasting.

The Rise and Fall of Keynesian Macroeconomic Theory and Structural Forecasting

Some important forecasting situations involve conditional forecasts; that is, forecasts of one or more variables conditional upon maintained assumptions regarding, for example, the behavior of policymakers. Conditional forecasts require structural models. Structural econometrics, and hence structural macroeconomic forecasting, makes use of macroeconomic theory, which implies that developments in structural forecasting naturally lag behind developments in theory. The first major wave of twentieth century macroeconomic theory was the Keynesian theory of the 1930s and 1940s, and it was followed by major advances in structural macroeconomic forecasting.

When Keynes’s General Theory was published in 1936, theory was distinctly ahead of measurement. Measurement soon caught up, however, in the form of the systems of equations associated with Klein’s (1946) Keynesian Revolution and Klein and Goldberger’s (1955) Econometric Model of the United States: 1929–1952. Indeed, the period following the publication of the General Theory was one of unprecedented and furious intellectual activity directed toward the construction, estimation and analysis of Keynesian structural econometric models. The statistics side of the structural econometrics research was fueled by the advances of Fisher, Neyman, Pearson, and many others earlier in the century. The economics side, of course, was driven by Keynes’s landmark contribution, which spoke eloquently to the severe economic problems of the day and seemed to offer a workable solution.

The intellectual marriage of statistics and economic theory was reflected in the
growth of the Econometric Society and its journal, *Econometrica*, and beautifully distilled in the work of the Cowles Commission for Research in Economics at the University of Chicago in the 1940s and early 1950s. The intellectual focus and depth of talent assembled there were unprecedented in the history of economics. Cowles researchers included T.W. Anderson, K. Arrow, G. Debreu, T. Haavelmo, L. Hurwicz, L.R. Klein, T. Koopmans, H. Markowitz, J. Marshak, F. Modigliani, H. Simon, A. Wald, and many others. A central part (although by no means the only part) of the Cowles research program was identification and estimation of systems of stochastic difference equations designed to approximate the postulated decision rules of Keynesian macroeconomic theory.

Just as the blending of mathematical statistics and economics associated with the Cowles Commission was historically unprecedented, so too was the optimism about solving pressing macroeconomic problems. Early on, the macroeconomic system-of-equations research program appeared impressively successful, and structural econometric forecasting blossomed in the late 1950s and 1960s, the heyday of the large-scale Keynesian macroeconomic forecasting models. There was strong consensus regarding the general paradigm, even if there was disagreement on details such as the relative slopes of IS and LM curves, and the models were routinely used for forecasting and policy analysis in both academia and government.

But cracks in the foundation, which began with intellectual dissatisfaction with the underpinnings of Keynesian macroeconomic systems of equations, started to appear in the late 1960s and early 1970s. First, economists became dissatisfied with the lack of foundations for the disequilibrium nature of the Keynesian model. A new and still ongoing research program began which sought microfoundations for Keynesian macroeconomic theory, particularly for the central tenets of sticky wages and prices. Many key early contributions appear in the classic Phelps et al. (1970) volume, and more recent contributions are collected in Mankiw and Romer (1991).

Second, just as macroeconomists became increasingly disenchanted with the ad hoc treatment of sticky prices in traditional models, they became similarly disenchanted with ad hoc treatment of expectations. Building on early work by Muth (1960, 1961), who introduced the idea of rational expectations and showed that schemes such as adaptive expectations were rational only in unlikely circumstances, the "rational expectations revolution" quickly took hold; Sargent and Wallace (1975) is an important and starkly simple early paper.

Third, and most generally, economists became dissatisfied not only with certain parts of the Keynesian macroeconomic program, such as the assumptions about price behavior and expectations formation, but rather with the overall modeling approach embodied in the program. The approach was dubbed the "system-of-equations" approach by Prescott (1986), in reference to the fact that it concentrated on the estimation of parameters of equation systems representing ad hoc postulated decision rules ("consumption functions," "investment functions," and so on) as opposed to more fundamental parameters of economic theory. Newly emerging macroeconomic work in the late 1960s and early 1970s, in contrast, was firmly grounded in tastes and technology. Lucas and Prescott (1971) and Lucas (1978) remain classic examples. Work in the tastes-and-technology tradition accelerated rapidly following Lucas's (1976) formal critique of the system-of-equations approach, based on the insight that analysis based on decision rules is a fundamentally defective paradigm for producing conditional forecasts, because the parameters of decision rules will generally change when policies change.

Finally, if the cracks in the foundation of Keynesian structural forecasting began as intellectual dissatisfaction, they were widened by the economic facts of the 1970s; in particular, the simultaneous presence of high inflation and unemployment, which naturally led economists to question the alleged inflation/unemployment tradeoff embedded in the Keynesian systems of equations. In addition, a series of studies published in the early 1970s revealed that simple statistical extrapolations, making no assumptions at all about economic structure, often forecasted macroeconomic activity just as well as large-scale Keynesian macroeconomic models; Nelson (1972) remains a classic. Keynesian macroeconomics soon declined, and Keynesian structural econometric forecasting followed suit.

**Nonstructural Forecasting**

By the late 1970s, it was clear that Keynesian structural macroeconomic forecasting, at least as traditionally implemented, was receding. One response was to augment the traditional system-of-equations econometrics in attempts to remedy its defects. Important work along those lines was undertaken by R. Fair and J. Taylor (for example, Fair, 1984, 1994; Taylor, 1995), who developed methods for incorporating rational-expectations into econometric models, as well as methods for rigorous assessment of model fit and forecasting performance. Models in the Fair-Taylor spirit are now in use at a number of leading policy organizations, including the Federal Reserve Board and the International Monetary Fund, and those, for example, in Brayton et al. (1997). They are an important step forward, even if the theory on which they are built remains largely in the system-of-equations tradition.

Another response, involving a more radical change of direction, was to explore alternative, nonstructural forecasting methods. Many forecasting chokes involve unconditional, rather than conditional, forecasts—that is, interest often centers on the likely future path of the economy when policy remains unchanged, so that the Lucas critique is not relevant—and unconditional forecasting does not require a structural model. That insight, together with the emerging discontent with Keynesian macroeconomic theory and the lack of a well-developed alternative, produced tremendous interest in nonstructural econometric forecasting in the 1970s. The title of an important paper by Sargent and Sims (1977), "Business Cycle Modeling Without Pretending to Have too Much a Priori Theory," nicely summarizes the spirit of the times.

The impressive intellectual development of nonstructural forecasting spans many decades; it predates the Keynesian episode and continues to the present.

---

1 For a concise history of the Chicago days of the Cowles Commission, see Hildebrath (1986, ch. 3).
Macroeconomists and econometricians didn't pay much attention at first to in the 1970s, as the world economic scene was changing. However, by the end of the decade, the need for more sophisticated forecasting methods became apparent. The advent of econometric models allowed for more accurate predictions of economic variables. These models are based on statistical techniques and are used to forecast future economic conditions.

Key developments in econometric forecasting during the 1970s included the introduction of the Box-Jenkins methodology. This approach, developed by George Box and Gwilym Jenkins, was a significant advancement in the field of econometrics. The Box-Jenkins methodology is a statistical approach for time series analysis that allows for the modeling and forecasting of non-stationary time series data. It involves the identification, estimation, and validation of an appropriate model for a time series. The methodology is widely used in economics, finance, and many other fields where time series data are analyzed.

Another important development during this period was the introduction of the Vector Autoregressive (VAR) model. The VAR model is a generalization of the univariate autoregressive model that allows for the modeling of the dynamic interrelationships among multiple time series. This model has been widely used in macroeconomics to analyze the interactions between different economic variables.

In conclusion, the 1970s were a period of significant advancement in econometric forecasting. The introduction of the Box-Jenkins methodology and the VAR model, among other developments, contributed to the evolution of econometric forecasting methods and their application in various fields. These advancements continue to be relevant and are still widely used today.
nous" or "exogenous." In vector autoregressions, in contrast, all variables are endogenous.

The mechanics of vector autoregressions are straightforward. Recall that we approximate dynamics with a univariate autoregression by regressing a variable on its own past values. In a vector autoregression, by way of logical extension, we regress each of a set of variables on past values of itself and past values of every other variable in the system. Cross-variable linkages are automatically incorporated because we include lags of all variables in each equation, and because we allow for correlations among the disturbances of the various equations. It turns out that one-equation-at-a-time least squares estimation of vector autoregressions is statistically efficient, in spite of the potential correlation of disturbances. Moreover, it is relatively simple and numerically stable, in contrast to the tedious numerical optimization required for estimation of multivariate ARMA models.

Many multivariate extensions of the Box-Jenkins paradigm are conveniently implemented in the vector autoregressive framework. Here we introduce a few to help convey a feel for the breadth of modern time-series econometrics and forecasting. The discussion is necessarily brief; for a more detailed introduction to modern time series forecasting, see Diebold (1998).

Granger (1969) and Sims (1972) made important early multivariate contributions, providing tools for exploring causal patterns in multivariate systems. The Granger-Sims causality notion is predictive; we say that \( x \) Granger-Sims causes \( y \) if the history of \( x \) is useful for forecasting \( y \), over and above the history of \( y \). We commonly use Granger-Sims causality tests to help identify and understand the patterns of cross-linkages and feedback in vector autoregressions.

The dynamic factor model of Sargent and Sims (1977) and Geweke (1977) is another important early multivariate contribution. The essential idea of dynamic factor models is that some economic shocks are common across sectors and others are idiosyncratic, so that large sets of variables may depend heavily on just a few common underlying sources of variation, a common feature of economic models and evidently also of economic data. The common shocks, or "factors," are stochastic conomovements and facilitate parsimonious modeling and forecasting of large numbers of variables. Dynamic factor models have proved particularly useful with the emergence of macroeconomic panel datasets, including cross-country, cross-region, and cross-state data. Important recent contributions include Stock and Watson (1988), Quah and Sargent (1993), Forni and Reichlin (1997), and Stock and Watson (1997).

Granger (1981) and Engle and Granger (1987) develop the related idea of cointegration. We say that two series are cointegrated if each contains a stochastic trend, yet there exists a linear combination of the two trends that does not. Thus, for example, each of two series \( x \) and \( y \) may contain stochastic trend, but the spread

---

4 Ben McCallum notes in private communication that in an important sense, the situation was even worse: the endogenous-exogenous labeling was arguably not arbitrary, but rather systematic, with policy variables labeled as "exogenous" on the grounds that they could have been managed exogenously by policymakers if they had been unorthodox enough to do so.

5 For a good exposition of econometrics in the LSE tradition, see Hendry (1995).

6 For overviews, see for example Moore (1983) and Zarnowitz (1992).
the current regime (say, expansion or contraction). In the observed indicator models of Tong (1990) and Granger and Terásvirta (1993), the indicator variable is some aspect of the history of an observable variable. For example, the current regime may be determined by the sign of last period’s growth rate. In contrast, Hamilton (1989) argues that models with unobserved regime indicators may be more appropriate in many business, economic and financial contexts. In Hamilton’s widely-applied model, sometimes called a “Markov-switching” or “hidden-Markov” model, the regime is governed by an unobserved indicator.

The future of nonstructural economic forecasting will be more of the same—steady progress—fueled by cheap, fast computing, massive storage, and increased sophistication of numerical and simulation techniques. Such techniques are rapidly allowing us to estimate complicated models not amenable to treatment with standard methods, and to dispense with the unrealistic assumptions often invoked in attempts to quantify forecast uncertainty. Efron and Tibshirani (1993) and Gouriéroux and Monfort (1996), for example, provide good examples of recent developments. The future of nonstructural macroeconomic forecasting will likely also involve combining aspects of the linear and nonlinear traditions, as for example, with vector autoregressive models that allow for factor structure and regime switching (Diebold and Rudebusch, 1996; Kim and Nelson, 1998a, b).

A New Wave of Macroeconomic Theory—and Structural Forecasting

Nonstructural models are unrestricted reduced-form models. As such, they are useful for producing unconditional forecasts in a variety of environments ranging from firm-level business forecasting to economy-wide macroeconomic forecasting. Again, however, in macroeconomics we often want to analyze scenarios that differ from the conditions presently prevailing, such as the effects of a change in a policy rule or a tax rate. Such conditional forecasts require structural models.

As we have seen, an early wave of structural econometrics followed the development of Keynesian theory. But the Keynesian theory was largely based on postulated decision rules, rather than the economic primitives of taste and technology; the system-of-equations approach to structural econometric forecasting inherited that defect and hence, wasn’t really structural. Ultimately the system-of-equations approach to both theory and forecasting declined in the 1970s. Progress toward a new and truly structural macroeconomic forecasting had to await a new wave of powerful theory developed in the 1970s and 1980s. The new theory has its roots in the dissatisfaction, percolating in the late 1960s and early 1970s, with the system-of-equations approach. In many respects, the essence of the new approach is methodological and reflects a view of how macroeconomics should be done. Lucas (1972), in particular, paved the way for a new macroeconomics based on dynamic stochastic model economies with fully-articulated preferences, technologies, and rules of the game. Hence the descriptively accurate name: dynamic stochastic general equilibrium (DSGE) modeling. The key innovation is that DSGE models are built on a foundation of fully-specified stochastic dynamic optimization, as opposed to reduced-form decision rules, and are therefore not subject to the Lucas critique. But ultimately the “new” theory is neither new nor radical; rather, it is very much in the best tradition of neoclassical economics.

The new research program has sought from the outset to make clear that dynamic stochastic general equilibrium models can address practical, empirical questions. Early on, for example, Kydland and Prescott (1982) used DSGE models to argue that a neoclassical model driven purely by real technology shocks could explain a large fraction of U.S. business cycle fluctuations. Hence the early name “real business cycle” models. Later work, however, broadened the approach to allow for rich demographic structures, imperfect competition and sticky prices (and hence real effects of monetary shocks), and much else; the papers collected in Cooley (1995) offer a good overview. Ultimately, again, the essence of the new approach is not about whether the shocks that drive the cycle are real or monetary, whether prices are flexible or sticky, or whether competition is perfect or imperfect, but rather about the way macroeconomic questions should be approached.

The earliest and still rapidly developing strand of the dynamic stochastic general equilibrium literature makes use of simple “linear-quadratic” models, in which agents with quadratic preferences make optimizing decisions in environments with linear production technologies. Linear-quadratic models are surprisingly more flexible than a superficial assessment might indicate; they nest a variety of popular and useful preference and technology structures. Linear-quadratic models are also convenient, because a large literature provides powerful methods for solving, analyzing and forecasting with them. Moreover, it turns out that optimizing behavior within linear-quadratic economic models implies decision rules, such as those that govern consumption or investment behavior, that are stochastic linear functions of other variables. In particular, the decision rules conform to the great workhorse nonstructural model, the vector autoregression, subject to restrictions arising from theory. The result is a marvelous union of modern macroeconomic theory and nonstructural time-series econometrics, paving the way for a new structural econometrics.

Maximum likelihood methods are central to linear-quadratic DSGE modeling and trace to the important early work of Hansen and Sargent (1980); the modern approach is to construct and maximize the likelihood function using a state space representation in conjunction with the Kalman filter. Initially, maximum likelihood estimation was challenging in all but the simplest cases, but recent improvements in numerical algorithms and computing power have begun to make estimation and forecasting with linear-quadratic DSGE models workable for routine analysis and forecasting. Hansen and Sargent (1998) provide a powerful overview, synthesis, and extension of linear-quadratic DSGE modeling.

Kydland and Prescott (1982) started a distinct, but intimately related and equally important, strand of the DSGE literature. Two key features differentiate their product. First, Kydland and Prescott do not require that preferences be quadratic and technology be linear; instead, they use non-linear-quadratic models that are (arguably) more natural. Non-linear-quadratic models are challenging to solve,
and the Kydland- Prescott program spurred a great deal of important research on numerical and computational aspects of model solutions. One interesting outcome of that research is that, although non-linear-quadratic models don't have tidy vector autoregressive systems of decision rules, they nevertheless often have decision rules that can be accurately approximated by vector autoregressions. Second, Kydland and Prescott acknowledge from the outset that their models, like all models, are false, and they recognize that traditional econometric estimation procedures such as Gaussian maximum likelihood may lose some of their appeal in such situations. Partly for that reason, and partly because of the sheer difficulty, non-linear-quadratic DSGE models often eschew formal estimation in favor of less structured "calibration" methods, as described in this journal in Kydland and Prescott (1996). Calibration means different things to different people, but the central idea is learning about the properties of a complicated DSGE model and attempting to assess its agreement with the data, based on simulations of the model economy. The parameters underlying the simulated model economy are typically set informally, sometimes by statistical considerations such as generating realistic amounts of volatility in observed variables, sometimes by economic considerations such as producing "reasonable" steady state behavior, and sometimes by appealing to previous empirical studies.

Calibration is the natural response of economic theory to the computer age; hence the commonly used synonym "quantitative economic theory." Calibration, however, fails to provide a complete and probabilistic assessment of agreement between model and data and therefore, fails to deliver the goods necessary for forecasting with DSGE models. Econometric discontent based on recognition of that fact has been simmering for some time and is expressed forcefully by Sims (1996) in the Winter 1996 symposium in this journal on calibration and econometrics. The growing list of such symposia includes a special issue of Journal of Applied Econometrics (see the introduction by Pagan, 1994) and an Economic Journal "Controversy" section (see the Introduction by Quah, 1995).

7 See, for example, Rust (1990) and Judd (1988), who describe and contribute to the impressive advances being made for solving non-linear-quadratic stochastic dynamic programming problems.

8 The reasoning is straightforward. Loosely speaking, under correct specification, Gaussian maximum likelihood estimates converge to the true parameter values as the sample size grows; hence the estimated model converges to the true model, which is the best model to use for any purpose. Under misspecification, however, the parameters can't converge to the "true" values, because an incorrect model has been fitted. Instead, the parameters converge to values that make the fitted model the best approximation to the data, where the measure of goodness of approximation is induced by the estimation procedure.

9 The key insight is that, under misspecification, the best approximation for one purpose may differ from the best approximation for another purpose. The measure of goodness of approximation associated with Gaussian maximum likelihood estimation is 1-step-ahead mean squared forecast error. Thus, if the model is to be used for 1-step-ahead forecasting, and if mean squared error is the relevant loss function, Gaussian maximum likelihood estimation is a logical choice. If, on the other hand, the model is to be used for another purpose, such as 4-step-ahead forecasting, Gaussian maximum likelihood estimation is less appealing.

10 Important exceptions exist, however, such as McGee's, Rogerson, and Wright (1997), who estimate non-linear-quadratic DSGE models by maximum likelihood methods.

11 See also Hansen and Heekman (1996), in the same symposium, the lead paper in which is Kydland and Prescott (1996).
procedures that enable us to assess agreement between model and data predictability at various horizons of interest.

If structural modeling and forecasting have come a long way, they still have a long way to go; in closing this section, it is tempting to comment on a few aspects of their likely future development. DSGE theory will continue to improve and will begin to take certain aspects of reality, such as heterogeneity, more seriously. In particular, recent work (Geweke, 1985; Kirman, 1995; Altissimo, 1997) highlights the fact that aggregators functions may not be structural with respect to policy interventions, which suggests that current-vintage representative-agent DSGE models may not fully address the Lucas critique.12

The stochastic dynamics of driving variables, such as technology shocks, will be similarly enriched to reflect recent developments in nonstructural modeling, such as the possibility of regime switching, and to allow for multiple sources of uncertainty, including measurement error. The resulting models will have approximate representations as vector autoregressions with factor structure, possibly involving cointegration, as in King, Plosser, Stock and Watson (1991), and possibly with regime switching, as in Diebold and Rudebusch (1996) and Kim and Nelson (1998a, 1998b). Formal econometric procedures will be used to diagnose possible model inadequacies, as in Chow and Kwan (1997).

One might expect the scale of DSGE forecasting models to grow over time. That is likely to happen, and current models that determine, for example, three or four variables in equilibrium, are likely to evolve into richer models that determine, say, eight or ten variables in equilibrium.13 But the expansion in scale is likely to stop there, for two reasons. First, the demise of the large-scale models heightened professional awareness of the fact that bigger models are not necessary better, an idea memorably enshrined in Zellner's (1992) KISS principle: Keep It Simple. Second, in contrast to models in the system-of-equations tradition, which are typically estimated equation-by-equation and then assembled in modular fashion, the nature of DSGE models requires that their parameters be jointly estimated, which limits the complexity of the models that can be entertained.

And last not least, shrinkage will likely emerge as a key component of estimation techniques for DSGE forecasting models. Shrinkage refers to the idea of coaxing, or "shrinking," parameter estimates in certain directions. Shrinkage can be implemented using Bayesian methods to coax parameter estimates away from the likelihood maximum and toward the prior mean. It seems obvious that shrinkage in a "correct" direction will likely improve forecast performance. Less obvious, but equally true, is the insight that even shrinkage in "incorrect" directions can improve forecast performance, by drastically reducing forecast error variance at the potentially low price of a small increase in bias.

Shrinkage has a long history of productive use in nonstructural modeling and forecasting. For example, it has long been known that vector autoregressions estimated using Bayesian shrinkage techniques produce forecasts drastically superior to those from unrestricted vector autoregressions. The "Minnesota prior," a simple vector random walk, remains widely used.14 Shrinkage has an equally bright future in the new structural modeling and forecasting. Shrinkage is a potentially tailor-made device for estimating potentially misspecified DSGE forecasting models, because, as we have seen, DSGE theory essentially amounts to restrictions on vector autoregressions. At one extreme, we can ignore the theory and forecast with an estimated unrestricted vector autoregression: no shrinkage, loosely corresponding to a Bayesian analysis with a diffuse prior. At the other extreme, we can directly impose the theory and forecast with a restricted vector autoregression: complete shrinkage, loosely corresponding to a Bayesian analysis with a "spiked" prior. Intermediate cases, corresponding to forecasting with vector autoregressions estimated with various informative, but not spiked, priors are potentially more interesting. First, we may use statistically-oriented priors, such as the familiar Minnesota prior, which shrinks toward a vector random walk. Second, we may use statistically-oriented, but theory-inspired, priors, such as one corresponding to factor structure. Third, we may use DSGE theory-based priors, as in Ingram and Whiteman (1994), to coax the estimates in directions implied by an explicit economic theory.

Concluding Remarks

In a recent New York Times article entitled "The Model Was Too Rough: Why Economic Forecasting Became a Sideshow," economics writer Peter Passell noted: "Americans held unrealistic expectations for forecasting in the 1960's—as they did for so many other things in that optimistic age, from space exploration to big government . . . ." Our expectations for forecasting were quite appropriately revised downward in the 1970s and 1980s, and the ensuing era of humility has been good for all. The new humility, however, is not symptomatic of failure, just as the bravado of the 1960s was not symptomatic of success.

As the 1990s draw to a close, we find ourselves at a critical and newly optimistic juncture, with the futures of structural and nonstructural forecasting very much intertwined. The ongoing development of nonstructural forecasting, together with the recent developments in dynamic stochastic general equilibrium theory and associated structural estimation methods, bodes well for the future of macroeconomic forecasting. Only time will tell whether linear-quadratic or nonlinear-quadratic approximations to the macroeconomy are the best approach for practical macroeconomic forecasting, but regardless, the seeds have been

---

12 See also Oliner, Rudebusch, and Sichel (1996), who document the instability of empirical Euler equations for investment.
13 The work of Sims and his coauthors has already reached that point.
14 For an extensive discussion, see Doan, Litterman and Sims (1984). The Bayesian vector autoregressive tradition continues to progress, as for example with the work of Sims and Zha (1997), who develop methods applicable to large systems.
sown for a new structural econometrics and structural econometric forecasting, a modern and thorough implementation of the Cowles vision. The new structural econometrics is emerging more slowly than did the earlier wave following Keynes, because the baby was almost thrown out with the 1970s bathwater; the flawed econometrics that Lucas criticized was taken in some circles as an indictment of all econometrics. It has taken some time to get on with macroeconomic work, but progress is evident.

The hallmark of macroeconomic forecasting over the next 20 years will be a marriage of the best of the nonstructural and structural approaches, facilitated by advances in numerical and simulation techniques that will help macroeconomists to solve, estimate, simulate, and yes, forecast with rich models. Moreover, related developments will occur in a variety of fields well beyond macroeconomics. It's already happening and in some cases, progress has been underway for years, as evidenced by example from the recent literatures in industrial organization (Erickson and Pakes, 1995), labor economics (Eckstein and Wolpin, 1989; Stock and Wise, 1990; Rust, 1994), public economics (Rios-Rull, 1995), agricultural economics (Rosen, Murphy and Scheinkman, 1994), health economics (Gilleskie, 1997), development economics (Rosenzweig and Wolpin, 1993), environmental economics (Rotthell and Rust, 1995), and international economics (Backus, Kehoe and Kydland, 1994).

Helpful input was provided by the panel members and audience at the American Economic Association's 1997 New Orleans roundtable, "Monetary and Fiscal Policy: The Role for Structural Macroeconomic Models," as well as participants at seminars at the Federal Reserve Bank of Atlanta, the International Monetary Fund, and the 1997 International Symposium on Forecasting. I am also grateful to Brad De Long, Robert Fildes, Lusti Kilaian, Alan Krueger, Ben McCullum, Timothy Taylor, Ken Wolpin, Mike Woodford, and seminar participants at the Federal Reserve Bank of San Francisco, especially Tim Cogley, Kent Keane, Glenn Rudebusch and Tom Sargent. I, however, bear full responsibility for all remaining flaws. The National Science Foundation, the Sloan Foundation, and the University of Pennsylvania Research Foundation provided research support. This paper is available on the World Wide Web at (http://www.ssc.upenn.edu/~diebold/).

