Bayesian Cross-Sectional Analysis of the Conditional Distribution of Earnings of Men in the United States, 1967-1996

John Geweke and Michael Keane
Departments of Economics and Statistics, University of Iowa
john-geweke@uiowa.edu
Department of Economics, Yale University
michael.keane@yale.edu

March 31, 2005

Abstract

This study develops practical methods for Bayesian nonparametric inference in regression models. The emphasis is on extending a nonparametric treatment of the regression function to the full conditional distribution. It applies these methods to the relationship of earnings of men in the United States to their age and education over the period 1967 through 1996. Principal findings include increasing returns to both education and experience over this period, rising variance of earnings conditional on age and education, a negatively skewed and leptokurtic conditional distribution of log earnings, and steadily increasing inequality with asymmetric and changing impacts on high- and low-wage earners. These results are insensitive to several alternative nonparametric specifications of distribution of earnings conditional on age and education.

Acknowledgement 1 *Grant R01-HD37060-01 from the National Institutes of Health provided financial support for this work.*

Much of applied statistics and econometrics is concerned with the measurement and interpretation of conditional distributions. In statistics the core curriculum devotes substantial time to regression, a topic that practically defines elementary econometrics and is the main point of departure in advanced treatments. Typically the conditional distribution is that of a univariate random variable y conditional on a random vector $\mathbf{x}$. Regression, narrowly defined, is concerned only with $E(y \mid \mathbf{x})$, yet this topic alone is the basis of a huge literature in mathematical statistics and theoretical econometrics. The strong assumption that regression is linear in $\mathbf{x}$, made in

elementary treatments, is rarely justified on theoretical grounds and as an empirical matter often can be overturned. Therefore much of this literature has taken nonparametric or semiparametric approach to regression. The textbooks by Hardle (1989) and Green and Silverman (1994) provide introductory yet comprehensive treatments of non-Bayesian approaches; for Bayesian treatments, see, for example Erkanli and Gopalan (1994), Wong and Kohn (1996), Smith and Kohn (1996), Koop and Tobias (2004) and Koop and Poirier (2004).

In general, substantive interest in the conditional distribution goes well beyond conditional expectations to embrace the entire distribution $p(y \mid \mathbf{x})$. Well-known examples include the pricing of derivatives of financial assets and the study of inequality and mobility of earnings across individuals. Nonparametric approaches are much less common at this level of generality, due mainly to well-understood practical problems, but even semiparametric treatments of both the regression function and the conditional distribution simultaneously are rare. In this study we take nonparametric and semiparametric Bayesian approaches to this question. The specific statistical question we address is Bayesian inference for a functional of $p(y \mid \mathbf{x})$. This includes not only the expectation of y corresponding to given value of $\mathbf{x}$, but also functionals of the conditional distribution such as the coefficient of kurtosis and measures of inequality like the Gini coefficient. The emphasis is on methods that can be carried out quickly and reliably using software that is widely available. We achieve our objectives by combining, alternatively, polynomial basis functions or Wiener process smoothness priors for regression, with normal mixture modeling for regression residuals $y - E(y \mid \mathbf{x})$. Section 2 describes our methods in detail. The closest precedent for our approach appears to be that of Smith and Kohn (1996). That study, however, is concerned with outliers and confines consideration to scale mixture of normal distributions. The findings in Section 3 show that assumption clearly would be inappropriate in our application.

1 Earnings and the PSID data

Our substantive concern is with the distribution of the earnings of individual men conditional on age and education. To introduce notation used throughout this study, we have available for each of 30 years $t = 1967, \ldots, 1996$ a sample of the age a_{ti}, education e_{ti} and logarithm of earnings y_{ti} for each of $n(t)$ men $(i = 1, \ldots, n(t))$. We are interested in functionals of the density $p(y_{ti} \mid \mathbf{x}_{ti})$, where $\mathbf{x}_{ti} = (a_{ti}, e_{ti})'$. Due to the structure of the data, described in Section 1.2, we treat each year as a separate sample.

1.1 Modelling earnings

There is a long and well established literature studying the relationship between earnings and the determinants of earnings suggested by life-cycle human capital models.

Going back at least to the seminal work of Mincer (1958) the essence of these models is that an individual's productivity, or human capital, is an increasing function of formal education and work experience. Heckman et al. (2003) review this work. By far the most common measure of formal education is years of schooling, and the most common measure of experience is age for individuals who have limited or no spells of labor force nonparticipation following the completion of their formal education. Since our study is limited to men, the latter assumption is reasonable.

A fully specified relationship between earnings, one the one hand, and age and education, on the other, has several properties that are of substantive interest. One is the impact of education on earnings, sometimes termed returns to education. We use that terminology here, recognizing that returns so measured do not distinguish between the impact of education on the earnings of a given individual and the fact that individuals for whom further education is more likely to provide economic benefits are more likely to choose more years of schooling. Another is the distribution of earnings over the life cycle. As individuals age the relative benefits of working and leisure change because of increasing financial wealth (on average), a changing planning horizon, and changing family circumstances.

This relationship need not be constant over time. On the demand side, changes in technology can change the relative productivity of more and less highly educated individuals, and may make it more advantageous for a given number of hours of work over the life cycle to be more concentrated in fewer years or spread out over more years. On the supply side, demographic changes driven by changes in fertility and immigration shift the age distribution of the labor force, and to the extent that older and younger workers are imperfect substitutes in the workforce their relative wages will change for this reason as well. Changes in economic policy including the taxation of earnings, public subsidies for higher education, unemployment benefits, and public pension benefits, will also affect wages and decisions about hours of work, thereby changing the relationship. There is no reason that the effect of these changes should be limited to expectations of earnings (or log earnings) conditional on age and education: the entire distribution may be affected. The extent to which this distribution has changed, and especially the implications for inequality in the distribution of earnings, has been a topic of intense interest in both the academic literature and public policy forums. In this study we characterize the distribution and its change over the period 1967 through 1996.

1.2 Data

Our findings are all based on the panel study of income dynamics (PSID). The PSID is a household-based panel that has collected information on earnings and other aspects of household economic activity. From 1968 through 1997 the survey was fielded annually, collecting data pertaining to the previous year. We use these data, identifying each wave of the panel by the year previous to the interview since that is the year

to which the information pertains. The survey was not fielded in 1998, and data has been conducted biannually since 1999. Death or divorce can lead to one or two new households, and in these cases the PSID tracks the new households. It also follows households formed eventually by children in households, and from time to time the sample is refreshed with new households as required to preserve the stratification of the sample. We restrict our study to male household heads between the ages of 25 and 65, inclusive, who are labor force participants as indicated by earnings of $1,000 or more in a year. Black males and the 1989-1993 Latino source sample are excluded. For each year we assemble data on age, education and earnings of these individuals. We model the distribution of earnings conditional on age and education separately for each year. Thus we do not exploit the panel structure of the data in this study.

Figure 1 conveys some properties of the sample that are important in understanding the findings reported in Section 3. Because the PSID follows households the sample has grown steadily since its inception. Our data from 1994 through 1996 is pre-release and does not include new, young households. Through the period 1967-1996 white male labor force participants became steadily better educated, as a group. There has been a strong tendency to leave the labor force earlier and, since 1976, a tendency to enter later. Conclusions about the distribution of earnings conditional on age and education will, therefore, most strongly reflect the data for men between the ages of 30 and 50 with 12 or more years of education. At the other extreme, if education is less than 8 years, then conclusions will strongly reflect prior information about the similarity of the conditional distribution for poorly educated men to the conditional distribution for well educated men, together with the information in the data about well educated men. These differences are conveyed in greater posterior uncertainty about the distribution of earnings conditional on combinations of age and education that are poorly represented in the sample, as compared with combinations that are well represented.

2 Methodology

We are interested in learning about $p_t(y_{ti} \mid \mathbf{x}_{ti})$. This notation reflects the fact that we model each year separately. We do not exploit the panel features of the data. It is important to bear in mind that earnings are not conditionally independent across years. Beginning with the textbook Bayesian linear model we first weaken the assumption that $E_t(y_{ti} \mid \mathbf{x}_{ti})$ is a linear function of $\mathbf{x}_{ti}$ (Section 2.1) and then weaken the assumption that the conditional distribution of y_{ti} given $\mathbf{x}_{ti}$ is conditionally Gaussian (Section 2.2).

2.1 Nonlinear regression

We begin by weakening the assumption that the regression function is linear in $\mathbf{x}_t$ in favor of the specification that it is a smooth function of $\mathbf{x}_t$, while maintaining

the assumption that $y_{ti} - E_t(y_{ti} \mid \mathbf{x}_{ti})$ is i.i.d. Gaussian. As with all subjective conditions, "smoothness" can be characterized in different ways. This section takes up two different approaches. Each leads to a posterior distribution identical to that of a normal model linear in unknown coefficients, but with a different matrix of covariates and with a different interpretation of the coefficient vector in each case. That nonlinear regression is thus isomorphic to linear regression has two desirable consequences. On the practical level, many Bayesian computational methods for the normal linear model can be applied in normal nonlinear regression. On the conceptual level, many of the rich elaborations of the normal linear model that have been taken up in Bayesian analysis can be applied directly in nonlinear regression. These include the extension to non-normality in Section 2.2.

2.1.1 Polynomial basis functions

A sequence of normal linear models A_j $(j = 1, 2, \ldots)$ for each of the period t models captures the essentials of nonlinear regression with basis functions. In model A_j,

$$y_{ti} = f_{tj}(\mathbf{x}_{ti}) + \varepsilon_{ti} = \sum_{\ell=1}^{k_j} \beta_{tj\ell}\phi_{j\ell}(\mathbf{x}_{ti}) + \varepsilon_{ti} = \beta_{tj}'\phi_j(\mathbf{x}_{ti}) + \varepsilon_{ti};\ \ \varepsilon_{ti} \overset{i.i.d.}{\sim} N\left(0, h^{-1}\right). \quad (1)$$

Model A_j specifies basis functions ϕ_{ji} $(i = 1, \ldots, k_j; j = 1, 2, \ldots)$ so that as j increases, the function f_{tj} is, loosely speaking, more flexible. We take the basis functions to be monomials; other possibilities include Fourier functions (Gallant (1981)) and the Muntz-Szatz series (Barnett and Jonas (1983)). In the *interactive polynomial model* $\phi_{j\ell}(\mathbf{x}_{ti})$ consists of all terms of the form $a_{ti}^m e_{ti}^n$ $(m = 0, \ldots, k_a(j);\ n = 0, \ldots, k_e(j))$. In the *separable polynomial model* $\phi_{j\ell}(\mathbf{x}_{ti})$ consists of all terms of the form a_{ti}^m or e_{ti}^n $(m = 0, \ldots, k_a(j);\ n = 1, \ldots, k_e(j))$. In both cases the relation between j and $(k_a(j), k_e(j))$, for $j = 1, \ldots, 9$ is

	$k_e = 0$	$k_e = 1$	$k_e = 2$
$k_a = 0$	$j = 1$	$j = 3$	$j = 6$
$k_a = 1$	$j = 2$	$j = 4$	$j = 8$
$k_a = 2$	$j = 5$	$j = 7$	$j = 9$

and the pattern continues for larger values of j. The Weirstrass Theorem guarantees that the interactive polynomial model provides an arbitrarily close approximation of $E(y_{ti} \mid \mathbf{x}_{ti})$ for j sufficiently large, but at the cost of a vector of basis functions of high dimension. There is no such justification for the separable polynomial model, but separability is of substantive interest.

In formulating prior distributions of the coefficient vector β_{tj}, it is useful to think in terms of the function f. This is especially important in comparing variants with different numbers of basis functions, because it ensures comparable priors. For example, the prior distribution consisting of the components $f(\mathbf{x}_i^*) \overset{iid}{\sim} N(\mu, \sigma^2)$ for selected

points $\mathbf{x}_i^*$ $(i = 1, \ldots, n)$ implies the prior distribution consisting of the components $\beta_j' \phi_j(\mathbf{x}_i^*) \overset{iid}{\sim} N(\mu, \sigma^2)$ $(i = 1, \ldots, n)$ when the order of expansion is j. If J is the highest order of expansion considered and $n \geq k_J$ points are chosen appropriately, this approach will provide comparable and proper prior distributions for the coefficients in all orders of expansion considered. It also facilitates comparison of polynomial models with models based on smoothness priors described in the next section.

For year t we take the points $\mathbf{x}_i^*$ to be the observed covariates $\mathbf{x}_{ti}$ and choose $\mu = 10.5$, which is close to the sample mean of y_{ti} for all years. Then denoting the matrix of covariates in year t and order of expansion j by $\mathbf{X}_{tj}$, the prior distribution of the $k_j \times 1$ vector β_{tj} is

$$\beta_{tj} \sim N\left[\theta, \lambda n(t) \left(\mathbf{X}_{tj}' \mathbf{X}_{tj}\right)^{-1}\right],$$

where $\theta' = (\mu, 0, \ldots, 0)$ and λ is a hyperparameter to be chosen. The prior distribution of h in (1) is gamma,

$$\underline{s}^2 h \sim \chi^2(\underline{\nu}).$$

Experimentation with the three hyperparameters λ, $\underline{s}^2$ and $\underline{\nu}$ showed that findings are robust over changes of nearly an order of magnitude. Based on marginal likelihoods for several years and several variants of the models, we settled on $\lambda = 200$, $\underline{s}^2 = 12$ and $\underline{\nu} = 1$. Marginal likelihoods were then computed corresponding to all orders of polynomial expansion k_a and k_e for both the interactive polynomial and separable polynomial models. For the thirty separate regressions corresponding to the 30 years in the sample, average marginal likelihood was maximized with the choice $k_a = 4$, $k_e = 1$ in the interactive polynomial model, and the choice $k_a = 4$, $k_e = 2$ in the separable polynomial model. These values are used in all of the results reported subsequently.

2.1.2 Wiener process smoothness priors

There is a complementary approach when the regression function is separable, so that

$$y_{ti} = f_{ta}(a_{ti}) + f_{te}(e_{ti}) + \varepsilon_{ti}; \;\; \varepsilon_{ti} \sim N\left(0, h^{-1}\right). \tag{2}$$

Models of this form have been widely studied in the non-Bayesian semiparametric estimation literature. In the Bayesian literature the same problems have been addressed by smoothness priors. We take a Bayesian smoothness prior approach that emphasizes two properties. The first is that the prior distribution should be capable of allowing the investigator to study the behavior of f_{ta} or f_{te} at points not in the data set: for example, in some years t not all levels of education are represented in our data set as indicated in Section 1.2. The second is that the prior distribution should be formulated in a fashion that assures comparability between models using smoothness priors and models using basis functions, so that Bayes factors are not driven by arbitrary assumptions that are implicit in priors but opaque to the investigator.

The essentials of nonlinear regression with smoothness priors are captured in the simpler model with a single covariate,

$$y_t = f(x_t) + \varepsilon_t, \varepsilon_t \overset{i.i.d.}{\sim} N\left(0, h^{-1}\right) \quad (t = 1, \ldots, T). \tag{3}$$

The function $f(\tau)$ is defined on a closed interval $\tau \in [\tau_1, \tau_2]$. The prior must incorporate the idea that f is a smooth function, in the sense that it is differentiable and $df(\tau)/d\tau$ changes slowly with τ. By formulating a prior that pertains to all points $\tau \in [\tau_1, \tau_2]$ we guarantee coherence if the investigator decides to incorporate a point that is of interest but for which there is no data.

A convenient and powerful analytic tool for expressing these beliefs is the Wiener process $W(\tau)$, defined on $\tau \in [0, \infty)$ with $W(0) = 0$. A standard representation is $W(\tau) = \int_0^\tau dW(u)$, it being understood that the orthogonal increments $dW(u)$ are normally distributed. One important property of a Wiener process is $W(\tau + s) - W(\tau) \sim N(0, s)$, for all $\tau \geq 0$ and all $s > 0$: this limits the rapidity with which W can move as a function of τ, and this feature can in turn be controlled by appropriate scaling of W. Another important property is that any pair of increments $W(\tau + s) - W(\tau)$ and $W(\tau' + s') - W(\tau')$ has a bivariate normal distribution. Each increment has mean zero. If $[\tau, \tau + s]$ and $[\tau', \tau' + s']$ do not overlap then the increments are uncorrelated, whereas if the intervals do overlap then their covariance is the length of the overlap: in general,

$$cov\left[W(\tau + s) - W(\tau),\ W(\tau' + s') - W(\tau')\right] = \int_0^\infty I_{[\tau, \tau+s]}(u)\, I_{[\tau', \tau'+s']}(u)\, du.$$

for all positive τ, τ', s and s'.

