2SLS Using Weak Instruments: Implications for Estimating the Frisch Labor Supply Elasticity

By Michael Keane and Timothy Neal*

There is a long standing controversy over the magnitude of the Frisch labor supply elasticity. Macro economists using DSGE models often calibrate it to be large, while numerous micro data studies estimate it is near zero. A large literature has emerged that attempts to reconcile the micro and macro results. We offer a new and simple explanation: Most micro studies estimate the Frisch using a 2SLS regression of hours changes on income changes. But the available instruments are typically "weak." In that case, it is an inherent property of 2SLS that estimates of the Frisch will (spuriously) appear more precise when they are more shifted in the direction of the OLS bias, which is negative. As a result, Frisch elasticities near zero will (spuriously) appear to be precisely estimated, while large estimates will appear to be very imprecise. This will naturally bias micro data studies toward concluding the Frisch is small. We show how the use of a weak instrument robust hypothesis test, the Anderson-Rubin test, leads us to conclude the Frisch elasticity is large and significant in the NLSY97 data. In contrast, a conventional 2SLS t-test would lead us to conclude it is not significantly greater then zero.

Keywords: *Frisch elasticity, labor supply, weak instruments, 2SLS, Anderson-Rubin test*

JEL: *J22, D15, C12, C26*

I. Introduction

The elasticity of labor supply with respect to predictable wage changes – known as the Frisch elasticity – plays a key role in many economic policy debates. The Frisch elasticity is special because predictable wage changes have pure substitution effects. Conesa, Kitao and Krueger (2009) show that the higher is this elasticity, the higher is the optimal tax rate on capital income (relative to that on labor). And macro models where real shocks play a key role in business cycles require the Frisch to be large in order to match observed fluctuations in work hours over the cycle – see Prescott (2006). Because of its importance, a large literature attempts to estimate the Frisch elasticity, as exemplified by classic papers by MaCurdy (1981) and Altonji (1986) and surveyed in Keane (2011, 2021).

* CEPAR & School of Economics, UNSW. Corresponding author: m.keane@unsw.edu.au

The basic idea behind most of this literature is as follows: Given panel data on individual workers observed over time, one may run an OLS regression of changes in log work hours on changes in log wages. This estimates an elasticity of hours with respect to wage changes. But it does not correspond to the elasticity concept of interest – the elasticity of hours with respect to *predictable* wage changes – as some wage changes are predictable while others are surprises. Instead, the approach pioneered by MaCurdy (1981) involves running a two stage least squares (2SLS) regression where one instruments for the change in log wages using an instrument with two properties: First, it predicts wage growth at the individual level. Second, it is known at the start of the time period over which changes in wages are calculated (so it is uncorrelated with surprise wage changes).

Why does this 2SLS regression estimate the Frisch elasticity? Imagine we have annual data. Then the first stage of 2SLS is a regression in which the dependent variable is the change in a worker's log wage from one year to the next, while the independent variable is the instrument. Because the instrument is designed to be something the worker already knew during the first year, the fitted values from this regression give us predictable changes in wages. The second stage of 2SLS regresses the change in log hours from one year to the next on these predictable changes in log wages. This delivers an estimate of the elasticity of hours with respect to predictable wage changes, which is what we are looking for.

Unfortunately, the literature on estimating the Frisch elasticity has been hampered by weak instrument problems. This is because it is hard to find variables that are both known in advance and are good predictors of a person's wage growth during the next year. In other words, the lion's share of annual wage growth at the individual level appears to be idiosyncratic or unpredictable. We now provide a concrete illustration of how this leads to weak instrument problems in estimating the Frisch elasticity, along with an idea for dealing with these problems.

II. Estimating the Frisch Labor Supply Elasticity

We estimate the Frisch elasticity using data from the National Longitudinal Survey of Youth 1997 (NLSY97). The NLSY97 follows a sample of American youth born in 1980-84. The 8,984 respondents were aged 12-17 when first interviewed in 1997.[1] We use data from rounds 11 through 15, which contain data on labor income and work hours in 2005 to 2010. The regression we run is:

$$\Delta lnH_{it} = \alpha + \beta \Delta lnW_{it} + \boldsymbol{\gamma}\mathbf{C}_{it} + \epsilon_{it} \tag{1}$$

where H_{it} is annual hours worked for respondent i in year t, W_{it} is the wage, and $\mathbf{C}_{it}$ is a vector of control variables which includes year dummies (to capture business cycle effects on hours worked) as well as respondent age and race/ethnicity.