References


Granger, C. W. J. and R. Joyeux, "An Intro-


Ingram, B. and C. Whitefearn, “Supplanting the ‘Minnesota’ Prior: Forecasting Macroecono-


Judd, K., Numerical Methods in Economics. Cam-


Kim, C.-J. and C. R. Nelson, Business Cycle Turning Points: A New Coincident Index, and Test of Duration Dependence Based on a Dy-


Kim, C.-J. and C. R. Nelson, Dynamic Time Series Models and Markov Switching: Classical and Gibbs Sampling Approaches. Manuscript, Department of Economics, University of Washing-

ton, 1998b.


Klein, L. R. and A. E. Goldberger, An Econo-
mic Model of the United States, 1929–1932. Am-


Time Series Analysis, Cointegration, and Applications†

By Clive W. J. Granger*

The two prize winners in Economics this year would describe themselves as “Econometricians,” so I thought that I should start by explaining that term.

One can begin with the ancient subject of Mathematics which is largely concerned with the discovery of relationships between deterministic variables using a rigorous argument. (A deterministic variable is one whose value is known with certainty.) However, by the middle of the last millennium it became clear that some objects were not deterministic, they had to be described with the use of probabilities, so that Mathematics grew a substantial subfield known as “Statistics.” This later became involved with the analysis of data and a number of methods have been developed for data having what may be called “standard properties.”

However, in some areas of application, the data that they generated were found to be not standard, and so special sub-subfields needed to be developed. For example, Biology produced Biometrics, Psychology gave us Psychometrics, and Economics produced Econometrics.

There are many types of economic data, but the type considered by Robert Engle and myself is known as time series. Consider the measurement of unemployment rates which is an important measure of the health of the economy. Figures are gathered by a government agency and each month a new number is announced. Next month there will be another value, and so forth. String these values together in a simple graph and you get a “time series.”

Rather than show a diagram, I would rather you use internal visualization (I think that you learn more that way). Suppose that you have a loosely strung string of pearls which you throw down, gently, onto a hard table top with the string of pearls roughly stretched out. You will have created a time series with time represented by the distance down the table, the size of the variable as the distance from the bottom edge of the table to a point, and the set of pearls giving the points in the series. As the placement of one pearl will impact where the next one lies, because they are linked together, this series will appear to be rather smooth, and will not have big fluctuations in value from one term to the next.

Time series can vary in many ways: some are gathered very often, others less frequently. Values for many important financial variables are known not merely daily, but can be found within seconds, if they change, such as highly traded stock prices or exchange rates. These are called “high-frequency data” and are the input for Robert Engle’s studies.

At the other extreme, some aspects of the overall, or “macro,” economy, such as national income, consumption, and investment, may be available only quarterly for many countries, and only annually for others. Population data is also available only annually or less frequently. Many of these series are rather smooth, moving with local trends or with long swings, but the swings are not regular. It is this relative smoothness that makes them unsuitable for analysis with standard statistical procedures, which assumes data to have a property known as “stationarity.” Many series in economics, particularly in finance and macroeconomics, do not have this property and can be called “integrated” or, sometimes incorrectly, “nonstationary.” However, when expressed in terms of changes or rates of returns, these derived series appear closer to being stationary. The string of pearls would be “integrated,” as it is a smooth series.

Methods to analyze a single integrated series had been proposed previously by George E. P. Box and Gwilym M. Jenkins (1970) and others, but the joint analysis of pairs, or more, of such series was missing an important feature. It turns out that the difference between a pair of inte-

† This article is a revised version of the lecture Clive W. J. Granger delivered in Stockholm, Sweden, on December 8, 2003, when he received the Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel. The article is copyright © The Nobel Foundation 2003 and is published here with the permission of the Nobel Foundation.

* Department of Economics, University of California-San Diego, La Jolla, CA 92093 (e-mail: cgranger@ucsd.edu).
grated series can be stationary, and this property is known as "cointegration." More formally, two smooth series, properly scaled, can move and turn, but slowly, in similar but not identical fashions, but the distance between them can be stationary.

Suppose that we had two similar chains of pearls and we threw each on the table separately, but for ease of visualization, they do not cross one another. Each would represent a smooth series but would follow different shapes and have no relationship. The distance between the two sets of pearls would also give a smooth series if you plotted it.

However, if the pearls were set in small but strong magnets, it is possible that there would be an attraction between the two chains, and that they would have similar, but not identical, smooth shapes. In that case, the distance between the two sets of pearls would give a stationary series and this would give an example of cointegration.

For cointegration, a pair of integrated, or smooth, series must have the property that a linear combination of them is stationary. Most pairs of integrated series will not have the property, so that cointegration should be considered as a surprise when it occurs. In practice, many pairs of macroeconomic series seem to have the property, as is suggested by economic theory.

Once we know that a pair of variables has the cointegration property, it follows that the variables have a number of other interesting and useful properties. For example, they must both be cointegrated with the same hidden common factor. Further, they can be considered to be generated by what is known as the "error-correction model," in which the change of one of the series is explained in terms of the lag of the difference between the series, possibly after scaling, and lags of the differences of each series. The other series will be represented by a similar dynamic equation. Data generated by such a model are sure to be cointegrated. The error-correction model has been particularly important in making the idea of cointegration practically useful. It was invented by the well-known econometrician Dennis Sargent, who took some famous equations from the theory of economic growth and made them stochastic.

The early development of the cointegration idea was helped greatly by colleagues and friends in the Scandinavian countries, including Søren Johansen and Katerina Juselius in Copenhagen, who developed and applied sophisticated testing procedures; Svend Hylleberg in Århus, who extended the theory to seasonal data; and Eiliv Jansen and his colleagues at the Bank of Norway, who successfully applied it to a large econometric model of Norway. To complete the set, Timo Teräsvirta, who is from Finland but now lives in Stockholm, helped develop models that were useful in nonlinear formulations of cointegration. I am delighted that they are all here as my guests.

The modern macro economy is large, diffuse, and difficult to define, measure, and control. Economists attempt to build models that will approximate it—that will have similar major properties so that one can conduct simple experiments on them, such as determining the impacts of alternative policies or the long-run implications of some new institution. Economic theorists do this using constraints suggested by the theory, whereas the econometrician builds empirical models using what is hopefully relevant data and which captures the main properties of the economy in the past. All models simply assume that the model is correct and extrapolate from there, but hopefully with an indication of uncertainty around future values.

Error-correction models have been a popular form of macro model in recent years, and cointegration is a common element. Applications have been considered using almost all major variables including investment, taxes, consumption, employment, interest rates, government expenditure, and so forth.

It is these types of equations that central banks, the Federal Reserve Bank, and various model builders have found useful for policy simulations and other considerations.

A potentially useful property of forecasts based on cointegration is that when extended some way ahead, the forecasts of the two series will form a constant ratio, as is expected by some asymptotic economic theory. It is this asymptotic result that makes this class of models of some interest to economic theorists who are concerned with "equilibrium." Whether the form of equilibria suggested by the models is related to that discussed by the theorists is unclear.

These ideas and models are fairly easily extended to many variables. Once the idea of cointegration (a name that was my own inven-
tion, incidentally) became public, it was quickly picked up and used by many other econometricians and applied economists. There are now numerous citations and applications based on it. Robert Engle and I quickly realized that the concept of cointegration and its extensions could be used to explain and remove a variety of difficulties that we had observed in our own research and that of others. It seemed to be the missing piece in our approach to modeling groups of series.

An example is a problem known as "spurious regressions." It had been observed, by Paul Newbold and myself in a small simulation published in 1974, that if two independent integrated series were used in a regression, one chosen as the "dependent variable" and the other the "explanatory variable," the standard regression computer package would very often appear to "find" a relationship, whereas in fact there was none. That is, standard statistical methods would find a "spurious regression." This observation led to a great deal of reevaluation of empirical work, particularly in macroeconomics, to see if apparent relationships were correct or not. Many editors had to look again at their list of accepted papers. Putting the analysis in the form of an error-correction model resolves many of the difficulties found with spurious regression.

I am often asked how the idea of cointegration came about; was it the result of logical deduction or a flash of inspiration? In fact, it was rather more prosaic. A colleague, David Hendry, stated that the difference between a pair of integrated series could be stationary. My response was that it could be proved that he was wrong, but in attempting to do so, I showed that he was correct, and generalized it to cointegration, and proved the consequences such as the error-correction representation. I do not always agree with the philosopher Karl Popper, but in his book "The Logic of Scientific Discovery," according to Malachi Haim Hacohen (2000, p. 244), Popper believed that "discovery was not a matter of logic" but rather the application of methodology, which fits the discovery of cointegration. This insight intrigues me partly because Popper's book appeared almost exactly at the time of my birth, in September 1934. At this same time Popper was debating Heisenberg on the relevance of probability theory in physics. It happens to be the case that echoes of that debate still persist, but relating to economics. My position is that it is clear that we can best describe many of the movements of economic variables, and the resulting data, using probabilistic concepts. I should also point out that 1934 was also the year that John M. Keynes finished the first draft of "The General Theory of Employment, Interest, and Money," although it was very many years before I became aware of this book.

As an aside, I wrote this lecture while visiting the Department of Economics of the University of Canterbury in New Zealand, where Karl Popper also spent some years after World War II.

Before considering the usefulness of the new methods of analysis, I would like to take a personal detour. This Prize has climaxed a year which started with my being named a Distinguished Fellow of the American Economic Association. Previously in my career, I have been Chair of two economics departments, yet I have received very little formal training in economics. One-third of my first year as an undergraduate at the University of Nottingham was in economics, with introductions to micro and in national accounts, and that was it. Whatever other knowledge I have, it has come from living among economists for about 40 years, by osmosis, attending seminars, having discussions with them, and general reading. My question is: does this say something about me, or something about the field of economics? I think it is true to say that I am not the first Nobel Prize winner in Economics to have little formal training in economics. I wonder if economics has less basic core material than is necessary for fields such as mathematics, physics, or chemistry. Economics does have a multitude of different aspects, applications, and viewpoints which has to each form their own basis, at least in practice. Economic theory does seem to maintain common concepts and features but these may be quite simplistic and are not necessarily realistic.

Possibly because it is not tied down by too many central concepts but certainly because economics involves a myriad of topics, both theoretical and applied, it is a hotbed of new ideas, concepts, approaches, and models. The availability of more powerful computing at low cost has just increased this activity even more.

In my reading I came across a statement (unfortunately, I have forgotten who the author
was) noting that "economics is a decision science, concerned with decision makers, such as consumers, employers, investors, and policy makers, in various forms of government, institutions, and corporations." I fully accept this viewpoint as it follows that the "purpose of economics is to help decision makers make better decisions." That statement is useful because it gives us a foundation with which we can compare and evaluate specific pieces of economic analysis. We can ask "how will a decision maker find this result useful?"

As I stated before, the main uses for the economic techniques that I helped develop, such as cointegration, were to build statistical models linking major economic variables that both fit the available data better and agree with the preconceptions of the model constructors about what the construction should look like. A major use of these models has been to provide short- and medium-term forecasts for important macro variables, such as consumption, income, investment, and unemployment, all of which are integrated series. The derived growth rates are found to be somewhat forecastable. Much less forecastable are inflation rates and returns from speculative markets, such as stocks, bonds, and exchange rates.

There are a number of stages to the forecasting process: getting the central forecast and then uncertainty bounds around it to give some idea of the risks involved in using this forecast. Finally, previous forecasts have to be gathered and evaluated. Hopefully any tendencies, trends, or swings in the errors can be detected so that one can learn and produce better forecasts in the future. The process of forecast evaluation, plus the use of combinations of forecasts from different series, is an ongoing research project at the University of California, San Diego.

Forecasts do not just come from time series, but also from panels of data, which can be thought of as a group of series of a similar nature measured from different sources. An example would be monthly inflation rates from each of the Scandinavian countries. Once one is clear about the purpose of the analysis, suitable techniques can be formulated.

I have recently been involved in such a project where the purpose is to study the future of the Amazon rainforest in Brazil. This forest covers an area larger than all the countries in the European Union put together, but it is being cut down quite rapidly. I was one of five authors who produced a report (Lykke Anderson et al., 2003) which includes a model that could forecast the decline of the forest under various policy scenarios. The forest is not being cut down for its timber, but to get at the land that the timber stands upon to produce food. Unfortunately, unlike European ex-forest land, its useful life span is often rather short, often becoming "fallow" within five years of being deforested.

The advantage of being an academic econometrician is the possibility of working on data from many areas. I have run pricing experiments in real supermarkets, I have analyzed data from stock markets, commodity prices—particularly gold and silver prices—and interest rates. I also considered electricity demand in small regions, female labor force participation, river flooding, and even sun spots. All data present their own unique problems and I continue to find data analysis fascinating, particularly in economics.

There are plenty of disadvantages to being a statistician working with economic data without very much training in the area, but occasionally it allows one to approach a problem from a different direction than that considered by most economists. As a statistician I am intrigued by the pure magnitude of some of the major economies, although economists pay little attention to this aspect of the real world. For example, in the United States there are about one hundred million households, so total consumption is the sum of all these households' consumptions. The sum over such a large number of families should have very special statistical properties, given various well-known limit theorems. If these properties are not observed, this also has important implications. I think that these, and many other topics concerning aggregation, are worth further study.

An earlier concept that I was concerned with was that of causality. As a postdoctoral student in Princeton in 1959–1960, working with Professors John Tukey and Oskar Morgenstem, I was involved with studying something called the "cross-spectrum," which I will not attempt to explain. Essentially one has a pair of interrelated time series and one would like to know if there are a pair of simple relations, first from the variable X explaining Y and then from the variable Y explaining X. I was having difficulty seeing how to approach this question when I
met Dennis Gabor, who later won the Nobel Prize in Physics in 1971. He told me to read a paper by the eminent mathematician Norbert Weiner, which contained a definition that I might want to consider. It was essentially this definition, somewhat refined and rounded out, that I discussed, together with proposed tests in the mid-1960’s. The statement about causality has just two components:

1. The cause occurs before the effect; and
2. The cause contains information about the effect that is unique, and is in no other variable.

A consequence of these statements is that the causal variable can help forecast the effect variable after other data have first been used. Unfortunately, many users concentrated on this forecasting implication rather than on the original definition.

At that time, I had little idea that so many people had very fixed ideas about causation, but they did agree that my definition was not “true causation” in their eyes; it was only “Granger causation.” I would ask for a definition of true causation, but no one would reply. However, my definition was pragmatic and any applied researcher with two or more time series could apply it, so I got plenty of citations. Of course, many ridiculous papers appeared.

When the idea of cointegration was developed, over a decade later, it became clear immediately that if a pair of series was cointegrated then at least one of them must cause the other. There seems to be no special reason why these two quite different concepts should be related; it is just the way that the mathematics turned out.

As a brief aside for those of you with more technical training, what I have been telling you about so far mostly has been for concepts using linear models. Everything can be generalized to the nonlinear situation and recently efforts have been pushing into using similar concepts in conditional distributions, which is a very general form. It appears that causality will play a basic role in the generalization of the error-correction model, but that is still a work in progress.

I am not sure if the empirical studies on causation have proved to be so useful, although the debate relating to money supply and prices was interesting. The concept does help with the formulation of dynamic models in more useful ways.

I started this lecture talking about econometrics. We econometricians love numbers, so let me end with a few. The first two Nobel Prizes in Economics were to econometricians, Ragnar Frisch and Jan Tinbergen, for which we are very proud. There are now eight of us with the Prize, representing 15 percent of the Economics winners. However, in the current millennium, we represent about 44 percent of the winners, which I view as a healthy local trend.

Over my career and before today, I have met 21 Nobel Laureates: one in Physics (Dennis Gabor, 1970), one in Peace (Phillip Noel Baker, 1959), one in Chemistry (Harold Urey, 1934), plus 18 Prize winners in Economics. Without exception, I have found them to be very fine scholars and also having excellent personalities—willing to help a younger, inexperienced worker when seeking their advice or meeting them socially. I hope that I am able to live up to their very high standard.

REFERENCES


183

\[
\frac{1}{T} \sum_{t=1}^{T} \Delta Y_t^2 \to \sigma^2.
\]

Under the assumption that \( Y_0 = 0 \), the first term in Equation (14.22) can be written

\[
\frac{1}{T} \sum_{t=1}^{T} \Delta Y_t = \frac{1}{T} \frac{1}{2} \sum_{t=1}^{T} \Delta Y_t = \frac{1}{T} \frac{1}{2} \sum_{t=1}^{T} Z_t^2,
\]

which in turn obeys the central limit theorem; that is, \( Y_T^2 / \sqrt{T} \to N(0, \sigma^2) \). Thus

\[
(Y_T^2 / \sqrt{T})^2 - \frac{1}{T} \sum_{t=1}^{T} \Delta Y_t^2 \to \sigma^2 (Z^2 - 1),
\]

where \( Z \) is a standard normal random variable. Recall, however, that the square of a standard normal distribution has a chi-squared distribution with one degree of freedom. It therefore follows from Equation (14.22) that, under the null hypothesis, the numerator in Equation (14.21) has the limiting distribution

\[
\frac{1}{T} \sum_{t=1}^{T} \Delta Y_t \to \frac{\sigma^2}{2} (Z^2 - 1).
\]

(14.23)

The large-sample distribution in Equation (14.23) is different than the usual large-sample normal distribution when the regressor is stationary. Instead, the numerator of the OLS estimator of the coefficient on \( Y_t \) in this Dickey-Fuller regression has a distribution that is proportional to a chi-squared distribution with one degree of freedom, minus one.

This discussion has only considered the numerator of \( T \hat{\theta} \). The denominator also behaves unusually under the null hypothesis: because \( Y_t \) follows a random walk under the null hypothesis, \( \frac{1}{T} \sum_{t=1}^{T} Y_t^2 \) does not converge in probability to a constant. Instead, the denominator in Equation (14.21) is a random variable, even in large samples: under the null hypothesis, \( \frac{1}{T} \sum_{t=1}^{T} Y_t^2 \) converges in distribution jointly with the numerator. The unusual distributions of the numerator and denominator in Equation (14.21) are the source of the nonstandard distribution of the Dickey-Fuller test statistic and the reason that the ADF statistic has its own special table of critical values.

### 14.4 Cointegration

Sometimes two or more series have the same stochastic trend in common. In this special case, referred to as cointegration, regression analysis can reveal long-run relationships among time series variables, but some new methods are needed.

#### Cointegration and Error Correction

Two or more time series with stochastic trends can move together so closely over the long run that they appear to have the same trend component, that is, they appear to have a common trend. For example, two interest rates on U.S. government debt are plotted in Figure 14.2. One of the rates is the interest rate on 90-day U.S. Treasury bills, at an annual rate (R90); the other is the interest rate on a one-year U.S. Treasury bond (R1Y); these interest rates are discussed in Appendix 14.1. The interest rates exhibit the same long-run tendencies or trends: both were low in the 1960s, both rose through the 1970s to peaks in the early 1980s, then both fell through the 1990s. Moreover, the difference between the two series, \( R1Y - R90 \), which is called the "spread" between the two interest rates and is also plotted in Figure 14.2, does not appear to have a trend. That is, subtracting the 90-day interest rate from the one-year interest rate appears to eliminate the trends in both of the individual rates. Said differently, although the two interest rates differ, they appear to share a common stochastic trend: because the trend in each individual series is eliminated by subtracting one series from the other, the two series must have the same trend, that is, they must have a common stochastic trend.

![Figure 14.2](image-url)
Two or more series that have a common stochastic trend are said to be **cointegrated**. The formal definition of cointegration (due to Granger, 1983) is given in Key Concept 14.5. In this section, we introduce a test for whether cointegration is present, discuss estimation of the coefficients of regressions relating cointegrated variables, and illustrate the use of the cointegrating relationship for forecasting. The discussion initially focuses on the case that there are only two variables, \(X_t\) and \(Y_t\).

**Vector error correction model.** Until now, we have eliminated the stochastic trend in an \(I(1)\) variable \(Y_t\) by computing its first difference, \(\Delta Y_t\); the problems created by stochastic trends were then avoided by using \(\Delta Y_t\) instead of \(Y_t\) in time series regressions. If \(X_t\) and \(Y_t\) are cointegrated, however, another way to eliminate the trend is to compute the difference \(Y_t - \theta X_t\). Because the term \(Y_t - \theta X_t\) is stationary, it too can be used in regression analysis.

In fact, if \(X_t\) and \(Y_t\) are cointegrated, the first differences of \(X_t\) and \(Y_t\) can be modeled using a VAR, augmented by including \(\Delta y_t = \theta \Delta X_t\) as an additional regressor:

\[
\begin{align*}
\Delta Y_t &= \beta_{10} + \beta_{11} \Delta Y_{t-1} + \cdots + \beta_{1p} \Delta Y_{t-p} + \gamma_{11} \Delta X_{t-1} + \cdots + \\
&\quad + \gamma_{1p} \Delta X_{t-p} + \alpha_1 (Y_{t-1} - \theta X_{t-1}) + u_t \quad (14.24) \\
\Delta X_t &= \beta_{20} + \beta_{21} \Delta Y_{t-1} + \cdots + \beta_{2p} \Delta Y_{t-p} + \gamma_{21} \Delta X_{t-1} + \cdots + \\
&\quad + \gamma_{2p} \Delta X_{t-p} + \alpha_2 (Y_{t-1} - \theta X_{t-1}) + u_{2t} \quad (14.25)
\end{align*}
\]

The term \(Y_t - \theta X_t\) is called the **error correction term**. The combined model in Equations (14.24) and (14.25) is called a **vector error correction model** (VECM). In a VECM, past values of \(Y_t - \theta X_t\) help to predict future values of \(\Delta Y_t\) and/or \(\Delta X_t\).

**How Can You Tell Whether Two Variables Are Cointegrated?**

There are three ways to decide whether two variables can plausibly be modeled as cointegrated: use expert knowledge and economic theory, graph the series and see whether they appear to have a common stochastic trend, and perform statistical tests for cointegration. All three methods should be used in practice.

First, you must use your expert knowledge of these variables to decide whether cointegration is in fact plausible. For example, the two interest rates in Figure 14.2 are linked together by the so-called expectations theory of the term structure of interest rates. According to this theory, the interest rate on January 1 on the one-year Treasury bond is the average of the interest rate on a 90-day Treasury bill for the first quarter of the year and the expected interest rates on future 90-day Treasury bills issued in the second, third, and fourth quarters of the year; if not, then investors could expect to make money by holding either the one-year Treasury note or a sequence of four 90-day Treasury bills, and they would bid up prices until the expected returns are equalized. If the 90-day interest rate has a random walk stochastic trend, this theory implies that this stochastic trend is inherited by the one-year interest rate and that the difference between the two rates, that is, the spread, is stationary. Thus, the expectations theory of the term structure implies that, if the interest rates are \(I(1)\), then they will be cointegrated with a cointegrating coefficient of \(\theta = 1\) (Exercise 14.2).

Second, visual inspection of the series helps to identify cases in which cointegration is plausible. For example, the graph of the two interest rates in Figure 14.2 shows that each of the series appears to be \(I(1)\) but that the spread appears to be \(I(0)\), so that the two series appear to be cointegrated.

Third, the unit root testing procedures introduced so far can be extended to test for cointegration. The insight on which these tests are based is that if \(Y_t\) and \(X_t\) are cointegrated with cointegrating coefficient \(\theta\), then \(Y_t - \theta X_t\) is stationary; otherwise, \(Y_t - \theta X_t\) is nonstationary (is \(I(1)\)). The hypothesis that \(Y_t\) and \(X_t\) are not cointegrated (that is, that \(Y_t - \theta X_t\) is \(I(1)\)) therefore can be tested by testing the null hypothesis that \(Y_t - \theta X_t\) has a unit root; if this hypothesis is rejected, then \(Y_t\) and \(X_t\) can be modeled as cointegrated. The details of this test depend on whether the cointegrating coefficient \(\theta\) is known.

**Testing for cointegration when \(\theta\) is known.** In some cases expert knowledge or economic theory suggests values of \(\theta\). When \(\theta\) is known, the Dickey-Fuller and
DF-GLS unit root tests can be used to test for cointegration by first constructing the series \( z_t = Y_t - \theta X_t \), then testing the null hypothesis that \( z_t \) has a unit autoregressive root.

**Testing for cointegration when \( \theta \) is unknown.** If the cointegrating coefficient \( \theta \) is unknown then it must be estimated prior to testing for a unit root in the error correction term. This preliminary step makes it necessary to use different critical values for the subsequent unit root test.

Specifically, in the first step the cointegrating coefficient \( \theta \) is estimated by OLS estimation of the regression

\[
Y_t = \alpha + \theta X_t + \varepsilon_t.
\]  
(14.26)

In the second step, a Dickey-Fuller \( t \)-test (with an intercept but no time trend) is used to test for a unit root in the residual from this regression, \( \varepsilon_t \). This two-step procedure is called the Engle-Granger Augmented Dickey-Fuller test for cointegration, or \( \text{EG-ADF} \) (Engle and Granger, 1987).

Critical values of the \( \text{EG-ADF} \) statistic are given in Table 14.2.² The critical values in the first row apply when there is a single regressor in Equation (14.26), so that there are two cointegrated variables \((X_t, Y_t)\). The subsequent rows apply to the case of multiple cointegrated variables, which is discussed at the end of this section.

**Estimation of Cointegrating Coefficients**

If \( X_t \) and \( Y_t \) are cointegrated, then the OLS estimator of the coefficient in the cointegrating regression in Equation (14.26) is consistent. However, in general the OLS estimator has a nonnormal distribution, and inferences based on its \( t \)-statistics can be misleading whether or not those \( t \)-statistics are computed using HAC standard errors. Because of these drawbacks of the OLS estimator of \( \theta \), econometricians have developed a number of other estimators of the cointegrating coefficient.