A Wiener process has the properties ascribed to the function $f'(\tau) = df(\tau)/d\tau$, and thus we pursue the idea that in the prior distribution $f(\tau)$ is the integral of a such a process. The approach is that of Shiller (1984). Thus, $f(\tau) = \underline{h}^{-1/2} \int_0^\tau W(u)\, du$, where the precision hyperparameter $\underline{h}$ controls smoothness. For any two points τ and s,

$$f(\tau) = \underline{h}^{-1/2} \int_0^\tau W(u)\, du = \int_0^\tau \int_0^u dW(r)\, du, \tag{4}$$

$$f(s) = \underline{h}^{-1/2} \int_0^s W(v)\, dv = \int_0^s \int_0^v dW(p)\, dv, \tag{5}$$

have a joint normal distribution, with $E[f(\tau)] = E[f(s)] = 0$. If $s \geq \tau$ then from (4) and (5),

$$\begin{aligned} E[f(\tau) f(s)] &= \underline{h}^{-1} \int_0^\tau \int_0^s \min(u, v)\, dvdu = \underline{h}^{-1} \int_0^\tau \left[\int_0^u vdv + \int_u^s udv\right] du \\ &= \underline{h}^{-1} \int_0^\tau \left[\frac{u^2}{2} + u(s - u)\right] du = \frac{\underline{h}^{-1/2}\tau^2}{6}(3s - \tau). \end{aligned} \tag{6}$$

Since $var\left[f\left(\tau\right)\right] = \underline{h}^{-1/2}\tau^3/3$, the prior variance ascribed to $f\left(\tau\right)$ at a point $\tau = s_1$ will depend strongly on the idea that $\tau = 0$ is a special point at which it is known *a priori* that $f'\left(0\right) = 0$. This is an artificial assumption. It arises not from prior ideas about smoothness (the reason for introducing the Wiener process as a model for the prior) but rather from the analytical necessity of an initial condition for $f'\left(\tau\right)$. There is a similar problem with the slope of the function $f\left(\tau\right)$ between two points s_1 and s_2 $\left(s_2 > s_1\right)$, $\left[f\left(s_2\right) - f\left(s_1\right)\right]/\left(s_2 - s_1\right)$. From (6),

$$var\left\{\left[f\left(s_2\right) - f\left(s_1\right)\right]/\left(s_2 - s_1\right)\right\} =$$
$$\frac{1}{6\underline{h}}\begin{bmatrix} -\left(s_2 - s_1\right)^{-1} \\ \left(s_2 - s_1\right)^{-1} \end{bmatrix}'\begin{bmatrix} 2s_1^3 & s_1^2\left(3s_2 - s_1\right) \\ s_1^2\left(3s_2 - s_1\right) & 2s_2^3 \end{bmatrix}\begin{bmatrix} -\left(s_2 - s_1\right)^{-1} \\ \left(s_2 - s_1\right)^{-1} \end{bmatrix}$$
$$= \left(4s_1 + 2s_2\right)/6\underline{h} = \left[6s_1 + 2\left(s_2 - s_1\right)\right]/6\underline{h}, \tag{7}$$

which depends not only on the length of the interval $s_2 - s_1$, but also on the size of s_1. However for any three points $s_1 < s_2 < s_3$, we find that for the change in the slope of $f\left(\tau\right)$,

$$var\left[\frac{f\left(s_3\right) - f\left(s_2\right)}{s_3 - s_2} - \frac{f\left(s_2\right) - f\left(s_1\right)}{s_2 - s_1}\right] = \frac{\left(s_3 - s_1\right)}{3\underline{h}}, \tag{8}$$

which can be derived from (6) in the same way that (7) was derived. Thus the distribution of any change in slopes does not depend on the artifice of an initial condition for $f'\left(\tau\right)$. This fact is not surprising, given that $f'\left(\tau\right)$ is a Wiener process, and changes in the level of a Wiener process over an interval depend only on the length of the interval and not on the distance of the interval from the origin $\tau = 0$. Given $s_4 > s_3$,

$$cov\left[\frac{f\left(s_3\right) - f\left(s_2\right)}{s_3 - s_2} - \frac{f\left(s_2\right) - f\left(s_1\right)}{s_2 - s_1},\right.$$
$$\left.\frac{f\left(s_4\right) - f\left(s_3\right)}{s_4 - s_3} - \frac{f\left(s_3\right) - f\left(s_2\right)}{s_3 - s_2}\right] = \frac{\left(s_3 - s_2\right)}{6\underline{h}}. \tag{9}$$

If $s_6 > s_5 > s_4 \geq s_3 > s_2 > s_1$, then

$$cov\left[\frac{f\left(s_3\right) - f\left(s_2\right)}{s_3 - s_2} - \frac{f\left(s_2\right) - f\left(s_1\right)}{s_2 - s_1}, \frac{f\left(s_6\right) - f\left(s_5\right)}{s_6 - s_5} - \frac{f\left(s_5\right) - f\left(s_4\right)}{s_5 - s_4}\right] = 0. \tag{10}$$

Without loss of generality suppose that in a sample of size T there are m distinct values of $x_1, \ldots, x_T$. Denote the ordered distinct values by s_i $\left(i = 1, \ldots, m\right)$, and define $\mathbf{s} = \left(s_1, \ldots, s_m\right)'$ and

$$\beta = \left[f\left(s_1\right), \ldots, f\left(s_m\right)\right]'. \tag{11}$$

Then (3) may be written $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ with $x_{ti} = 1$ if $x_t = s_i$ and $x_{ti} = 0$ otherwise, and $\beta_i = f\left(s_i\right)$. The information in the smoothness prior of the form (8)-(10) may be expressed

$$\mathbf{R}\beta \sim N\left(\mathbf{0}, \mathbf{G}\right). \tag{12}$$

The matrix $\mathbf{R}$ is $(m-2)\times m$, with

$$r_{ii}=(s_{i+1}-s_i)^{-1}, r_{i,i+1}=-\left[(s_{i+1}-s_i)^{-1}+(s_{i+2}-s_{i+1})^{-1}\right],$$
$$r_{i,i+2}=(s_{i+2}-s_{i+1})^{-1} \quad (i=1,\ldots,m-2)$$

and all other elements 0. The matrix $\mathbf{G}$ is $(m-2)\times(m-2)$ with

$$g_{ii}=(s_{i+2}-s_i)/3\underline{h}, \quad g_{i,i+1}=g_{i+1,i}=(s_{i+2}-s_{i+1})/6\underline{h} \quad (i=1,\ldots,m-2)$$

and all other elements 0.

Because (12) provides a distribution of $m-2$ linear combinations of m coefficients, more information is needed for a proper prior distribution for β. To construct a proper prior that is comparable with those used in the previous section, we amend (3) slightly, by writing

$$y_t=\alpha_1+\alpha_2 x_t+f(x_t)+\varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} N\left(0,h^{-1}\right) \quad (t=1,\ldots,T). \tag{13}$$

Then (13) has the form $y=\mathbf{X}_1\alpha+\mathbf{X}_2\beta+\varepsilon$, for the suitably arranged $T\times 2$ matrix $\mathbf{X}_1$ and $T\times m$ matrix $\mathbf{X}_2=\mathbf{X}$. The two restrictions

$$\sum_{i=1}^{m} f(s_i)=0, \quad f(s_1)=f(s_m) \tag{14}$$

identify α_1, α_2 and f in (13) without imposing any additional restrictions. They can be imposed by writing

$$\beta=\mathbf{Q}\beta^*, \text{with } \mathbf{Q}'=\left[\iota_{m-2}(-1/2) \vdots \mathbf{I}_{m-2} \vdots \iota_{m-2}(-1/2)\right]. \tag{15}$$

The restrictions (15) plus the prior information (12) in $\mathbf{y}=\mathbf{X}_1\alpha+\mathbf{X}_2\beta+\varepsilon$ are equivalent to $\beta^*\sim N\left(\mathbf{0},\underline{\mathbf{H}}_2^{-1}\right)$, in $\mathbf{y}=\mathbf{X}_1\alpha+\mathbf{X}_2^*\beta^*+\varepsilon$ where $\mathbf{X}_2^*=\mathbf{X}_2\mathbf{Q}$ and $\underline{\mathbf{H}}_2=\mathbf{Q}'\mathbf{R}'\mathbf{G}^{-1}\mathbf{R}\mathbf{Q}$. If $\alpha\sim N\left(\underline{\alpha},\underline{\mathbf{H}}_1^{-1}\right)$ is independent of β in the prior distribution, then as $\underline{h}\to\infty$ in $\mathbf{G}$, the marginal likelihood of the model must approach that of the linear model $y_t=\alpha_1+\alpha_2 x_t+\varepsilon_t$. By using the same prior distribution for α and h that we used in the special case of the linear model in the polynomial expansions of the previous section, we guarantee comparability of those models with models based on smoothness priors.

Extension from the case of a single covariate (3) to the model at hand (2) is straightforward so long as the smoothness priors for f_{ta} and f_{te} are independent, which we take them to be. The extension may be expressed

$$y_{ti}=\beta_{t1}+\beta_{t2}a_{ti}+\beta_{t3}e_{ti}+f_{ta}(a_{ti})+f_{te}(e_{ti})+\varepsilon_{ti}; \ \ \varepsilon_{ti}\sim N\left(0,h^{-1}\right).$$

The prior distributions of h and $(\beta_{t1},\beta_{t2},\beta_{t3})'$ are the same as those for the particular case $j=3$ of the linear model. The prior distributions of f_{ta} and f_{te} are independent of each other and of these prior distributions, and are given by (12) and (14).

The hyperparameter $\underline{h}^{-1/2}$, which is the prior standard deviation of $f(\tau) - f(\tau - 1)$ controlling smoothness, is set to 0.01 in for f_{ta} and 0.10 for f_{ea}, values that produce large marginal likelihood values relative to most other choices in most of the samples. Experimentation showed that varying these values by a factor of 10 had little effect on the results that we report in Section 3.

2.1.3 Computation

Both of these approaches to nonlinear regression lead to regression models of the form

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \tag{16}$$

$$\varepsilon \sim N\left(\mathbf{0}, h^{-1}\mathbf{I}_T\right), \tag{17}$$

where the matrix of covariates $\mathbf{X}$ is $T \times k$, and independent prior distributions of the form

$$\beta \sim \mathbf{N}\left(\underline{\beta}, \underline{\mathbf{H}}^{-1}\right), \ \underline{s}^2 h \sim \chi^2\left(\underline{\nu}\right). \tag{18}$$

In both cases the number of observations T is the number in the sample for the year to which the regression model is being applied: 1751 in the smallest sample and 2698 in the largest. In the interactive polynomial model $k_a = 4$ and $k_e = 1$, so $k = 10$, and in the separable polynomial model $k_a = 4$ and $k_e = 2$, so $k = 7$. In the separable models with smoothness priors there are 41 distinct values of a_{ti} (ages 25 through 65); the corresponding vector β (see (11)) has 41 elements and β^* (see (15)) has 39 elements. There are 17 distinct values of e_{ti}, so the corresponding vector β^* has 15 elements. Thus $k = 57$ in the separable models with smoothness priors.

In the posterior distribution corresponding to (16)-(18),

$$\beta | \left(h, \mathbf{y}, \mathbf{X}\right) \sim N\left(\overline{\beta}, \overline{\mathbf{H}}^{-1}\right), \text{ with } \overline{\mathbf{H}} = \underline{\mathbf{H}} + h\mathbf{X}'\mathbf{X} \text{ and } \overline{\beta} = \overline{\mathbf{H}}^{-1}\left(\underline{\mathbf{H}}\underline{\beta} + h\mathbf{X}'\mathbf{y}\right);$$

$$\overline{s}^2 h | \left(\beta, \mathbf{y}, \mathbf{X}\right) \sim \chi^2\left(\overline{\nu}\right), \text{ with } \overline{s}^2 = \underline{s}^2 + \left(\mathbf{y} - \mathbf{X}\beta\right)'\left(\mathbf{y} - \mathbf{X}\beta\right) \text{ and } \overline{\nu} = \underline{\nu} + T.$$

The corresponding Gibbs sampling algorithm produces a sequence of draws $\left\{\beta^{(m)}, h^{(m)}\right\}$ from the posterior distribution with little or no detectable serial correlation. Computations were carried out using the Bayesian Analysis, Computation and Communications (BACC) software[1] running under Matlab 6.5. The results reported here are based on 1,000 iterations following 10 warm-up iterations. Computation time is less than one second for the polynomial models and less than five seconds for the separable models with smoothness priors.

2.2 Non-Gaussian conditional distributions

Based on our earlier work (Geweke and Keane (2001)) the assumption of Gaussian disturbances is bound to be poor in this context. In measuring the evolution of

[1]The Bayesian Analysis, Computation and Communication extension for Matlab, Gauss, and Splus; see http://www.cirano.qc.ac/~bacc.

inequality in earnings over the period 1967 through 1996, it is therefore essential to employ a more flexible distribution of earnings conditional on age and education. In measuring changes in returns to education, the age profile of earnings, and the conditional variance of earnings, a more flexible distribution should provide a more reliable indication of uncertainty about these changes.

2.2.1 Mixture of normals distributions

The normal mixture linear model begins with (16) and then introduces the latent state vector $\widetilde{\mathbf{s}} = (\widetilde{s}_1, \ldots, \widetilde{s}_T)'$. Conditional on $\mathbf{X}$, the $\widetilde{s}_t$ are i.i.d. with $P(\widetilde{s}_t = j) = \pi_j$, and thus

$$p(\widetilde{\mathbf{s}} | \mathbf{X}) = \prod_{t=1}^{T} \pi_{\widetilde{s}_t} = \prod_{j=1}^{m} \pi_j^{T_j} \tag{19}$$

where $T_j = \sum_{t=1}^{T} \delta(\widetilde{s}_t, j)$ is the number of observations t for which $\widetilde{s}_t = j$.

Corresponding to each of the m states j there is a mean parameter α_j and a positive precision parameter h_j; let $\alpha = (\alpha_1, \ldots, \alpha_m)'$, $\mathbf{h} = (h_1, \ldots, h_m)'$ and $\pi = (\pi_1, \ldots, \pi_m)'$. Conditional on $\widetilde{s}_t = j$, $\varepsilon_t \sim N\left[\alpha_j, (h \cdot h_j)^{-1}\right]$. Thus

$$\begin{aligned} p[y_t \mid \beta, h, \pi, \alpha, \mathbf{h}, \widetilde{s}_t = j, \mathbf{X}] &\propto (h \cdot h_j)^{1/2} \\ \cdot \exp\left[-h \cdot h_j (y_t - \alpha_j - \beta' \mathbf{x}_t)^2 / 2\right] &\quad (t = 1, \ldots, T). \end{aligned} \tag{20}$$

The disturbances ε_t are i.i.d. and follow a full discrete normal mixture distribution:

$$p(\varepsilon_t \mid h, \pi, \alpha, \mathbf{h}, \mathbf{X}) \propto h^{1/2} \sum_{j=1}^{m} \pi_j h_j^{1/2} \exp\left[-h \cdot h_j (\varepsilon_t - \alpha_j)^2 / 2\right].$$

The mixture of normals distribution is very flexible. Figure 2 provides several examples. For the special case in which the means α_j are all the same the normal mixture distribution is known as the scale mixture of normals distribution. That distribution is symmetric, unimodal, and must be leptokurtic, that is, the coefficient of kurtosis $K = E[\varepsilon_t - E(\varepsilon_t)]^4 / var(\varepsilon_t)^2 > 3$, its value if ε_t is normally distributed. Panels (a) and (f) of Figure 2 provide examples. If the means α_j are not all the same then the normal mixture distribution can be skewed, as illustrated in panels (c) and (d). It can also be platykurtic (i.e., $K < 3$), as is the case in panels (b) and (e). Of course, these distributions can be multimodal (panel (e)). With a sufficient number of components, the normal mixture distribution can mimic distributions that are quite different from the normal, like the uniform (panel (b)).

The conditionally conjugate prior densities in the normal mixture linear model are (18) for β and h. The prior distribution of π is Dirichlet,

$$p(\pi) = \Gamma(mr) \Gamma(r)^{-m} \prod_{j=1}^{m} \pi_j^{r-1}. \tag{21}$$

The components of $\mathbf{h}, \underset{\cdot}{\nu}^2 h_j \overset{i.i.d.}{\sim} \chi^2(\underset{\cdot}{\nu})$ $(j = 1, \ldots, m)$ have independent gamma distributions,

$$p(\mathbf{h}) = 2^{m\underset{\cdot}{\nu}/2}\Gamma(\underset{\cdot}{\nu}/2)^{-m/2}\left(\underset{\cdot}{\nu}^2\right)^{m\underset{\cdot}{\nu}/2}\prod_{j=1}^{m} h_j^{(-\underset{\cdot}{\nu}-2)/2}\exp\left(-\underset{\cdot}{\nu}^2 h_j/2\right). \tag{22}$$

Finally, $\alpha \mid h \sim N\left[\mathbf{0}, (\underline{h}_\alpha \cdot h)^{-1}\mathbf{I}_m\right]$, so that

$$p(\alpha \mid h) = (2\pi)^{-m/2}(\underline{h}_\alpha h)^{m/2}\exp(-\underline{h}_\alpha h\alpha'\alpha/2). \tag{23}$$

These prior distributions are conditionally conjugate, symmetric across states, and require the specification of just three hyperparameters: r, $\underset{\cdot}{\nu}$, and $\underline{h}_\alpha$. The specification $E(\alpha) = \mathbf{0}$ resolves the identification issues with respect to α and β. The prior variance in β conveys uncertainty about the location of the distribution of $\mathbf{y}$ given $\mathbf{X}$. The prior distribution of α is scale dependent on $h^{-1/2}$: that is, it states prior beliefs about the shape of the distribution. Keeping in mind that $E(\mathbf{h}) = \mathbf{e}_m$, a prior distribution with $\underline{h}_\alpha^{-1/2} = 5$ implies a prior probability of multimodality that is near 1, whereas $h_\alpha^{-1/2} = 1/5$ makes this probability negligibly small. Keeping in mind that $E(\alpha) = \mathbf{0}$, choice of $\underset{\cdot}{\nu}$ governs the prior probability of tail thickness in the mixture normal density relative to the normal. In the prior distribution the ratio $h_j/h_k \sim F(\underset{\cdot}{\nu}, \underset{\cdot}{\nu})$ for all $j \neq k$. If $\underset{\cdot}{\nu} = 1$ the prior probability of component variance ratios at least as great as those shown in Figure 2(f) is significant, whereas if $\underset{\cdot}{\nu} = 5$ it is negligible. With these considerations in mind, and after examining the implications of different choices for marginal likelihood in the three regression models and several samples, we settled on the choices $\underline{h}_\alpha^{-1/2} = 1.58$, $\underset{\cdot}{\nu} = 0.2$, and $r = 1$.

The states are unidentified in this model: that is, a relabelling that interchanges two or more states leaves the posterior distribution unaffected. In our application this lack of identification is harmless, because the states are simply a modeling device to provide a flexible conditional distribution and have no independent substantive interpretation. None of the questions of interest depend in any way on identification of the states. Moreover, leaving the states unidentified provides some advantages in computation.

2.2.2 Computation

Posterior inference in the normal mixture model utilizes five blocks: (α, β), h, π, $\mathbf{h}$, and $\widetilde{\mathbf{s}}$. It is useful to define

$$\underset{T\times m}{\widetilde{\mathbf{Z}}}(\widetilde{\mathbf{s}}) = \widetilde{\mathbf{Z}} = [\widetilde{\mathbf{z}}_1, \ldots, \widetilde{\mathbf{z}}_T]' = [\delta(\widetilde{s}_t, j)], \quad \underset{T\times(m+k)}{\widetilde{\mathbf{W}}} = \begin{bmatrix} \widetilde{\mathbf{Z}} & \mathbf{X} \end{bmatrix},$$

$$\underset{(m+k)\times 1}{\underline{\gamma}} = \begin{pmatrix} \underline{\mathbf{0}} \\ \underline{\beta} \end{pmatrix}, \quad \underset{(m+k)\times(m+k)}{\underline{\mathbf{H}}_\gamma(h)} = \underline{\mathbf{H}}_\gamma = \begin{bmatrix} \underline{h}_\alpha h\mathbf{I}_m & \mathbf{0} \\ \mathbf{0} & \underline{\mathbf{H}}_\beta \end{bmatrix},$$

and $\underset{T\times T}{\widetilde{\mathbf{Q}}}\left(\widetilde{\mathbf{s}}\right)=\widetilde{\mathbf{Q}}=diag\left(h_{\widetilde{s}_1},\ldots,h_{\widetilde{s}_T}\right)$. With this notation (20) is equivalent to

$$p\left(\mathbf{y}\mid\gamma,h,\pi,\mathbf{h},\widetilde{\mathbf{s}},\mathbf{X}\right)\propto h^{T/2}\left|\widetilde{\mathbf{Q}}\right|^{1/2} \tag{24}$$

$$\cdot\exp\left[-h\left(\mathbf{y}-\widetilde{\mathbf{W}}\gamma\right)'\widetilde{\mathbf{Q}}\left(\mathbf{y}-\widetilde{\mathbf{W}}\gamma\right)/2\right]. \tag{25}$$

The kernel of the conditional posterior density of γ is the product of (18), (23), and (25), from which the conditional posterior distribution is

$$\gamma\sim\mathbf{N}\left(\overline{\gamma},\overline{\mathbf{H}}_\gamma\right);\ \ \overline{\mathbf{H}}_\gamma=\underline{\mathbf{H}}_\gamma+h\widetilde{\mathbf{W}}'\widetilde{\mathbf{Q}}\widetilde{\mathbf{W}},\ \overline{\gamma}=\overline{\mathbf{H}}_\gamma^{-1}\left[\underline{\mathbf{H}}_\gamma\underline{\gamma}++h\widetilde{\mathbf{W}}'\widetilde{\mathbf{Q}}\widetilde{\mathbf{y}}\right]. \tag{26}$$

The conditional posterior density of h is the product of (18), (23) and (20). This kernel corresponds to the conditional posterior distribution

$$\left[\underline{s}^2+\underline{h}_\alpha\alpha'\alpha+\sum_{t=1}^{T}h_{(t)}\left(y_t-\alpha'\widetilde{\mathbf{z}}_t-\beta'\mathbf{x}_t\right)^2\right]h\sim\chi^2\left(\underline{\nu}+m+T\right). \tag{27}$$

The conditional posterior density kernel of π is the product of (19) and (21), $\prod_{j=1}^{m}p_j^{r+T_j-1}$, and thus the conditional posterior distribution is Dirichlet with parameters $r+T_j$ $(j=1,\ldots,m)$. The conditional posterior density kernel of $\mathbf{h}$ is the product of (22) and (20), which implies

$$\left[\underline{s}_\cdot^2+\sum_{t=1}^{T}\delta\left(s_t,j\right)\left(y_t-a_j-\beta'\mathbf{x}_t\right)^2\right]h_j\sim\chi^2\left(\underline{\nu}_\cdot+T_j\right)\quad(j=1,\ldots,m).$$

The conditional posterior density kernel for the state assignments $\widetilde{\mathbf{s}}$ is the product of (19) and (20) taken over $t=1,\ldots,T$. Thus the states $\widetilde{s}_t$ are conditionally independent, with

$$P\left(\widetilde{s}_t=j\right)\propto p_jh_j\exp\left[-h\cdot h_j\left(y_t-a_j-\beta'\mathbf{x}_t\right)^2/2\right]\quad(j=1,\ldots,m). \tag{28}$$

Draws from these multinomial distributions are straightforward.