[1] Of that, 6748 is a random sample of the birth cohort while 2236 is an over-sample of minority groups.

Our hours measure is "Total annual hours worked at all civilian jobs during the year in question" while our income measure is "Annual income from wages, salary, commissions, and tips before tax deductions." We obtain an annual wage measure by taking the ratio of annual income to annual hours. Regressions that involve percentage changes can be quite sensitive to measurement error and outliers, as these can generate extreme percentage changes. So, as is typical in this literature, we implement a number of sample screens designed to eliminate outliers.[2]

It is important to understand why OLS estimation of (1) does not deliver the elasticity we are interested in. In general, predictable and unpredictable components of wage growth have very different effects on labor supply. For example, a surprise wage increase makes a worker wealthier than he/she expected to be at the start of the year. Hence it generates an income effect that may induce the person to work less. Economic theory gives no clear prediction about what will happen. In contrast, a predictable wage increase doesn't make a worker wealthier (precisely because it was predictable) so basic economic theory says it should induce a pure substitution effect that increases labor supply. It is this substitution or "Frisch" effect of predictable wage changes that we want to estimate.

One can think of this as a measurement error problem that can be solved by IV in the usual way taught in basic econometrics: Actual wage changes are a noisy measure of predictable wage changes, as they sum both predictable and surprise changes. We want to estimate only the effect of predictable wage changes, while effects of surprise wage changes sit in the error term. Actual wage changes are obviously correlated with the surprise wage changes, as some actual wage changes are surprises. Hence actual wage changes are endogenous. To consistently estimate the Frisch effect we need to instrument for actual wage changes using a variable that is correlated with predictable wage changes but uncorrelated with the surprise wage changes that sit in the error.

Our key task then is to choose an instrument that is known to workers at the start of each year, and that generates predictable wage growth during the year. MaCurdy (1981) and many subsequent papers use education as the instrument for wage growth. The motivation is that annual wage growth tends to be faster for more educated workers.[3] We adopt a closely related approach: The NLS administered an aptitude test called the Armed Services Vocational Aptitude Battery (ASVAB) to respondents when they were 13 to 18 years old.[4] We find that the ASVAB percentile score is a much stronger predictor of wage growth than education, so we use that as our instrument. But the idea is similar: Not surprisingly, wage growth is predictably faster for higher ability workers.

[2]Observations were excluded if income was less than $\$3,000$, the annual wage was less than \$2.70 per hour worked, the total number of hours worked was less than 400 or above 4,160 (roughly 80 hours a week), or if the percentage change in wages from the last year was below -50% or above 70%.

[3]He also used interactions of education and age, to allow the effect of education to differ by age.

[4]The ASVAB measures aptitude in several areas including mathematics, general science, paragraph comprehension, and mechanical skills. It was administered in summer 1997 to spring 1998, when the youth were aged 13 to 18 (those aged 13 to 14 were given an easier version of the test). The NLS grouped respondent's into three-month age windows and calculated a youth's percentile rank within his age group.

We did the analysis separately for men and women, as prior literature has shown that their labor supply behavior differs in important ways. Interestingly, the ASVAB score is a much better predictor of wage growth for men than women.[5] For this reason, we decided to focus only on results for men. Our full data set has 5,931 annual observations on 2,100 young men aged 22 to 30 who we observe over 2 to 6 years (the average being 3.8 years).

III. NLS Estimates of the Frisch Elasticity

Table 1 shows the results from estimating regressions of changes in log hours on changes in log wages, as in equation (1). The first column shows OLS results. The coefficient on the log wage change is -0.42 and very highly significant, with a standard error of 0.011.[6] This implies that a 10% wage increase is associated with a 4.2% reduction in hours of work. There are two possible reasons for a negative response of hours to wage changes: As we already noted, surprise wage changes may generate income effects that reduce labor supply. But it is generally viewed as implausible that income effects alone could generate such a large negative effect.

Another important factor driving the OLS estimate negative is a phenomenon called "denominator bias" that plagues many labor supply studies. The problem is that the wage rate is measured as the ratio of earnings to hours. If the hours variable in the denominator of that ratio is measured with error, it causes a worker's measured wage to be too low precisely when his/her measured hours are too high. This induces an (artificial) negative covariance between measured hours and measured wages that drives the estimated elasticity negative. As a result, the OLS estimate cannot be interpreted causally. A second virtue of instrumenting for wage changes is that it also deals with this measurement error problem.