One such estimator of \( \theta \) that is simple to use in practice is the so-called **dynamic OLS (DOLS)** estimator (Stock and Watson, 1993). The DOLS estimator is based on a modified version of Equation (14.26) that includes past, present, and future values of the change in \( X_t \):

\[
Y_t = \beta_0 + \theta X_t + \sum_{j=0}^{p} \delta_j \Delta X_{t-j} + \varepsilon_t.
\]  
(14.27)

Thus, in Equation (14.27), the regressors are \( X_t, \Delta X_{t-1}, \ldots, \Delta X_{t-p} \). The DOLS estimator of \( \theta \) is the OLS estimator of \( \theta \) in the regression of Equation (14.27).

If \( X_t \) and \( Y_t \) are cointegrated, then the DOLS estimator is efficient in large samples. Moreover, statistical inferences about \( \theta \) and the \( \delta_j \)s in Equation (14.27) based on HAC standard errors are valid. For example, the \( t \)-statistic constructed using the DOLS estimator with HAC standard errors has a standard normal distribution in large samples.

One way to interpret Equation (14.27) is to recall from Section 13.3 that cumulative dynamic multipliers can be computed by modifying the distributed lag regression of \( Y_t \) on \( X_t \) and its lags. Specifically, in Equation (13.7), the cumulative dynamic multipliers were computed by regressing \( Y_t \) on \( \Delta X_t \), lags of \( \Delta X_t \), and \( X_{t-1} \); the coefficient on \( X_{t-1} \) in that specification is the long-run cumulative dynamic multiplier. Similarly, if \( X_t \) were strictly exogenous, then in Equation (14.27), the coefficient on \( X_t \) \( \theta \), would be the long-run cumulative multiplier, that is, the long-run effect on \( Y \) of a change in \( X \). If \( X_t \) is not strictly exogenous, then the coefficients do not have this interpretation. Nevertheless, because \( X_t \) and \( Y_t \) have a common stochastic trend if they are cointegrated, the DOLS estimator is consistent even if \( X_t \) is endogenous.

The DOLS estimator is not the only efficient estimator of the cointegrating coefficient. The first such estimator was developed by Søren Johansen (Johansen, 1988). For a discussion of Johansen's method and of other ways to estimate the cointegrating coefficient, see Hamilton (1994, Chapter 20).

---

²The critical values in Table 14.2 are taken from Fuller (1976) and Phillips and Ouliaris (1990). Following a suggestion by Hansen (1990), the critical values in Table 14.2 are chosen so that they apply whether or not \( X_t \) and \( Y_t \) have drift components.
Even if economic theory does not suggest a specific value of the cointegrating coefficient, it is important to check whether the estimated cointegrating relationship makes sense in practice. Because cointegration tests can be misleading (they can improperly reject the null hypothesis of no cointegration more frequently than they should, and frequently they improperly fail to reject the null), it is especially important to rely on economic theory, institutional knowledge, and common sense when estimating and using cointegrating relationships.

**Extension to Multiple Cointegrated Variables**

The concepts, tests, and estimators discussed here extend to more than two variables. For example, if there are three variables, $Y_1$, $X_{1t}$, and $X_{2t}$, each of which is $I(1)$, then they are cointegrated with cointegrating coefficients $\theta_1$ and $\theta_2$ if $\gamma_1 - \theta_1 X_{1t} - \theta_2 X_{2t}$ is stationary. When there are three or more variables, there can be multiple cointegrating relationships. For example, consider modeling the relationship among three interest rates: the three-month rate, the one-year rate, and the five-year rate ($R_{5y}$). If they are $I(1)$, then the expectations theory of the term structure of interest rates suggests that they will all be cointegrated. One cointegrating relationship suggested by the theory is $R_{90} - R_{1y}$, and a second relationship is $R_{90} - R_{5y}$. (The relationship $R_{1y} - R_{5y}$ is also a cointegrating relationship, but it contains no additional information beyond that in the other relationships because it is perfectly multicollinear with the other two cointegrating relationships.)

The EG-ADF procedure for testing for a single cointegration relationship among multiple variables is the same as for the case of two variables, except that the regression in Equation (14.26) is modified so that both $X_{1t}$ and $X_{2t}$ are regressors; the critical values for the EG-ADF test are given in Table 14.2, where the appropriate row depends on the number of regressors in the first-stage OLS cointegrating regression. The DOLS estimator of a single cointegrating relationship among multiple $X$'s involves including the level of each $X$ along with leads and lags of the first difference of each $X$. Tests for multiple cointegrating relationships can be performed using the system methods, such as Johansen's (1988) method, and the DOLS estimator can be extended to multiple cointegrating relationships by estimating multiple equations, one for each cointegrating relationship. For additional discussion of cointegration methods for multiple variables, see Hamilton (1994).

**A cautionary note.** If two or more variables are cointegrated then the error correction term can help to forecast these variables and, possibly, other related variables. However, cointegration requires the variables to have the same stochastic trends. Trends in economic variables typically arise from complex interactions of disparate forces, and closely related series can have different trends for subtle reasons. If variables that are not cointegrated are incorrectly modeled using a VECM, then the error correction term will be $I(1)$; this introduces a trend into the forecast that can result in poor out-of-sample forecast performance. Thus forecasting using a VECM must be based on a combination of compelling theoretical arguments in favor of cointegration and careful empirical analysis.

**Application to Interest Rates**

As discussed above, the expectations theory of the term structure of interest rates implies that, if two interest rates of different maturities are $I(1)$, then they will be cointegrated with a cointegrating coefficient of $\theta = 1$, that is, the spread between the two rates will be stationary. Inspection of Figure 14.2 provides qualitative support for the hypothesis that the one-year and three-month interest rates are cointegrated. We first use unit root and cointegration test statistics to provide more formal evidence on this hypothesis, then estimate a vector error correction model for these two interest rates.

**Unit root and cointegration tests.** Various unit root and cointegration test statistics for these two series are reported in Table 14.3. The unit root test statistics in the first two rows examine the hypothesis that the two interest rates, the three-month rate ($R_{90}$) and the one-year rate ($R_{1y}$), individually have a unit root. Two of the four statistics in the first two rows fail to reject this hypothesis at the 10% level, and three of the four fail to reject at the 5% level. The exception is the ADF statistic evaluated for the 90-day Treasury bill rate ($-2.96$), which rejects the unit root hypothesis at the 5% level. The ADF and DF-GLS statistics lead to

<table>
<thead>
<tr>
<th>Series</th>
<th>ADF Statistic</th>
<th>DF-GLS Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_{90}$</td>
<td>$-2.96^*$</td>
<td>$-1.88^*$</td>
</tr>
<tr>
<td>$R_{1y}$</td>
<td>$-2.22$</td>
<td>$-1.37$</td>
</tr>
<tr>
<td>$R_{1y} - R_{90}$</td>
<td>$-6.31^{**}$</td>
<td>$-5.59^{**}$</td>
</tr>
<tr>
<td>$R_{1y} - 1.046R_{90}$</td>
<td>$-6.97^{**}$</td>
<td></td>
</tr>
</tbody>
</table>

$R_{90}$ is the interest rate on 90-day U.S. Treasury bills, at an annual rate, and $R_{1y}$ is the interest rate on one-year U.S. Treasury bonds. Regressions were estimated using quarterly data over the period 1962:1–1996:IV. The number of lags in the unit root test statistic regressions were chosen by AIC (six lags maximum). Unit root test statistics are significant at the *10%, *5%, or **1% significance level.
different conclusions for this variable (the ADF test rejects the unit root hypothesis at the 5% level while the DF-GLS test does not), which means that we must exercise some judgment in deciding whether these variables are plausibly modeled as $I(1)$. Taken together, these results suggest that the interest rates are plausibly modeled as $I(1)$.

The unit root statistics for the spread, $R1y_{t} - R90_t$, test the further hypothesis that these variables are not cointegrated against the alternative that they are. The null hypothesis that the spread contains a unit root is rejected at the 1% level using both unit root tests. Thus we reject the hypothesis that the series are not cointegrated against the alternative that they are, with a cointegrating coefficient $\theta = 1$. Taken together, the evidence in the first three rows of Table 14.3 suggests that these variables plausibly can be modeled as cointegrated with $\theta = 1$.

Because in this application economic theory suggests a value for $\theta$ (the expectations theory of the term structure suggests that $\theta = 1$) and because the error correction term is $I(0)$ when this value is imposed (the spread is stationary), in principle it is not necessary to use the EG-ADF test, in which $\theta$ is estimated. Nevertheless, we compute the test as an illustration. The first step in the EG-ADF test is to estimate $\theta$ by the OLS regression of one variable on the other; the result is

$$\text{\hat{R1y}_{t}} = 0.361 + 1.046R90_{t}, \hat{R}^2 = 0.973. \quad (14.28)$$

The second step is computing the ADF statistic for the residual from this regression, $\hat{z}_t$. The result, given in the final row of Table 14.3, is less than the 1% critical value of $-3.96$ in Table 14.2, so the null hypothesis that $\hat{z}_t$ has a unit autoregressive root is rejected. This statistic also points towards treating the two interest rates as cointegrated. Note that no standard errors are presented in Equation (14.28) because, as previously discussed, the OLS estimator of the cointegrating coefficient has a nonnormal distribution and its t-statistic is not normally distributed, so presenting standard errors (HAC or otherwise) would be misleading.

A vector error correction model of the two interest rates. If $Y_t$ and $X_t$ are cointegrated, then forecasts of $\Delta Y_t$ and $\Delta X_t$ can be improved by augmenting a VAR of $\Delta Y_t$ and $\Delta X_t$ by the lagged value of the error correction term, that is, by computing forecasts using the VECM in Equations (14.24) and (14.25). If $\theta$ is known, then the unknown coefficients of the VECM can be estimated by OLS, including $z_{t-1} = Y_{t-1} - \theta X_{t-1}$ as an additional regressor. If $\theta$ is unknown, then the

VECM can be estimated using $\hat{z}_{t-1}$ as a regressor, where $\hat{z}_t = Y_t - \hat{\theta}X_t$, where $\hat{\theta}$ is an estimator of $\theta$.

In the application to the two interest rates, theory suggests that $\theta = 1$, and the unit root tests supported modeling the two interest rates as cointegrated with a cointegrating coefficient of 1. We therefore specify the VECM using the theoretically suggested value of $\theta = 1$, that is, by adding the lagged value of the spread, $\hat{R1y}_{t-1} - \hat{R90}_{t-1}$, to a VAR in $\Delta \hat{R1y}_t$ and $\Delta \hat{R90}_t$. Specified with two lags of first differences, the resulting VECM is

$$\text{\hat{R1y}_{t}} = 0.14 - 0.24\text{\hat{R90}_{t-1}} - 0.44\text{\hat{R90}_{t-2}} - 0.01\text{\Delta \hat{R1y}_{t-1}} \quad (14.29)$$

$$+ 0.15\Delta \hat{R1y}_{t-2} - 0.18(\text{\hat{R1y}_{t-1} - \hat{R90}_{t-1}}) \quad \begin{pmatrix} 0.17 \\ 0.32 \end{pmatrix} \quad \begin{pmatrix} 0.34 \\ 0.39 \end{pmatrix}$$

$$\text{\hat{R90}_{t}} = 0.36 - 0.14\text{\Delta \hat{R90}_{t-1}} - 0.33\text{\Delta \hat{R90}_{t-2}} - 0.11\text{\Delta \hat{R1y}_{t-1}} \quad (14.30)$$

$$+ 0.10\Delta \hat{R1y}_{t-2} - 0.52(\text{\hat{R1y}_{t-1} - \hat{R90}_{t-1}}) \quad \begin{pmatrix} 0.16 \\ 0.30 \end{pmatrix} \quad \begin{pmatrix} 0.29 \\ 0.35 \end{pmatrix}$$

In the first equation, none of the coefficients are individually significant at the 5% level and the coefficients on the lagged first differences of the interest rates are not jointly significant at the 5% level. In the second equation, the coefficients on the lagged first differences are not jointly significant, but the coefficient on the lagged spread (the error correction term), which is estimated to be $-0.52$, has a t-statistic of $-2.17$, so it is statistically significant at the 5% level. Although lagged values of the first difference of the interest rates are not useful for predicting future interest rates, the lagged spread does help to predict the change in the future Treasury bond rate. When the one-year rate exceeds the 90-day rate, the one-year rate is forecasted to fall in the future.

14.5 Conditional Heteroskedasticity

The phenomenon that some times are tranquil while others are not—that is, that volatility occurs in clusters—shows up in many economic time series. This section presents a pair of models for quantifying volatility clustering or, as it is also known, conditional heteroskedasticity.
Vector Autoregressions

James H. Stock and Mark W. Watson

Macroeconomists do four things: describe and summarize macroeconomic data, make macroeconomic forecasts, quantify what we do or do not know about the true structure of the macroeconomy, and advise (and sometimes become) macroeconomic policymakers. In the 1970s, these four tasks—data description, forecasting, structural inference and policy analysis—were performed using a variety of techniques. These ranged from large models with hundreds of equations to single-equation models that focused on interactions of a few variables to simple univariate time series models involving only a single variable. But after the macroeconomic chaos of the 1970s, none of these approaches appeared especially trustworthy.

Two decades ago, Christopher Sims (1980) provided a new macroeconometric framework that held great promise: vector autoregressions (VARs). A univariate autoregression is a single-equation, single-variable linear model in which the current value of a variable is explained by its own lagged values. A VAR is an n-equation, n-variable linear model in which each variable is in turn explained by its own lagged values, plus current and past values of the remaining n – 1 variables. This simple framework provides a systematic way to capture rich dynamics in multiple time series, and the statistical toolkit that came with VARs was easy to use and to interpret. As Sims (1980) and others argued in a series of influential early papers, VARs held out the promise of providing a coherent and credible approach to data description, forecasting, structural inference and policy analysis.

In this article, we assess how well VARs have addressed these four macroecono-

---

1. Readers interested in more detail than provided in this brief tutorial should see Hamilton’s (1994) textbook or Watson’s (1994) survey article.
2. The inflation data are computed as \( \pi_t = 400\text{log}(P_t/P_{t-1}) \), where \( P_t \) is the chain-weighted GDP price index and \( \pi_t \) is the civilian unemployment rate. Quarterly data on \( \pi_t \) and \( R_t \) are formed by taking quarterly averages of their monthly values.
3. Frequently, the Akaike (AIC) or Bayesian (BIC) information criteria are used; for a discussion, see Lütkepohl (1995, chapter 4).
macroeconomic applications—then the error terms in the reduced form model will also be correlated across equations.

A recursive VAR constructs the error terms in each regression equation to be uncorrelated with the error in the preceding equations. This is done by judiciously including some contemporaneous values as regressors. Consider a three-variable VAR, ordered as 1) inflation, 2) the unemployment rate, and 3) the interest rate. In the first equation of the corresponding recursive VAR, inflation is the dependent variable, and the regressors are lagged values of all three variables. In the second equation, the unemployment rate is the dependent variable, and the regressors are lags of all three variables plus the current value of the inflation rate. The interest rate is the dependent variable in the third equation, and the regressors are lags of all three variables, the current value of the inflation rate plus the current value of the unemployment rate. Estimation of each equation by ordinary least squares produces residuals that are uncorrelated across equations.4 Evidently, the results depend on the order of the variables: changing the order changes the VAR equations, coefficients, and residuals, and there are \( n! \) recursive VARs representing all possible orderings.

A structural VAR uses economic theory to sort out the contemporaneous links among the variables (Bernanke, 1986; Blanchard and Watson, 1986; Sims, 1986). Structural VARs require "identifying assumptions" that allow correlations to be interpreted causally. These identifying assumptions can involve the entire VAR, so that all of the causal links in the model are spelled out, or just a single equation, so that only a specific causal link is identified. This produces instrumental variables that permit the contemporaneous links to be estimated using instrumental variables regression. The number of structural VARs is limited only by the inventiveness of the researcher.

In our three-variable example, we consider two related structural VARs. Each incorporates a different assumption that identifies the causal influence of monetary policy on unemployment, inflation and interest rates. The first relies on a version of the "Taylor rule," in which the Federal Reserve is modeled as setting the interest rate based on past rates of inflation and unemployment.5 In this system, the Fed sets the federal funds rate \( R \) according to the rule

\[
R = \pi^* + 1.5(\pi - \pi^*) - 1.25(\bar{u} - u^*) + \text{lagged values of } \pi, \bar{u}, u + \varepsilon,
\]

where \( \pi^* \) is the desired real rate of interest, \( \pi \) and \( \bar{u} \) are the average values of inflation and unemployment rate over the past four quarters, \( \pi^* \) and \( u^* \) are the target values of inflation and unemployment, and \( \varepsilon \) is the error in the equation. This relationship becomes the interest rate equation in the structural VAR.

4 In the jargon of VARs, this algorithm for estimating the recursive VAR coefficients is equivalent to estimating the reduced form, then computing the Cholesky factorization of the reduced form VAR covariance matrix; see Lütkepohl (1993, chapter 2).
5 Taylor's (1993) original rule used the output gap instead of the unemployment rate. Our version uses Okun's Law (with a coefficient of 2.5) to replace the output gap with unemployment rate.

The equation error, \( \varepsilon \), can be thought of as a monetary policy "shock," since it represents the extent to which actual interest rates deviate from this Taylor rule. This shock can be estimated by a regression with \( R_t = 1.25 \pi_t + 1.25 \bar{u}_t \) as the dependent variable, and a constant and lags of interest rates, unemployment and inflation on the right-hand side.

The Taylor rule is "backward looking" in the sense that the Fed reacts to past information (\( \pi_t \) and \( \bar{u}_t \) are averages of the past four quarters of inflation and unemployment), and several researchers have argued that Fed behavior is more appropriately described by forward-looking behavior. Because of this, we consider another variant of the model in which the Fed reacts to forecasts of inflation and unemployment four quarters in the future. This Taylor rule has the same form as the rule above, but with \( \pi_t \) and \( u_t \) replaced by four-quarter ahead forecasts computed from the reduced form VAR.

Putting the Three-Variable VAR Through Its Paces

The different versions of the inflation-unemployment-interest rate VAR are put through their paces by applying them to the four macroeconomic tasks. First, the reduced form VAR and a recursive VAR are used to summarize the comovements of these three series. Second, the reduced form VAR is used to forecast the variables, and its performance is assessed against some alternative benchmark models. Third, the two different structural VARs are used to estimate the effect of a policy-induced surprise move in the federal funds interest rate on future rates of inflation and unemployment. Finally, we discuss how the structural VAR could be used for policy analysis.

Data Description

Standard practice in VAR analysis is to report results from Granger-causality tests, impulse responses and forecast error variance decompositions. These statistics are computed automatically (or nearly so) by many econometrics packages (RATS, Eviews, TSP and others). Because of the complicated dynamics in the VAR, these statistics are more informative than are the estimated VAR regression coefficients or R² statistics, which typically go unreported.

Granger-causality statistics examine whether lagged values of one variable help to predict another variable. For example, if the unemployment rate does not help predict inflation, then the coefficients on the lags of unemployment will all be zero in the reduced-form inflation equation. Panel A of Table 1 summarizes the Granger-causality results for the three-variable VAR. It shows the p-values associated with the F-tests for testing whether the relevant sets of coefficients are zero. The unemployment rate helps to predict inflation at the 5 percent significance level (the p-value is 0.02, or 2 percent), but the federal funds interest rate does not (the p-value is 0.27). Inflation does not help to predict the unemployment rate, but the federal funds rate does. Both inflation and the unemployment rates help predict the federal funds interest rate.
Table 1
VAR Descriptive Statistics for ($\pi$, $u$, $R$)

A. Granger-Causality Test

<table>
<thead>
<tr>
<th>Regressor</th>
<th>$\pi$</th>
<th>$u$</th>
<th>$R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi$</td>
<td>0.00</td>
<td>0.31</td>
<td>0.00</td>
</tr>
<tr>
<td>$u$</td>
<td>0.02</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$R$</td>
<td>0.97</td>
<td>0.01</td>
<td>0.00</td>
</tr>
</tbody>
</table>

B. Variance Decompositions from the Recursive VAR Ordered as $\pi$, $u$, $R$

B.i. Variance Decomposition of $\pi$

<table>
<thead>
<tr>
<th>Forecast Horizon</th>
<th>Forecast Standard Error</th>
<th>Variance Decomposition (Percentage Points)</th>
<th>$\pi$</th>
<th>$u$</th>
<th>$R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.96</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1.34</td>
<td>83</td>
<td>10</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>1.75</td>
<td>82</td>
<td>17</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>1.97</td>
<td>82</td>
<td>16</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

B.ii. Variance Decomposition of $u$

<table>
<thead>
<tr>
<th>Forecast Horizon</th>
<th>Forecast Standard Error</th>
<th>Variance Decomposition (Percentage Points)</th>
<th>$\pi$</th>
<th>$u$</th>
<th>$R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.25</td>
<td>1</td>
<td>99</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0.64</td>
<td>0</td>
<td>98</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>0.79</td>
<td>7</td>
<td>82</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>0.92</td>
<td>16</td>
<td>66</td>
<td>18</td>
<td>1</td>
</tr>
</tbody>
</table>

B.iii. Variance Decomposition of $R$

<table>
<thead>
<tr>
<th>Forecast Horizon</th>
<th>Forecast Standard Error</th>
<th>Variance Decomposition (Percentage Points)</th>
<th>$\pi$</th>
<th>$u$</th>
<th>$R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.85</td>
<td>2</td>
<td>19</td>
<td>79</td>
<td>79</td>
</tr>
<tr>
<td>4</td>
<td>1.84</td>
<td>9</td>
<td>50</td>
<td>41</td>
<td>41</td>
</tr>
<tr>
<td>8</td>
<td>2.44</td>
<td>12</td>
<td>60</td>
<td>28</td>
<td>28</td>
</tr>
<tr>
<td>12</td>
<td>2.63</td>
<td>16</td>
<td>59</td>
<td>25</td>
<td>25</td>
</tr>
</tbody>
</table>

Notes: $\pi$ denotes the rate of price inflation, $u$ denotes the unemployment rate and $R$ denotes the Federal Funds interest rate. The entries in Panel A show the $p$-values for tests that lags of the variable in the row labeled Regressor do not enter the reduced form equation for the column variable labeled Dependent Variable. The results were computed from a VAR with four lags and a constant term over the 1960:1-2000:I:IV sample period.

Impulse responses trace out the response of current and future values of each of the variables to a one-unit increase in the current value of one of the VAR errors, assuming that this error returns to zero in subsequent periods and that all other errors are equal to zero. The implied thought experiment of changing one error while holding the others constant makes most sense when the errors are uncorrelated across equations, so impulse responses are typically calculated for recursive and structural VARs.

The impulse responses for the recursive VAR, ordered $\pi$, $u$, $R$, are plotted in Figure 1. The first row shows the effect of an unexpected 1 percentage point increase in inflation on all three variables, as it works through the recursive VAR system with the coefficients estimated from actual data. The second row shows the effect of an unexpected increase of 1 percentage point in the unemployment rate, and the third row shows the corresponding effect for the interest rate. Also plotted are ±1 standard error bands, which yield an approximate 66 percent confidence interval for each of the impulse responses. These estimated impulse responses show patterns of persistent common variation. For example, an unexpected rise in inflation slowly fades away over 24 quarters and is associated with a persistent increase in unemployment and interest rates.

The forecast error decomposition is the percentage of the variance of the error made in forecasting a variable (say, inflation) due to a specific shock (say, the error term in the unemployment equation) at a given horizon (like two years). Thus, the forecast error decomposition is like a partial $R^2$ for the forecast error, by forecast horizon. These are shown in Panel B of Table 1 for the recursive VAR. They suggest considerable interaction among the variables. For example, at the 12-quarter horizon, 75 percent of the error in the forecast of the federal funds interest rate is attributed to the inflation and unemployment shocks in the recursive VAR.

Forecasting

Multistep-ahead forecasts, computed by iterating forward the reduced form VAR, are assessed in Table 2. Because the ultimate test of a forecasting model is its out-of-sample performance, Table 2 focuses on pseudo out-of-sample forecasts over the period from 1985:1 to 2000:IV. It examines forecast horizons of two quarters, four quarters and eight quarters. The forecast $h$ steps ahead is computed by estimating the VAR through a given quarter, making the forecast $h$ steps ahead, reestimating the VAR through the next quarter, making the next forecast and so on through the forecast period.6

As a comparison, pseudo out-of-sample forecasts were also computed for a univariate autoregression with four lags—that is, a regression of the variable on lags

6 Forecasts like these are often referred to as pseudo or "simulated" out-of-sample forecasts to emphasize that they simulate how these forecasts would have been computed in real time, although, of course, this exercise is conducted retrospectively, not in real time. Our experiment deviates slightly from what would have been computed in real time because we use the current data, which includes later revisions made to the inflation and unemployment data by statistical agencies, rather than the data available in real time.
of its own past values—and for a random walk (or "no change") forecast. Inflation rate forecasts were made for the average value of inflation over the forecast period, while forecasts for the unemployment rate and interest rate were made for the final quarter of the forecast period. Table 2 shows the root mean square forecast error for each of the forecasting methods. (The mean squared forecast error is computed as the average squared value of the forecast error over the 1985–2000 out-of-sample period, and the resulting square root is the root mean squared forecast error reported in the table.) Table 2 indicates that the VAR either does no worse than or improves upon the univariate autoregression and that both improve upon the random walk forecast.