Computations follow a Gibbs sampling algorithm based on these five blocks. Since the states are left unidentified, the Markov chain is characterized by label switching. We observed this regularly with mixtures of three components and occasionally in models with two components, using 20,000 iterations in each case. The results presented in the next section are all based on every tenth value of 20,000 iterations following 1,000 warm-up iterations. Figure 3 illustrates MCMC output using the 1986 data, the two component mixture model, and two of the regression functions. The illustration labels as state 1 the state with the smaller probability: the top panels

display π_1, the middle panels show $\alpha_2 - \alpha_1$, and the lower panels $(h_1/h_2)^{1/2}$. The MCMC output is nearly serially uncorrelated in all cases: relative numerical efficiencies are between 0.65 and 0.75 for the polynomial models (left-hand panels), and between 0.25 and 0.50 for the separable smooth function models. All computations were carried out using BACC running under Matlab 6.5, and require about 4 minutes for each polynomial model and 10 minutes for each separable smooth function model.

3 Findings

This section presents the evidence on the substantive questions of interest motivating our application. We begin by summarizing the relative performance of the alternative approaches to nonlinear regression and the distribution of regression residuals (Section 3.1). Next we take up returns to education and the age profile of earnings, which are functionals of the regression (Section 3.3). Lastly we turn to measures of inequality in the distribution of earnings, which are functionals of the entire conditional distribution (Section 3.4). The principal vehicles for conveying the results are conditional medians and interquartile ranges for these functionals for each of the thirty years in the sample. Tables in an appendix provide corresponding posterior means and standard deviations.

3.1 Model comparison

The conclusion of Section 2.1.1 describes the selection of orders of polynomial expansion and prior hyperparameters in the basis function approach to regression, the conclusion of Section 2.1.2 indicates how the smoothness hyperparameters were chosen in the Wiener process prior approach, and the end of Section 2.2.1 does the same for the three prior distribution hyperparameters for the normal mixture models of regression residuals. Given these choices, we now compare these approaches to modeling $p(y \mid \mathbf{x})$ using Bayes factors.

Figure 4 provides histograms for the posterior probabilities across the three alternative specifications of the regression function assuming a Gaussian distribution for disturbances, in the three left panels. It does the same assuming a mixture of two normal distributions in the three right panels. In both cases interactive polynomials are favored in the early samples (through about 1980) but not thereafter. In the later samples (after about 1980) the models with Wiener process priors are more highly favored if one assumes Gaussian residuals. Given a mixture of normals specification for the residuals, the separable polynomial specification is favored more often than the Wiener process priors after about 1980. Perhaps more important than any of these results, however, is the fact that no one model captures a disproportionate share of the posterior probability in most years – note that the figures convey small positive probabilities for quite a few models and years.

Appendix Table 1 provides log Bayes factors for the models, all taken relative to an interactive polynomial model of order $k_a = 3$ in age and $k_e = 1$ in education, with Gaussian residuals. (We take this as a benchmark because it is the closest to a consensus of earnings model specifications in the literature.) The most important information conveyed by this table is that models with mixtures of two normal distributions completely dominate the Gaussian models. For any combination of specifications of the regression function the *log* Bayes factor in favor of a mixture of two normals model versus a Gaussian model is never less than 170 and exceeds 300 in some cases. Comparison of this table with Figure 1 shows that the log Bayes factors comparing normal mixture and Gaussian models are roughly proportional to sample size, as might be expected. Bayes factors comparing mixtures of two normals with mixtures of three normals (not presented) weakly favor a mixture of two normals. The reason for this will become apparent when we inspect conditional distributions in detail in Section 3.4.

Because no one mixture of normals model dominates in many of the years, it is necessary to average across models for those functionals whose posterior distribution is sensitive to the specification of the regression function. In Sections 3.3 and 3.4 we report results for all three specifications, and it turns out that these distributions are insensitive to specification. In the case of conditional means and variances, we report findings for all six specifications (the product of three for regressions and two for residual specification) in order to assess the sensitivity of the posterior distribution to the incorrect assumption of normality of residuals.

3.2 Model evaluation

Bayes factors provide relative comparisons of alternative models, but they do not reveal the adequacy or inadequacy of these models in describing interesting aspects of the data. We address this question through the structured but less formal method of posterior predictive analysis. To describe this method succinctly, denote the parameters of the model by θ, the observables in the sample by $\mathbf{y}$, and the entire observed data set by $\mathbf{y}^o$. Let $g(\mathbf{y})$ be an interesting scalar feature that can be observed in the data, such as a sample moment or quantile. Then the posterior distribution of θ induces a distribution on $g(\mathbf{y})$ by means of the model density $p(\mathbf{y} \mid \theta)$: it is the predictive distribution for $g(\mathbf{y})$ corresponding to a hypothetical repetition of the experiment of drawing a sample of the same size from the same population. A finding that with extremely high probability $g(\mathbf{y})$ would exceed or fall short of the observed $g(\mathbf{y}^o)$ in this hypothetical experiment casts doubt on the specification of the model. This idea goes back to the notion of surprise discussed by Good (1953) and its essentials were further developed by Rubin (1984) in what he termed "model monitoring by posterior predictive checks." The mechanics of model evaluation entail generating an artificial sample $\mathbf{y}^{(m)}$ corresponding to each draw $\theta^{(m)}$ from the posterior simulator and computing $g\left(\mathbf{y}^{(m)}\right)$ $(m = 1, \ldots, M)$, and then finding the position of $g(\mathbf{y}^o)$

relative to the empricial c.d.f., which can be expressed

$$p_g^* = M^{-1} \sum_{m=1}^{M} I_{(-\infty, g(\mathbf{y}^o)]} \left[g\left(\mathbf{y}^{(m)}\right) \right].$$

If p_g^* is close to zero, then the posterior distribution overpredicts $g(\mathbf{y})$; if it is close to one, then the posterior distribution underpredicts. For further details see Lancaster (2004, Section 2.5) or Geweke (2005, Section 8.3).

We evaluate the Gaussian and mixture models using six different features of the data $g(\mathbf{y})$. We include the Gaussian models in spite of their inferior performance in the model comparison exercise because we wish to see if the mixture models perform at least as well as the Gaussing models with respect to all these features. Figures 5 through 10 provide the results of this analysis. Each figure corresponds to a different function g, to be described shortly. The six panels in each figure correspond to the three Gaussian and three mixture models, just as in Figure 4. The value p_g^* was computed for each of the 30 samples corresponding to the years 1967-1996, and each panel provides a histogram of the p_g^* values over the 30 samples. Values of p_g^* are sorted into five bins of equal size, except that if $g(\mathbf{y}^o) < g\left(\mathbf{y}^{(m)}\right)$ for all iterations, then p_g^* is assigned to a bin just below zero (see, for example, Figure 9) and if $g(\mathbf{y}^o) > g\left(\mathbf{y}^{(m)}\right)$ for all iterations, then p_g^* is assigned to a bin just above one (see, for example, Figure 10); all results are based on $M = 2,000$ iterations. Thus if a histogram is concentrated to the left, the posterior distribution tends to overpredict the observed feature of the data, and if concentrated to the right tends to underpredict. Appendix Tables 2 through 7 provide the values p_g^* for the individual years.

Figures 5, 6 and 7 provide evidence on the ability of the models to describe systematic differences in earnings by age and education. In Figure 5 the function g is the difference in average log earnings between men with 16 years of education and men with 12 years of education, in each of the 30 sample years. In Figure 6 it is this difference computed for men age 45 and age 25, and in Figure 7 it is for men age 60 and men age 45. There is a mild tendency for all models to underpredict the observed returns to college education (Figure 5). The tendency is slightly more pronounced for mixture models than for Gaussian models. The difficulty is most pronounced in the mixture models with interactive polynomials, in which the sample returns to education fell below the posterior median in 29 out of 30 years, and in one year (1977) $p_g^* = 1.0$. All models do better in capturing average log earnings by age: Figures 6 and 7 show little, if any, tendency for models to overpredict or underpredict. While the difficulties in capturing average log earnings between men with 16 years of education and men with 12 years of education in the sample are not extreme, we find them somewhat puzzling in view of the fact that the evidence supports a very nearly linear relationship between education and log earnings in the regression function. The fact that there are many more men with exactly 16 years and exactly 12 years of education than there are men of any one age (an in particular,

ages 25, 45 and 60) means that there is less sampling variation in observed returns to education, and this may contribute to the evident mild difficulties in fitting these returns.

Figures 8, 9 and 10 show how well each of the models describes aspects of the distribution of log earnings conditional on age and education. In the case of Figure 8, the function $g(\mathbf{y})$ was computed by finding least squares estimates of the regression function (e.g., a polynomial of order 4 in age and 2 in education in the case of the separable polynomial models), and then computing the usual least squares estimate of the standard deviation of the disturbance. In Figure 9 the procedure was the same except that the final object $g(\mathbf{y})$ is the coefficient of skewness of the least squares residuals, and in Figure 10 it is the coefficient of kurtosis of these residuals. The first of these figures stands in marked contrast to the latter two. For the Gaussian models, the sample conditional standard deviation $g(\mathbf{y}^o)$ in Figure 8 is always close to the median of the distribution. This reflects the well-known fact that a normal linear model recovers the population linear projection and the variance about that projection whether the model is correctly specified or not; see Geweke (2005), Example 3.4.3. For similar reasons $g(\mathbf{y}^o)$ is also near the median in the mixture models. Figure 9 dramatically demonstrates the failure of the Gaussian models to describe the negative skewness in the distribution of log earnings conditional on age, and Figure 10 does the same for the leptokurtosis in this distribution. The mixture models account for these features rather well, although there is a mild tendency for the posterior distribution to underpredict the degree of conditional kurtosis observed in the sample.

Posterior predictive analysis can be conducted with other interesting functions of the observables $g(\mathbf{y})$. Appendix Tables 8, 9 and 10 report results related to aspects of conditional distributions studied below in Section 3.4.2. In many of these cases the failure of the normal models was as dramatic as that in Figures 9 and 10, and in all cases the mixture models had stronger posterior predictive performance.

3.3 Conditional means and variances

Each of the six models provides the functionals $E(y \mid \mathbf{x})$ and $var(y \mid \mathbf{x})$; the former depends on $\mathbf{x}$ but the latter does not. Figure 11 shows posterior medians and interquartile ranges for $f_1(a) = E_{1986}(y \mid a, e = 12)$ and Figure 12 does the same for $f_2(e) = E_{1986}(y \mid a = 40, e)$. A change in the conditioning $e = 12$ or $a = 40$ produces only level shifts in the medians and interquartile ranges for the separable polynomial and Wiener process prior specifications, whereas for the interactive polynomial regressions both the shapes of the curves and the sizes of the interquartile ranges will be affected. The values of $f_1(a)$ for different values of a are highly dependent, via the polynomial restriction for the first two specifications and the Wiener process prior for the third. The posterior medians and interquartile ranges in Figure 11, and the posterior means and standard deviations in Appendix Table 11, pertain to the

marginal posterior distribution of $f_1(a)$ at each value of a, and therefore do not provide an indication of uncertainty about the regression function as a whole. A similar precaution is in order for the interpretation of Figure 12 and Appendix Table 12.

The posterior medians and means of $f_1(a)$ are remarkably insensitive to the specification of the model. Differences in posterior means are always substantially less than corresponding posterior standard deviations, and differences in posterior medians are less than the corresponding interquartile ranges. The same cannot be said of $f_2(e)$. We observe a similar congruence across specifications for those values of e on which the sample is concentrated, $e = 12$ and above (recall Figure 1). For lower values of e discrepancies markedly increase, especially for the interactive polynomial model in which $f_2(e)$ is linear. The models have substantially different implications for $f_2(e)$, $e \leq 4$, but because of the dearth of data for such low levels of education different implications have very little impact on the posterior distribution: these are cases of almost pure extrapolation.

The sizes of the posterior interquartile ranges and standard deviations tell a more complex, but quite informative, story. Consider first the case of $f_2(e)$, which is somewhat simpler. For all specifications the posterior interquartile range and standard deviation increase as e decreases. This is due to the lack of observations on poorly educated men, and is familiar from elementary econometrics. For any given specification of the regression function and value of a, the interquartile ranges and posterior standard deviations of $f_2(e)$ are smaller for most mixture of normal specifications of the disturbance than for the Gaussian specification. (This feature is somewhat more transparent in Appendix Table 12 than it is in Figure 12.) This is due to the fact that under the mixture of normals specification the conditional distribution is substantially leptokurtic, as described below in Section 3.4.1. The Gaussian and mixture of normals specification lead to similar inferences about conditional variance, a fact familiar from the analysis of misspecification of regression residual distributions in linear regression and demonstrated below in Section 3.3.3 for this application. In this situation the mixture of normals distribution is more informative about the regression function than is the Gaussian distribution due to its higher concentration of residuals near the mean of the distribution, illustrated below in Section 3.4.1 for this application. This is perhaps most easily seen in the limiting case in which kurtosis increases without bound while variance remains constant: Chebychev's inequality implies that the probability density function of the residuals collapses about the point zero, and as this limit is approached the posterior distribution conveys the value of $E(y \mid \mathbf{x})$ with certainty given a fixed specification of the model and prior.

Interquartile ranges and posterior standard deviations for $f_1(a)$ increase as a increases, due to the fact that sample size is a decreasing function of age in 1986. For the same specification of the regression function, the ratio of the posterior standard deviation in the mixture of normals specification to that in the Gaussian specification is a monotone increasing function of age, as may be confirmed from Appendix Table 11. The same is true of the interquartile range, although this admittedly requires a

very keen eye inspecting Figure 11. This can be traced to the fact that in the mixture of normals model, draws of s_t from (28) occur more often from the high-variance state than from the low-variance state for older men relative to the frequency for younger men. This reflects a misspecification of the model, which implies that the distribution of these states should be the same at all ages. We return to this point when discussing directions for future research in the concluding section.

3.3.1 Returns to education

Changes in education imply changes in expected log earnings. We focus here on the difference in expected log earnings between college graduates ($e = 16$) and high school graduates ($e = 12$). In the separable polynomial and Wiener process prior models these differences do not depend on age. For the interactive polynomial regression specification we condition on age 40, at which point expected log earnings approach their peak over the life cycle. Figure 13 provides the posterior medians, upper and lower quartiles of the posterior distribution of this functional for all six models and each of the 30 data sets for the years $1967, \ldots, 1996$, and Appendix Table 13 indicates the corresponding posterior means and standard deviations. Recall that posterior distributions are constructed separately for each year, and there is no functional or prior dependence between years in Figure 13 (or in any of the succeeding figures) as there was between years of education in Figure 11 or ages in Figure 12. The results are not independent across years, since that data are taken from a panel. While our models do not address this dependence, regarding results as independent across years is probably a good approximation in interpreting these figures and similar figures that follow.

Posterior means and medians for returns to education are sensitive to the specification of the regression function, and almost completely insensitive to the specification of the distribution of the regression residuals. However, there is substantially less posterior uncertainty about returns in the mixture of normals models, for the reasons just discussed in considering $f_2(e)$: posterior standard deviations and interquartile ranges are about 25% smaller. Across regression specifications and for the same specification of the residual distribution, models that have the highest posterior probability (Figure 4) tend to provide the smallest posterior standard deviations and interquartile ranges.

The separable regression functions indicate stagnating returns to college education early in the sample, declining from about 0.35 in 1967 to about 0.25 in 1979. By 1990 returns have nearly doubled, to about 0.48, where they remain through 1996. In the context of the mixture of normals specification the early downward movement amounts to about four posterior standard deviations, and the later increase is the equivalent of about eight posterior standard deviations. Conditional on these models, there is substantial confidence in the "U" shape of returns to college education over the period 1967-1996. Given the interactive polynomial specification, returns to college

change substantially less over the period. They remain in the range 0.30 to 0.35, which has roughly the same length as the interquartile range, through 1979. They then increase to about 0.48, but not until the early 1990's. All models identify 1979 through 1996 as a period in which returns to college education increased substantially, with an upward shift on the order of 0.20.

Recall that the interactive polynomial specification was favored in many of the years prior to 1980 (Figure 4), although the log-odds ratio never exceeds about 15 (Appendix Table 1). In the interactive polynomial models, returns to education can depend on age. Figure 14 and Appendix Table 14 show that in these models, returns to college education are lower for younger men than for older men, in roughly the first half of the sample. Through about 1985, returns to college are between 0.1 and 0.2 for men age 25, as opposed to the substantially higher levels for men age 40 or 60. After 1985 returns to college are similar for men of all ages, consistent with the Bayes factors in favor of the separable models for those years. In all specifications the sample provides considerably more information about returns to college at age 40 than at age 25 or 60. This is consistent with the distribution of age and education in the sample (Figure 1). In particular, in the last years of the sample there are very few very young men, and this is reflected in large interquartile ranges in Figure 14.

3.3.2 Age profile of earnings

The hump displayed in the age profile of earnings for 1986 in Figure 11 is characteristic of all years in the sample, but the shape changes markedly over the period 1967-1996. We measure the change by considering two functionals of $E(y \mid \mathbf{x})$,

$$E(y \mid a = 45, e = 12) - E(y \mid a = 25, e = 12)$$

and

$$E(y \mid a = 60, e = 12) - E(y \mid a = 45, e = 12).$$

The first reflects mainly returns to experience, since virtually all men are labor market participants between the ages of 25 and 45, and most of them work full-time during almost all of these years. The second reflects mainly changes in labor market participation that accelerate between the ages of 45 and 60. Conditioning on 12 years of education matters only in the interactive polynomial models of regression.

As was the case with returns to education, posterior means and medians for the age profiles shown in Figures 15 and 17 and Appendix Tables 15 and 17 depend on the model for the regression function, but are insensitive to whether the specification of the regression residual is Gaussian or a mixture of normals. On the other hand, posterior interquartile ranges and standard deviations are smaller in the mixture models, for reasons discussed above. Across the years changes in the size of posterior interquartile ranges and standard deviations are due to changes in the size of the sample. In the case of returns to experience between the ages of 25 and 45 these sizes increase substantially in 1994, 1995 and 1996, due to the dearth of 25 to 30 year olds

in the samples for those years, above and beyond the decline in overall sample size indicated in Figure 1.