Next we look at the 2SLS results. The second column of Table 1 shows the first stage of 2SLS, where we regress log wages changes on the ASVAB percentile score to construct predictable wage changes. The coefficient is 0.039 and highly significant (standard error 0.012). This means a male worker in the 100th percentile of ability is predicted to have annual wage growth 3.9% higher than a male worker in the 1st percentile. The heteroskdasticity robust F-test for significance of ability in the first stage regression is 10.12, which gives a p-value of 0.002. This implies significance at much better than the 1% level.

It is important to note, however, that the R^2 of the first stage regression is only .007, implying a correlation between our predictions and actual wage changes of .084. In fact the partial R^2 that shows the fraction of wage variation explained by the ASVAB test alone is .002, implying a partial correlation of only .041. This illustrates the point that annual wage growth is very hard to predict. It is important to emphasize, however, that a higher R^2 in the first stage is not

[5] It is not clear if this is because wages grow relatively faster for high ability men than for high ability women, or because the ASVAB is not as good a proxy for labor market skills of women.

[6] All standard errors and F-statistics reported in this paper are heteroskedasticity robust.

TABLE 1—ELASTICITY ESTIMATES - NLSY97

	OLS	2SLS 1^{st} Stage	2SLS 2^{nd} Stage	Reduced Form
Dependent Variable	ΔH	ΔW	ΔH	ΔH
Change in Wages	-0.416		0.597	
	(0.015)		(0.403)	
	[0.015]		[0.363]	
Ability Score		0.039		0.024
		(0.012)		(0.011)
		[0.011]		[0.010]
F-Statistic (Robust)		10.12		4.47
p-value		0.002		0.035
F-Statistic (Clustered)		12.23		5.64
p-value		0.001		0.018
R^2	0.210	0.007		0.009

Note: Robust standard errors are in parentheses and clustered standard errors (by individual) are in square brackets. All regressions controls for year effects, age, and race/ethnicity. $N = 5,931$

necessarily a good thing. Measured wage changes contain both unpredictable and measurement error components that we specifically want to filter out, so we actually want the R^2 of the first stage regression to be much less than one.

Now we look at the second stage 2SLS results, where we regress log hours changes on log predictable wage changes to obtain an estimate of the Frisch elasticity. This is reported in the third column of Table 1. Strikingly the estimate is 0.597, implying that a 10% predictable wage increase generates a 6% *increase* in work hours. So the use of 2SLS actually flips the sign of the coefficient.

This 2SLS estimate is clearly more reasonable: Economic theory predicts a positive Frisch elasticity, as a predictable wage increase should have a positive substitution effect on labor supply. And a Frisch elasticity of 0.6 is in the middle of the range of estimates surveyed in Keane (2011, 2021).

Notice however, that the standard error on the 2SLS estimate is a substantial .403, giving a t-statistic of only 1.48 and a p-value of .138. So, while the estimated Frisch elasticity is a substantial 0.6, it is not even significantly different from zero at the 10% level. This imprecision leaves us in a quandry over what we ought to conclude from the analysis.

The imprecision in our 2SLS estimate is a direct consequence of the fact that the ASVAB score (the instrument) only explains a small part of the variance of wage changes. The weaker is the correlation of the instrument with the en-

dogeneous variable, the higher will be the 2SLS standard error. In the present case the standard error goes up by a factor of 25 when we go from OLS to 2SLS because the partial correlation between the instrument (the ASVAB score) and the endogenous variable (wage changes) is .041, and $1/.041 \approx 25$ (a useful rule of thumb to remember). This imprecision in 2SLS estimates has plagued most of the literature on estimating the Frisch elasticity.

IV. An Example of the Weak Instrument Problem

The situation we see here, which is very typical of attempts to estimate the Frisch elasticity, is a classic example of the "weak instrument" problem. This refers to a situation where the instrument is statistically significant in the first stage of 2SLS, but it only explains a small part of the variance in the endogenous variable. In the present case, the ASVAB score is highly significant in the first stage (p=.002), but it it only explains a small part of the variance in wage changes (partial correlation = 0.04). It is statistically significant because even small effects tend to be significant when sample size is this large (N = 5,931).

Unfortunately, 2SLS results can be very unreliable when instruments are weak. In particular, 2SLS t-tests may be unreliable, and 2SLS estimates may be biased towards OLS. The important paper by Bound, Jaeger and Baker (1995) made applied economists acutely aware of these problems. This in turn led to an explosion of theoretical work on the "weak instrument problem." In this work theorists seek to find criteria that instruments should satisfy for 2SLS results to be reliable.