**Structural Inference**

What is the effect on the rates of inflation and unemployment of a surprise 100 basis-point increase in the federal funds interest rate? Translated into VAR jargon,

**Table 2**

<table>
<thead>
<tr>
<th>Forecast Horizon</th>
<th>Inflation Rate</th>
<th>Unemployment Rate</th>
<th>Interest Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RW</td>
<td>AR</td>
<td>VAR</td>
</tr>
<tr>
<td>2 quarters</td>
<td>0.20</td>
<td>0.70</td>
<td>0.68</td>
</tr>
<tr>
<td>4 quarters</td>
<td>0.20</td>
<td>0.65</td>
<td>0.68</td>
</tr>
<tr>
<td>8 quarters</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Note: Entries are the root mean squared error of forecasts computed recursively for univariate and vector autoregressions (each with four lags) and a random walk ("no change") model. Results for the random walk and univariate autoregressions are shown in columns labeled RW and AR, respectively. Each model was estimated using data from 1964I through the beginning of the forecast period. Forecasts for the inflation rate are for the average value of inflation over the period. Forecasts for the unemployment rate and interest rate are for the final quarter of the forecast period.

this question becomes: What are the impulse responses of the rates of inflation and unemployment to the monetary policy shock in a structural VAR?

The solid line in Figure 2 plots the impulse responses computed from our model with the backward-looking Taylor rule. It shows the inflation, unemployment, and real interest rate responses to a 1 percentage point shock in the nominal federal funds rate. The initial rate hike results in the real interest rate exceeding 50 basis points for six quarters. Although inflation is eventually reduced by approximately 0.5 percentage points, the lags are long, and most of the action occurs in the third year after the contraction. Similarly, the rate of unemployment rises by approximately 0.2 percentage points, but most of the economic slowdown is in the third year after the rate hike.

How sensitive are these results to the specific identifying assumption used in this structural VAR—that the Fed follows the backward-looking Taylor rule? As it happens, very sensitive. The dashed line in Figure 2 plots the impulse responses computed from the structural VAR with the forward-looking Taylor rule. The impulse responses in real interest rates are broadly similar under either rule. However, in the forward-looking model the monetary shock produces a 0.5 percentage point increase in the unemployment rate within a year, and the rate of inflation drops sharply at first, fluctuates, then leaves a net decline of 0.5 percentage points after six years. Under the backward-looking rule, this 100 basis-point rate hike produces a mild economic slowdown and a modest decline in inflation several years hence; under the forward-looking rule, by this same action the Fed wins a major victory against inflation at the cost of a swift and sharp recession.

**Policy Analysis**

In principle, our small structural VAR can be used to analyze two types of policies: surprise monetary policy interventions and changing the policy rule, like shifting from a Taylor rule (with weight on both unemployment and inflation) to an explicit inflation targeting rule.
Impulse Responses of Monetary Policy Shocks for Different Taylor Rule Identifying Assumptions

Response of Inflation

Response of Unemployment

Response of Real Interest Rates

Note: The solid line is compared with the backward-looking Taylor rule; the dashed line, with the forward-looking Taylor rule.

If the intervention is an unexpected movement in the federal funds interest rate, then the estimated effect of this policy on future rates of inflation and unemployment is summarized by the impulse response functions plotted in Figure 2. This might seem somewhat odd policy, but the same mechanics can be used to evaluate a more realistic intervention, such as raising the federal funds rate by 50 basis points and sustaining this increase for one year. This policy can be engineered in a VAR by using the right sequence of monetary policy innovations to hold the federal funds interest rate at this sustained level for four quarters, taking into account that in the VAR, actions on interest rates in earlier quarters affect those in later quarters (Sims, 1982; Waggoner and Zha, 1999).

Analysis of the second type of policy—a shift in the monetary rule itself—is more complicated. One way to evaluate a new policy rule candidate is to ask what would be the effect of monetary and nonmonetary shocks on the economy under the new rule. Since this question involves all the structural disturbances, answering it requires a complete macroeconomic model of the simultaneous determination of all the variables, and this means that all of the causal links in the structural VAR must be specified. In this case, policy analysis is carried out as follows: a structural VAR is estimated in which all the equations are identified, then a new model is formed by replacing the monetary policy rule. Comparing the impulse responses in the two models shows how the change in policy has altered the effects of monetary and nonmonetary shocks on the variables in the model.

How Well Do VARs Perform the Four Tasks?

We now turn to an assessment of VARs in performing the four macroeconomic tasks, highlighting both successes and shortcomings.

Data Description

Because VARs involve current and lagged values of multiple time series, they capture comovements that cannot be detected in univariate or bivariate models. Standard VAR summary statistics like Granger-causality tests, impulse response functions and variance decompositions are well-accepted and widely used methods for portraying these comovements. These summary statistics are useful because they provide targets for theoretical macroeconomic models. For example, a theoretical model that implied that interest rates should Granger-cause inflation but unemployment should not would be inconsistent with the evidence in Table 1.

Of course, the VAR methods outlined here have some limitations. One is that the standard methods of statistical inference (such as computing standard errors for impulse responses) may give misleading results if some of the variables are highly persistent. Another limitation is that, without modification, standard VARs miss nonlinearities, conditional heteroskedasticity and drifts or breaks in parameters.

Forecasting

Small VARs like our three-variable system have become a benchmark against which new forecasting systems are judged. But while useful as a benchmark, small VARs of two or three variables are often unstable and thus poor predictors of the future (Stock and Watson, 1996).

State-of-the-art VAR forecasting systems contain more than three variables and allow for time-varying parameters to capture important drifts in coefficients (Sims, 1995). However, adding variables to the VAR creates complications, because the number of VAR parameters increases as the square of the number of variables; a nine-variable, four-lag VAR has 335 unknown coefficients (including the inter-

\* Bootstrap methods provide some improvements (Elliott, 1999) for inference about impulse responses, but treatment of this problem that are fully satisfactory theoretically are elusive (Stock, 1997; Wright, 2000).
cepts). Unfortunately, macroeconomic time series data cannot provide reliable estimates of all these coefficients without further restrictions.

One way to control the number of parameters in large VAR models is to impose a common structure on the coefficients, for example using Bayesian methods, an approach pioneered by Litterman (1980) (six variables) and Sims (1993) (nine variables). These efforts have paid off, and these forecasting systems have solid real-time track records (McNees, 1990; Zarnowitz and Braun, 1993).

Structural Inference

In our three-variable VAR in the previous section, the estimated effects of a monetary policy shock on the rates of inflation and unemployment (summarized by the impulse responses in Figure 2) depend on the details of the presumed monetary policy rule followed by the Federal Reserve. Even modest changes in the assumed rule result in substantial changes in these impulse responses. In other words, the estimates of the structural impulse responses hinge on detailed institutional knowledge of how the Fed sets interest rates.  

Of course, the observation that results depend on assumptions is hardly new. The operative question is whether the assumptions made in VAR models are any more compelling than in other econometric models. This is a matter of heated debate and is thoughtfully discussed by Leeper, Sims and Zha (1996), Christiano, Eichenbaum and Evans (1999), Cochrane (1998), Rudebusch (1998) and Sims (1998). Below are three important criticisms of structural VAR modeling.

First, what really makes up the VAR "shocks"? In large part, these shocks, like those in conventional regression, reflect factors omitted from the model. If these factors are correlated with the included variables, then the VAR estimates will contain omitted variable bias. For example, officials at the Federal Reserve might scoff at the idea that they mechanically followed a Taylor rule, or any other fixed-coefficient mechanical rule involving only a few variables; rather, they suggest that their decisions are based on a subtle analysis of very many macroeconomic factors, both quantitative and qualitative. These considerations, when omitted from the VAR, end up in the error term and (incorrectly) become part of the estimated structural "shocks" used to estimate an impulse response. A concrete example of this in the VAR literature involves the "price puzzle." Early VARs showed an odd result: inflation tended to increase following monetary policy tightening. One explanation for this (Sims, 1992) was that the Fed was looking forward when it set interest rates and that simple VARs omitted variables that could be used to predict future inflation. When these omitted variables intimated an increase in inflation, the Fed tended to increase interest rates. Thus, these VAR interest rate shocks pressed increases in inflation. Because of omitted variables, the VAR mistakenly labeled these increases in interest rates as monetary shocks, which led to biased impulse responses. Indeed, Sims's explanation of the price puzzle has led to the practice of including commodity prices in VARs to attempt to control for predicted future inflation.

Second, policy rules change over time, and formal statistical tests reveal widespread instability in low-dimensional VARs (Stock and Watson, 1996). Constant parameter structural VARs that miss this instability are improperly identified. For example, several researchers have documented instability in monetary policy rules (for example, Bernanke and Blinder, 1992; Bernanke and Mihov, 1998; Clarida, Gali and Gertler, 2000; Boivin, 2000), and this suggests misspecification in constant coefficient VAR models (like our three-variable example) that are estimated over long sample periods.

Third, the timing conventions in VARs do not necessarily reflect real-time data availability, and this undercuts the common method of identifying restrictions based on timing assumptions. For example, a common assumption made in structural VARs is that variables like output and inflation are sticky and do not respond "within the period" to monetary policy shocks. This seems plausible over the period of a single day, but becomes less plausible over a month or quarter.

In this discussion, we have carefully distinguished between recursive and structural VARs: recursive VARs use an arbitrary mechanical method to model contemporaneous correlation in the variables, while structural VARs use economic theories to associate these correlations with causal relationships. Unfortunately, in the empirical literature the distinction is often murky. It is tempting to develop economic "theories" that, conveniently, lead to a particular recursive ordering of the variables, so that their "structural" VAR simplifies to a recursive VAR, a structure called a "Wold causal chain." We think researchers yield to this temptation far too often. Such cobbled-together theories, even if superficially plausible, often fall apart on deeper inspection. Rarely does it add value to regarbage a recursive VAR and sell it as structural.

Despite these criticisms, we think it is possible to have credible identifying assumptions in a VAR. One approach is to exploit detailed institutional knowledge.

An example of this is the study by Blanchard and Perotti (1999) of the macroeconomic effects of fiscal policy. They argue that the tax code and spending rules impose tight constraints on the way that taxes and spending vary within the quarter, and they use these constraints to identify the exogenous changes in taxes and spending necessary for causal analysis. Another example is Bernanke and Mihov (1998), who use a model of the reserves market to identify monetary policy shocks. A different approach to identification is to use long-run restrictions to identify shocks; for example, King, Plosser, Stock and Watson (1991) use the long-run neutrality of money to identify monetary shocks. However, assumptions based on the infinite future raise questions of their own (Faust and Leeper, 1997).

A constructive approach is to explicitly consider the uncertainties in the assumptions that underlie structural VAR analysis and see what inferences, or range of inferences, still can be made. For example, Faust (1998) and Uhlig (1999).
discuss inference methods that can be applied using only inequality restrictions on the theoretical impulse responses (for example, monetary contractions do not cause booms).

**Policy Analysis**

Two types of policies can be analyzed using a VAR: one-off innovations, in which the same rule is maintained; and changes in the policy rule. The estimated effect of one-off innovations is a function of the impulse responses to a policy innovation, and potential pitfalls associated with these have already been discussed. Things are more difficult if one wants to estimate the effect of changing policy rules. If the true structural equations involve expectations (a.k.a., an expectational Phillips curve), then the expectations will depend on the policy rule; thus, in general, all VAR coefficients will depend on the rule. This is just a version of the Lucas (1976) critique. The practical importance of the Lucas critique for this type of VAR policy analysis is a matter of debate.

**After Twenty Years of VARs**

VARs are powerful tools for describing data and for generating reliable multivariate benchmark forecasts. Technical work remains, most notably extending VARs to higher dimensions and richer nonlinear structures. Even without these improvements, however, VARs have made lasting contributions to the macroeconomist's toolkit for tackling these two tasks. Whether 20 years of VARs have produced lasting contributions to structural inference and policy analysis is more debatable. Structural VARs can capture rich dynamic properties of multiple time series, but their structural implications are only as sound as their identification schemes. While there are some examples of thoughtful treatments of identification in VARs, far too often in the VAR literature the central issue of identification is handled by ignoring it. In some fields of economics, such as labor economics and public finance, identification can be obtained credibly using natural experiments that permit some exogenous variation to be teased out of a relationship otherwise fraught with endogeneity and omitted variables bias. Unfortunately, these kinds of natural experiments are rare in macroeconomics.

Although VARs have limitations when it comes to structural inference and policy analysis, so do the alternatives. Calibrated dynamic stochastic general equilibrium macroeconomic models are explicit about causal links and expectations and provide an intellectually coherent framework for policy analysis. But the current generation of these models do not fit the data well. At the other extreme, single-equation models, for example, regressions of inflation against lagged interest rates, are easy to estimate and sometimes can produce good forecasts. But if it is difficult to distinguish correlation and causality in a VAR, it is even more so in single-equation models, which can, in any event, be viewed as one equation pulled from a larger VAR. Used wisely and based on economic reasoning and institutional detail, VARs both can fit the data and, at their best, can provide sensible estimates of some causal connections. Developing and melding good theory and institutional detail with flexible statistical methods like VARs should keep macroeconomists busy well into the new century.

We thank Jean Boivin, Olivier Blanchard, John Cochrane, Charles Evans, Ken Kuttner, Eric Leeper, Glenn Rudebusch, Chris Sims, John Taylor, Tao Zha and the editors for useful suggestions. This research was funded by NSF grant SBR-9730489.

**References**


James H. Stock and Mark W. Watson 115


of the $j$th orthogonalized

ring of the variables.

$E(y_{t+n}) = \mathbf{\mu}$: the unconditional calculation of the portion

of the $j$th.

Alternatively, recalling that $a_j \sqrt{\text{Var}(u_j)}$ is equal to $p_j$, the $j$th column of

the Cholesky factor $P$, result [11.6.6] can equivalently be written as

$$MSE(y_{t+n}) = \sum_{j=1}^{n} \left( a_j p_j + \mathbf{\psi}_j p_j \mathbf{\psi}_j^t + \mathbf{\psi}_j^t p_j \mathbf{\psi}_j \right) + \ldots + \mathbf{\psi}_{t-1} p_j \mathbf{\psi}_j$$  \hspace{1cm} [11.5.7]

11.6. Vector Autoregressions and Structural Econometric Models

Pitfalls in Estimating Dynamic Structural Models

The vector autoregression was introduced in Section 10.1 as a statistical description of the dynamic interrelations between $n$ different variables contained in the vector $y$. This description made no use of prior theoretical ideas about how these variables are expected to be related, and therefore cannot be used to test our theories or interpret the data in terms of economic principles. This section explores the relation between VARs and structural econometric models.

Suppose that we would like to estimate a money demand function that expresses the public's willingness to hold cash as a function of the level of income and interest rates. The following specification was used by some early researchers:

$$M_t - P_t = \beta_0 + \beta_1 Y_t + \beta_2 I_t + \beta_3 (M_{t-1} - P_{t-1}) + v_t^p. \hspace{1cm} [11.6.1]$$

Here, $M_t$ is the log of the nominal money balances held by the public at date $t$, $P_t$ is the log of the aggregate price level, $Y_t$ is the log of real GNP, and $I_t$ is a nominal interest rate. The parameters $\beta_1$ and $\beta_2$ represent the effect of income and interest rates on desired cash holdings. Part of the adjustment in money balances to a change in income is thought to take place immediately, with further adjustments coming in subsequent periods. The parameter $\beta_3$ characterizes this partial adjustment. The disturbance $v_t^p$ represents factors other than income and interest rates that influence money demand.

It was once common practice to estimate such a money demand equation with Cochrane-Orcutt adjustment for first-order serial correlation. The implicit assumption behind this procedure is that

$$v_t^p = \rho v_{t-1}^p + u_t^p. \hspace{1cm} [11.6.2]$$

where $u_t^p$ is white noise. Write equation [11.6.2] as $(1 - \rho L)v_t^p = u_t^p$ and multiply both sides of [11.6.1] by $(1 - \rho L)$:

$$M_t - P_t = (1 - \rho)\beta_0 + \beta_1 Y_{t-1} + \beta_2 I_{t-1} + (\beta_3 + \rho)(M_{t-1} - P_{t-1}) - \beta_3 (M_{t-2} - P_{t-2}) + u_t^p. \hspace{1cm} [11.6.3]$$

Equation [11.6.3] is a restricted version of

$$M_t - P_t = \alpha_0 + \alpha_1 Y_t + \alpha_2 Y_{t-1} + \alpha_3 I_t + \alpha_4 I_{t-1} + \alpha_5 (M_{t-1} - P_{t-1}) + \alpha_6 (M_{t-2} - P_{t-2}) + u_t^p, \hspace{1cm} [11.6.4]$$

where the seven parameters ($\alpha_0$, $\alpha_1$, $\ldots$, $\alpha_6$) are restricted in [11.6.3] to be nonlinear functions of the underlying five parameters ($\rho$, $\beta_0$, $\beta_1$, $\beta_2$, $\beta_3$). The assumption of [11.6.2] can thus be tested by comparing the fit of [11.6.3] with that from unconstrained estimation of [11.6.4].

By definition, $v_t^p$ represents factors influencing money demand for which the researcher has no explicit theory. It therefore seems odd to place great confidence in a detailed specification of its dynamics such as [11.6.2] without testing this assumption against the data. For example, there do not seem to be clear theoretical grounds for ruling out a specification such as

$$v_t^p = \rho_1 v_{t-1}^p + \rho_2 v_{t-2}^p + u_t^p,$$

or, for that matter, a specification in which $v_t^p$ is correlated with lagged values of $Y$ or $I$.

Equation [11.6.1] further assumes that the dynamic multiplier relating money demand to income is proportional to that relating money demand to the interest rate:

$$\frac{\partial (M_{t+1} - P_{t+1})}{\partial Y_t} = \beta_1 \beta_3$$

$$\frac{\partial (M_{t+1} - P_{t+1})}{\partial I_t} = \beta_2 \beta_3.$$

Again, it seems a good idea to test this assumption before imposing it, by comparing the fit of [11.6.1] with that of a more general dynamic model. Finally, inflation may have effects on money demand that are not captured by nominal interest rates. The specification in [11.6.1] incorporates very strong assumptions about the way nominal money demand responds to the price level.

To summarize, a specification such as [11.6.1] and [11.6.2] implicitly imposes many restrictions on dynamics for which there is little or no justification on the basis of economic theory. Before relying on the inferences of [11.6.1] and [11.6.2], it seems a good idea to test that model against a more general specification such as

$$M_t = k_1 + \beta_{11} P_t + \beta_{12} Y_t + \beta_{13} I_t + \beta_{14} M_{t-1} + \beta_{15} P_{t-1} + \beta_{16} Y_{t-1} + \beta_{17} I_{t-1} + \beta_{18} M_{t-2} + \beta_{19} P_{t-2} + \beta_{20} Y_{t-2} + \beta_{21} I_{t-2} + \ldots + \beta_{1p} M_{t-p} + \beta_{1p} P_{t-p} + \beta_{1p} Y_{t-p} + \beta_{1p} I_{t-p} + u_t^p. \hspace{1cm} [11.6.5]$$

Like equation [11.6.1], the specification in [11.6.5] is regarded as a structural money demand equation; $\beta_{1p}$ and $\beta_{1p}$ are interpreted as the effects of current income and the interest rate on desired money holdings, and $u_t^p$ represents factors influencing money demand other than inflation, income, and interest rates. Compared with [11.6.1], the specification in [11.6.5] generalizes the dynamic behavior for the error term $v_t^p$, the partial adjustment process, and the influence of the price level on desired money holdings.

11.6. Vector Autoregressions and Structural Econometric Models
Although [11.6.5] relaxes many of the dubious restrictions on the dynamics implied by [11.6.1], it is still not possible to estimate [11.6.5] by OLS, because of simultaneous equations bias. OLS estimation of [11.6.5] will summarize the correlation between money, the price level, income, and the interest rate. The public’s money demand adjustments are one reason these variables will be correlated, but not the only one. For example, each period, the central bank may be adjusting the interest rate \( r \) to a level consistent with its policy objectives, which may depend on current and lagged values of income, the interest rate, the price level, and the money supply:

\[
I = \kappa_1 + \beta_{12}^{(1)} M + \beta_{13}^{(1)} P_t + \beta_{14}^{(1)} Y_t + \beta_{1i}^{(1)} M_{t-i} + \beta_{12}^{(1)} P_{t-i} + \beta_{13}^{(1)} Y_{t-i} + \beta_{14}^{(1)} I_{t-i} + \ldots
\]  

[11.6.6]

Here, for example, \( \beta_{12}^{(1)} \) captures the effect of the current price level on the interest rate that the central bank tries to achieve. The disturbance \( u_t^1 \) captures changes in policy that cannot be described as a deterministic function of current and lagged money, the price level, income, and the interest rate. If the money demand disturbance \( u_t^p \) is unusually large, this will make \( M_t \) unusually large. If \( \beta_{12}^{(1)} > 0 \), this would cause \( I_t \) to be unusually large as well, in which case \( u_t^1 \) would be positively correlated with the explanatory variable \( I_t \) in equation [11.6.5]. Thus, [11.6.5] cannot be estimated by OLS.

Nor is central bank policy and endogeneity of \( I_t \) the only reason to be concerned about simultaneous equations bias. Money demand disturbances and changes in central bank policy also have effects on aggregate output and the price level, so that \( Y_t \) and \( P_t \) in [11.6.5] are endogenous as well. An aggregate demand equation, for example, might be postulated that relates the level of output to the money supply, price level, and interest rate:

\[
Y_t = \kappa_3 + \beta_{31}^{(2)} M + \beta_{32}^{(2)} P_t + \beta_{33}^{(2)} I_t + \beta_{31}^{(2)} M_{t-2} + \beta_{33}^{(2)} P_{t-2} + \beta_{33}^{(2)} I_{t-2} + \ldots
\]  

[11.6.7]

with \( u_t^1 \) representing other factors influencing aggregate demand. Similarly, an aggregate supply curve might relate the aggregate price level to the other variables being studied. The logical conclusion of such reasoning is that all of the date \( t \) explanatory variables in [11.6.5] should be treated as endogenous.

**Relation Between Dynamic Structural Models and Vector Autoregressions**

The system of equations [11.6.5] through [11.6.7] (along with an analogous aggregate supply equation describing \( P_t \)) can be collected and written in vector form as

\[
B_t y_t = k_t + B_{1y} y_{t-1} + B_{2y} y_{t-2} + \ldots + B_{p} y_{t-p} + u_t,
\]  

[11.6.8]

where

\[
y_t = (M_t, P_t, Y_t, L_t)^
\]

\[
u_t = (u_t^1, u_t^2, u_t^3, u_t^4)^
\]

\[
B_0 = \begin{bmatrix}
-\beta_{12}^{(1)} & -\beta_{13}^{(1)} & -\beta_{14}^{(1)} & 0 \\
-\beta_{31}^{(2)} & -\beta_{32}^{(2)} & 0 & 0 \\
-\beta_{31}^{(2)} & -\beta_{32}^{(2)} & 0 & 0 \\
-\beta_{41}^{(1)} & -\beta_{42}^{(1)} & -\beta_{44}^{(1)} & 0
\end{bmatrix}
\]

\[
k = (k_1, k_2, k_3, k_4)
\]

and \( B_0 \) is a \( (4 \times 4) \) matrix whose row \( i \), column \( j \) element is given by \( \beta_{ij}^{(s)} \) for \( s = 1, 2, \ldots, p \). A large class of structural models for an \((n \times 1)\) vector \( y \), can be written in the form of [11.6.8].

Generalizing the argument in [11.6.3], it is assumed that a sufficient number of lags of \( p \) are included and the matrices \( B_0 \) are defined so that \( u \) is vector white noise. If instead, say, \( u_t \) followed an \( r \)-th order VAR, with

\[
u_t = F_1 u_{t-1} + F_2 u_{t-2} + \ldots + F_r u_{t-r} + \epsilon_t,
\]

then we could premultiply [11.6.8] by \((I_r - F_1 L^1 - F_2 L^2 - \ldots - F_r L^r)\) to arrive at a system of the same basic form as [11.6.8] with \( p \) replaced by \((p + r)\) and with \( u_t \) replaced by the white noise disturbance \( \epsilon_t \).

If each side of [11.6.8] is premultiplied by \( B_0^{-1} \), the result is

\[
y_t = c + \Phi_{1} y_{t-1} + \Phi_{2} y_{t-2} + \ldots + \Phi_{p} y_{t-p} + \epsilon_t,
\]  

[11.6.9]

where

\[
c = B_0^{-1} k
\]  

[11.6.10]

\[
\Phi_{s} = B_0^{-1} F_s \quad \text{for } s = 1, 2, \ldots, p
\]  

[11.6.11]

\[
\epsilon_t = B_0^{-1} \epsilon_t
\]  

[11.6.12]

Assuming that [11.6.8] is parameterized sufficiently richly that \( u \) is vector white noise, then \( \epsilon_t \) will also be vector white noise and [11.6.9] will be recognized as the vector autoregressive representation for the dynamic structural system [11.6.8]. Thus, a VAR can be viewed as the reduced form of a general dynamic structural model.

**Interpreting Impulse-Response Functions**

In Section 11.4 we calculated the impulse-response function

\[
\frac{\delta y_{i,j}}{\delta \epsilon_j}
\]  

[11.6.13]

This magnitude describes the effect of an innovation in the \( j \)th variable on future values of each of the variables in the system. According to [11.6.12], the VAR innovation \( \epsilon_t \) is a linear combination of the structural disturbances \( u_t \). For example,
it might turn out that
\[ \varepsilon_t = 0.3u_t^p - 0.6u_t^s + 0.1u_t^A - 0.5u_t^C. \]

In this case, if the cash held by the public is larger than would have been forecast using the VAR (\(\varepsilon_t\) is positive), this might be because the public's demand for cash is higher than is normally associated with the current level of income and interest rate (that is, \(u_t^p\) is positive). Alternatively, \(\varepsilon_t\) might be positive because the central bank has chosen to ease credit (\(u_t^C\) is negative), or a variety of other factors. In general, \(\varepsilon_t\) represents a combination of all the different influences that matter for any variables in the economy. Viewed this way, it is not clear why the magnitude [11.6.13] is of particular interest.