Turning to Figure 15 and Appendix Table 15, all models indicate that returns to experience rose between 1967 and 1975, but the amount depends on the model: the increase is about 0.15 for the separable regression functions but only about 0.06 in the interactive polynomial regression function. From 1975 through 1991 return to the 20 years of experience between the ages of 25 and 45 remains about 0.40, in all of the models. Over the period 1991 to 1996, the polynomial regression functions show a sharp increase in these returns, reaching about 0.55 by 1996. The smoothed separable models, on the other hand, indicate only a modest increase, from about 0.40 to about 0.43. We believe this difference can be ascribed to the thinning of the sample of very young men in the later years, which implies that expected earnings for 25-year-olds in these samples is largely extrapolation. Since the polynomial regression functions and the Wiener process prior approach extrapolation in fundamentally different ways, the different conclusions are not surprising. This difference in conclusions should be resolved when we update the pre-release sample with the full data.

Figure 16 and Appendix Table 16 indicate how returns to experience varies by level of education in the interactive polynomial models. Overall, the expected difference in log earnings between ages 45 and 25 is greater, the higher the level of education, but this difference steadily narrowed throughout the thirty-year period and practically vanished by 1996. Through 1985, the difference for poorly educated men is only about half what it is for high school graduates. After 1985 there is evidence that returns to experience for these men rises to returns for high school graduates, but by the 1990's the point is moot since there are almost no poorly educated young men in the sample. This accounts for the huge interquartile ranges in the top two panels of Figure 16 in the 1990's. College graduates average about 15% greater growth in earnings than do high school graduates, through about 1985. After that, earnings growth drops for the college graduates and increases for the high school graduates. The scarcity of young men in the last years of the sample makes it hard to assess returns to experience in the 1990's.

Figure 17 and Appendix Table 17 exhibit findings about the difference in expected log earnings at age 60 and age 45. As expected, the difference is always negative. All of the models indicate that the algebraic difference increased by about .10 to .15 between the late 1960's and 1988, after which it declined sharply to the values of the late 1960's. These changes are all modest absolutely and in comparison with changes over the 30-year period found for returns to education and to experience between ages 25 and 45. A striking feature of these results is that the difference in expected earnings between men of ages 60 and 45 is systematically more negative in the Gaussian models than it is in the mixture models. We believe that this can be traced to the evident misspecification of the assumption that state probabilities are independent of age. Older men are classified more frequently into the less probable state. As detailed in Section 3.4.1, the less probable normal distribution has a lower

mean and a higher standard deviation than the more probable state. With more men classified into the less probable state at age 60 than at age 45, some of the decline in expected log earnings between ages 45 and 60 is accounted for by the high probability of a normal mixture component with a lower mean at age 60 than at age 45, leaving less to be explained by the regression function. We suggest ways that future research can address this issue in the concluding section.

In the interactive polynomial models, the expected difference in log earnings between ages 60 and 45 can depend on the level of education. Figure 18 and Appendix Table 18 show that education does not have quite as strong an impact on expected earnings growth between the ages of 45 and 60 as it does between the ages of 25 and 45, but this impact changed in a systematic way between 1967 and 1996. Early in the sample, expected log earnings fell most rapidly between the ages of 45 and 60 for the most poorly educated men. For example, in 1970 the difference is about -0.30 for men with 8 years of education, -0.15 for high school graduates, and about zero for college graduates. In 1977 the pattern reversed, and the reversal became stronger over time. By the 1990's, men with 8 years of education have almost no change in earnings between the ages of 45 and 60, whereas expected log earnings decline by about 0.10 for high school graduates and 0.20 for college graduates. It is important to keep in mind that in the latter years, there is considerable uncertainty about the impact of education on the age profile of earnings, as indicated by the Bayes factors summarized in Figure 4. We conjecture these changes over the years, like many of the others discussed in this Section, can be assessed more accurately in a longitudinal model in which the entire sample can be studied at once.

3.3.3 Conditional variances

The standard deviation of log earnings conditional on age and education increased sharply over the period, from about 0.6 in the late 1960's to about 0.75 in the early 1990's: see Figure 19 and Appendix Table 19. In the misspecified Gaussian models the standard deviation is still the standard deviation of the actual distribution of regression residuals, so it is not surprising that the posterior means and medians of the standard deviation of this distribution are about the same: those for the mixture models are on the order of 1% to 2% higher than those in the Gaussian models. Differences across the specification of the regression function are even smaller.

Not surprisingly, posterior interquartile ranges and standard deviations of the regression residual standard deviations are substantially larger for the normal mixture model than for the Gaussian model, in most years by about a factor of two. Nevertheless, the change in the standard deviation of the conditional distribution over the 1967-1996 period amounts to about six normal mixture model posterior standard deviations.

3.4 Conditional distributions

As discussed in Section 2.2.1, the normal mixture models accommodate a rich variety of distributions. With separate models for each of the 30 years in the sample, we can track changes in the conditional distribution in the same way that changes in the regression function were tracked in Section 3.3. This can be done through a wide variety of functionals. We concentrate here on functionals familiar from distribution theory (Section 3.4.1) and those related to the measurement of inequality (Section 3.4.2).

First, however, we present some evidence on the distribution of log earnings regression residuals. Each panel of Figure 20 provides the posterior mean of a mixture of normals density (heavier line) together with the posterior mean of a normal density with the same mean (zero) and standard deviation (about 0.68) as the mixture of normals density. The posterior means are taken with respect to the 1986 sample. Comparing top, middle and bottom panels, mixture of normals densities are insensitive to the specification of the regression function. Comparing the left and right panels, the mixture of normals density with three components is very similar to that with two components. A mixture of three normals requires three more parameters than a mixture of two normals. Since both distributions provide about the same fidelity to the data, marginal likelihoods favor a mixture of two normals over a mixture of three normals. This happens for all years, and hence we conduct all further analysis using the two-component mixture models. Given the similarity of the conditional mixture of normals densities across regression functions, all further results in the paper pertain to the Wiener process prior separable regression function. Detailed results for other regression functions may be found in Appendix Tables 20 through 25.

These mixture of normal distributions are strongly negatively skewed: the mode is almost one-half standard deviation larger than the mean and the left tail is much thicker than the right tail. The excess kurtosis of the mixture of normals distribution is evident in the fact that the mode of its density is substantially higher than the mode of the corresponding normal density. Because the skewness coefficient is negative, the thicker left tail of the normal mixture density is evident in Figure 20 whereas the thicker right tail is not.

Figure 21 exhibits the two-component normal mixture probability density function posterior means for six years evenly spaced through the sample. The increasing standard deviation, documented in Section 3.3.3, is evident. The shape as well as the scale the distribution changes, a feature that we document more closely in the next subsection.

3.4.1 The evolution of the conditional distribution of earnings

There are many ways of summarizing the shape of a probability density. Two of the most familiar are skewness and kurtosis. The top panels of Figure 22, and Appendix

Table 20, show how these measures have changed for the distribution of log earnings conditional on age and education over the sample. Skewness in the distribution has diminished, moving from about -1.25 in the late 1960's to around -1.15 in the 1970's to about -1 by 1990. Thus the difference between the left and right tail of the distribution has diminished; but, of course, the tails have also expanded with the growth in standard deviation, and we return to the implications of both movements for low earnings in Section 3.4.2. Kurtosis declined sharply through the mid-1980's, after which it rose again, returning to 1970 levels by 1996. We conclude that the shape of the distribution has evolved slowly and systematically, even as standard deviation was increasing. However these changes are small relative to the absolute magnitude of skewness and excess kurtosis.

The shape may also be summarized in terms of the components of the normal mixture model. In all years, the component of the mixture model with the higher probability has a positive mean and a smaller standard deviation; the component with the lower probability has a negative mean and a larger standard deviation. The ratio of the larger to the smaller standard deviation declined until about 1980 and then rose again beginning about 1990. The pattern is similar to the movement in kurtosis, and accounts for most of its change. The spread in the means of the components was about the same in the first and last years in the sample, but declined somewhat in the years before 1983 and rose again thereafter. The fact that the difference in the component means was about the same in the early 1990's as in the late 1960's, together with the increase in the standard deviation, accounts for movement in the skewness coefficient toward zero over the thirty-year period. The probabilities of the mixture components change somewhat over the period: the component with the negative mean and larger standard deviation has probability roughly 0.2 in the earlier and later years, but is between 0.25 and 0.3 during the 1980's.

The lower right panel of Figure 22 tracks the proportion of the probability density that is within one standard deviation of the mean, in each year. (See also Appendix Table 22.) For a Gaussian density this proportion is 0.68. Note that the change in this probability closely follows the change in kurtosis shown in the top right panel: for the same standard deviation, the more leptokurtic the distribution, the more probability must be concentrated near the mean in order to preserve the standard deviation. Since 1976, changes in this probability have been small.

3.4.2 The evolution of inequality in earnings

Inequality in the distribution of earnings, or of any other non-negative variable associated with a population of individuals, can be measured in a variety of ways. Perhaps the most familiar is the Gini coefficient, which derives from the Lorenz curve. The Lorenz curve $L(p)$ is defined on the interval $[0, 1]$ and is the fraction of total earnings accruing to individuals in earnings quantile p or lower. If all individuals have the same earnings then $L(p) = p$ and in general $L(p) < p$. The Gini coefficient is

$G = 2\left[1 - \int_0^1 L(p)\right] dp$; $G \in [0,1]$, with $G = 1$ if and only if all individuals have the same earnings and $G = 0$ if and only if all earnings accrue to one individual. We consider two other measures of inequality: P, the fraction of men with earnings less than one-half of median earnings, and R, the fraction of earnings accruing to men in the top decile of the earnings distribution.

These measures depend on the population, and we consider two such populations. The first is a hypothetical population of men of the same age and education, which leads to measures of inequality conditional on age and education. Given the specifications of our models, this measure of inequality will be the same for all ages and education, but it changes from year to year. For example, it will be affected by the change in the conditional standard deviation conveyed in Figure 19, but it will not be affected by the increasing return to education identified in Figure 13. The second population we consider is the distribution of age and education in our sample for 1986, selected only as a convenient benchmark, and leads to measures of inequality unconditional on age and education. Measures of inequality for this population will be affected by changes in the regression of log earnings on age and education, such as the increasing return to education. We expect unconditional measures of inequality of earnings to be greater than conditional measure of inequality. (If inequality were measured by variance, this would be a consequence of the Rao-Blackwell Theorem.) It is important to keep in mind that all of the changes considered here condition on a fixed distribution of age and education, and abstract from the important demographic changes that are evident in Figure 1.

Figure 23 shows the evolution of each measure of inequality over the 1967-1996 period; unconditional measures are in the left panels, and conditional measure are in the right panels. Each panel provides a measure of inequality based on the mixture model by means of the darker lines and, for comparison, the measure based on the Gaussian model by means of the lighter lines. Appendix Tables 23, 24 and 25 provide detail for posterior means and standard deviations of these measures based on the mixture models.

The discrepancy between inequality measures based on the Gaussian and mixture distributions is striking. In all cases, the Gaussian specification leads to a higher measure of inequality than does the mixture specification. This occurs because the Gaussian specification reliably captures the mean and variance in the leptokurtic mixture distributions, but in so doing moves most points in the distribution farther from the mean. This is documented in the lower right panel of Figure 22, which shows that the mixture specification places about 78% of the probability within one standard deviation of the mean, whereas for a Gaussian distribution this fraction is always 68%. The difference is largest in the fraction of earnings accruing to men in the top decile, and lowest in the fraction of men with earnings below half the median. This occurs because the mixture distributions are negatively skewed, meaning that the overdispersion in the Gaussian specification is less for earnings below the conditional mean than above.

Most measures of inequality have risen steadily through the period, and in every case the increase is large relative to posterior interquartile ranges or standard deviations. Inequality arises primarily from the distribution of earnings conditional on age and education, rather than from differences in expected earnings for different levels of age and education, as indicated by the fact that the conditional measure is nearly as large as the unconditional measure in each case.

The rise in the Gini coefficient has been steady, with plateaus in the 1970's and 1980's and a sharp jump between 1981 and 1982. The pattern of changes in the unconditional and conditional Gini coefficients are quite similar. The change in the Gini coefficient is similar to that in the standard deviation of the conditional distribution of log earnings (Figure 19). The change in the Gini coefficient can be ascribed primarily to the increase in these standard deviations.

The fraction of men with earnings below half the median rose from 11.5% to 13.5% between 1967 and 1980, and jumped to 16.5% by 1982 where it remained to the end of the sample. Conditional on age and education the pattern is similar. The movement in the standard deviation documented in Figure 19 accounts in part for these changes. But the impact of the increase in standard deviations is tempered by the decline in the probability of the low-mean state, documented in the lower left panel of Figure 22, and this accounts for the plateau in this measure of inequality in the late 1980's and early 1990's.

The fraction of earnings accruing to men in the top decile shows the same slow increase through the 1970's followed by a sharp increase over the 1980-1982 period, but then it continues to rise steadily through the rest of the sample period. In fact, this measure grows nearly twice as much after 1982 as it does before 1980. Comparison of Figure 19 and Appendix Table 19 with the lower panels of Figure 23 and Appendix Table 25 indicates that like the Gini coefficient, the movement in this measure is driven by the change in scale of the distribution of log earnings conditional on age and education.

4 Conclusions and directions for future research

We undertook this study using a Bayesian approach because we find it natural. It allows us to address questions the way they are asked, conditioning on data and assumptions, and to answer them by providing probability distributions for interesting unobservables like returns to education and measures of inequality. The results are exact and can be traced to explicit assumptions, which can be varied deliberately in order to assess the sensitivity of conclusions to different assumptions. We have done that for several important aspects of the specification of the model.

The study has advanced this methodology in two ways.

1. It shows that state of the art Bayesian methods permit the simultaneous nonparametric modeling of the regression function, employing either basis functions

or smoothness priors, and the conditional distribution, using mixture of normals distributions.

2. The computations required are modest, both absolutely and in comparison with what is often required in non-Bayesian nonparametric methods. They amount to special cases of simple models already incorporated in extensions of popular mathematical applications software.

We began by examining the evidence in the data on the suitability of alternative modelling assumptions, and reached two main conclusions.

1. Three alternative expansions of the regression function—interactive polynomials, separable polynomials, and separable functions with Wiener process smoothness priors—are highly competitive in the sense that different formulations are favored in different years, and the Bayes factors in any one year are rarely extreme. Moreover, substantive conclusions are generally insensitive to choice among these three specifications and consequently averaging over specifications is not essential.

2. The traditional specification of a Gaussian conditional distribution is decidedly inferior to a mixture of two normals. Mixing three normals does not improve the fit.

In the context of this specification we reached several conclusions about the evolution of the distribution of the earnings of men in the U.S. over the thirty-year period 1967 through 1996.

1. The ratio of expected earnings of college graduates to high school graduates declined from about 1.4 in 1967 to around 1.3 in 1979, and then rose to approximately 1.6 by 1990 where it remained through 1996.

2. The ratio of expected earnings at age 45 to those at age 25 grew from about 1.4 in 1967 to almost 1.6 by 1975, where it remained until 1991, after which there is some evidence of further growth in this ratio.

3. The conditional variance of earnings increased steadily over the period. The standard deviation rose from about 0.6 in 1967 to around 0.75 by the mid-1990's.

4. The conditional distribution of earnings was negatively skewed, but the coefficient of skewness rose from -1.25 to -1 over the sample period. The coefficient of kurtosis is between 5 and 7 for most of the sample

5. Inequality in earnings rose steadily through the period. Conditional on age and education, the Gini coefficient steadily rose from about 0.25 to 0.35. Returns to education and experience contribute further to inequality, and the unconditional Gini coefficient steadily rose from about 0.30 to 0.40 over the sample period. This increase in inequality is reflected in growth in the proportion of men with low earnings in the first half of the sample, and in the fraction of income accruing to the top decile in the latter half of the sample.

This study is part of our ongoing research on the evolution of earnings in the U.S. We note several extensions of this work.

1. The analysis here can be repeated, organizing by age rather than by year. That is, we can construct 41 samples of a-year-olds, and examine $p_a\left(y_{ai} \mid t_{ni}, e_{ni}\right)$. This approach explicitly models the impact of the evolution of earnings, and drops any assumption about smoothness in age. One could organize by education or cohort, as well, but these do not lead to cross-sectional analyses.

2. Our work here strongly suggests that the density $p_t\left(y_{ti} \mid a_{ti}, e_{ti}\right)$ depends on a_{ti} and e_{ti} through more than just the regression function. A natural extension of the approach taken here is to permit the state probabilities to depend on covariates, and we are currently pursuing this approach.

3. Ultimately, it should be possible to develop a practical, fully nonparametric longitudinal model in which the conditional distributions of all shocks are potentially non-Gaussian and sensitive to covariates.

References

Barnett, W.A., and A, Jonas (1983), The Muntz-Szatz demand system: An application of a globally well balanced series expansion. Economics Letters 11: 337-342.

Erkanli, A. and R. Bopalan (1994), Bayesian nonparametric regression: Smoothing using Gibbs sampling, in: D. Berry, K. Chaloner and J. Geweke (eds.), Bayesian Statistics and Econometrics: Essays in honor of Arnold Zellner. Wiley, New York.

Gallant, A.R. (1981), On the bias in flexible functional forms and an essentially unbiased form: The Fourier flexible form. Journal of Econometrics 15: 211-245.

Geweke, J. (2005). Contemporary Bayesian Econometrics and Statistics. New York: Wiley.

Geweke, J. and M. Keane (2000), An empirical analysis of earnings dynamics among men in the PSID: 1968-1989. Journal of Econometrics 96: 293-356.

Good, I.J. (1956), The surprise index for the multivariate normal distribution. Annals of Mathematical Statistics 27: 1130-1135.

Green, P. and B. Silverman (1994), Nonparametric regression and generalized linear models. Chapman and Hall, London.

Hardle, W. (1989), Applied Nonparametric Regression. Cambrdge University Press, Cambridge.

Heckman, J.J., L.J. Lochner and P. Todd (2003), Fifty years of Mincer earnings regressions. IZA Discussion Paper No. 775.

Koop, G. and D.J. Poirier (2004), Bayesian variants of some classical semiparametric regression techniques. Journal of Econometrics, forthcoming.

Koop, G. and J.L. Tobias, Semiparametric Bayesian regression in smooth coefficient models. Journal of Econometrics, forthcoming.

Lancaster, T (2004). An Introduction to Modern Bayesian Econometrics. Malden MA: Blackwell Publishing.

Mincer, J. (1958), Investment in human capital and personal income distribution. Journal of Political Economy 66: 281-302.

Rubin, D.B. (1984), Bayesianly justifiable and relevant frequency calculations for the applied statistician. Annals of Statistics 12: 1151-1172.

Shiller, R.J. (1984), Smoothness priors and nonlinear regression. Journal of the American Statistical Association 79: 609-615.

Smith, M. and R. Kohn (1996), Nonparametric regression using Bayesian variable selection. Journal of Econometrics 75: 317-344.

Wong, C. and R. Kohn (1996), A Bayesian approach to additive semiparametric regression. Journal of Econometrics 74: 209-236.