The key insight of the weak instrument literature is that the quality of 2SLS estimates depends crucially on the size of the first stage partial F-statistic that tests significance of the instrument, where bigger is better. It is useful to recall the basic relationship that $F = NR^2/(1-R^2)$. Properties of 2SLS do not depend on N or first-stage R^2 *per se*, but only how they combine to form F. So a large sample size alone is not sufficient to ensure that 2SLS will deliver reliable results.

In an important paper, Staiger and Stock (1997) studied behavior of the 2SLS estimator at different levels of instrument strength. They developed the well-known "Staiger-Stock" rule of thumb, which says that the first-stage F should be at least 10 before we have confidence in 2SLS results. This $F > 10$ advice has been widely adopted in practice and presented in textbooks. For example, Stock and Watson (2015, p.490) say: "One simple rule of thumb is that you do need not to worry about weak instruments if the first stage F-statistic exceeds 10."[7]

In our application to estimating the Frisch elasticity, the first stage partial F-statistic for testing significance of the ASVAB instrument is 10.12. So we are right on the borderline between a "weak" and an acceptably strong instrument. Should we trust the 2SLS results in this case? Is the 2SLS t-test, which tells us that the Frisch elasticity is not significantly different from zero, really reliable?

[7]Stock and Yogo (2005) proposed critical values for F based on maximal size distortion in t-tests one is willing to tolerate. $F > 16.4$ ensures that a two-tailed 5% t-test will reject at a 10% rate or less.

V. The Anderson-Rubin Approach

In the early days of instrumental variable methods, Anderson and Rubin (1949) developed another method we can use to test if our estimate of the Frisch elasticity is significant. The Anderson-Rubin (AR) test relies on the so-called "reduced form regression." This is the regression of the outcome of interest on the instrument itself, along with the control variables. In our case this means a regression of the change in log hours on the ASVAB score itself, along with the control variables (time, age, race). The AR test judges the Frisch elasticity estimate to be significant if the ASVAB score is significant in the reduced form regression.

The logic of the AR test is simple: A fundamental assumption of the IV method is that the instrument only affects the outcome of interest indirectly through its effect on the endogenous variable. Hence, if the instrument is significant in the reduced form, it implies that the endogenous variable has a causal impact on the outcome of interest. In our case, if the ASVAB score is significant in the reduced form, it means that predictable wage changes influence work hours.[8]

The last column of Table 1 reports the reduced form results. Here, the ASVAB score is clearly significant, with a t-stat of 2.18 (p-value 0.035). So we are left with a quandry: The AR test says our 2SLS estimate of the Frisch elasticity is significant, while the t-test says it isn't. Which result should we believe?

The AR test is recommended by econometric theorists as clearly superior to the t-test when instruments are weak, and no worse when instruments are strong. This is because the AR test has two major advantages: First, it is "robust" to weak instrument problems, which means a 5% level AR test will reject a true null hypothesis at the correct 5% rate regardless of the strength or weakness of the instruments. In contrast, the t-test is unreliable: If instruments are weak, a 5% t-test may reject a true hypothesis at rates far above or below 5%, depending on details of the situation. Second, Moreira (2009) shows that in the case of a single instrument (which is what we have in our Frisch elasticity application) the AR test is the most powerful robust test. This means that if the null hypothesis is false, the AR test will reject the null hypothesis, and conclude the parameter of interest is significant, more frequently than any other robust test.[9]

Despite its clear advantages, the AR test has been generally neglected by applied researchers. In fact, as far as we know, it has never been adopted in the vast literature on estimating labor supply elasticities. In our Frisch elasticity application, given that the first-stage F statistic is only slightly above 10, conventional

[8] One could argue, instead, that the ASVAB score somehow affects hours growth directly. But in that case the ASVAB score is not a valid instrument, so the 2SLS results are completely invalid anyway. The very assumptions that make the IV approach valid in the first place also make the AR test valid.

[9] Andrews, Stock and Sun (2019) argue that the AR test should be widely adopted by applied researchers. They state its advantages more formally: "In just-identified models ... Moreira (2009) shows that the AR test is uniformly most powerful unbiased. ... Thus, the AR test has (weakly) higher power than any other size-α unbiased test no matter the true value of the parameters. In the strongly identified case, the AR test is asymptotically efficient in the usual sense and so does not sacrifice power relative to the conventional t-test. ... Since AR confidence sets are robust to weak identification and are efficient in the just-identified case, there is a strong case for using these procedures in just-identified settings."

wisdom says we are in a borderline case where weak instruments may or may not be a concern. Clearly the AR test should be viewed as more reliable than the t-test in this context. In the next section we present a numerical experiment based on our data that show just how superior the AR test is in practice.