By contrast, if we were able to calculate
\[ \frac{\partial y_{t+1}}{\partial u_t^C}, \quad [11.6.14] \]
this would be of considerable interest. Expression [11.6.14] identifies the dynamic consequences for the economy if the central bank were to tighten credit more than usual and is a key magnitude for describing the effects of monetary policy on the economy.

Section 11.4 also discussed calculation of an orthogonalized impulse-response function. For \(\Omega = E(\varepsilon_t\varepsilon_t')\), we found a lower triangular matrix \(\Lambda \) and a diagonal matrix \(D\) such that \(\Omega = ADA'\). We then constructed the vector \(A^{-1}\varepsilon\), and calculated the consequences of changes in each element of this vector for future values of \(y\).

Recall from [11.6.12] that the structural disturbances \(u\) are related to the VAR innovations \(\varepsilon\), by
\[ u_t = B_0\varepsilon_t. \quad [11.6.15] \]

Suppose that it happened to be the case that the matrix of structural parameters \(B_0\) was exactly equal to the matrix \(A^{-1}\). Then the orthogonalized innovations would coincide with the true structural disturbances:
\[ u_t = B_0\varepsilon_t = A^{-1}\varepsilon_t. \quad [11.6.16] \]

In this case, the method described in Section 11.4 could be used to find the answers to important questions such as [11.6.14].

Is there any reason to hope that \(B_0\) and \(A^{-1}\) would be the same matrix? Since \(A\) is lower triangular, this clearly requires \(B_0\) to be lower triangular. In the example [11.6.8], this would require that the current values of \(P\), \(Y\), and \(I\) do not influence money demand, that the current value of \(M\) but not that of \(Y\) or \(I\) enters into the aggregate supply curve, and so on. Such assumptions are rather unusual, though there may be another way to order the variables such that a recursive structure is more palatable. For example, a Keynesian might argue that prices respond to other economic variables only with a lag, so that the coefficients on current variables in the aggregate supply equation are all zero. Perhaps money and interest rates influence aggregate demand only with a lag, so that their current values are excluded from the aggregate demand equation. One might try to argue further that the interest rate affects desired money holdings only with a lag as well. Because most central banks monitor current economic conditions quite closely, perhaps all the current values should be included in the equation for \(I\). These assumptions suggest ordering the variables as \(y = (P, Y, M, I)^T\), for which the structural model would be
\[
\begin{pmatrix}
P_t \\ Y_t \\ M_t \\ I_t \\
\end{pmatrix}
= 
\begin{bmatrix}
B_{t1}^{(1)} & B_{t2}^{(1)} & B_{t3}^{(1)} & B_{t4}^{(1)} \\
B_{t1}^{(2)} & B_{t2}^{(2)} & B_{t3}^{(2)} & B_{t4}^{(2)} \\
B_{t1}^{(3)} & B_{t2}^{(3)} & B_{t3}^{(3)} & B_{t4}^{(3)} \\
\end{bmatrix}
\begin{bmatrix}
P_{t-1} \\ Y_{t-1} \\ M_{t-1} \\ I_{t-1} \\
\end{bmatrix}
+ 
\begin{bmatrix}
\beta_{t1}^{(1)} & \beta_{t2}^{(1)} & \beta_{t3}^{(1)} & \beta_{t4}^{(1)} \\
\beta_{t1}^{(2)} & \beta_{t2}^{(2)} & \beta_{t3}^{(2)} & \beta_{t4}^{(2)} \\
\beta_{t1}^{(3)} & \beta_{t2}^{(3)} & \beta_{t3}^{(3)} & \beta_{t4}^{(3)} \\
\end{bmatrix}
\begin{bmatrix}
u_t^{(1)} \\ \cdots \\ \cdots \\ \cdots \\
\end{bmatrix}
+ 
\begin{bmatrix}
u_t^{(2)} \\ \cdots \\ \cdots \\ \cdots \\
\end{bmatrix}
\quad [11.6.17]
\]
Suppose there exists such an ordering of the variables for which \(B_0\) is lower triangular. Write the dynamic structural model [11.6.8] as
\[ B_0y_t = -\Gamma x_t + u_t, \quad [11.6.18] \]
where
\[ -\Gamma = \begin{bmatrix}
\mathbf{B}_1 & \mathbf{B}_2 & \cdots & \mathbf{B}_s
\end{bmatrix}, \quad [11.6.19] \]

This corresponds to ordering the variables as \(x = (\varepsilon_t, \cdots, \varepsilon_{t+p})\). The first column of \(\mathbf{B}_1\) gives the long-run coefficients on \(\varepsilon\).

Suppose, furthermore, that the disturbances in the structural equations are serially uncorrelated and uncorrelated with each other:
\[ E(u_t'u_t) = \begin{bmatrix} D & 0 \\ 0 & I \end{bmatrix}, \quad [11.6.20] \]

where \(D\) is a diagonal matrix. The VAR is the reduced form of the dynamic structural model [11.6.18] and can be written as
\[ y_t = \Pi x_{t-1} + \varepsilon_t, \quad [11.6.21] \]

where
\[ \Pi = -B_{t-1}'\mathbf{T}, \quad [11.6.22] \]

\[ \varepsilon_t = B_{t-1}'u_t, \quad [11.6.23] \]

Letting \(\Omega\) denote the variance-covariance matrix of \(\varepsilon_t\), [11.6.22] implies
\[ \Omega = E(\varepsilon_t\varepsilon_t') = B_{t-1}'E(u_t'u_t)(B_{t-1})' = B_{t-1}'D(B_{t-1})'. \quad [11.6.23] \]
Note that if the only restrictions on the dynamic structural model are that $B_n^\top$ is lower triangular with unit coefficients along the principal diagonal and that $D$ is diagonal, then the structural model is just identified. To see this, note that these restrictions imply that $B_n^{-1}$ must also be lower triangular with unit coefficients along the principal diagonal. Recall from Section 4.4 that given any positive definite symmetric matrix $\Omega$, there exist a unique lower triangular matrix $A$ with 1's along the principal diagonal and a diagonal matrix $D$ with positive entries along the principal diagonal such that $\Omega = ADA^\top$. Thus, unique values $B_n^{-1}$ and $D$ of the required form can always be found that satisfy (11.6.23). Moreover, any $B_n$ matrix of this form is nonsingular, so that $\Gamma$ in (11.6.21) can be calculated uniquely from $B_n$ and $\Omega$ as $\Gamma = -B_n^{-1}D$. Thus, given any allowable values for the reduced-form parameters $(\Pi$ and $\Omega$), there exist unique values for the structural parameters $(B_n, \Gamma, D)$ of the specified form, establishing that the structural model is just identified.

Since the model is just identified, full-information maximum likelihood (FIML) estimates of $(B_n, \Gamma, D)$ can be obtained by first maximizing the likelihood function with respect to the reduced-form parameters $(\Pi$ and $\Omega$) and then using the unique mapping from reduced-form parameters to find the structural parameters. The maximum likelihood estimates of $\Pi$ are found from OLS regressions of the elements of $\gamma$, on $x$, and the MLE of $\Omega$ is obtained from the variance-covariance matrix of the residuals from these regressions. The estimates $B_n^{-1}$ and $D$ are then found from the triangular factorization of $\Omega$. This, however, is precisely the procedure described in calculating the orthogonalized innovations in Section 11.4. The estimate $\hat{\Pi}$ described there is thus the same as the FIML estimate of $B_n^{-1}$. The vector of orthogonalized residuals $u = \Pi^{-1}e$, would correspond to the vector of structural disturbances, and the orthogonalized impulse-response coefficients would give the dynamic consequences of the structural events represented by $u$, provided that the structural model is lower triangular as in (11.6.17).

Nonrecursive Structural VARs

Even if the structural model cannot be written in lower triangular form, it may be possible to give a structural interpretation to a VAR using a similar idea to that in equation (11.6.23). Specifically, a structural model specifies a set of restrictions on $B_n$ and $D$, and we can try to find values satisfying these restrictions such that $B_n^{-1}D(B_n^{-1})^\top = \Omega$. This point was developed by Bernanke (1986), Blanchard and Watson (1986), and Sims (1986).

For illustration, consider again the model of supply and demand discussed in equations (9.3.2) and (9.3.3). In that specification, quantity ($q_t$) and price ($p_t$) were endogenous variables and weather ($w_t$) was exogenous, and it was assumed that both disturbances were i.i.d. The structural VAR approach to this model would allow quite general dynamics by adding $p$ lags of all three variables to equations (9.3.2) and (9.3.3), as well as adding a third equation to describe the dynamic behavior of weather. Weather presumably does not depend on the behavior of the market, so the third equation would for this example just be a univariate auto-regression. The model would then be

$$q_t = B_p + \beta_0^{(q)} q_{t-1} + \beta_1^{(q)} p_{t-1} + \beta_2^{(q)} w_{t-1} + \beta_3^{(q)} q_{t-2} + \beta_4^{(q)} p_{t-2} + \beta_5^{(q)} w_{t-2} + \cdots$$

$$+ \beta_6^{(q)} q_{t-3} + \beta_7^{(q)} p_{t-3} + \beta_8^{(q)} w_{t-3} + \cdots + \beta_9^{(q)} q_{t-p} + \beta_{10}^{(q)} p_{t-p} + \beta_{11}^{(q)} w_{t-p} + u_t$$

where $\omega_0^{(q)} q_{t-1} + \omega_1^{(q)} p_{t-1} + \cdots + \omega_p^{(q)} q_{t-p} + \omega_p^{(q)} p_{t-p} + \omega_p^{(q)} w_{t-p} + u_t$.

We could then take $(u_t', u_{t-1}', u_{t-2}')$ to be a white noise vector with diagonal variance-covariance matrix given by $D$. This is an example of a structural model (11.6.18) in which $\omega_i = 0$ for $i = 1, \ldots, p$.

There is no way to order the variables so as to make the matrix $B_n$ lower triangular. However, equation (11.6.22) indicates that the structural disturbances $u$ are related to the VAR residuals $e$, by $e_t = B_n^{-1}u_t$. Thus, if $B_n$ is estimated by maximum likelihood, then the impulse-response functions could be calculated as in Section 11.4 with $A$ replaced by $B_n^{-1}$, and the results would give the effects of each of the structural disturbances on subsequent values of variables of the system. Specifically,

$$\frac{\partial e_t}{\partial u_t} = B_n^{-1},$$

so that the effect on $e_t$ of the $j$th structural disturbance $u_t$ is given by $b_t^j$, the $j$th column of $B_n^{-1}$. Thus, we would calculate

$$\frac{\partial y_{t+j}}{\partial u_t} = \frac{\partial y_{t+j}}{\partial e_t} \frac{\partial e_t}{\partial u_t} = \Psi b_t^j,$$

for $\Psi$, the $(n \times n)$ matrix of coefficients for the $j$th lag of the $MA(\infty)$ representation (11.4.1).

FIML Estimation of a Structural VAR with Unrestricted Dynamics

FIML estimation is particularly simple if there are no restrictions on the coefficients $\Gamma$ on lagged variables in (11.6.18). For example, this would require including lagged values of $p_{t-1}$ and $q_{t-p}$ in the weather equation (11.6.26). Using (11.6.23), the log likelihood function for the system (11.6.18) can be written as

$$L(B, D, \Pi) = -(Tn/2) \log(2\pi) - (T/2) \log|B_0^{-1}D(B_0^{-1})^\top| - (1/2) \sum_{t=1}^{T} \left( y_t - \Pi x_t \right)' \left( B_0^{-1}D(B_0^{-1})^\top \right)^{-1} \left( y_t - \Pi x_t \right).$$

If there are no restrictions on lagged dynamics, this is maximized with respect to $\Pi$ by OLS regression of $y_t$ on $x_t$. Substituting this estimate into (11.6.28) as in
[11.1.25] produces
\[ \mathcal{L}(\mathbf{B}_n, \mathbf{D}, \mathbf{H}) = -(7n/2) \log(2\pi) - (7/2) \log(\mathbf{B}_n^{-1} \mathbf{D}(\mathbf{B}_n^{-1}))']' \]
\[ - (1/2) \sum_{t=1}^T \hat{\epsilon}_t [\mathbf{B}_n^{-1} \mathbf{D}(\mathbf{B}_n^{-1})']' \hat{\epsilon}_t. \]  \[ 11.6.29 \]

But
\[ \sum_{t=1}^T \hat{\epsilon}_t [\mathbf{B}_n^{-1} \mathbf{D}(\mathbf{B}_n^{-1})']' \hat{\epsilon}_t = \sum_{t=1}^T \text{trace}[\hat{\epsilon}_t [\mathbf{B}_n^{-1} \mathbf{D}(\mathbf{B}_n^{-1})']' \hat{\epsilon}_t] \]
\[ = \sum_{t=1}^T \text{trace}[\mathbf{B}_n^{-1} \mathbf{D}(\mathbf{B}_n^{-1})']' \hat{\epsilon}_t \hat{\epsilon}_t' \]
\[ = \text{trace}[\mathbf{B}_n^{-1} \mathbf{D}(\mathbf{B}_n^{-1})']' \mathbf{H} \]
\[ = T \times \text{trace}[\mathbf{B}_n^{-1} \mathbf{D}(\mathbf{B}_n^{-1})']' \mathbf{H}. \]  \[ 11.6.30 \]

Furthermore,
\[ \log \mathcal{L}(\mathbf{B}_n, \mathbf{D}, \mathbf{H}) = \log(\mathcal{L}(\mathbf{B}_n^{-1}, \mathbf{D}, \mathbf{B}_n^{-1})) = -\log(\mathbf{B}_n) + \log(\mathbf{D}). \]  \[ 11.6.31 \]

Substituting [11.6.31] and [11.6.30] into [11.6.29], FIML estimates of the structural parameters are found by choosing \( \mathbf{B}_n \) and \( \mathbf{D} \) so as to maximize
\[ \mathcal{L}(\mathbf{B}_n, \mathbf{D}, \mathbf{H}) = -(7n/2) \log(2\pi) + (7/2) \log(\mathbf{B}_n) - (7/2) \log(\mathbf{D}) \]
\[ - (7/2) \text{trace}(\mathbf{B}_n^{-1} \mathbf{D}(\mathbf{B}_n^{-1})') \mathbf{H}. \]  \[ 11.6.32 \]

Using calculations similar to those used to analyze [11.1.25], one can show that if there exist unique matrices \( \mathbf{B}_n \) and \( \mathbf{D} \) of the required form satisfying \( \mathbf{B}_n^{-1} \mathbf{D}(\mathbf{B}_n^{-1})' = \mathbf{H} \), then maximization of [11.6.32] will produce estimates \( \mathbf{B}_n \) and \( \mathbf{D} \) satisfying
\[ \mathbf{B}_n^{-1} \mathbf{D}(\mathbf{B}_n^{-1})' = \hat{\mathbf{H}}. \]  \[ 11.6.33 \]

This is a nonlinear system of equations, and numerical maximization of [11.6.32] offers a convenient general approach to finding a solution to this system of equations.

**Identification of Structural VARs**

The existence of a unique maximum of [11.6.32] requires both an order condition and a rank condition for identification. The order condition is that \( \mathbf{B}_n \) and \( \mathbf{D} \) have no more unknown parameters than \( \mathbf{H} \). Since \( \mathbf{H} \) is symmetric, it can be summarized by \( n(n+1)/2 \) distinct values. If \( \mathbf{D} \) is diagonal, it requires \( n \) parameters, meaning that \( \mathbf{B}_n \) can have no more than \( n(n-1)/2 \) free parameters. For the supply-and-demand example of [11.6.24] through [11.6.26], \( n = 3 \), and the matrix \( \mathbf{B}_n \) in [11.6.27] has \( 3(3-1)/2 = 3 \) free parameters (\( \beta \), \( \gamma \), and \( \delta \)), plus one that is fixed for identification.

Even if the order condition is satisfied, the model may still not be identified. For example, suppose that
\[ \mathbf{B}_n = \begin{bmatrix} 1 & -\beta & 0 \\ 1 & -\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix}. \]

Even though this specification satisfies the order condition, it fails the rank condition, since the value of the likelihood function will be unchanged if \( \beta \) and \( \gamma \) are switched along with \( \sigma_\beta^2 \) and \( \sigma_\gamma^2 \).

To characterize the rank condition, suppose that there are \( n_\theta \) elements of \( \mathbf{B}_0 \) that must be estimated; collect these in an \( (n_\theta \times 1) \) vector \( \theta_\theta \). The identifying assumptions can be represented as a known \( (n^2 \times n_\theta) \) matrix \( \mathbf{S}_\theta \) and a known \( (n^2 \times 1) \) vector \( \mathbf{s}_\theta \) for which
\[ \text{vec}(\mathbf{B}_0) = \mathbf{S}_\theta \mathbf{\theta}_\theta + \mathbf{s}_\theta. \]

For example, for the dynamic model of supply and demand represented by [11.6.27],
\[ \text{vec}(\mathbf{B}_n) = \begin{bmatrix} 1 \\ 1 \\ 0 \\ -\beta \\ -\gamma \\ 0 \\ 0 \\ -h \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \]
\[ \mathbf{\theta}_\theta = \begin{bmatrix} \beta \\ \gamma \\ h \end{bmatrix}. \]

\[ \text{vec}(\mathbf{D}) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \]
\[ \mathbf{S}_\theta = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \]

Similarly, collect the unknown elements of \( \mathbf{D} \) in an \( (n_\theta \times 1) \) vector \( \mathbf{\theta}_D \), with
\[ \text{vec}(\mathbf{D}) = \mathbf{S}_D \mathbf{\theta}_D + \mathbf{s}_D. \]

for \( \mathbf{S}_D \) an \( (n^2 \times n_\theta) \) matrix and \( \mathbf{s}_D \) an \( (n^2 \times 1) \) vector. For the supply-and-demand example,
\[ \text{vec}(\mathbf{D}) = \begin{bmatrix} \sigma_\beta^2 \\ \sigma_\gamma^2 \\ \sigma_\delta^2 \\ \sigma_\beta^2 \\ \sigma_\gamma^2 \\ \sigma_\delta^2 \\ \sigma_\beta^2 \\ \sigma_\gamma^2 \\ \sigma_\delta^2 \\ \sigma_\beta^2 \\ \sigma_\gamma^2 \\ \sigma_\delta^2 \end{bmatrix}, \]
\[ \mathbf{S}_D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \]
Since [11.6.33] is an equation relating two symmetric matrices, there are \( n^* = n(n + 1)/2 \) separate conditions, represented by
\[
\text{vech}(\Omega) = \text{vech} \left( [B_{\alpha}(\theta_\alpha)]^{-1} D(\theta_\alpha) [A(\theta_\alpha)]^{-1} \right).
\]  
[11.6.36]
Denote the right side of [11.6.36] by \( f(\theta_\alpha, \theta_\beta) \), where \( f(\mathbb{R}^n \times \mathbb{R}^{m*}) \rightarrow \mathbb{R}^{n*} \):
\[
\text{vech}(\Omega) = f(\theta_\alpha, \theta_\beta).
\]  
[11.6.37]
Appendix 11.B shows that the \([n^* \times (n^*_p + n^*_q)]\) matrix of derivatives of this function is given by
\[
J = \begin{bmatrix}
\frac{\partial \text{vech}(\Omega)}{\partial \theta_\alpha} & \frac{\partial \text{vech}(\Omega)}{\partial \theta_\beta}
\end{bmatrix}
\]  
[11.6.38]
where \( D^*_\alpha \) is the \((n^*_p \times n^*_p)\) matrix defined in [11.1.45].
Suppose that the columns of the matrix in [11.6.38] were linearly dependent; that is, suppose there exists a nonzero \([n^*_p + n^*_q] \times 1\) vector \( \lambda \) such that \( J \lambda = 0 \). This would mean that if a small multiple of \( \lambda \) were added to \((\theta_\alpha', \theta_\beta')\), the model would imply the same probability distribution for the data. We would have no basis for distinguishing between these alternative values for \((\theta_\alpha', \theta_\beta')\), meaning that the model would be unidentified.

Thus, the rank condition for identification of a structural VAR requires that the \((n^*_p + n^*_q)\) columns of the matrix \( J \) in [11.6.38] be linearly independent. The order condition is that the number of rows of \( J \) \((n^* = n(n + 1)/2)\) be at least as great as the number of columns.

To check this condition in practice, the simplest approach is usually to make a guess as to the values of the structural parameters and check \( J \) numerically.
Giannini (1992) derived an alternative expression for the rank condition and provided computer software for checking it numerically.

**Structural VAR with Restrictions on \( \Pi \)**
The supply-and-demand example of [11.6.24] to [11.6.26] did not satisfy the assumptions behind the derivation of [11.6.32], [11.6.26] imposed the restriction that lagged values of \( p \) and \( q \) did not belong in the weather equation. Where such restrictions are imposed, it is no longer the case that the \( 
\text{FIML} \) estimates of \( \Pi \) are obtained by OLS, and system parameters would have to be estimated as described in Section 11.3. As an alternative, OLS estimation of [11.6.24] through [11.6.26] would still give consistent estimates of \( \Pi \), and the variance-covariance matrix of the residuals from these regressions would provide a consistent estimate \( \hat{\Pi} \). One could still use this estimate in [11.6.32], and the resulting maximization problem would give reasonable estimates of \( B_{\Pi} \) and \( D\).

**Structural VARs and Forward-Looking Behavior**
The supply-and-demand example assumed that lagged values of price and quantity did not appear in the equation for weather. The spirit of VARs is that such assumptions ought to be tested before being imposed. What should we conclude if, contrary to our prior expectations, the price of oranges turned out to Granger-cause the weather in Florida? It certainly cannot be that the price is a cause of the weather. Instead, such a finding would suggest forward-looking behavior on the part of buyers or sellers of oranges; for example, it may be that if buyers anticipate bad weather in the future, they bid up the price of oranges today. If this should prove to be the case, the identifying assumption in [11.6.24] that demand depends on the weather only through its effect on the current price needs to be reexamined. Proper modeling of forward-looking behavior can provide an alternative way to identify VARs, as explored by Flavin (1981), Hansen and Sargent (1981), and Keating (1990), among others.

**Other Approaches to Identifying Structural VARs**
Identification was discussed in previous subsections primarily in terms of exclusion restrictions on the matrix of structural coefficients \( B_{\Pi} \). Blanchard and Diamond (1989, 1990) used a priori assumptions about the signs of structural parameters to identify a range of values of \( B_{\Pi} \), consistent with the data. Shapiro and Watson (1988) and Blanchard and Quah (1989) used assumptions about long-run multipliers to achieve identification.

**A Critique of Structural VARs**
Structural VARs have appeal for two different kinds of inquiry. The first potential user is someone who is primarily interested in estimating a structural equation such as the money demand function in [11.6.1]. If a model imposes restrictions on the dynamics of the relationship, it seems good practice to test these restrictions against a more general specification such as [11.6.5] before relying on the restricted model for inference. Furthermore, in order to estimate the dynamic consequences of, say, income on money demand, we have to take into account the fact that, historically, when income goes up, this has typically been associated with future changes in income and interest rates. What time path for these explanatory variables should be assumed in order to assess the consequences for money demand at time \( t + s \) of a change in income at time \( t \)? A VAR offers a framework for posing this question—we use the time path that would historically be predicted for those variables following an unanticipated change in income.

A second potential user is someone who is interested in summarizing the dynamics of a vector \( y \), while imposing as few restrictions as possible. Insofar as this summary includes calculation of impulse-response functions, we need some motivation for what the statistics mean. Suppose we find that there is a temporary rise in income following an innovation in money. One is tempted to interpret this finding as suggesting that expansionary monetary policy has a positive but temporary effect on output. However, such an interpretation implicitly assumes that the orthogonalized "money innovation" is the same as the disturbance term in a description of central bank policy. Insofar as impulse-response functions are used to make statements that are structural in nature, it seems reasonable to try to use an orthogonalization that represents our understanding of these relationships as well as possible. This point has been forcefully argued by Cooley and LeRoy (1985), Leamer (1985), Bernanke (1986), and Blanchard (1988), among others. Even so, it must be recognized that convincing identifying assumptions are hard to come by. For example, the ordering in [11.6.17] is clearly somewhat arbitrary, and the exclusion restrictions are difficult to defend. Indeed, if there were compelling identifying assumptions for such a system, the fierce debates among
11.7. Standard Errors for Impulse-Response Functions

**Standard Errors for Nonorthogonalized Impulse-Response Function Based on Analytical Derivatives**

Section 11.4 discussed how the matrix of impulse-response coefficients at lag $s$, would be constructed from knowledge of the autoregressive coefficients. In practice, the autoregressive coefficients are not known with certainty but must be estimated by $OLS$ regressions. When the estimated values of the autoregressive coefficients are used to calculate $\Psi_s$, it is useful to report the implied standard errors for the estimates $\hat{\Psi}_s$.\(^{13}\)

Adopting the notation from Proposition 11.1, let $k = np + 1$ denote the number of coefficients in each equation of the VAR and let $\pi = \text{vec}(\Pi)$ denote the $(nk \times 1)$ vector of parameters for all the equations; the first $k$ elements of $\pi$ give the constant term and autoregressive coefficients for the first equation, the next $k$ elements of $\pi$ give the parameters for the second equation, and so on. Let $\psi_i = \text{vec}(\Psi_i)$ denote the $(n^2 \times 1)$ vector of moving average coefficients associated with lag $s$. The first $n$ elements of $\psi_i$ are given by the first row of $\Psi_i$, and identify the response of $y_{1,t-s}$ to $e_t$. The next $n$ elements of $\psi_i$ are given by the second row of $\Psi_i$ and identify the response of $y_{2,t-s}$ to $e_t$, and so on. Given the values of the autoregressive coefficients in $\pi$, the VAR can be simulated to calculate $\hat{\Psi}_s$. Thus, $\hat{\psi}_i$ could be regarded as a nonlinear function of $\pi$, represented by the function $\hat{\psi}_i(\pi)$.\(^{13}\)

The impulse-response coefficients are estimated by replacing $\pi$ with the $OLS$ estimates $\hat{\pi}_T$, generating the estimate $\hat{\Psi}_{s,T} = \hat{\psi}_i(\hat{\pi}_T)$. Recall that under the conditions of Proposition 11.1, $\sqrt{T}(\pi - \pi) \rightarrow X$, where

$$X \sim N\left(0, (\Omega \otimes Q^{-1})\right).$$

[11.7.1]

Standard errors for the impulse-response coefficients can then be calculated by applying Proposition 7.4:

$$\sqrt{T}(\hat{\psi}_{s,T} - \hat{\psi}_i) \rightarrow G X,$$

where

$$G_i = \frac{\partial \hat{\psi}_i(\pi)}{\partial \pi}.$$  \[[11.7.2]\]

That is,

$$\sqrt{T}(\hat{\psi}_{s,T} - \hat{\psi}_i) \sim N\left(0, G_i (\Omega \otimes Q^{-1}) G_i'\right).$$

[11.7.3]

Standard errors for an estimated impulse-response coefficient are given by the square root of the associated diagonal element of $(1/T)\hat{G}_{s,T}(\hat{\Omega}_T \otimes \hat{Q}_T^{-1})\hat{G}_{s,T}$, where

$$\hat{G}_{s,T} = \frac{\partial \hat{\psi}_i(\pi)}{\partial \pi} \bigg|_{\pi = \hat{\pi}_T},$$

$$\hat{Q}_T = (1/T) \sum_{t=1}^{T} x_t x_t',$$

with $x_t$ and $\hat{\Omega}_T$ as defined in Proposition 11.1.