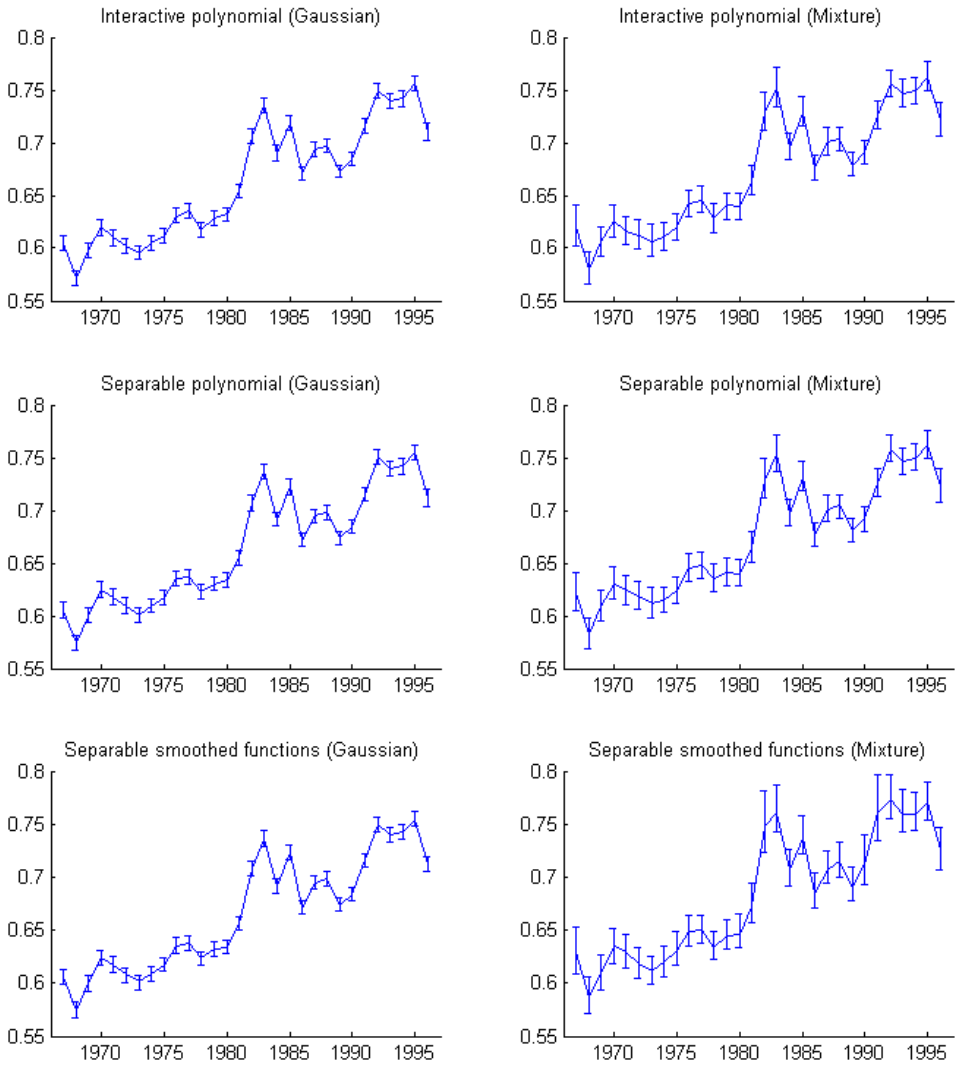

Figure 19: Posterior medians, upper and lower quartiles for the standard deviation of the distribution of log earnings conditional on age and education.

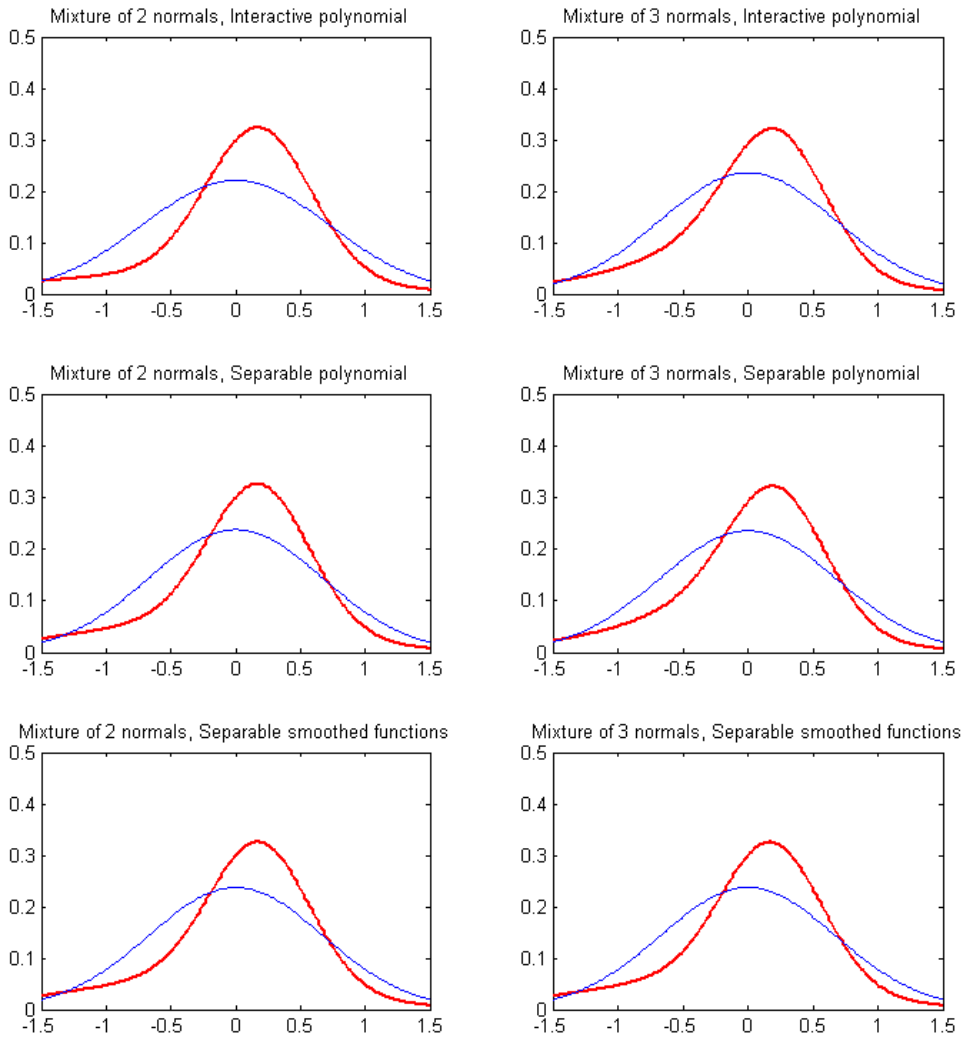

Figure 20: Posterior mean of the mixture of normals conditional p.d.f. for the 1985 sample (heavy line) together with the posterior p.d.f. of the corresponding normal p.d.f. (light line).

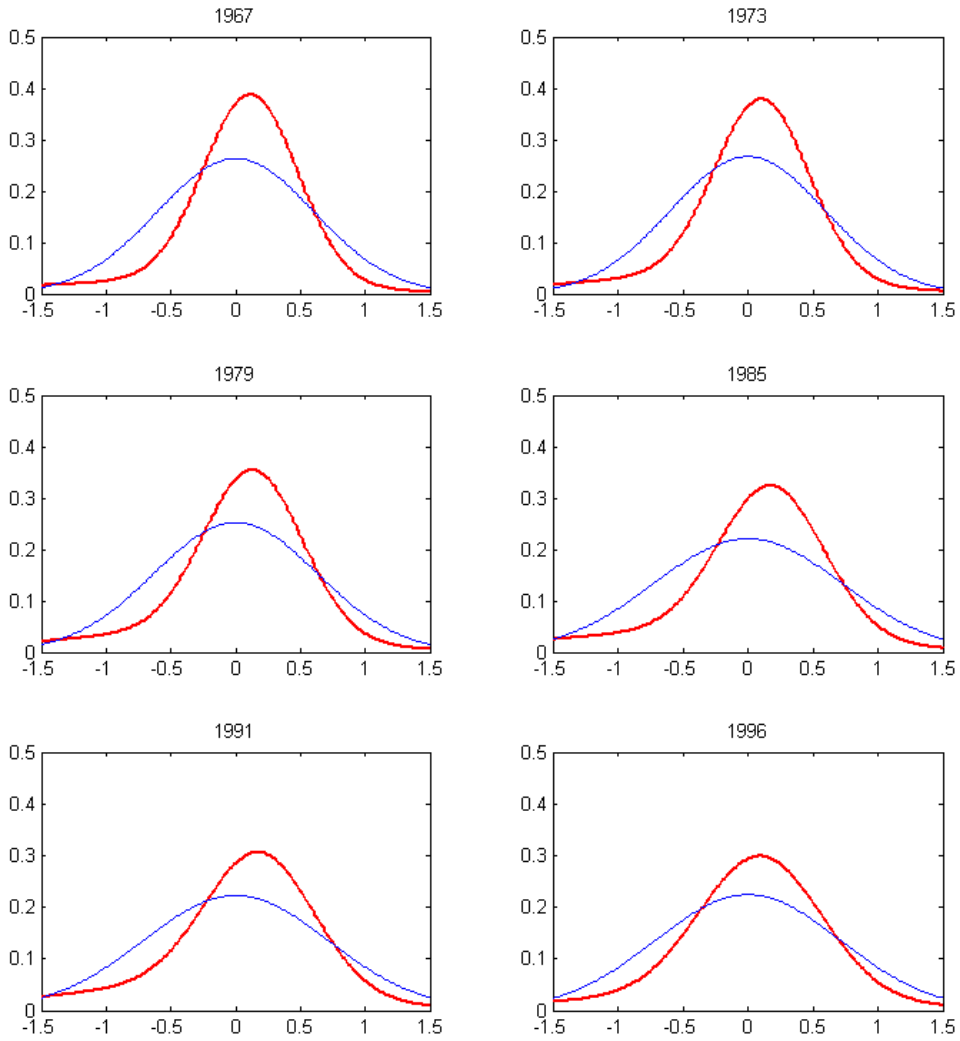

Figure 21: Posterior mean of the two-component mixture of normals conditional p.d.f. for each of several samples (heavey line) together with the posterior mean of the corresponding normal p.d.f. (light line).

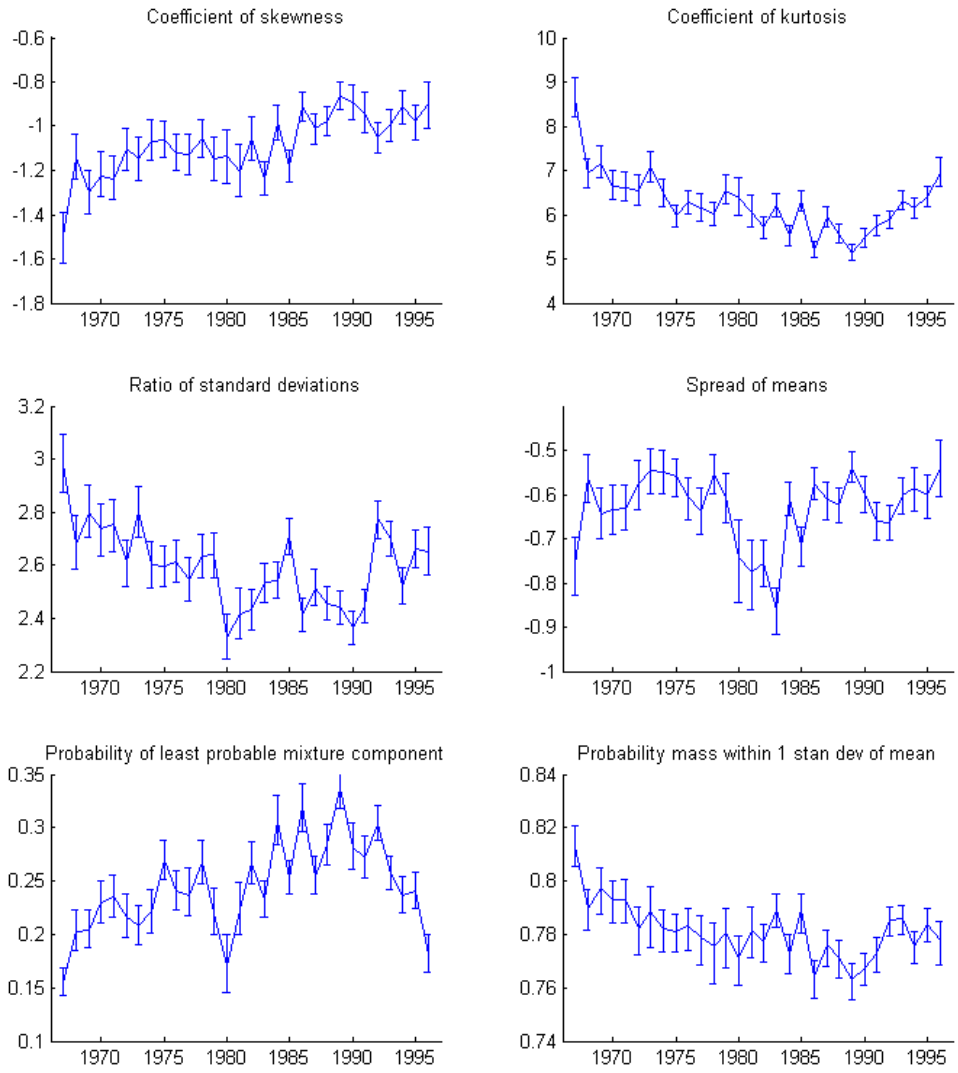

Figure 22: Posterior medians, upper and lower quartiles for several aspects of the two-component mixture of normals distribution of log earnings conditional on age and education.

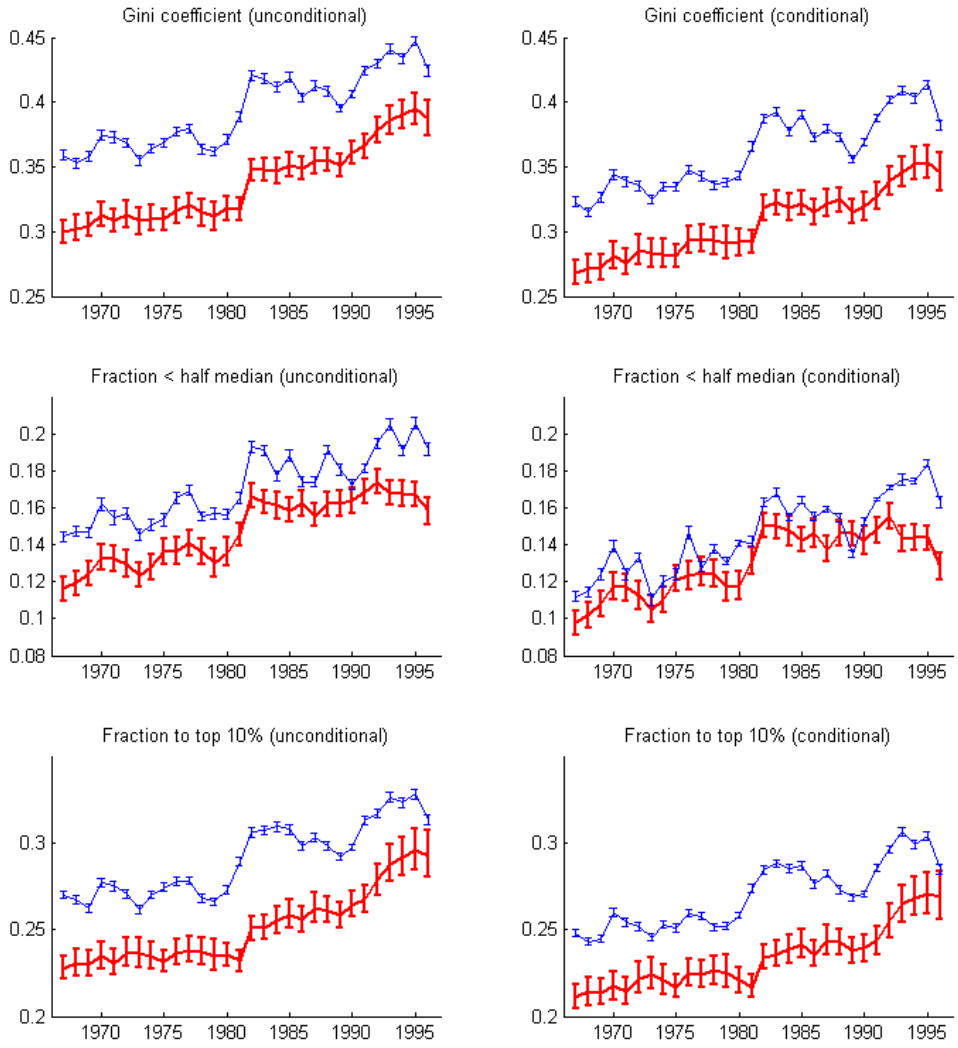

Figure 23: Posterior medians, upper and lower quartiles for several measures of inequality, using a mixture of normals distribution (heavy line) and a Gaussian distribution (light line) of the residuals

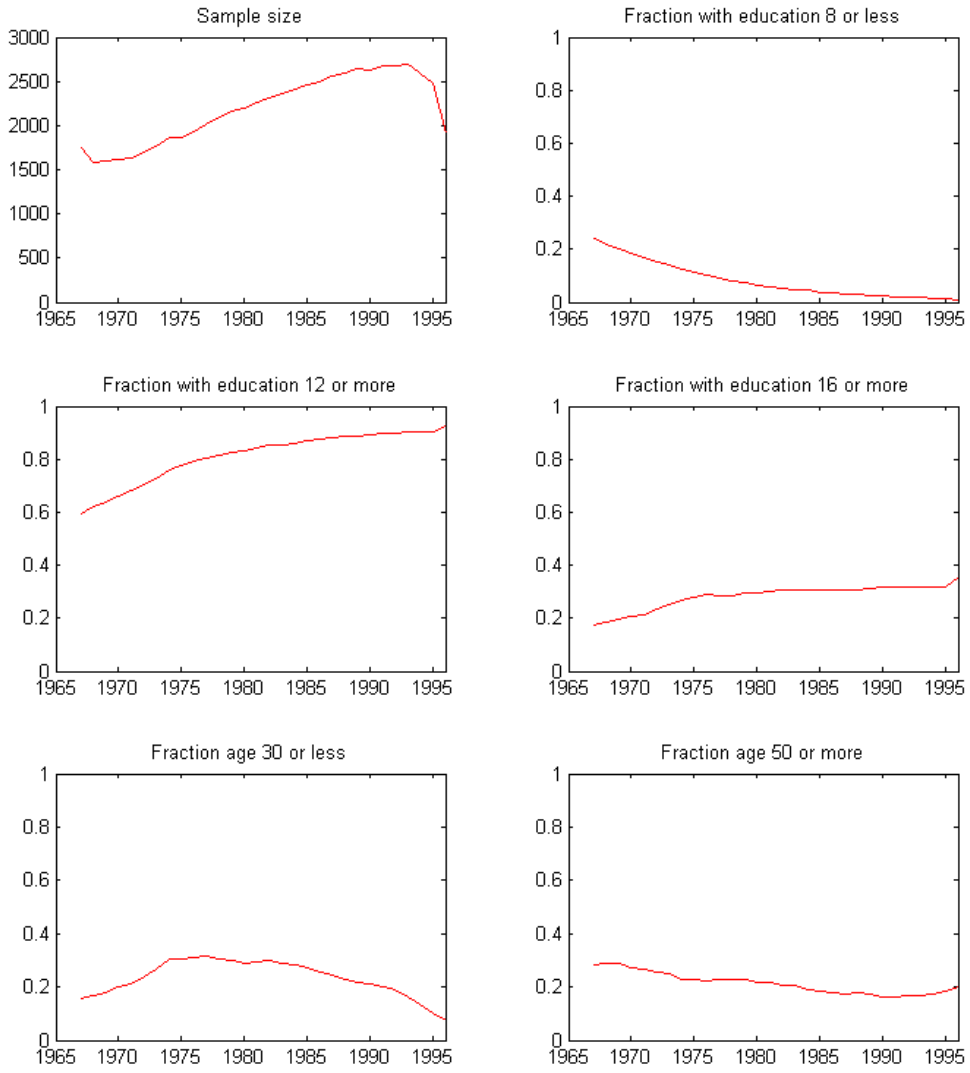

Figure 1: Some aspects of the PSID data sets

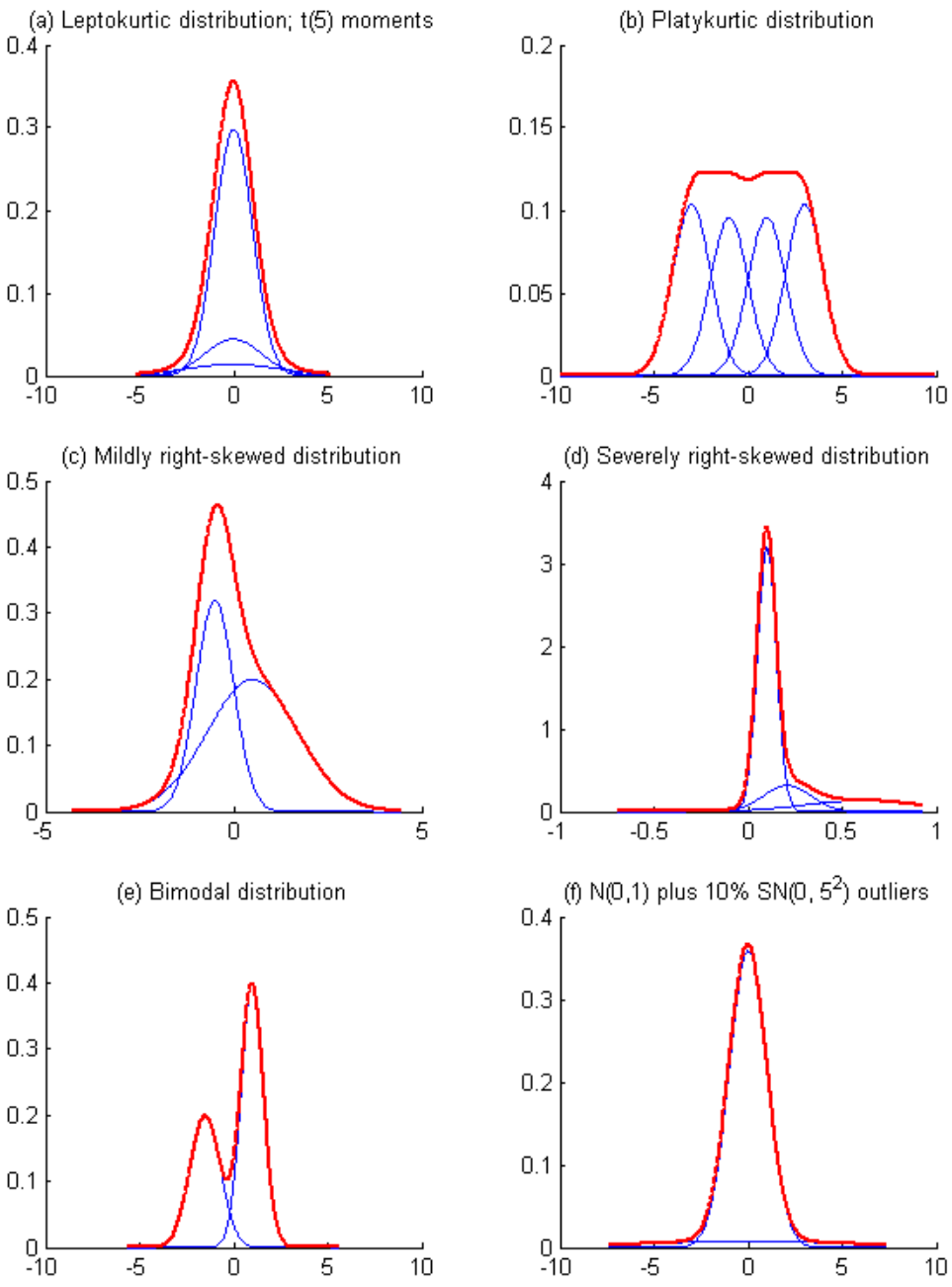

Figure 2: Several mixture of normals probability density functions

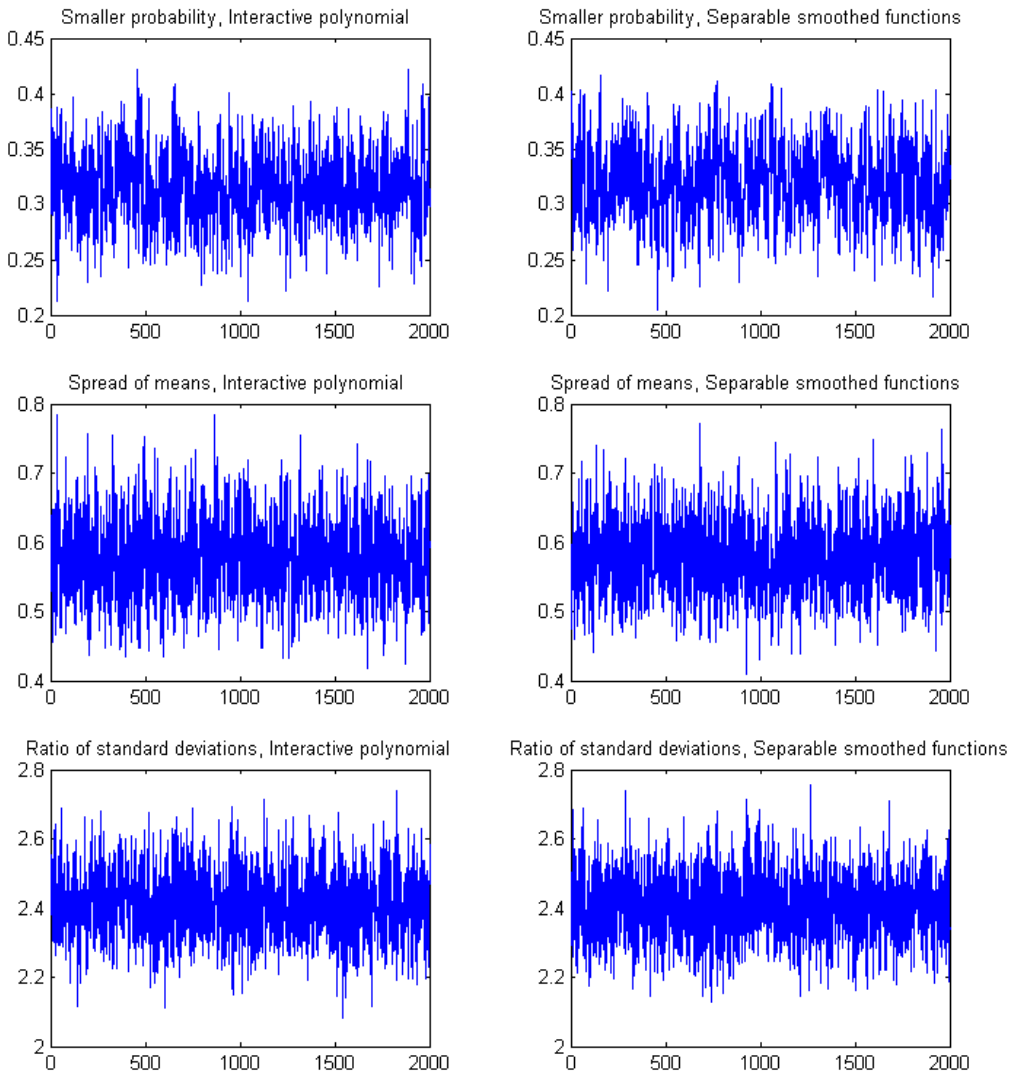

Figure 3: Markov chain Monte Carlo for model with mixture of two normals, 1985 sample

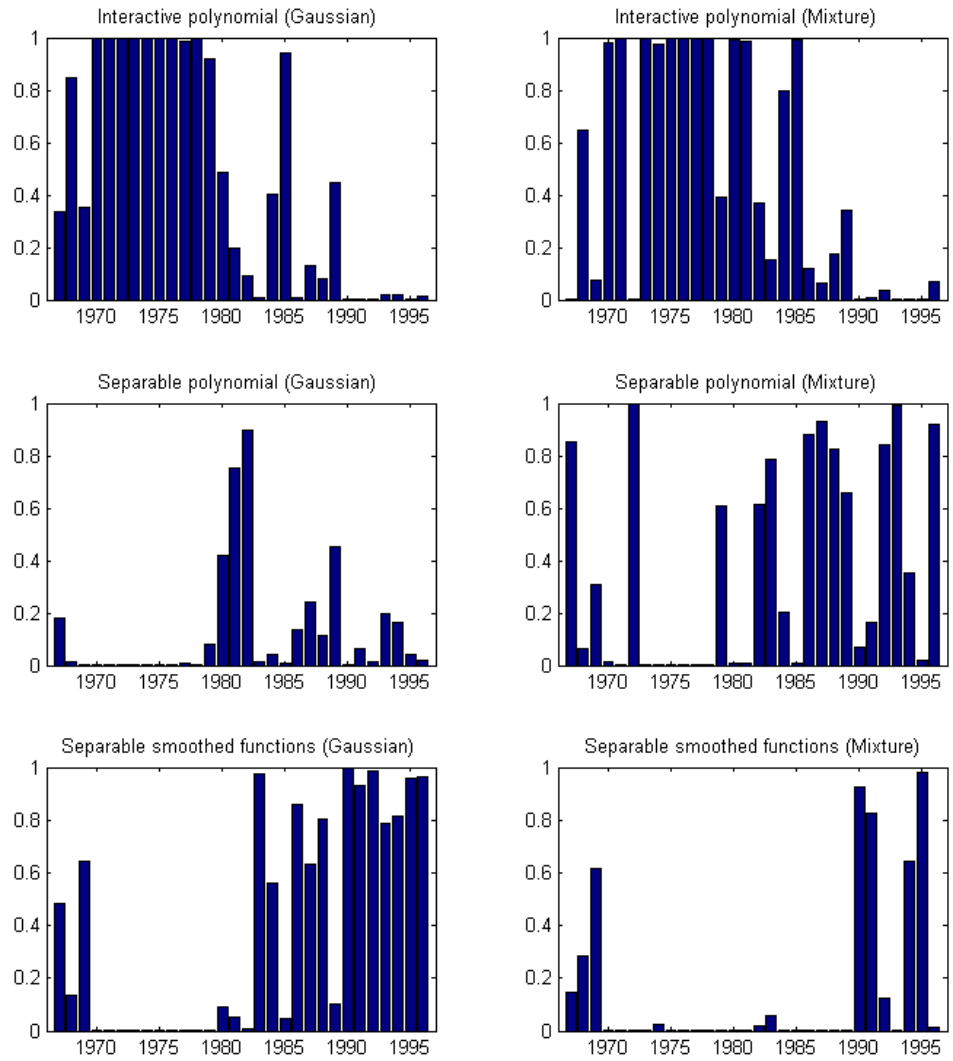

Figure 4: The left [right] panels show posterior probabilities of regression functions given Gaussian [mexture of normals] regression residuals.

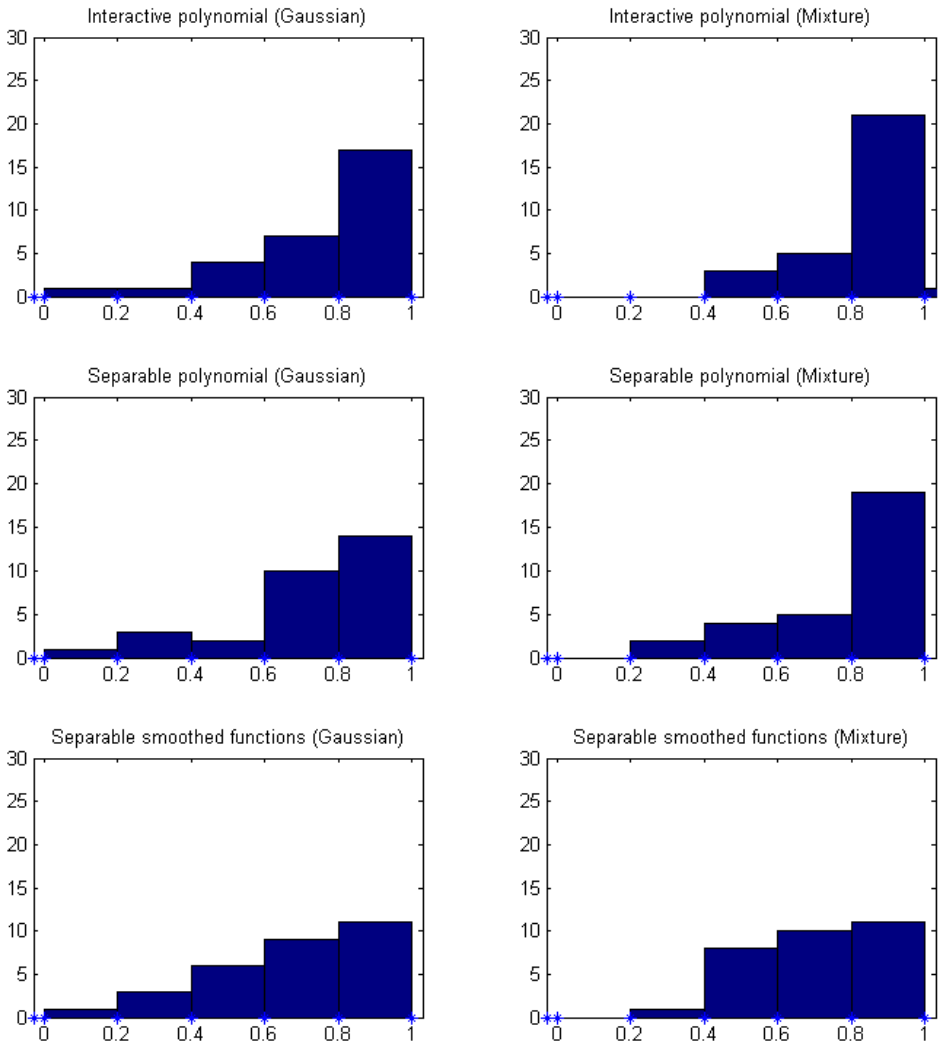

Figure 5: Posterior predictive p_g^* for difference between average sample log earnings of men with 16 and 12 years of education

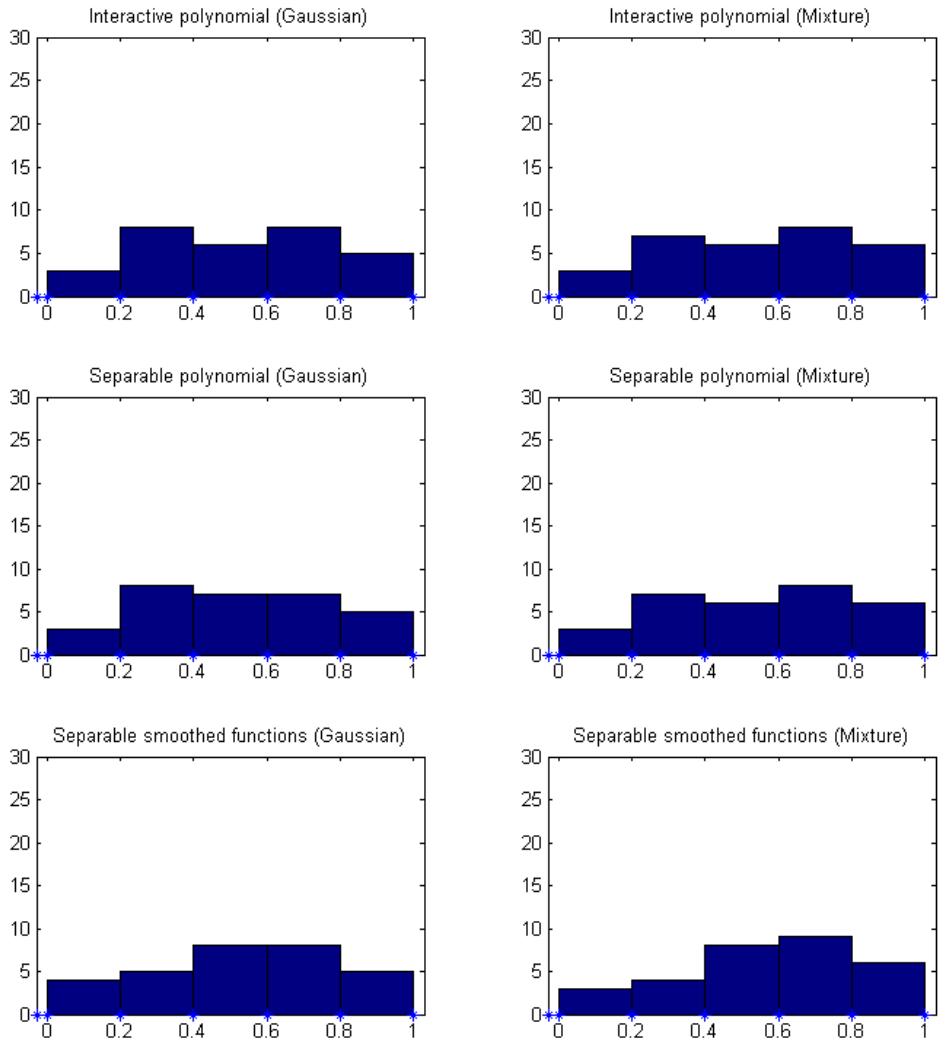

Figure 6: Posterior predictive p_g^* for difference between average sample log earnings of men age 45 and men age 25