To apply this result, we need an expression for the matrix $G_i$ in [11.7.2]. Appendix 11.B to this chapter establishes that the sequence $(G_{s,T})_{s=1}^{\infty}$ can be calculated by iterating on

$$G_0 = [I \otimes (0_{n \times n}) \Psi_0 \cdots \Psi_{s-1} \Psi_s \cdots \Psi_{s-p}] + \left(\Phi_1 \otimes (I_l)\right) G_{s-1} + \cdots + \left(\Phi_{s-p} \otimes (I_l)\right) G_{s-p},$$

[11.7.4]

Here $0_{n \times n}$ denotes an $(n \times n)$ vector of zeros. The iteration is initialized by setting $G_0 = G_{s-1} = \cdots = G_{s-p+1} = 0_{n \times n}$. It is also understood that $\Psi_0 = I_l$ and $\Psi_s = 0_n$ for $s < 0$. Thus, for example,

$$G_1 = [I \otimes (0_{n \times n}) I_l 0_n \cdots 0_n],$$

$$G_2 = [I \otimes (0_{n \times n}) \Psi_1 I_l \cdots 0_n] + (\Phi_1 \otimes (I_l)) G_1.$$

A closed-form solution for [11.7.4] is given by

$$G_s = \sum_{i=1}^{s} [\Psi_{s-i} \otimes (0_{n \times n}) \Psi_{s-i} \cdots \Psi_{s-i-p+1}] + (\Phi_1 \otimes (I_l)) G_{s-1}.$$ \[[11.7.5]\]

**Alternative Approaches to Calculating Standard Errors for Nonorthogonalized Impulse-Response Function**

The matrix of derivatives $G_i$ can alternatively be calculated numerically as follows. First we use the $OLS$ estimates $\hat{\pi}$ to calculate $\hat{\psi}_i(\hat{\pi})$ for $s = 1, 2, \ldots, m$. We then increase the value of the $i$th element of $\pi$ by some small amount $\Delta$, holding all other elements constant, and evaluate $\hat{\psi}_i(\hat{\pi} + e_i \Delta)$ for $s = 1, 2, \ldots, m$, where $e_i$ denotes the $i$th column of $I_n$. Then the $(n^2 \times 1)$ vector

$$\hat{\psi}_i(\hat{\pi} + e_i \Delta) - \hat{\psi}_i(\hat{\pi})$$

is given an estimate of the $i$th column of $G_i$. By conducting separate evaluations of the sequence $\hat{\psi}_i(\hat{\pi} + e_i \Delta)$ for each $i = 1, 2, \ldots, nk$, all of the columns of $G_i$ can be filled in.

Monte Carlo methods can also be used to infer the distribution of $\hat{\psi}_i(\hat{\pi})$. Here we would randomly generate an $(nk \times 1)$ vector drawn from a $N(\hat{\pi}, (1/T)\hat{\Omega}_T \otimes \hat{Q}_T^{-1})$ distribution. Denote this vector by $\pi^{(1)}$, and calculate $\hat{\psi}_i(\pi^{(1)})$. Draw a second vector $\pi^{(2)}$ from the same distribution and calculate $\hat{\psi}_i(\pi^{(2)})$. Repeat this for, say, $10,000$ separate simulations. If $9500$ of these simulations result in a value of the first element of $\psi_i$ that is between $\hat{\psi}_i$ and $\hat{\psi}_i$, then $(\hat{\psi}_i, \hat{\psi}_i)$ can be used as a $95\%$ confidence interval for the first element of $\hat{\psi}_i$.

Runkle (1987) employed a related approach based on bootstrapping. The idea behind bootstrapping is to obtain an estimate of the small-sample distribution of $\hat{\pi}$ without assuming that the innovations $e_t$ are Gaussian. To implement this procedure, first estimate the VAR and save the coefficient estimates $\hat{\pi}$ and the fitted residuals $(\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_T)$. Then consider an artificial random variable $u$ that has probability $(1/T)$ of taking on each of the particular values $(\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_T)$. The
The Dynamic Effects of Aggregate Demand and Supply Disturbances

By OLIVIER JEAN BLANCHARD and DANNY QUAH

We interpret fluctuations in GNP and unemployment as due to two types of disturbances: disturbances that have a permanent effect on output and disturbances that do not. We interpret the first as supply disturbances, the second as demand disturbances. Demand disturbances have a hump-shaped mirror-image effect on output and unemployment. The effect of supply disturbances on output increases steadily over time, peaking after two years and reaching a plateau after five years.

It is now widely accepted that GNP is reasonably characterized as a unit root process: a positive innovation in GNP should lead one to revise upward one's forecast on GNP for all horizons. Following the influential work of Charles Nelson and Charles Plosser (1982), this statistical characterization has been recorded and refined by numerous authors including John Campbell and N. Gregory Mankiw (1987a), Peter Clark (1987, 1988), John Cochrane (1988), Francis Diebold and Glenn Rudebusch (1988), George Evans (1987), and Mark Watson (1986).

How should this finding affect one's views about macroeconomic fluctuations? Were there only one type of disturbance in the economy, then the implications of these findings would be straightforward. That disturbance would affect the economy in a way characterized by estimated univariate-moving average representations, such as those offered by Campbell and Mankiw. The problem would simply be to find out what this disturbance was, and why its effects had the shape that they did. The way to proceed would be clear.

However, if GNP is affected by more than one type of disturbance, as is likely, the interpretation becomes more difficult. In that case, the univariate-moving average representation of output is some combination of the dynamic response of output to each of the disturbances. The work in Stephen Beveridge and Nelson (1981), Andrew Harvey (1985), and Watson (1986) can be viewed as early attempts to get at this issue.

To proceed, given the possibility that output may be affected by more than one type of disturbance, one can impose a priori restrictions on the response of output to each of the disturbances, or one can exploit information from macroeconomic variables other than GNP. In addition to the work noted above, Clark (1987) has also used the first approach. This paper adopts the second, and considers the joint behavior of output and unemployment. Campbell and Mankiw (1987b), Clark (1988), and Evans (1987) have also taken this approach. One of the main results here is that changes in the size of disturbances may initially increase unemployment. This is followed by a decline in unemployment, with a slow return over time to its original value.

While this dynamic characterization is fairly sharp, the data are not as specific as to the relative contributions of demand and supply disturbances to output fluctuations. On the one hand, we find that the time-series of demand and supply disturbances to output fluctuations is a time-series of output constructed by putting all supply disturbance realizations equal to zero, has peaks and troughs which coincide with most of the NBER troughs and peaks. But, when we turn to variance decompositions of output at various horizons, we find that the respective contributions of supply and demand disturbances are not precisely estimated. For instance, at a forecast horizon of four quarters, we find that, under alternative assumptions, the contribution of demand disturbances ranges from 40 percent to over 95 percent.

The rest of the paper is organized as follows. Section I analyzes identification, and Section II discusses our economic interpretation of the disturbances. Section III discusses estimation, and Section IV characterizes the dynamic effects of demand and supply disturbances to output and unemployment. Section V characterizes the relative contributions of demand and supply disturbances to fluctuations in output and unemployment.

1. Identification

In this section, we show how our assumptions characterize the process followed by output and unemployment, and how this process can be recovered from the data.

We make the following assumptions. There are two types of disturbances in the economy, one affecting output and the other unemployment. The first has a long-run effect on either unemployment or output, but not both. The second has a long-run effect on unemployment, but may have a long-run effect on output. Finally, these two disturbances are correlated at all leads and lags. These restrictions in effect define the two disturbances. As indicated in the introduction, and discussed at length in the next section, we will refer to the first as demand disturbances, and to the second as supply disturbances. How we name the disturbances however is irrelevant for the argument of this section.

The demand and supply components described above are permitted to be serially correlated. Under regularity conditions, each of these components can always be uniquely represented as an invertible distributed lag of serially uncorrelated disturbances. Thus, we can refer to the associated serially uncorrelated disturbances as the demand and supply disturbances themselves: this is without ambiguity or loss of generality. We will then

---

*Both authors are with the Economics Department, MIT, Cambridge MA 02139, and the NBER. We thank Stanley Fischer, Juto Rotemberg, Mark Watson for helpful discussions, and the NSF for financial assistance. We are also grateful for the comments of two anonymous referees and of participants at an NBER Economic Fluctuations meeting, and for the hospitality of the MTT Statistics Center.

1As will become clear, our work differs from those in that we wish to examine the dynamic effects of disturbances that have permanent effects; such issues cannot be addressed by studies that restrict the permanent component to be a random walk. In either work, one of us has characterized the effects of different parametric specifications (such as lag length restrictions, a rational form for the log distribution) for the question of the relative importance of permanent and transitory components. See Quah (1988)."
also require a further technical condition: the innovations in the bivariate Wold decomposition of output growth and unemployment are linear combinations of the underlying demand and supply disturbances.

We now derive the joint process followed by output and unemployment implied by our assumptions. Let \( Y \) and \( U \) denote the logarithm of GNP and the level of the unemployment rate, respectively, and let \( \epsilon_{y} \) and \( \epsilon_{u} \) be the two disturbances. Let \( X \) be the vector \((\Delta Y, U)'\) and \( e \) be the vector of disturbances \((\epsilon_{y}, \epsilon_{u})'\). The assumptions above imply that \( X \) follows a stationary process given by:

\[
X(t) = A(0)e(t-1) + A(1)e(t-2) + \cdots
\]

\[
= \sum_{j=0}^{\infty} A(j)e(t-j),
\]

\[
\text{Var}(e) = \Omega.
\]

This moving average representation is unique and can be obtained by first estimating and then inverting the vector autoregressive representation of \( X \) in the usual way.

Comparing equations (1) and (2) we see that \( \lambda \), the vector of innovations, and \( \epsilon_{y} \), the vector of original disturbances, are related by \( \lambda = A(0)e(t) \), and that \( A(j) = C(j)A(0), \) \( j = 0, 1, \ldots, \) for all \( j \). Thus knowledge of \( A(0) \) allows one to recover \( e \) from \( \lambda \), and similarly to obtain \( A(j) \) from \( C(j) \).

Is \( A(0) \) identified? An informal argument suggests that it is. Equations (1) and (2) imply that \( A(0) \) satisfies \( A(0)A(0)' = \Omega \), and that the upper left-hand entry in \( \Sigma_{\epsilon_{y}e}A(0) = \Sigma_{\epsilon_{u}e}A(0) \) is 0. Given \( \Omega \), the first relation imposes three restrictions on the four elements of \( A(0) \); given \( \Sigma_{\epsilon_{y}e}A(0) \), the other implication imposes a fourth restriction. This informal argument is indeed correct. A rigorous and constructive proof, which we actually use to obtain \( A(0) \), is as follows: Let \( \Sigma \) denote the unique lower triangular Choleski factor of \( \Omega \). Any matrix \( A(0) \) such that \( A(0)A(0)' = \Omega \) is an orthonormal transformation of \( S \). The restriction that the upper left-hand entry in \( \Sigma_{\epsilon_{y}e}A(0) \) be equal to 0 is an orthogonality restriction that then uniquely determines this orthonormal transformation.\footnote{Notice that identification is achieved by a long-run restriction. This raises a heavy technical issue. Without precise prior knowledge of base lengths, inference and restrictions on the kind of long-run behavior we are interested in here is delicate. See for instance Christopher Sims (1972), where we are extrapolating here from Sims’s results which assume exactly exogenous regression. Similar problems may arise in the VAR case, although the results of Kenneth Burt (1974) suggest otherwise. Nevertheless, we can generalize our long-run restriction to one that applies to some neighborhood of frequency zero, instead of just a restriction at the point zero. Under appropriate regularity conditions, we can show that our results are the limit of those from that kind of restriction, as the neighborhood shrinks to zero.}

In summary, our procedure is as follows. We first estimate a vector autoregressive representation for \( X \), and invert it to obtain (2). We then construct the matrix \( A(0); \) and use this to obtain \( A(j) = C(j)A(0), \) \( j = 0, 1, \ldots, \) and \( \epsilon_{y} = A(0)'\epsilon_{u} \). This gives output and unemployment as functions of current and past demand and supply disturbances.

II. Interpretation

Interpreting residuals in small dimensional systems as "structural" disturbances is always perilous, and our interpretation of disturbances as supply and demand disturbances is no exception. We discuss various issues in turn.

Our interpretation of disturbances with permanent effects as supply disturbances, and of disturbances with transitory effects as demand disturbances is motivated by a traditional Keynesian view of fluctuations. For illustrative purposes, as well as to focus the discussion below, we now provide a simple model which delivers those implications. The model is a variant of that in Stanley Fischer (1977):

\[
Y(t) = M(t) = M(t-1) + \epsilon_{y}(t),
\]

\[
\theta(t) = \theta(t-1) + \epsilon_{u}(t),
\]

where \( \epsilon_{y} \) and \( \epsilon_{u} \) are the serially uncorrelated and pairwise orthogonal demand and supply disturbances. Define unemployment \( U \) to be \( N = \bar{N} \); solving for unemployment and output growth then gives:

\[
\Delta Y = \epsilon_{y} - \epsilon_{u} - \epsilon_{y}(t-1)
\]

\[
+ \epsilon_{u}(t-1) + \epsilon_{y}(t-1) + \epsilon_{u}(t),
\]

\[
\Delta Y = \epsilon_{y} - \epsilon_{u} + \epsilon_{y}(t-1) - \epsilon_{u}(t-1) + \epsilon_{y}(t) - \epsilon_{u}(t).
\]

The variables \( Y, N, \) and \( \theta \) denote the log of output, employment, and productivity, respectively. Full employment is represented by \( \bar{N} \) and \( \bar{W} \). Here the log of the price level, the nominal wage, and the money supply.

Equation (3) states that aggregate demand is a function of real balances and productivity. Notice that productivity is allowed to affect aggregate demand directly; it can do so through investment demand for example, in which case \( \alpha > 0 \). Equation (4) is the production function; it relates output, employment, and productivity, and assumes a constant returns-to-scale technology. Equation (5) describes price-setting behavior, and gives the price level as a function of the nominal wage and of productivity. Finally the last equation, (6), characterizes wage-setting behavior in the economy: the wage is chosen one period in advance, and is set so as to achieve (expected) full employment.

To close the model, we need to specify how \( M \) and \( \theta \) evolve. We assume that they follow:

\[
M(t) = M(t-1) + \epsilon_{y}(t),
\]

\[
\theta(t) = \theta(t-1) + \epsilon_{u}(t),
\]

These two equations clearly satisfy the restrictions in equation (1) of the previous section. Due to nominal rigidities, demand disturbances have short-run effects on output and unemployment, but these effects disappear over time. In the long run, only supply, that is, productivity disturbances here, affect output. Neither of the disturbances have a long-run impact on unemployment.

This model is clearly only illustrative. More complex wage and price dynamics, such as in John Taylor (1980), will also satisfy the long-run properties embodied in equation (1). This model is nevertheless a useful vehicle to discuss the limitations of our interpretation of permanent and transitory disturbances.
Granting our interpretation of these disturbances as demand and supply disturbances, one may nevertheless question the assumption that the two disturbances are uncorrelated at all leads and lags. We think of this as a nonissue. The model makes clear that this orthogonality assumption does not eliminate for example the possibility that supply disturbances directly affect aggregate demand. Put another way, the assumption that the two disturbances are uncorrelated does not restrict the channels through which demand and supply disturbances affect output and unemployment again granting our interpretation of these disturbances as demand and supply disturbances, one may argue that even demand disturbances have a long-run impact on output: changes in the subjective discount rate, or changes in fiscal policy may well affect the savings rate, and subsequently the long-run capital stock and output. The presence of increasing return, and of learning by doing, also raise the probability that demand disturbances may have some long-run effects. Even if not, their effects through capital accumulation may be sufficiently long lasting to be indistinguishable from truly permanent effects in a finite data sample. We agree that demand disturbances may well have such long-run effects. However, we also believe that if so, those long-run effects are small compared to those of supply disturbances. To the extent that this is true, then our decomposition is “nearly correct” in the following sense: in a sequence of economies with the same size of disturbances, each with different dynamic effects on output and unemployment, that demand disturbances leave unaffected the dynamic relation between output and unemployment. That demand disturbances should leave the relation between output and unemployment nearly unaffected is highly plausible. Thus the subject to only one, or at least to one dominant, source of supply disturbances is more questionable. If there are many supply disturbances of roughly equal importance, and if, as is likely, each of them affects the dynamic relation between unemployment and output, our decomposition is likely to be meaningless.

In summary, our interpretation of the disturbances is subject to various caveats. Nevertheless, we believe that the interpretation to be reasonable and useful in understanding the results below. We now briefly discuss the relation of our paper to others on the same topic. We first examine how our approach relates to the business-cycle-versus-trend distinction.

Following estimation, we can construct two output series, a series reflecting only the effects of supply disturbances, obtained by setting all realizations of the demand disturbances to zero, and a series reflecting only the effects of demand disturbances, obtained by setting supply realizations to zero. By construction, the first series, the supply component, will be nonstationary while the second, the demand component, is stationary.

A standard distinction in describing output movements is the “business cycle versus trend” distinction. While there is no standard definition of these components, the trend is usually taken to be that part of output that would realize, were all prices perfectly flexible; business cycles are then taken to be the dynamics of actual output around its trend.

It is tempting to associate the first series with the “trend” component of output and the second series with the “business cycle” component. In our view, that association is unwarranted. If prices are in fact imperfectly flexible, deviations from trend will arise not only from demand disturbances, but also from supply disturbances: business cycles will occur due to both supply and demand disturbances. Put another way, supply disturbances will affect both the business cycle and the trend component. Identifying separately business cycles and trend is likely to be difficult, as the two will be correlated through their joint dependence on current and past supply disturbances.

With this discussion in mind, we now review the approaches to identification used by others.

Campbell and Mankiw (1987) assume the existence of two types of disturbances, "trend" and "cycle" disturbances, which are assumed to be uncorrelated. Their identifying restriction is then that trend disturbances do not affect unemployment. The discussion above suggests that this assumption of zero correlation between cycle and trend components is inappropriate; if these two disturbances are instead interpreted as supply and demand disturbances, respectively, the identifying restriction that supply disturbances do not affect unemployment is equally inappropriate.

Clark (1988) also assumes the existence of "trend" and "cycle" disturbances, and also assumes that "trend" disturbances do not affect unemployment but allows for contemporaneous correlation between trend and cycle disturbances. While this may be seen as an improvement over Campbell and Mankiw, it still severely constrains the dynamic effects of disturbances on output and unemployment in ways that are difficult to interpret.

The paper closest to ours is that of Evans (1987). Evans assumes two disturbance components, "unemployment" and "output" disturbances, which can be decomposed as supply and demand disturbances, respectively. By assuming the existence of a reduced form identical to equation (2) above, he also assumes that neither supply nor demand disturbances are uncorrelated, but that both may have a long-run effect on the level of output. However, instead of using the identification restriction that we use here, he assumes that supply disturbances have no contemporaneous effect on output and trend. Instead of using the identification restriction less appealing as a way of achieving identification, it should be clear however that our paper builds on Evans' work.

III. Estimation

We need to confront one final problem before estimation. The representation we use in Section I assumes that both the level of unemployment and the first difference of the logarithm of GNP are stationary around given levels. Postwar U.S. data however suggest instead both a small but steady increase in the average unemployment rate over the sample, as well as a decline in the average
growth rate of GNP since the mid-1970s.\textsuperscript{2} This raises two issues.

The first is that our basic assumptions may be wrong in fundamental ways. For instance, unemployment might in fact be nonstationary, and affected even in the long run by demand and supply disturbances. This is predicted by models with a "hysteresis" effect, as developed in Blanchard and Lawrence Summers (1986), and used by them to explain European unemployment. This property also obtains in some recent growth models with increasing returns to scale, where changes in the savings rate may affect not only the level but also the growth rate of output. While we cannot claim that such effects are not present here, we are willing to assume that their importance is minimal, for the period and the economy at hand.

Next, there is the issue of how to handle the apparent time trend in unemployment, and the apparent slowdown in growth since the mid-1970s. There is no clear solution for this, and we take an eclectic approach.\textsuperscript{3} To focus the discussion, we present as a base case the results from estimation allowing for a change in the growth rate of output, and for a secular increase in the unemployment rate, as captured by a fitted-linear time-trend regression line. There are three other cases of interest: (a) there is no change in the growth rate of output, but there is a secular change in the unemployment rate; (b) there is no secular trend in the unemployment rate, but there is a break in the average growth rate of output; and finally, (c) there is neither a change in the growth rate of output nor a secular change in the unemployment rate.

A VAR system in real GNP growth ($\Delta y$) and the unemployment rate ($u$), allowing for eight lags is estimated using observations from 1950:2 through 1974:4. The GNP data are quarterly; the monthly unemployment data are averaged to provide quarterly observations. Evans (1987) has estimated essentially the same bivariate VAR representation, although he uses instead the aggregate civilian unemployment rate. He has also tested the stationarity assumptions that we use here. The properties of the VAR representation and of the moving average representation found by direct inversion do not have any meaning within our framework, so we do not discuss those further here.

The mean growth rates for output are 3.62 percent and 2.43 percent, at an annual rate, over 1948:2 through 1972:4, and 1974:1 through 1974:4, respectively. This break point is chosen to coincide with the first OPEC oil shock. The fitted-trend regression coefficient for the unemployment rate series is 0.019, which implies a secular increase of 2.97 percentage points over the sample period. When we allow for a change in the output growth, we simply remove the different sample means before estimating the vector autoregression; similarly when we allow for a secular change in the unemployment rate, the fitted-trend line is removed before VAR analysis.

It turns out that the results for cases (a)–(c) are qualitatively similar to those for the base case. More precisely, the moving average responses to demand and supply disturbances are sufficiently close to those of the base case in their main features; the principal differences lie in the magnitudes of the responses. These differences are notable only in forecast error variance decompositions; we will therefore present four such decomposition tables for the different cases below. Because of the similarity in the other qualitative features however, and to conserve space, we will present results for the impulse responses and historical decompositions only for the base case.\textsuperscript{4}

We turn next to the dynamic effects of demand and supply disturbances.

\section*{IV. Dynamic Effects of Demand and Supply Disturbances}

The dynamic effects of demand and supply disturbances are reported in Figures 1 and 2. The vertical axes in Figures 1 and 2 denote simultaneously the log of output and the rate of unemployment; the horizontal axis denotes time in quarters. Figures 3–6 provide the same information, but now with one standard deviation bands around the point estimates.\textsuperscript{5}

Demand disturbances have a hump-shaped effect on output and unemployment. Their effects peak after two to four quarters. The effects of demand then decline to vanish after about three to five years. The responses in output and unemployment are mirror images of each other; we return to this aspect of the results below after discussing the effects of supply disturbances.

The output response is smallest when the raw data are used, without allowing for a break or a secular change in unemployment (case c, not shown); it also decays the most rapidly in this case. Once a change in the average growth rate of output is allowed, the treatment of possible secular changes in unemployment seems to be relatively unimportant for the responses to demand disturbances.

These dynamic effects are consistent with a traditional view of the dynamic effects of aggregate demand on output and unemployment, in which movements in aggregate demand build up until the adjustment of prices and wages leads the economy back to equilibrium.

Supply disturbances have an effect on the level of output which cumulates steadily over time. In the base case, the peak response is about eight times the initial effect and takes place after eight quarters. The effect decreases to stabilize eventually. For good statistical reasons, the long-run impact is imprecisely estimated. The dynamic response in unemployment is quite different: a positive supply disturbance (that is, a supply disturbance that has a positive long-run effect on output) initially increases unemploy-
ment slightly. Following this increase, the effect is reversed after a few quarters, and unemployment slowly returns to its original steady-state value. The dynamic effects of a supply disturbance on unemployment are largely over by about five years.

The qualitative results are similar across all alternative treatments of breaks and trends. The only significant difference appears in the initial unemployment response to demand disturbances: in the case when neither break nor trend is permitted, the response is initially negative rather than positive as in the base case. The one standard deviation band does however include positive values.