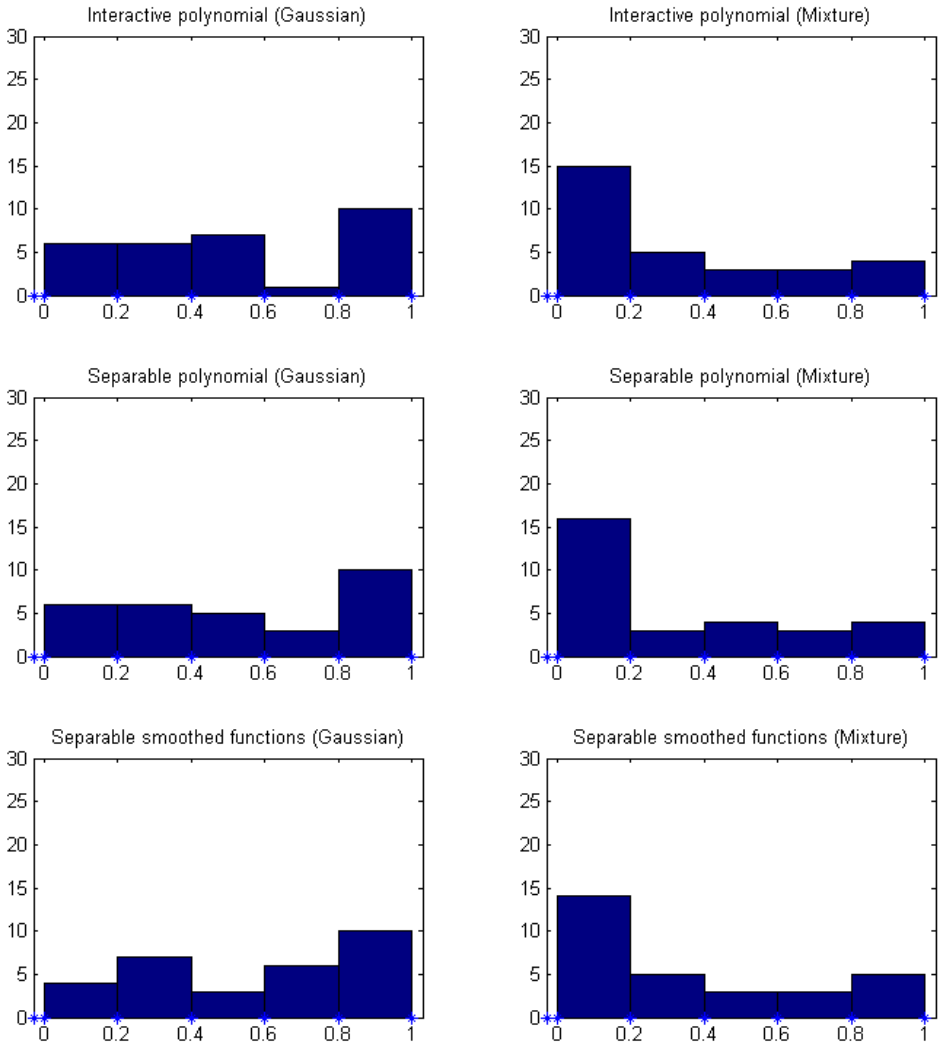

Figure 7: Posterior predictive p_g^* for difference between average sample log earnings of men age 60 and men age 45

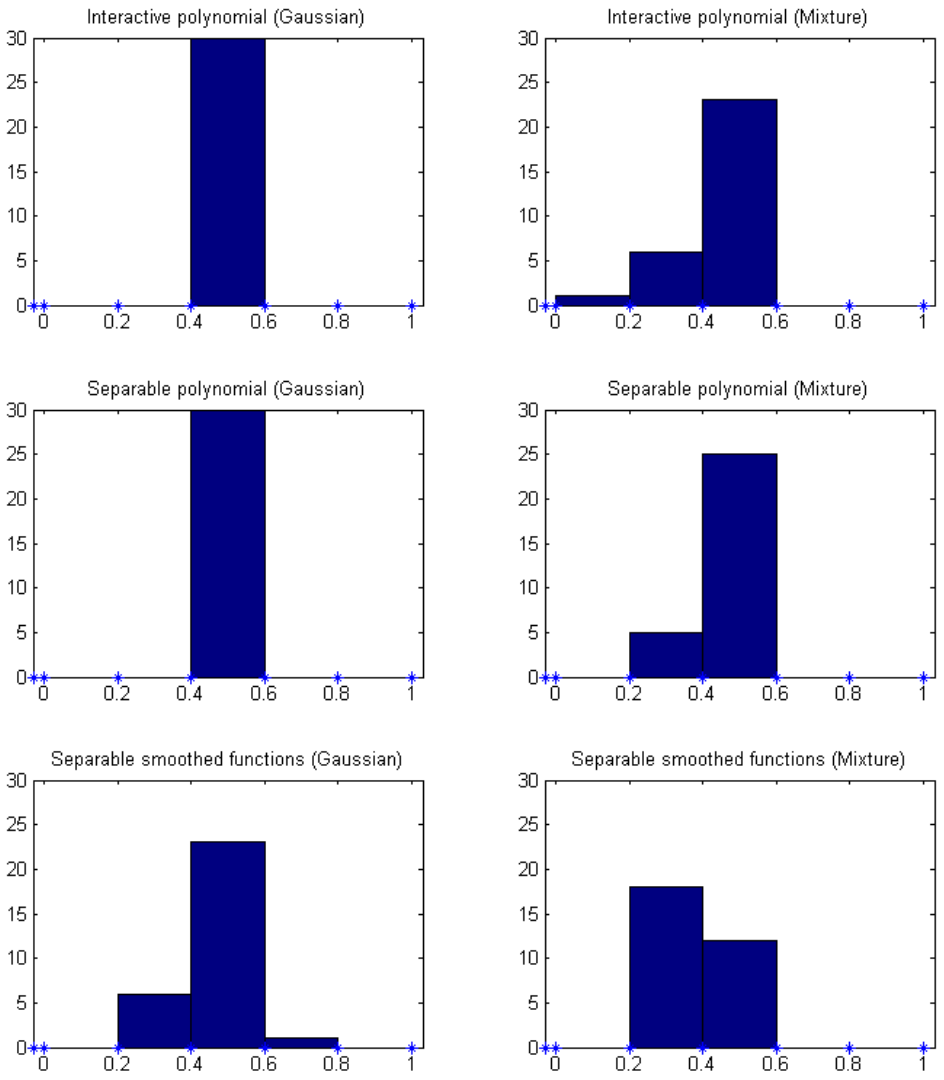

Figure 8: Posterior predictive p_g^* for standard deviation of least squares residuals

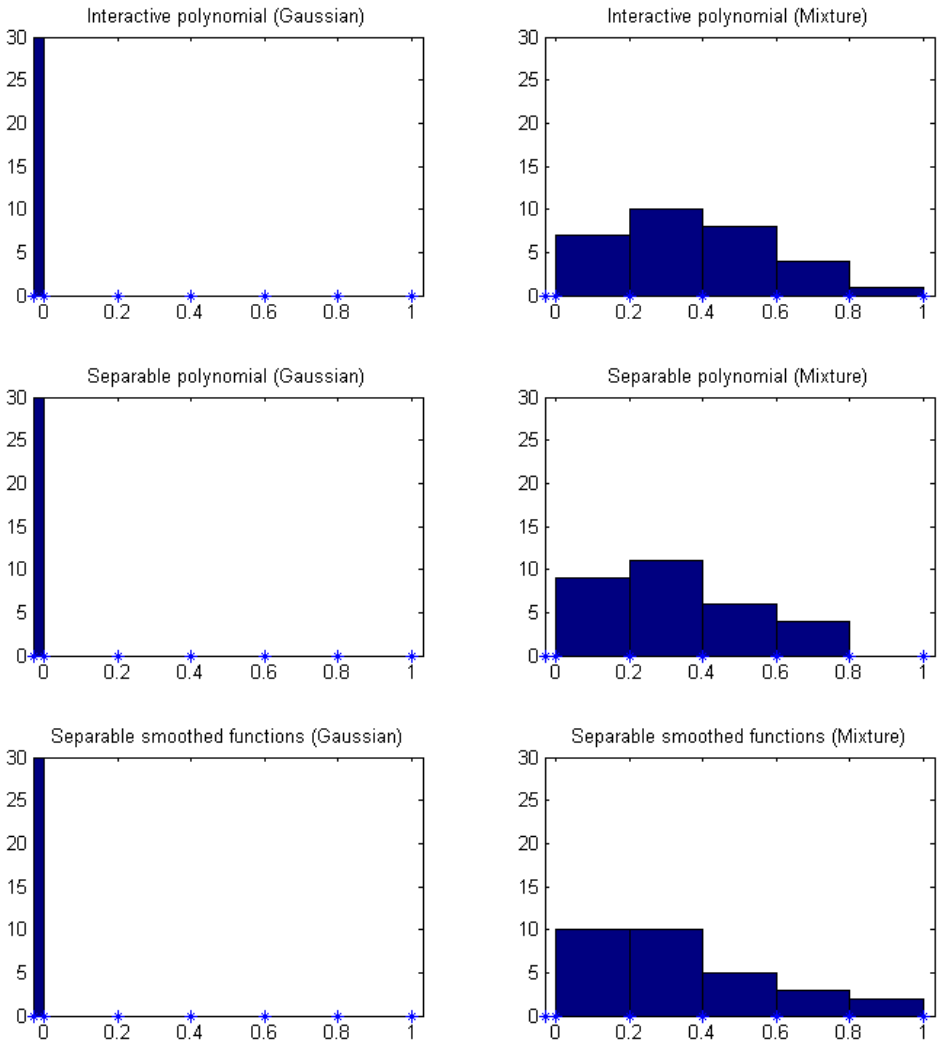

Figure 9: Posterior predictive p_g^* for coefficient of skewness of least squares residuals

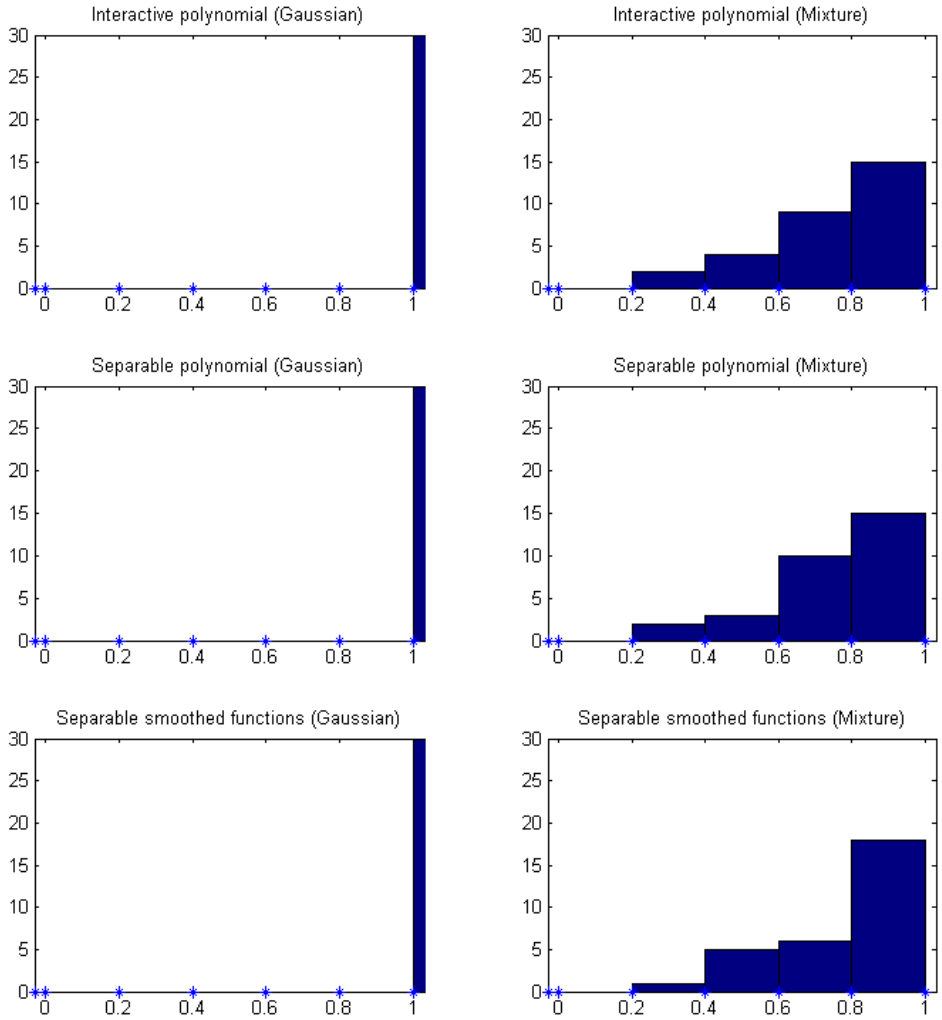

Figure 10: Posterior predictive p_g^* for coefficient of kurtosis of least squares residuals

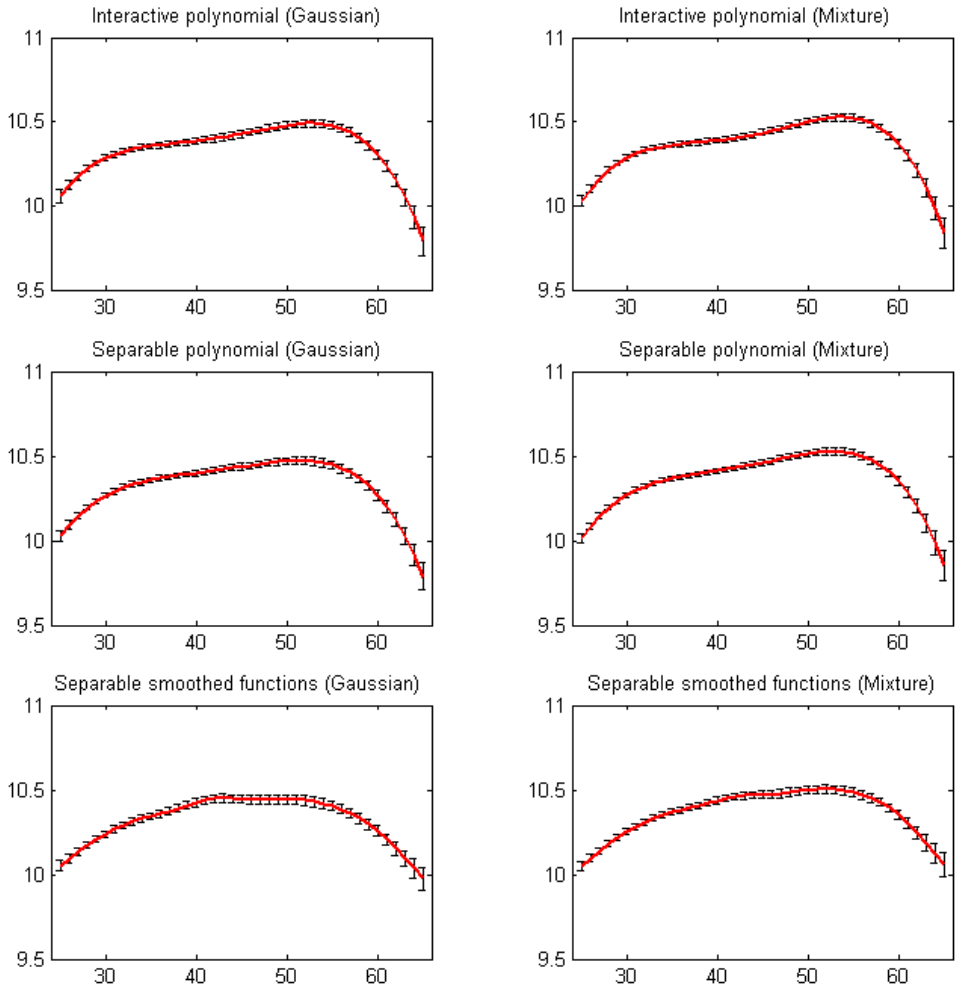

Figure 11: Posterior medians, upper and lower quartiles for expected log earnings conditional on age and 12 years of education, 1986 sample

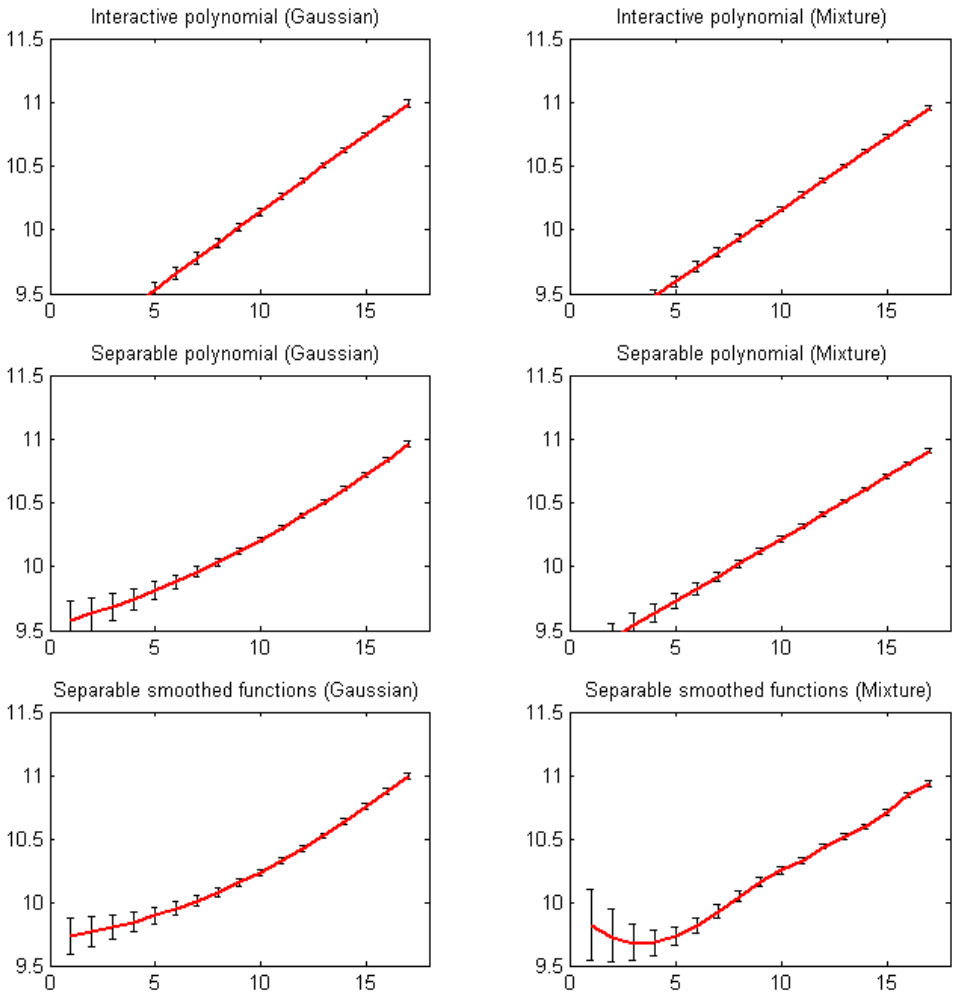

Figure 12: Posterior medians, upper and lower quartiles for expected log earnings conditional on education at age 40, 1986 sample

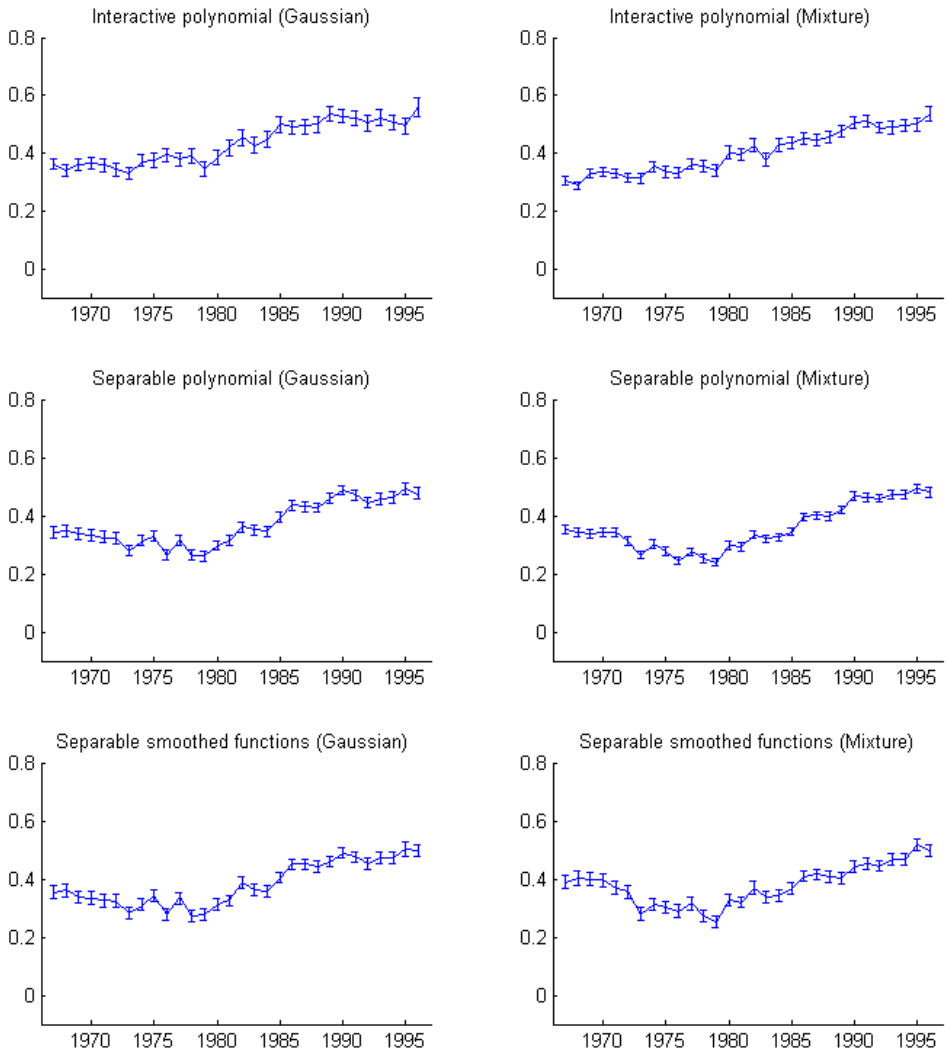

Figure 13: Posterior medians, upper and lower quartiles for the difference in expected log earnings given 16 years of education versus 12 years of education at age 40.

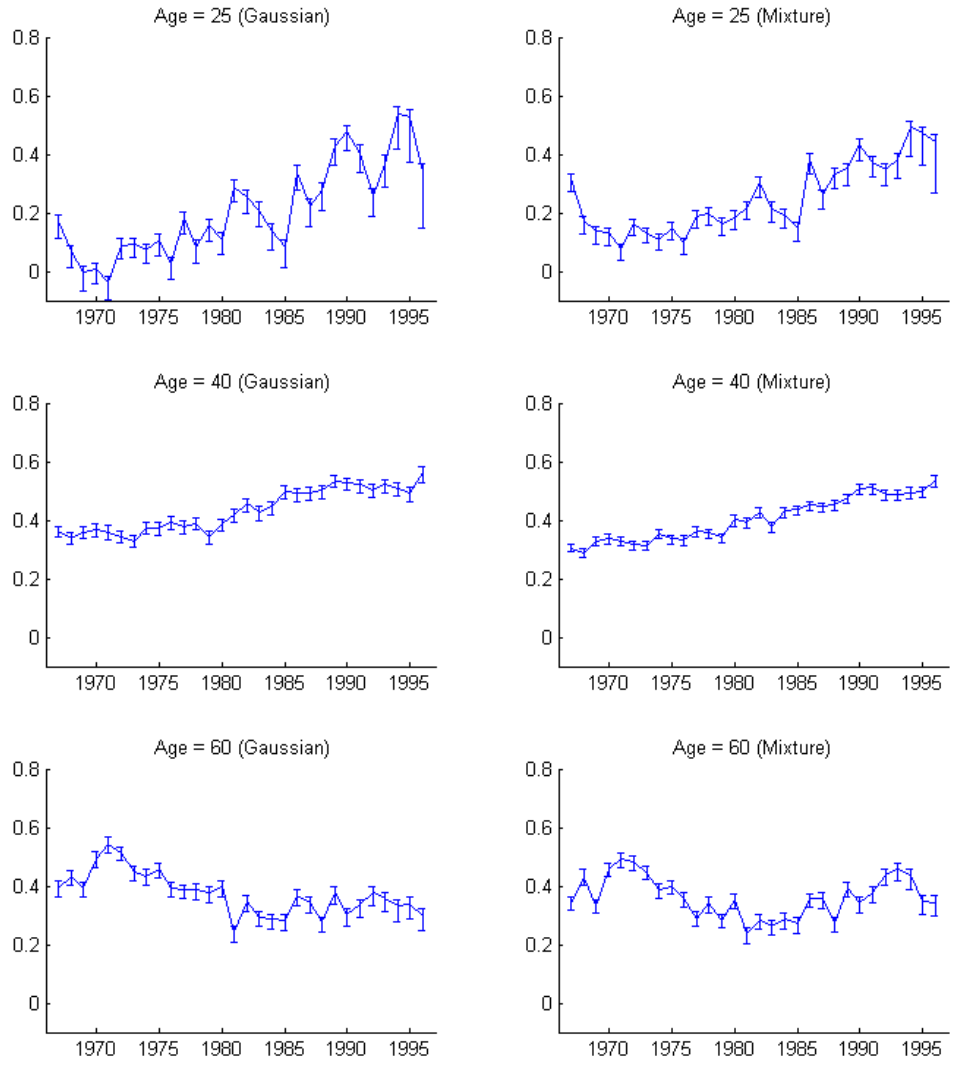

Figure 14: Posterior medians, upper and lower quartiles for the difference in expected log earnings given 16 years of education versus 12 years of education, conditional on alternative ages, interactive polynomials models

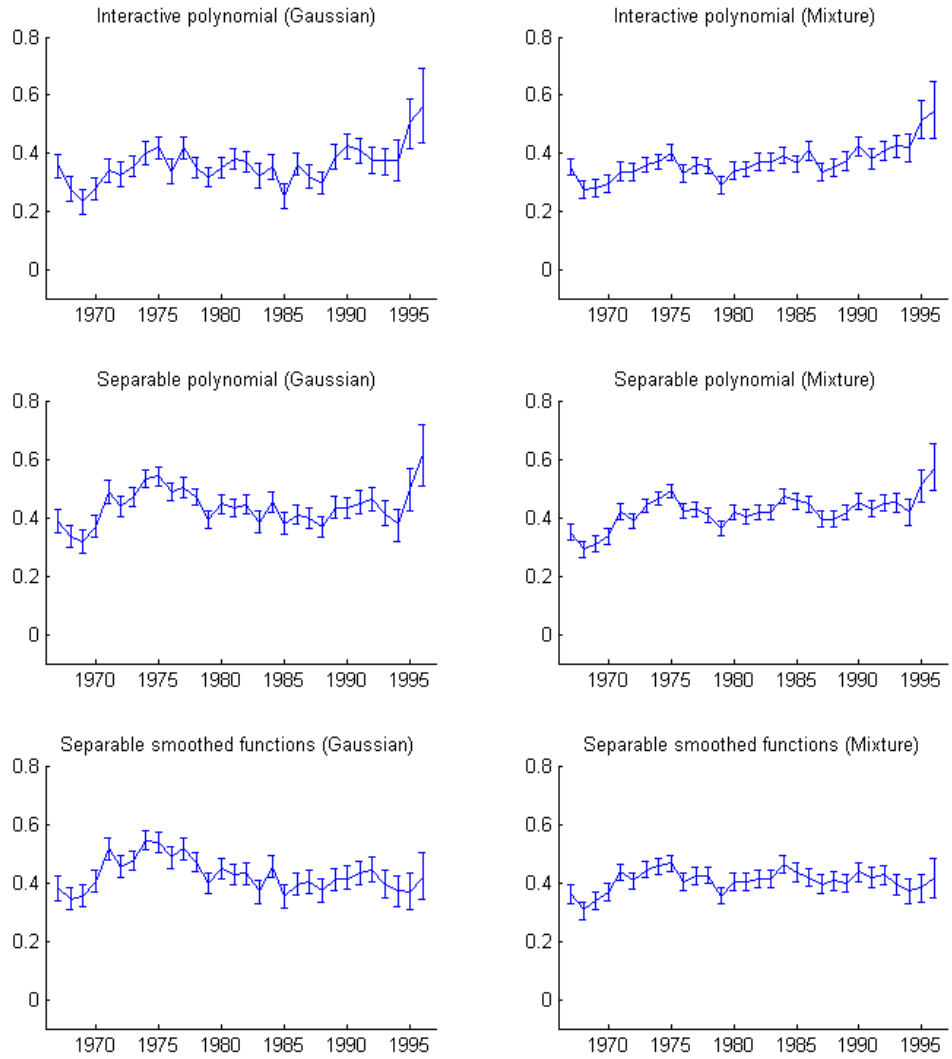

Figure 15: Posterior medians, upper and lower quartiles for the difference in expected log earnings at age 45 versus age 25, given 12 years of education

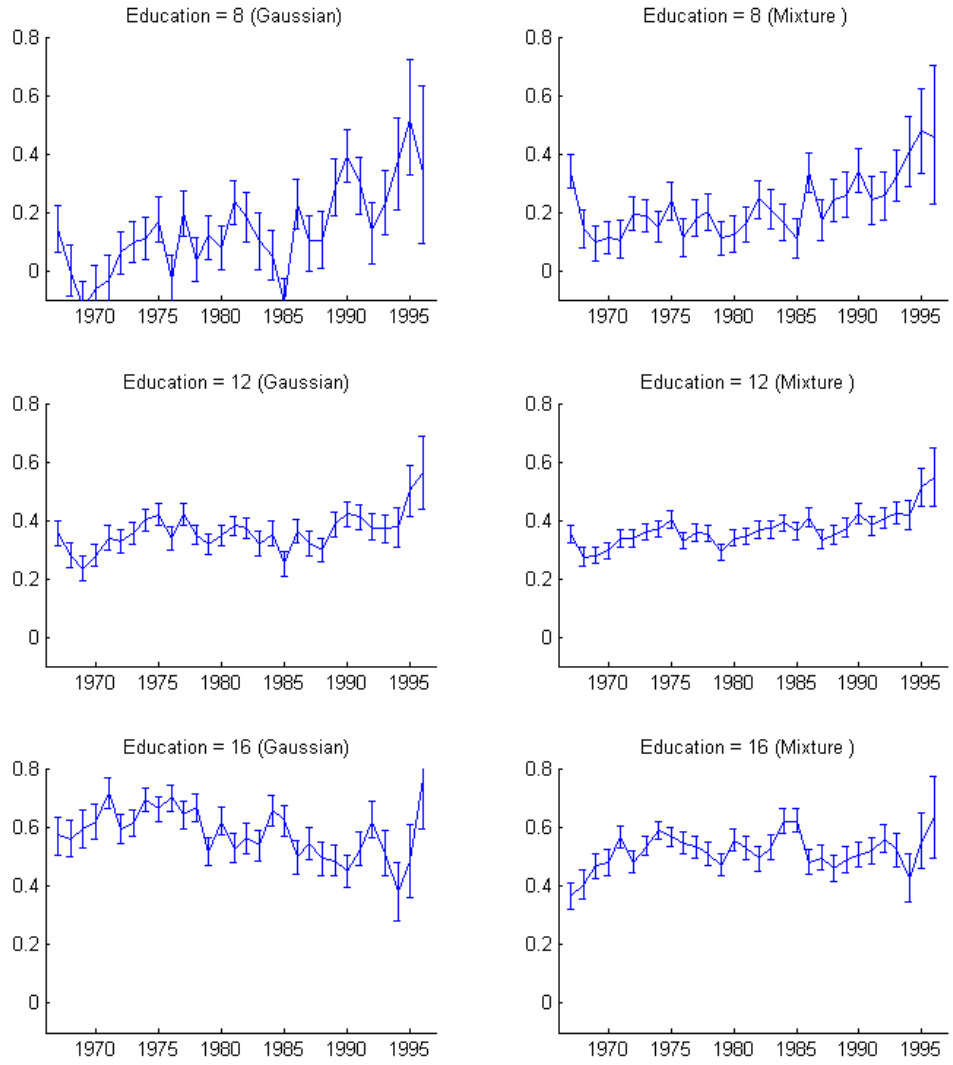

Figure 16: Posterior medians, upper and lower quartiles for the difference in expected log earnings at age 45 versus age 25, conditional on alternative levels of education, interactive polynomials models

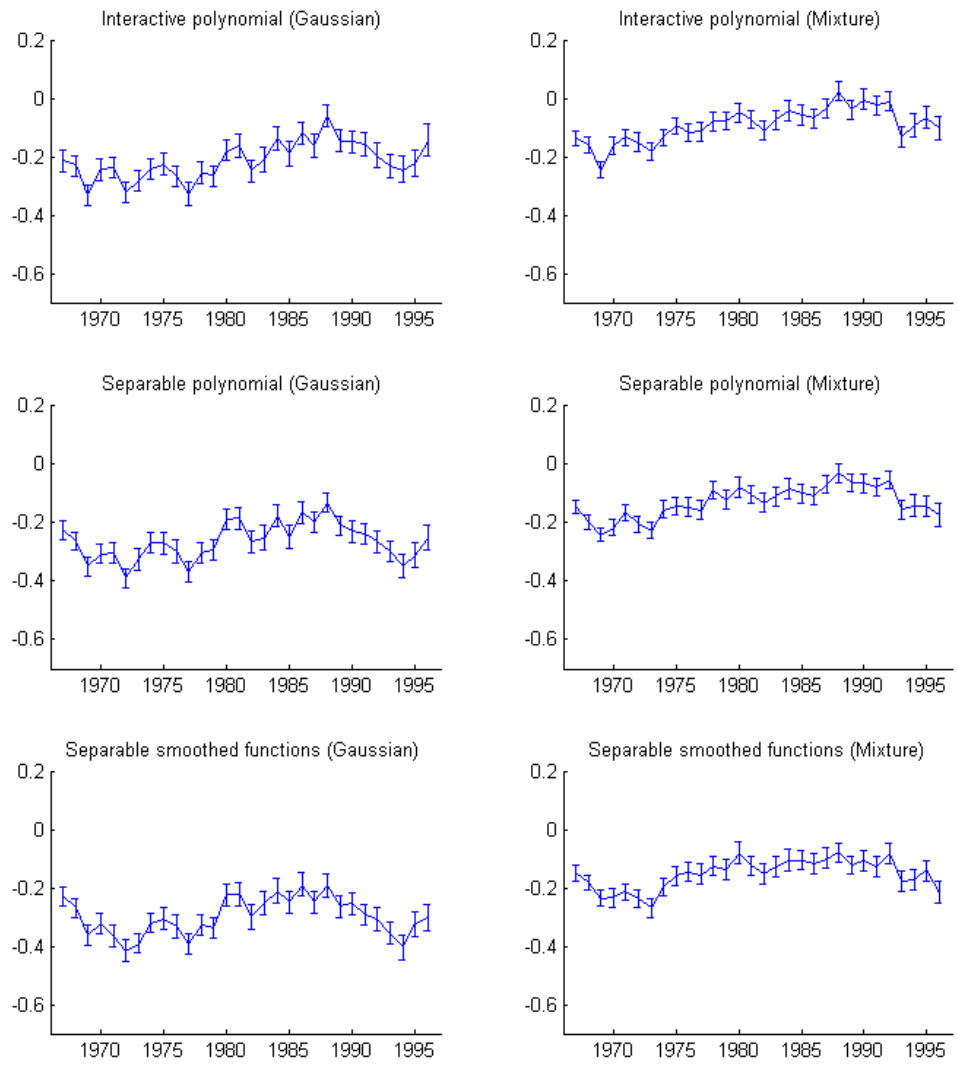

Figure 17: Posterior medians, upper and lower quartiles for the difference in expected log earnings at age 60 versus age 45, given 12 years of education.

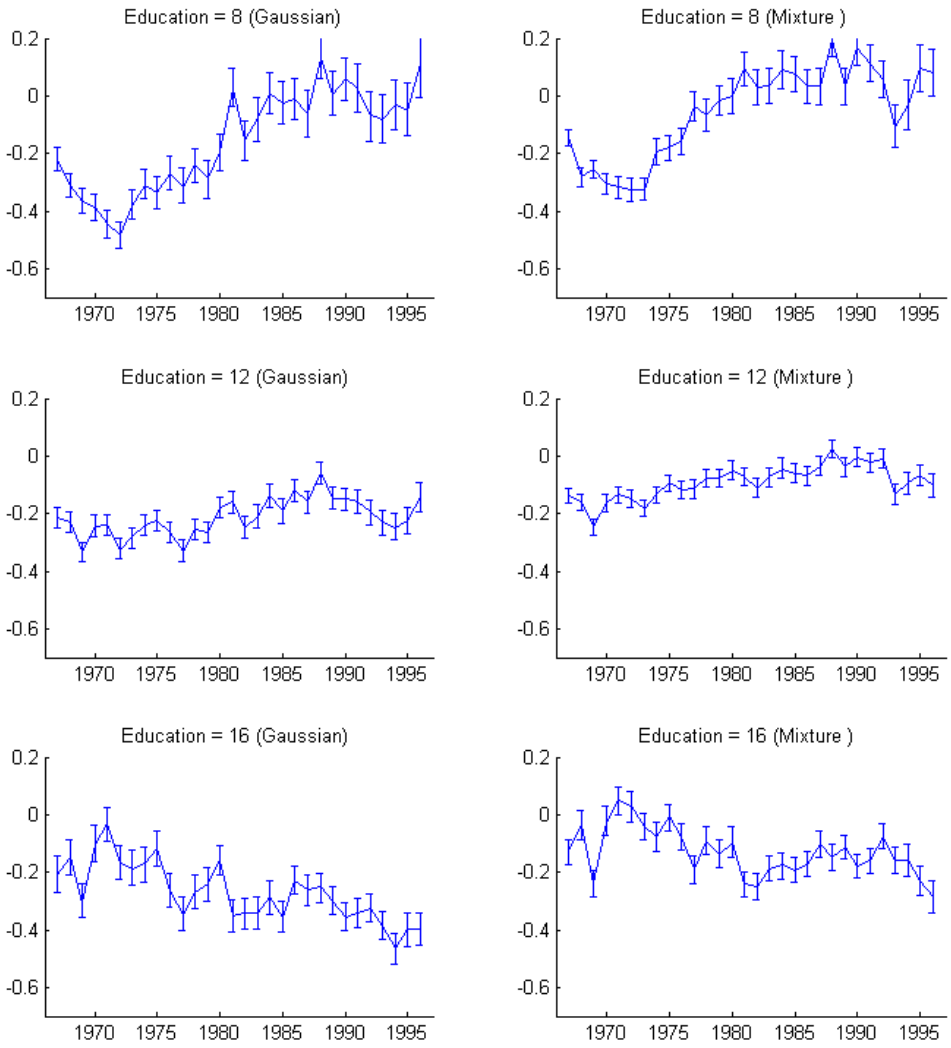

Figure 18: Posterior medians, upper and lower quartiles for the difference in expected log earnings at age 60 versus age 45, conditional on alternative levels of education, interactive polynomials models