The response of unemployment and output are suggestive of the presence of rigidities, both nominal and real. Nominal rigidities can explain why in response to a positive supply shock, say an increase in productivity, aggregate demand does not initially increase enough to match the increase in output needed to maintain constant unemployment; real-wage rigidities can explain why increases in productivity can lead to a decline in unemployment after a few quarters which persists until real wages have caught up with the new higher level of productivity.

Figure 1 and 2 also shed interesting light on the relation between changes in unemployment and output known as Okun's law. The textbook value of Okun's coefficient is about 2.5. Under our interpretation, this coefficient is a mongrel coefficient, as the joint behavior of output and unemployment depends on the type of disturbance affecting the economy. In the case of demand disturbances, Figure 1 suggests that there is indeed a tight relation between output and unemployment. At the peak responses, the graph suggests an implied coefficient between output and unemployment that is slightly greater than 2. In the case of supply disturbances,

there is no such close relation between output and unemployment. In the short run, output increases, unemployment may rise or fall; in the long run, output remains higher whereas—by assumption—unemployment returns to its initial value. In the intervening period, unemployment and output deviations are of opposite sign. At the peak responses, Figure 2 suggests an implied coefficient slightly exceeding four, higher in absolute value than Okun's coefficient. That the absolute value of the coefficient is higher for supply disturbances than for demand disturbances is exactly what we expect. Supply disturbances are likely to affect the relation between output and employment, and to increase output with little or no change in employment.

V. Relative Contributions of Demand and Supply Disturbances.

Having shown the dynamic effects of each type of disturbance, the next step is to assess their relative contribution to fluctuations in output and unemployment. We do this in two ways. The first is informal, and entails a comparison of the historical time-series of the demand component of output to the NBER chronology of business cycles. The second examines variance decompositions of output and unemployment in demand and supply disturbances at various horizons.

A. Demand Disturbances and NBER Business Cycles

From estimation of the joint process for output and unemployment, and our identifying restrictions, we can form the "demand components" of output and unemployment. These are the time paths of output and unemployment that would have obtained in the absence of supply disturbances. Similarly, by setting demand innovations to zero, we can generate the time-series of "supply components" in output and unemployment. From the identifying restriction that demand disturbances have no long-run effect on output, the resulting series of the demand component in the level of output is stationary. By the same token, both the demand and supply components of unemployment are stationary.

The time-series for these components are presented in Figures 7 through 10. Superimposed on these time series are the NBER peaks and troughs. Peaks are shown as vertical lines above the horizontal axis, troughs as vertical lines below the axis.

The peaks and troughs of the demand component in output match closely the NBER peaks and troughs. The two recessions of 1974–75 and 1979–1980 deserve special mention. Our decomposition attributes them in about equal proportions to adverse supply and demand disturbances. This is best shown by giving the estimated values of the supply and demand innovations over these periods. These are collected in Table 1. The recession of 1974–75 is therefore explained by an initial string of negative supply disturbances, and then of
negative demand disturbances. Similarly, the 1979–80 recession is first dominated by a large negative supply disturbance in the second quarter of 1979, and then a large negative demand disturbance a year later. Without appearing to interpret every single residual, we find these estimated sequences of demand and supply disturbances consist with less formal descriptions of these episodes.  

36Formal evidence of a slightly different nature is also available. In Blanchard and Watson (1986), evidence from four time-series is used to decompose fluctuations into supply and demand disturbances. There the recession of 1975 is attributed in roughly equal proportions to adverse demand and supply disturbances, that of 1980 mostly to demand disturbances. To see how much our characterization of the dynamic effects of demand and supply disturbances depend on the 1973–76 episode, we reestimated the model, leaving out 1973–1 to 1976–4. The estimated dynamic effects of both demand and supply disturbances are nearly identical to those described above.  

37By construction, the supply component of unemployment is close to actual unemployment for the first few observations in the sample. Thus, the large decrease from 1950 to 1952 in the supply component simply reflects the actual movement in unemployment in this period. In light of this, we reestimated the model from 1955–3 through the end of our sample. We found little change in the empirical results.

### Table 1: Demand and Supply Innovations

<table>
<thead>
<tr>
<th>Quarter</th>
<th>Demand (Percent)</th>
<th>Supply (Percent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1973–3</td>
<td>-0.8</td>
<td>1.9</td>
</tr>
<tr>
<td>1973–4</td>
<td>0.2</td>
<td>0.4</td>
</tr>
<tr>
<td>1974–1</td>
<td>-0.7</td>
<td>0.6</td>
</tr>
<tr>
<td>1974–2</td>
<td>0.5</td>
<td>1.5</td>
</tr>
<tr>
<td>1974–3</td>
<td>-1.8</td>
<td>1.0</td>
</tr>
<tr>
<td>1974–4</td>
<td>-0.7</td>
<td>1.1</td>
</tr>
<tr>
<td>1975–1</td>
<td>-1.3</td>
<td>2.8</td>
</tr>
<tr>
<td>1976–1</td>
<td>-0.1</td>
<td>0.3</td>
</tr>
<tr>
<td>1976–2</td>
<td>-0.4</td>
<td>1.7</td>
</tr>
<tr>
<td>1977–3</td>
<td>0.3</td>
<td>0.6</td>
</tr>
<tr>
<td>1977–4</td>
<td>-0.8</td>
<td>0.2</td>
</tr>
<tr>
<td>1980–1</td>
<td>0.7</td>
<td>0.4</td>
</tr>
<tr>
<td>1980–2</td>
<td>-1.9</td>
<td>3.0</td>
</tr>
</tbody>
</table>

Notes:  
1The estimated innovations for the other cases follow the same pattern as above.

### Table 2A: Variance Decomposition of Output and Unemployment (No Dummy Break, Time Trend in Unemployment)

<table>
<thead>
<tr>
<th>Horizon (Quarters)</th>
<th>Output</th>
<th>Unemployment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>99.0</td>
<td>51.9</td>
</tr>
<tr>
<td>2</td>
<td>(78, 99.7)</td>
<td>(53, 67.6)</td>
</tr>
<tr>
<td>3</td>
<td>98.6</td>
<td>63.9</td>
</tr>
<tr>
<td>4</td>
<td>(76, 99.6)</td>
<td>(50, 80.0)</td>
</tr>
<tr>
<td>5</td>
<td>97.0</td>
<td>82.7</td>
</tr>
<tr>
<td>6</td>
<td>(71, 98.5)</td>
<td>(49, 79.0)</td>
</tr>
<tr>
<td>7</td>
<td>81.7</td>
<td>81.1</td>
</tr>
<tr>
<td>8</td>
<td>(46, 87.0)</td>
<td>(36, 65.9)</td>
</tr>
<tr>
<td>12</td>
<td>67.2</td>
<td>86.2</td>
</tr>
<tr>
<td>40</td>
<td>(30, 71.9)</td>
<td>(25, 92.1)</td>
</tr>
<tr>
<td>50</td>
<td>39.3</td>
<td>85.6</td>
</tr>
<tr>
<td></td>
<td>(7, 39.3)</td>
<td>(5, 35.6)</td>
</tr>
</tbody>
</table>

### Table 2B: Variance Decomposition of Output and Unemployment (No Dummy Break, Time Trend in Unemployment)

<table>
<thead>
<tr>
<th>Horizon (Quarters)</th>
<th>Output</th>
<th>Unemployment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>93.8</td>
<td>70.2</td>
</tr>
<tr>
<td>2</td>
<td>(59, 93.9)</td>
<td>(53, 92.0)</td>
</tr>
<tr>
<td>3</td>
<td>87.5</td>
<td>79.5</td>
</tr>
<tr>
<td>4</td>
<td>(62, 90.4)</td>
<td>(50, 95.9)</td>
</tr>
<tr>
<td>5</td>
<td>93.4</td>
<td>93.9</td>
</tr>
<tr>
<td>6</td>
<td>(58, 93.3)</td>
<td>(61, 97.5)</td>
</tr>
<tr>
<td>7</td>
<td>78.9</td>
<td>95.7</td>
</tr>
<tr>
<td>8</td>
<td>(53, 90.0)</td>
<td>(63, 98.2)</td>
</tr>
<tr>
<td>9</td>
<td>52.5</td>
<td>88.9</td>
</tr>
<tr>
<td>12</td>
<td>(31, 68.6)</td>
<td>(63, 94.5)</td>
</tr>
<tr>
<td>20</td>
<td>37.8</td>
<td>79.7</td>
</tr>
<tr>
<td>40</td>
<td>(21, 31.4)</td>
<td>(58, 90.3)</td>
</tr>
<tr>
<td>50</td>
<td>18.7</td>
<td>72.9</td>
</tr>
<tr>
<td></td>
<td>(7, 23.5)</td>
<td>(5, 86.8)</td>
</tr>
</tbody>
</table>

### B. Variance Decompositions

While the above empirical evidence is suggestive, a more formal statistical assessment can be given by computing variance decompositions for output and unemployment at various horizons. Tables 2, and 2A–C give this variance decomposition for the different cases. The table has the following interpretation. Define the k quarter-ahead forecast error in output as the difference between the actual value of output and its forecast from equation (2) as of k quarters earlier. This forecast error is due to both unanticipated demand and supply disturbances in the last k quarters. The number for output at horizon k, k = 1, ..., 40 gives the percentage of variance of the k-quarter ahead forecast error due to demand. The contribution of supply, not reported, is given by 100 minus that number. A similar interpretation holds for the numbers for unemployment. The numbers in parentheses are one standard deviation bands, surrounding the point estimate.

13Again, these bands are asymmetric and obtained as described above.
Our identifying restrictions impose only one restriction on the variance decompositions, namely that the contribution of supply disturbances to the variance of output trends to unity as the horizon increases. All other aspects are unconstrained.

Two principal conclusions emerge from these tables. First, the data do not give a precise answer as to the relative contribution of demand and supply disturbances to movements in output at short and medium-term horizons. The results vary across alternative treatments of break and trend. In the base case, the relative contribution of demand disturbances to output fluctuations, at a four quarters horizon, is 98 percent. This contribution falls to 79 percent when no break is allowed but there is a time trend in unemployment, remains about the same when a break is allowed in output growth but there is no trend in the unemployment rate. When neither a break nor a trend is permitted, it is only 39 percent. Next, the standard error bands are quite large in each case, ranging from 71 to 99 percent in the base case, 54 to 90 percent in case A, 73 to 59 percent in case B, and 17 to 72 percent in case C. Evidently when a break is permitted in output growth, the treatment of the trend in unemployment appears to be quite unimportant. These cases are also when the demand contribution is more precisely estimated. Despite the differences across estimates, and the uncertainty associated with each set, we view the results as suggesting an important role for demand disturbances in the short run.

Second, estimates of the relative contribution of the different disturbances to unemployment do not appear to vary a great deal across alternative treatments of break and trend. The contribution of demand disturbances, four quarters ahead, to unemployment fluctuations varies from 80 to 96 percent. In the base case, the one standard error band ranges from 50 to 90 percent with a point estimate of 80 percent. In all cases, the demand disturbance appears to be quite important for unemployment fluctuations at all horizons.

VI. Conclusion and Extensions

We have assumed the existence of two types of disturbances generating unemployment and output dynamics, the first type having permanent effects on output, the second having only transitory effects. We have argued that these two types of disturbances could usefully be interpreted as supply and demand shocks. Under that interpretation, we have concluded that demand disturbances have a hump-shaped effect on output and unemployment which disappears after approximately two to three years, and that supply disturbances have an effect on output which accumulates over time to reach a plateau after five years. We have also concluded that demand disturbances make a substantial contribution to output fluctuations at short- and medium-term horizons; however, the data do not allow us to quantify this contribution with great precision.

While we find this simple exercise to have been worthwhile, we also believe that further work is needed, especially to validate and refine our identification of shocks as supply and demand shocks. We have in mind two specific extensions. The first is to examine the co-movements of what we have labeled the demand and supply components of GNP with a larger set of macroeconomic variables. Preliminary results appear to confirm our interpretation of shocks. We find in particular the supply component of GNP to be positively correlated with real wages at high to medium frequencies, while no such correlation emerges for the demand component. The second extension is to enlarge the system to one in four variables, unemployment, output, prices, and wages. This would also allow examination of different questions from an alternative perspective, as in Blanchard (1989). As one might expect, wage and price data will help identify more explicitly supply and demand disturbances. Research by Jordi Gali (1988), Sung-in Jun (1988), and Matthew Shapiro and Watson (1988) has already extended this work in that particular direction.

Technical Appendix

This technical appendix discusses further and establishes the claims made in the section on interpretation.

First, we asserted in the text that our identification scheme is approximately correct even when both disturbances have permanent effects on the level of output, provided that the long-run effect of demand on output is small. We now prove this.

The first element of the model, output growth, has the moving average representation in demand and supply disturbances:

\[ \Delta Y_t = \alpha_1(L)\epsilon_t + \alpha_2(L)\eta_t, \]

where \( \alpha_1(L) \) is the cumulative effect on the level of output \( Y \) of the disturbance \( \epsilon_t \). The moving average representation \( C(L) \), together with the innovation covariance matrix \( \Omega \), is related to our desired interpretable representation through some identifying ma...
trix $S$, such that:

$$ SS' = \mathbf{0}, \quad \text{and} \quad A(L) = C(L)S. $$

The model is identified by choosing a unique identifying matrix $S$. In the paper, we selected the unique matrix $S$ such that $s_{11}(1) = 0$.

Let the long-run effect of the demand disturbance be $\delta$ instead, where $\delta > 0$ without loss of generality. For each $\delta$, this implies a different identifying matrix $S(\delta)$. Let $[S(\delta) - S(0)] = \max_{S} (S(\delta) - S(0))^2$; this measures the deviation in the implied identifying matrix from that which we use. Since the approximation is thus seen to be a finite-dimensional problem, any matrix norm will induce the same topology, which is all that is needed to study the continuity properties of our identification scheme. All of the empirical results vary continuously in $S$ relative to this topology. Thus, it is sufficient to show that

$$ [S(\delta) - S(0)] \rightarrow 0 \quad \text{as} \quad \delta \rightarrow 0. $$

In words, if an economy has long-run effects in demand that are small but different from zero, our identifying scheme which incorrectly assumes the long-run effects to be zero never will be approximately the correct point estimates.

**Proof:**

We prove this as follows. Since both $S(0)$ and $S(\delta)$ are matrix square roots of $\mathbf{0}$, there exists an orthogonal matrix $V(\delta)$ such that:

$$ S(\delta) = S(0)V(\delta), \quad \text{where} \quad V(\delta)V(\delta)' = \mathbf{1}. $$

Then the long-run effect of demand is the $(1,1)$ element in the matrix:

$$ A(1; \delta) = C(1)S(\delta)C(1)V(\delta)' = \mathbf{1}. $$

But recall that the elements of the first row of $C(1)\mathbf{0}(0)$ are respectively, the long-run effects of demand and of supply on the level of output, when the long-run effect of demand is restricted to be zero. Thus for any $V(\delta)$, the new implied long-run effect of demand is simply the long-run effect of supply.

**Theorem:**

Let $X$ be a bivariate stochastic sequence generated by

$$ (i) \quad \mathbf{Y}_{t} = B(L)\mathbf{f}_{t}, $$

$$ (ii) \quad \mathbf{f}_{t} = (f_{1t}, f_{2t})'_{t}, \quad \text{with} \quad f_{1t}\mathbf{p}_{1}, \quad f_{2t}\mathbf{p}_{2}; $$

$$ (iii) \quad C_{t} = (c_{1t}, c_{2t})_{t}, \quad \text{with} \quad c_{1t} f_{1t}, \quad c_{2t} f_{2t}, $$

$$ (iv) \quad C_{t} = (c_{1t}, c_{2t})_{t}, \quad \text{with} \quad c_{1t} f_{1t}, \quad c_{2t} f_{2t}, $$

$$ (v) \quad B(1) = (1 - \theta)B_{1}(1), $$

$$ (vi) \quad B_{1}(1) = (1 - \theta)B_{1}(1). $$

It is straightforward to verify that the matrix covariogram implied by this moving average matches that of the true underlying model. Further, the unique zero of the determinant is 2, and consequently lies outside the unit circle. Therefore this moving average representation is, as asserted, obtained from the vector autoregressive representation of the true model.

However, this moving average does not satisfy our identifying assumption that the "demand" disturbance has only transitory effects on the level of output. We therefore apply our identifying transformation to obtain:

$$ X_{t} = \begin{pmatrix} [1 - L] & \frac{1}{2}(3 - L) & \mathbf{c}_{1t} \\ -1 & 1 & \mathbf{c}_{2t} \end{pmatrix}. $$

This moving average representation is what we would recover if in fact the data are generated by the three disturbances $(f_{1t}, f_{2t}, f_{3})$. Notice that while the supply disturbance affects output and unemployment only, the demand disturbance affects both output and unemployment. To clarify the issues involved, we provide an explicit example where our procedure produces misleading results. Suppose that there is only one supply disturbance and two demand disturbances: $f_{1t} = (f_{1t}, f_{2t}, f_{3})$. Suppose further that the first demand disturbance affects only output, while the second demand disturbance affects only unemployment. The supply disturbance affects both output and unemployment. Formally, assume that the model is:

$$ X_{t} = \begin{pmatrix} \mathbf{Y}_{t} \\ \mathbf{u}_{t} \end{pmatrix} = \begin{pmatrix} [1 - L] & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} f_{1t} \\ f_{2t} \end{pmatrix}. $$

An unrestricted VAR representation corre-
The second part of the theorem establishes necessary and sufficient conditions on the underlying model such that the bivariate identification procedure does not inappropriately confound demand and supply disturbances. In words, correct identification is possible if and only if the individual distributed lag responses in output growth and unemployment are sufficiently similar across the different demand disturbances, and across the different supply disturbances. This does not mean that the dynamic responses in output growth and unemployment across demand disturbances must be identical or proportional, simply that they differ up to a scalar lag distribution.

Thus even though in general a bivariate procedure is misleading, there are important and reasonable sets of circumstances under which our technique provides the "correct" answers. For instance, suppose that there is only one supply disturbance but multiple demand disturbances. Suppose further that each of the demand components in the level of output has the same distributed lag relation with the corresponding demand component in unemployment. This assumption is consistent with our "production function"-based interpretations below. Then our procedure correctly distinguishes the dynamic effects of demand and supply components in output and unemployment.

**PROOF OF THEOREM:** By (1)–(4), the matrix spectral density of $X$ is given by $S_X(u) = \sum_{k=-1}^{\infty} \beta_k B_k$. By reasoning analogously to that in pp. 44–48 in Rozanov (1967), there exists a $2 \times 2$ matrix function $C$, each of whose elements are analytic, with det $C \neq 0$ for $|z| \leq 1$, and $S_Z = CC^{*}$ on $|z| = 1$. This represents $X$ as a moving average in unit variance orthogonal white noise, obtainable from $S_Y$ in a V.A.R. mean square approximation. However, such a moving average $C$ need not satisfy the condition that the first (demand) disturbance have only transitory impact on the level of output (condition (iv)). Form the $2 \times 2$ orthogonal matrix $M$ whose second column is the transpose of the first row of $C$ evaluated at $z = 1$, normalized to have length 1, as a sector. Then $A = CM$ provides the moving average representation satisfying (vii)–(x). For the second part of the theorem, notice that $A$ has been constructed so that $|z| = 1$:

$$1 - z^{-1} \alpha_0 B_1 + \alpha_2 z^{-2}$$

$$1 - z^{-1} \beta_1 + \beta_2 z^{-2}$$

$$+ B_1 \beta_2^*$$

$$(1 - z) \alpha_1 \alpha_2^* + \alpha_1 \alpha_2$$

$$= (1 - z) \beta_1 \beta_2^*$$

$$+ B_1 \beta_2^*$$

$$|\alpha_1^2| + |\alpha_2^2| = B_2 \beta_2^*$$

$$+ B_2 \beta_2^*.$$  

For $\epsilon_2$ to be orthogonal to $f_0$, and $\epsilon_1$ to be orthogonal to $f_1$, at all lags and lags, it is necessary and sufficient that on $|z| = 1$:

(a) $|\alpha_1^2| = B_1 \beta_1^*$;

(b) $|\alpha_2^2| = B_1 \beta_1^*$;

(c) $|\alpha_1 \alpha_2| = B_2 \beta_2^*$;

(d) $a_1 \alpha_2^* = B_1 \beta_1^*$;

(e) $|\alpha_1 \alpha_2| = B_2 \beta_2^*$;

(f) $|\alpha_2^2| = B_2 \beta_2^*$;

Consider relations (a), (c), (d), (e), Denoting complex conjugation of $B$, by $B$, the triangular inequality implies that:

$$|B_{11} B_{22}| = \sum_{k=1}^{\infty} |B_{11} B_{22}|^2 \leq \sum_{k=1}^{\infty} |B_{11} B_{22}|^2,$$

where the inequality is strict unless $B_{11} B_{22}$ is a complex scalar multiple of $B_{11}$ for each $z$ on $|z| = 1$. Next, by the Cauchy-Schwarz inequality:

$$\sum_{k=1}^{\infty} |B_{11} B_{22}|^2 \leq \left( \sum_{k=1}^{\infty} |B_{11}|^2 \right)^{1/2} \left( \sum_{k=1}^{\infty} |B_{22}|^2 \right)^{1/2},$$

again with strict inequality except when $B_{22}$ is a complex scalar multiple of $B_{22}$ for each $z$ on $|z| = 1$. Therefore:

$$|\alpha_1 \alpha_2^*| \leq |\alpha_1^2| - |\alpha_2^2|,$$

where the inequality is strict except when $B_{22}$ is a complex scalar multiple of $B_{22}$ for each $z = 1$. But the strict inequality is a contradiction as $a_1$ and $a_2$ are just scalar functions. Thus (a), (c), and (e) can be simultaneously satisfied if and only if there exists some complex scalar function $\gamma(z)$ such that $B_{22} = \gamma_0 B_{22}$. A similar argument applied to (b), (d), and (f) shows that they hold simultaneously if and only if there exists some complex scalar function $\gamma(z)$ such that $B_{22} = \gamma_2 B_{22}$. This establishes the theorem.

**REFERENCES**


Detecting shift-contagion in currency and bond markets

Toni Gravelle a, Maril Kichian b, James Morley c,⁎

a International Monetary Fund, United States
b Bank of Canada, Canada
c Washington University, St. Louis, United States

Received 7 February 2005; received in revised form 27 June 2005; accepted 6 July 2005

Abstract

This paper investigates why financial market crises often increase the interdependence between assets associated with different countries. Two sources of increased co-movement in asset returns are considered: (i) larger common shocks operating through standard cross-country linkages and (ii) changes in the structural transmission of shocks across countries, referred to as "shift-contagion". To examine this issue, we develop a method for detecting shift-contagion with three notable features. First, parameters corresponding to the structural transmission of shocks across countries are identified in the presence of changing volatility regimes for the shocks. Second, the timing of changes in volatility is endogenously estimated instead of being exogenously assigned. Third, the countries in which crises originate need not be known or even included in the analysis. We apply the method to currency returns for developed countries and bond returns for emerging-market countries.

© 2005 Elsevier B.V. All rights reserved.

Keywords: Contagion; Financial market crises; Regime switching; Asset markets; Structural transmission

JEL classification: F42; G15; C12

1. Introduction

The spillover of crises from one financial market to another is loosely referred to as contagion, but precise definitions of contagion are many. One is that contagion occurs whenever asset returns associated with different countries co-vary in excess of what would be implied by the usual commercial, financial, and institutional links between countries. Another more narrow definition is that contagion occurs whenever shocks spread through herding behaviour in financial markets. A third broader definition refers to any increased co-movement between asset returns during crises as contagion. A fourth definition, referred to as "shift-contagion", suggests that to qualify as contagion, increased co-movement between asset returns during crises must be driven by changes in the structural transmission of shocks across countries, rather than just a change in size of underlying shocks.

Given this multiplicity of definitions, it is not surprising to find widely varying opinions as to which crisis events cause or have caused contagion. Early tests for a shift in the way shocks are transmitted across countries suggested the existence of contagion. For example, King and Warns (1990) modeled contagion as a spillover of volatility from one financial market to another in the presence of imperfect information availability across markets. They considered a number of tests for volatility spillovers, including estimating time-varying correlations between equity returns in international markets. They found correlations increased significantly just after the October 1987 stock market crash. Other studies, including Bennett and Kehl (1986) and Lee and Kim (1993) reached similar conclusions. However, Forbes and Rigobon (2002) argued that the conclusions from these and similar studies might be misleading due to the simultaneous nature of financial interactions and the presence of heteroskedasticity in equity returns. For example, in the case of heteroskedasticity, they pointed out that when the variances of two assets increase (as they typically do during periods of crises), their correlation also increases regardless of whether or not the structural transmission of shocks between these assets changes. Taking such econometric concerns into account, a number of recent studies have concluded that there is, in fact, little or no evidence of contagion in financial markets. For example, Forbes and Rigobon (2002) and Rigobon (2001) found little evidence of shift-contagion in equity and bond markets during the Mexican, Asian, and Russian crises of the 1990s. Similarly, Rigobon (2003a) concluded that no shift-contagion occurred between 1994 and 1999 in the Brady bond markets of Argentina and Mexico.

In this paper we develop a method for detecting shift-contagion with three notable features that are designed to address concerns about the previous empirical literature. First, the parameters related to the structural transmission of common shocks across countries are identified in the presence of regime-switching volatility in the common shocks. In particular, structural identification occurs as long as the heteroskedasticity in idiosyncratic structural shocks is not perfectly synchronous with the heteroskedasticity in the common structural shocks. This is an example of "identification through heteroskedasticity" (see Sentana and Fiorentino, 2001; Rigobon, 2003a). In terms of testing for shift-contagion, if the change in structural impact coefficients is proportional given a change in volatility regime, it suggests a change in the size of

⁎ Corresponding author. Department of Economics, Box 1208, Eliot Hall 204, Washington University, One Brookings Drive, St. Louis, MO 63130-4899, United States. Tel.: +1 314 935 4437; fax: +1 314 935 4156.
E-mail addresses: morley@wustl.edu (J. Morley).
0022-1996/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
shocks, but not in the structural transmission of shocks. Importantly, this approach addresses the econometric concerns raised by Forbes and Rigobon (2002) about early studies of contagion. Second, the timing of changes in volatility is endogenously estimated, instead of being exogenously assigned. There is considerable evidence of Markov regime switching in asset markets. Furthermore, the estimated regimes tend to match with events such as the Mexican, Asian and Russian crises of the 1990s. While endogenously estimating regimes obviously lowers the power of any test of contagion versus exogenously assigning the correct timing based on an ex post dating of crises, it is not always clear a priori what the correct timing should be. Meanwhile, results for tests of contagion can be quite sensitive to even small changes in the dating of crises (see Rigobon, 2003). Third, the countries in which crises originate need not be known or even included in the analysis. While crises often can be linked to specific countries, there are sometimes more general crises driven by large common shocks such as movements in U.S. macroeconomic variables, changes to international demand conditions, liquidity shocks, or changes in attitudes towards risk (see Calvo, 2002). Methods that assume crises are always caused by large idiosyncratic shocks in one country and look for spillovers to other countries will mistakenly find evidence of contagion whenever common shocks affect more than one country simultaneously. A related drawback of such methods is that the country generating the crisis may not always be known with certainty. For example, it is difficult to assign the instability in the European monetary system in the early 1990s to only one country, as many of them were experiencing a crisis at the same time. Our method detects common shocks, regardless of their source, and produces evidence of shift-contagion only if the structure of their simultaneous transmission to any pair of countries is fundamentally altered by crises.

We apply our method to investigate the presence of shift-contagion in currency markets for developed countries and bond markets for emerging-market countries. A priori, one might expect evidence of shift-contagion would be harder to detect in the currency markets for developed countries than bond markets for Latin American countries, in part because the currency markets are relatively less volatile. However, confirming the power of our test, we reject the null hypothesis of no shift-contagion for a number of currency returns, especially for European countries. Meanwhile, confirming the results in Rigobon (2003a), we find little evidence of shift-contagion in Latin American bond markets.

The paper proceeds as follows. Section 2 presents our model and explains the assumptions required for its identification. Section 3 describes the data and reports diagnostic tests that provide general support for our modeling assumptions. Section 4 reports model estimates and results for a number of diagnostic tests of the model. Section 5 reports results for our test of shift-contagion and provides some interpretation of the results. Section 6 concludes with a brief discussion of policy implications.

2 Model

In this section, we develop an empirical model of changing interdependence between two assets (e.g., two foreign currencies or two foreign bonds). Let \( s_{it} \) and \( s_{jk} \) denote continuously compounded returns on the two assets. Conceptually, the returns can be decomposed as follows:

\[
\eta_{it} = \mu_{i} + \epsilon_{it} \tag{1}
\]

where \( \mu_{i} \) is the expected return on asset \( i \), with \( i = 1, 2 \) throughout the remainder of the paper, and \( \epsilon_{it} \) is a forecast error reflecting unexpected news about the asset. Corresponding to an assumption of serially uncorrelated returns, the expected return is constant and the forecast error has mean zero and is uncorrelated across time (i.e., \( E[\epsilon_{it}|\epsilon_{t-k}] = 0 \) for all \( k > 0 \)). Meanwhile, for two assets in the same general market (e.g., the foreign exchange market or the bond market), the forecast errors are contemporaneously correlated \( E[\epsilon_{it}|\epsilon_{jkt}] \neq 0 \).

Contemporaneous correlation between forecast errors implies the existence of common structural shocks to the asset returns. In particular, we may decompose each forecast error into common and idiosyncratic structural shocks:

\[
\epsilon_{it} = \sigma_{it} z_{it} + \sigma_{it} x_{it} \tag{2}
\]

where \( z_{it} \) is the common shock, \( x_{it} \) is an idiosyncratic shock, and \( \sigma_{it} \) and \( \sigma_{it} \) determine the impact of the structural shocks on the asset returns. In particular, the variance of the \( z_{it} \) shocks are normalized to unity, giving the \( \sigma_{it} \) impact coefficients the interpretation of standard deviations of the structural shocks. As such, we normalize the impact coefficients to be positive, except for \( \sigma_{jk} \), which can be positive or negative in order to allow for a positive or negative correlation between \( z_{it} \) and \( z_{jk} \). The shocks have mean zero and are uncorrelated both across time and with each other \( E[z_{it}|z_{jk}] = 0 \) for all \( k > 0 \) and \( E[z_{it}|z_{jk}] = 0 \) for \( j \neq k \), where \( j = 1, 2 \) throughout the remainder of the paper.

We use the impact coefficients on the common shocks to investigate why the interdependence between two assets changes over time. For example, suppose that an increased co-movement between asset returns during financial market crises reflects larger common shocks operating through standard market linkages. Then, both \( \sigma_{ij} \) and \( \sigma_{jk} \) will be larger during crises. However, they will both increase in proportion to the larger size of the common shocks. That is, the ratio \( \sigma_{ij}/\sigma_{jk} \) will remain unchanged before and after the onset of a crisis. By contrast, suppose that crises produce a change in the transmission of common shocks to the two countries under consideration, as would be the case under shift-contagion. Then, the ratio \( \sigma_{ij}/\sigma_{jk} \) will be different during crises than in normal times. Thus, we can test for shift-contagion by estimating the impact coefficients for the common shocks and determining whether or not their ratio changes during crises.

Given the setup in (1)–(2), the main challenge is how to estimate the impact coefficients for the common shocks. Note that the variance–covariance matrix for the forecast errors can be represented in terms of the \( \sigma \) coefficients:

\[
\Sigma = \begin{bmatrix}
\sigma_{11} & \sigma_{12} \\
\sigma_{12} & \sigma_{22}
\end{bmatrix}
\begin{bmatrix}
\sigma_{11} & \sigma_{11} & \sigma_{12} & \sigma_{12} \\
\sigma_{11} & \sigma_{11} & \sigma_{12} & \sigma_{12}
\end{bmatrix}
\begin{bmatrix}
\sigma_{11} & \sigma_{12} \\
\sigma_{12} & \sigma_{22}
\end{bmatrix}
\tag{3}
\]

As discussed in the next section, the assumption of serially uncorrelated returns is more tenable for the currency returns than for the bond market returns. Thus, for the bond market returns, we consider a specification of time-varying expected returns, which, for simplicity of presentation, we describe in the next section.
Therefore, we can use estimates of the variance–covariance matrix \( \Sigma \) to make inferences about \( \sigma_{ei,1} \) and \( \sigma_{ei,2} \) and to test for shift-contagion. It is not immediately obvious that \( \sigma_{ei,1} \) and \( \sigma_{ei,2} \) will be identified given \( \Sigma \). Indeed, if the variances of the structural shocks remain constant, the impact coefficients will not be identified. In particular, there are only three moments corresponding to the forecast error variances and covariance, while there are four structural parameters:

\[
\begin{align*}
\text{var}(u_{11}) &= \sigma_{e1}^2 + \sigma_{e2}^2, \\
\text{var}(u_{22}) &= \sigma_{e2}^2 + \sigma_{e3}^2, \\
\text{cov}(u_{12}, u_{22}) &= \sigma_{e1} \sigma_{e2}. 
\end{align*}
\]

Of course, with constant variances, there would be no change in the interdependence between the two assets over time. Therefore, there would be no shift-contagion, by definition. On the other hand, in the presence of regime switching in the volatility of the structural shocks, there can be changes in the interdependence between assets. In this case, the structural impact coefficients may be identified.

For our model, we assume that each type of structural shock switches between low volatility and high volatility regimes, although only regime switching in the common shocks is necessary for identification of \( \sigma_{ei,1} \) and \( \sigma_{ei,2} \). In terms of the structural impact coefficients in (2), the regime switching can be represented as follows:

\[
\begin{align*}
\sigma_{ei} &= \sigma_{ei} (1 - S_i) + \sigma_{ei}^* S_i, \\
\sigma_{e2} &= \sigma_{e2} (1 - S_i) + \sigma_{e2}^* S_i, 
\end{align*}
\]

where the state variables \( S_i = (0, 1) \). We normalize the regimes such that an asterisk corresponds to higher volatility (i.e., \( \sigma^* > \sigma \)). Then, to see identification, consider the following moments related to a high volatility regime for each structural shock:

\[
\begin{align*}
\text{var}(u_{11}|S_i = 1) &= \sigma_{e1}^2 + \sigma_{e2}^2, \\
\text{var}(u_{22}|S_i = 1) &= \sigma_{e2}^2 + \sigma_{e3}^2, \\
\text{cov}(u_{12}, u_{22}|S_i = 1) &= \sigma_{e1} \sigma_{e2}^*, \\
\text{var}(u_{11}|S_i = 1) &= \sigma_{e1}^2 + \sigma_{e1}^2, \\
\text{var}(u_{22}|S_i = 1) &= \sigma_{e2}^2 + \sigma_{e2}^2.
\end{align*}
\]

Combined with the three moments in (4)–(6) corresponding to low variance regimes, these five moments are sufficient to identify the eight structural parameters in (7)–(8). This approach is an example of "identification through heteroskedasticity".

We complete the model by specifying how the volatility regimes evolve over time. In order to allow for sudden jumps between high and low volatility, we assume that the volatility regimes are Markov switching:

\[
\begin{align*}
\Pr[S_i = 0|S_{i-1} = 0] &= \psi_1, \\
\Pr[S_i = 1|S_{i-1} = 1] &= \psi_2.
\end{align*}
\]

Under this specification, the timing of changes in volatility is endogenously estimated instead of being exogenously assigned ex post. This approach stands in contrast to a number of papers looking at shift-contagion, including Rigobon (2001, 2003a,b), Forbes and Rigobon (2001), and Fevrem and Giavazzi (2002), in which crises are identified ex post based on "conventional wisdom" or as occurring whenever a shock is two or three standard deviations greater than average. The complete model is given by (1)–(2) (7)–(8) and (14)–(15). Then, given data on asset returns and an assumption of Normality for the underlying structural shocks, we can estimate the parameters via maximum likelihood using the techniques for Markov-switching models presented in Hamilton (1989).

3. Data

In principle, the model presented in the previous section can be applied to any pair of assets. In this paper we examine two categories of assets: currencies for Australia, Germany, Japan, Norway, Sweden and Switzerland, and Brady bonds for Mexico, Brazil, Venezuela and Argentina. Exchange rates for the currencies are quoted relative to the US dollar at weekly frequency and extend from the week of January 2, 1985 to the week of June 6, 2001. We did not use a systematic method in choosing these currencies except that we wanted to use the Deutsche Mark as a proxy for the Euro. Thus, we excluded other countries that are part of the Euro-zone from the set of foreign exchange data. The bond data are weekly spread-yields on the EMBI index constructed by JP Morgan and are US-dollar-denominated. The bond yields extend from the week of January 2, 1991 to the week of September 19, 2001 for all cases except Argentina where they start the week of May 5, 1993. For both currencies and bonds, we calculate continuously compounded returns by taking natural logs, differencing, and multiplying by 100 to get percentages.

Table 1 reports results for a number of diagnostic tests for the asset return data. In particular, we test for serial correlation, Normality, autoregressive conditional heteroskedasticity (ARCH),

\footnote{There are additional moments corresponding to the simultaneous occurrence of high volatility regimes for more than one type of structural shock. Thus, the structural parameters are potentially overidentified by sample moments, although these additional moments may not be particularly relevant if the high volatility regimes for different structural shocks do not coincide very often in the sample.}

\footnote{Cayuela, Cipolletta, and Syngelakis (2005) endogenously estimate regimes in a test of shift-contagion in East Asian stock markets during the period of June 1997–June 1998. However, instead of using Markov switching, they determine the timing of crises as corresponding to the dates that would maximize the value of their test statistics for shift-contagion. The computational burden of their approach (including determining the bootstrap distribution of their test statistics) constrains their analysis to consider only one possible crisis during a relatively short period of time. By contrast, endogenous estimation via the assumption of Markov-switching volatilities allows us to consider much longer sample periods that potentially include multiple crises.}
Table 1
Diagnostic tests for asset returns

<table>
<thead>
<tr>
<th>Q(1)</th>
<th>LM(1)</th>
<th>Q(4)</th>
<th>LM(4)</th>
<th>Jarque-Bera</th>
<th>ARCH(8)</th>
<th>ARCH(4)</th>
<th>LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Currency returns, developed countries</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Australia 2.04 2.03 17.72** 15.27*** 707.17*** 59.05*** 66.09*** 133.37***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Germany 0.15 0.15 4.19 3.95 99.99*** 0.93 26.88*** 55.02***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Japan 0.38 0.38 10.14** 9.41** 480.39*** 8.88** 11.18** 83.97***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Norway 0.16 0.16 2.80 2.72 143.84*** 3.53* 24.91*** 85.50***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sweden 0.35 0.35 2.92 2.98 600.97*** 0.01 10.74*** 75.70***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Switzerland 0.04 0.04 5.29 5.23 56.25*** 0.24 13.70** 32.65***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bond returns, emerging-market countries</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Argentina 5.16** 3.15** 27.37** 25.60** 1640.01*** 4.45* 17.20** 165.50***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brazil 12.86*** 12.86*** 43.25*** 36.53*** 7107.99*** 2.64 22.16*** 282.16***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mexico 1.79 1.80 14.49*** 14.15*** 394.54*** 39.72** 72.90*** 173.88***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Venezuela 1.22 1.22 16.37** 17.27** 1834.99*** 1.05 41.17** 142.71***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Q(1) refers to the Ljung–Box test for no serial correlation up to lag 1, LM(1) is the Breusch–Godfrey Lagrange Multiplier test for no serial correlation up to lag 1, Q(4) is the Jarque-Bera test for the null of Normality, ARCH(8) is the Lagrange Multiplier test for no ARCH effects of order 8, and LR is the likelihood ratio statistic for the null of no Markov switching in the variance. ** denotes significance at 1% level, * denotes significance at 5% level, and denotes significance at 10% level. All test statistics have a $\chi^2$ distribution under the null hypothesis, except for the LR test. The significance for the likelihood ratio test of Markov-switching is based on Garcia (1998).

Table 2
Estimates of impact coefficients for common shocks

<table>
<thead>
<tr>
<th>$\sigma_{12}$</th>
<th>$\sigma_{12}^*$</th>
<th>$\sigma_{21}$</th>
<th>$\gamma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Currency returns, developed countries</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Australia/Germany 0.106 (0.054) 1.249 (0.225) 1.393 (0.211) 2.082 (0.254) 7.88</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Australia/Japan 0.362 (0.215) -0.327 (0.188) 1.096 (0.115) 0.992 (0.147) 1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Australia/Norway 0.248 (0.135) 0.698 (0.243) 1.415 (0.157) 1.768 (0.211) 2.30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Australia/Sweden 0.223 (0.061) 0.822 (0.072) 1.426 (0.162) 1.598 (0.174) 3.10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Australia/Switzerland 0.103 (0.054) 0.963 (0.086) 1.426 (0.188) 1.803 (0.150) 7.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Germany/Japan 0.743 (0.406) 0.901 (0.504) 1.355 (0.414) 2.096 (0.199) 1.11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Germany/Norway 1.290 (0.047) 1.109 (0.038) 1.931 (0.110) 1.831 (0.062) 1.12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Germany/Sweden 1.264 (0.054) 1.217 (0.054) 1.728 (0.065) 1.274 (0.048) 1.31</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Germany/Switzerland 1.383 (0.044) 1.510 (0.051) 2.212 (0.174) 2.053 (0.159) 1.21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Japan/Norway 0.841 (0.033) 0.665 (0.055) 2.113 (0.150) 1.175 (0.097) 1.42</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Japan/Sweden 0.822 (0.056) 0.663 (0.084) 2.044 (0.155) 1.094 (0.096) 1.51</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Japan/Switzerland 0.834 (0.049) 0.949 (0.090) 2.116 (0.189) 1.092 (0.027) 1.42</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Norway/Sweden 1.023 (0.037) 1.020 (0.033) 1.394 (0.173) 1.814 (0.110) 1.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Norway/Switzerland 1.087 (0.080) 1.375 (0.056) 1.851 (0.164) 1.835 (0.134) 1.28</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sweden/Switzerland 1.017 (0.037) 1.490 (0.052) 1.969 (0.163) 1.915 (0.161) 1.46</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bond returns, emerging-market countries</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Argentina/Brazil 4.598 (0.473) 4.592 (0.556) 17.445 (2.779) 17.631 (6.979) 1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Argentina/Mexico 3.483 (0.321) 5.456 (0.369) 8.824 (1.088) 14.599 (2.825) 1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Argentina/Venezuela 5.096 (0.239) 3.732 (0.260) 21.943 (7.113) 17.535 (6.000) 1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brazil/Mexico 3.249 (0.226) 4.007 (0.239) 14.194 (5.181) 12.695 (3.129) 1.38</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brazil/Venezuela 4.064 (0.566) 3.201 (0.527) 18.823 (9.206) 15.049 (2.813) 1.01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mexico/Venezuela 5.096 (0.214) 2.998 (0.261) 6.897 (2.281) 15.875 (4.542) 3.62</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Standard errors are reported in parentheses.

In terms of the apparent serial correlation in bond returns, we speculate that the short-horizon predictability could reflect a risk premium that varies with the level of volatility in the bond market. Specifically, for our model of bond returns only, we replace the assumption of a constant expected return in (1) with an assumption that the expected return is time-varying and depends on the state of the common shocks:

$$\mu_q = \mu_1 (1 - S_{t-1}) + \mu_2 S_{t-1}. \quad (16)$$

Note that we do not relate the expected return to the idiosyncratic shocks, which are uncorrelated with the common shocks and, therefore, provide diversifiable risk that should not be priced according to basic financial theory (e.g., the CAPM). In the next section, we present diagnostic tests to support this assumption for the bond returns.

4. Estimates

Table 2 reports estimates of model parameters related to common structural shocks. While the impact coefficients in the low volatility regime, $\sigma_{12}$ and $\sigma_{21}$, vary greatly across countries, they are of similar magnitude for a given country. For example, the standard deviation of common shocks in the low volatility regime is small for Australia, ranging between 0.103 and 0.362. By contrast, it is large for Mexico, ranging from 4.007 and 5.456. Meanwhile, the impact coefficients in the high volatility regime, $\sigma_{12}'$ and $\sigma_{21}'$, are more similar across countries, at least within the currency and bond markets. For currency returns, the standard deviations range between 0.992 and 2.212. For bond returns, they range between 6.837 and 21.945. Consistent with the idea of shift-contagion, the implication of these results is that common shocks have disparate effects on different countries in normal times, but even very different countries get lumped together during crises.

In order to consider the implications of the estimates for shift-contagion, we also report a ratio of the estimated impact coefficients in Table 2. The ratio reveals whether impact coefficients in the high volatility regime are proportional to the impact coefficients in the low volatility regime. Again, the coefficients would be proportional if crises correspond to a change in the size of shocks, but not a change in the structural transmission of the shocks across countries. Our statistic, denoted $\gamma$, is constructed as follows:

$$\gamma = \max \left( \frac{\sigma_{12}}{\sigma_{12}'}, \frac{\sigma_{21}}{\sigma_{21}'} \right). \quad (17)$$

In particular, $\gamma$ is the ratio of the impact coefficients in the high volatility regime to the ratio of the impact coefficients in the low volatility regime. For straightforward comparison across country pairs, we consider the absolute value of the ratios and we normalize the order of the
5. Test

The points estimate in this previous section suggest the possibility of shift-contraction in a number of countries. We formally test for this presence using a likelihood ratio statistic for the following hypothesis:

\[ H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2 \]

where the null hypothesis corresponds to no shift-contraction and the alternative hypothesis corresponds to shift-contraction. Given general support for our model specification, including the Marko-switching assumption, the likelihood ratio statistic under the null hypothesis is a \( F \) distribution with \( N-5 \) degrees of freedom, where \( N \) is the number of observations.

The likelihood ratio statistic is computed as:

\[ LR = -2 \left( \ln L_{\text{null}} - \ln L_{\text{alternative}} \right) \]

where \( L_{\text{null}} \) and \( L_{\text{alternative}} \) are the likelihoods under the null and alternative hypotheses, respectively.

In terms of the empirical results, we can only reject the null hypothesis of no shift-contraction when the likelihood ratio statistic is significant. The likelihood ratio statistic is computed for all of the Eurozone countries, except for Italy, which we exclude due to a lack of data.

The results of the likelihood ratio test are consistent with the previous test results. In terms of the Eurozone countries, we can only reject the null hypothesis of no shift-contraction when it is included in the country pair. For the other country pairs, the likelihood ratio statistic is not significant.
Table 4
Likelihood ratio test for shift-contagion

<table>
<thead>
<tr>
<th>Currency returns, developed countries</th>
<th>Australia</th>
<th>Germany</th>
<th>Japan</th>
<th>Norway</th>
<th>Sweden</th>
</tr>
</thead>
<tbody>
<tr>
<td>Germany</td>
<td>2.221</td>
<td>0.128</td>
<td>0.826</td>
<td>0.542</td>
<td>0.499</td>
</tr>
<tr>
<td>Japan</td>
<td>2.643</td>
<td>0.104</td>
<td>0.470</td>
<td>0.493</td>
<td>0.493</td>
</tr>
<tr>
<td>Norway</td>
<td>0.077</td>
<td>0.781</td>
<td>0.060</td>
<td>0.012</td>
<td>0.012</td>
</tr>
<tr>
<td>Sweden</td>
<td>0.481</td>
<td>0.498</td>
<td>0.000</td>
<td>0.011</td>
<td>0.011</td>
</tr>
<tr>
<td>Switzerland</td>
<td>0.665</td>
<td>0.427</td>
<td>0.533</td>
<td>0.033</td>
<td>0.006</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bond returns, emerging-market countries</th>
<th>Argentina</th>
<th>Brazil</th>
<th>Mexico</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazil</td>
<td>0.000</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Mexico</td>
<td>2.012</td>
<td>0.106</td>
<td>8.700</td>
</tr>
<tr>
<td>Venezuela</td>
<td>0.900</td>
<td>0.343</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Likelihood ratio statistic is for the null of no contagion against the alternative of contagion for the indicated country pair. The test statistic has a $X^2(1)$ distribution under the null hypothesis. $p$-values are reported in parentheses.

For the country pairs that include Mexico, the likelihood ratio statistics range from 2.012 for Argentina/Mexico to 8.780 for Brazil/Mexico.14

In order to interpret the findings from the likelihood ratio tests, it is useful to examine the probabilities of high volatility regimes for some of the country pairs. Fig. 1 displays probabilities of high volatility regimes for common shocks to European currency returns. The high volatility in 1991, 1992, and 1993 is present in every case and is likely related to difficulties over the Exchange Rate Mechanism (ERM) requiring foreign exchange intervention for countries seeking to join the common Euro currency. The crisis clearly affected all of the countries, including those outside of the ERM. However, while the structural links appear to change in many cases, the high volatility common shocks have the same proportionate effect as the low volatility common shocks for Norway/Sweden, suggesting no change in the structural links between these two countries. Meanwhile, Fig. 2 displays probabilities of high volatility regimes for common shocks to Latin American bond returns. There are common high volatility regimes in early 1994, early 1995, late 1997, the second-half of 1998, and early 1999. The 1994 regime likely corresponds to the Venezuelan banking and currency crisis at the time, the 1995 regime likely corresponds to the Mexican crisis, the 1997 regime likely corresponds to the Asian crisis, the 1998 regime likely corresponds to the Russian crisis, and the 1999 regime likely corresponds to the Brazilian devaluation. It is notable that all of these common high volatility regimes are present even when the country that is the source of a given crisis is not included in a given country pair. Also, it is perhaps surprising that the Argentinean crisis in late 2000 and 2001 does not generally show up as a common high volatility regime, meaning that any ex post dating procedure that includes this crisis in the common regime would distort inferences about shift-contagion. What is most notable for the bond returns is the lack of shift-contagion for Argentina/Brazil, Argentina/Mexico, Argentina/Venezuela, and Brazil/Venezuela. The only

14 For the bond returns, there is a direct correspondence between the point estimates in Table 3 and the likelihood ratio test results in Table 4, suggesting that the test has high power.

Fig. 1. The timing of high volatility regimes for common shocks to European currency returns. The five rows display the filtered probabilities of high volatility common shocks for Germany/Norway, Germany/Switzerland, Norway/Sweden, Norway/Switzerland, and Sweden/Switzerland, respectively.

6. Conclusions

One motivation of the contagion literature is to address how countries can reduce their vulnerability to external shocks during periods of heightened volatility. In this vein, it is important to understand whether a shock is transmitted across markets via channels that only appear during turbulent periods (crisis-contingent channels) or whether they are transmitted via links that exist in all states of the world. For Latin American countries, the empirical results presented in this paper suggest that shocks are generally transmitted via long-term linkages
Acknowledgement

We thank two anonymous referees for helpful comments. Pas-Lin Tien provided excellent research assistance. This paper represents the views of the authors and does not necessarily reflect those of the International Monetary Fund or the Bank of Canada.

References


between those countries, so that attempts at reducing their vulnerability to contagion via short-term or temporary strategies may be ineffective. On the other hand, for many of the developed countries, there is evidence to suggest that some shocks get transmitted only during turbulent periods. This would imply that certain short-term stabilizing policies, such as foreign exchange intervention or tighter monetary policy, could be warranted during periods of heightened volatility and crisis.