VI. Monte Carlo Experiment

In this section we compare the AR test and the t-test to see which is a more reliable guide to the statistical significance of our Frisch elasticity estimate. To do this, we conduct the following experiment: We start from the NLS sample of N=5,931 observations that we used to generate the estimates in Table 1. We can then "bootstrap" a new artificial dataset by sampling 5,931 observations with replacement from the original sample. We do this 10,000 times to form 10,000 artifical datasets. We then repeat the analysis of Table 1, applying OLS and 2SLS to all 10,000 datasets, and summarize the results in Table 2.

Table 2—Results from Monte Carlo Bootstrap Samples

	OLS		2SLS		First Stage	Reduced Form	
	$\hat{\beta}$	S.E.	$\hat{\beta}$	S.E.	F Statistic	$\hat{\pi}$	S.E.
Median	-0.4163	0.0146	0.5998	0.4013	10.1314	0.0238	0.0112
Mean	-0.4164	0.0146	0.7185	4.7202	11.0923	0.0237	0.0112
Std. Dev.	0.0148	0.0004	3.8636	251.4631	6.4577	0.0111	0.0003

Note: $N = 5,931$ *for each of the* $10,000$ *samples used to form the results.*

A. OLS Estimates and Standard Errors

The first thing to notice in Table 2 is that both the median and mean of the OLS estimates of the Frisch elasticity (across all 10,000 datasets) are equal (to three decimal places) to the (downward biased) value of -0.416 we obtained using the original NLS sample. This is exactly what we expect to see. Because we constructed our 10,000 artificial "bootstrap" datasets from the original NLS sample, they all have similar properties to the original – in terms of the covariances of the variables like hours and wages. So we expect that, on average, across the 10,000 samples, we will get results similar to those from the original sample.

The second thing to notice in Table 2 is how the estimates of the Frisch elasticity vary across the artificial samples. The third row of Table 2 reports the standard deviation of the estimates across the 10,000 samples. Note that the OLS estimates of the elasticity have a standard deviation of 0.015. Again, this is exactly what we expect to see. Because we are taking random samples of observations from the original dataset, the covariance of wages and hours will differ slightly across the samples, leading to slightly different OLS estimates.

These findings highlight a very special property of OLS: The standard deviation of the OLS estimates across the 10,000 samples (0.015) is exactly equal (to three decimal places) to the OLS standard error estimate we reported back in Table 1. This is a very useful property, as it means the estimated OLS standard error is a very good guide to how variable the estimated Frisch elasticity really is across different samples. In other words, the OLS standard error estimate is useful for making judgements about statistical significance.[10]

B. 2SLS Estimates and Standard Errors

Now we examine how the 2SLS estimates and standard errors behave. The first thing to note in Table 2 is that the median 2SLS estimate of the Frisch elasticity (across all 10,000 datasets) is 0.600, which is very close to the 2SLS estimate 0.597 we obtained using the original NLS dataset. Again, this is exactly what we expect. As all 10,000 of our artificial datasets were constructed from our original NLS sample, we can think of the NLS sample as the "population" from which all 10,000 datasets are drawn. In this population, 0.597 is in fact the true value of the Frisch elasticity. We see that the median 2SLS estimate accurately uncovers the true elasticity value (while of course OLS does not). This is precisely why 2SLS is a very useful statistical procedure.

The second notable point is that the standard deviation of the 2SLS estimates across the 10,000 data sets is 3.864. In contrast to OLS, this bears no resemblance to the 2SLS standard error estimate of 0.403 we reported back in Table 1. This is our first indication that the 2SLS standard error estimate is not a good guide to the actual variability of the 2SLS estimates across samples.[11] This in turn means that 2SLS t-statistics – which rely on those standard error estimates – will not be a useful guide to significance of 2SLS estimates.

To further explore the behavior of the 2SLS standard error, Figure 1 plots the 2SLS standard errors against the 2SLS estimates of the Frisch elasticity from each of the 10,000 samples. A very interesting aspect of the figure is the strong positive covariance between 2SLS estimates and their standard errors: In samples where the estimated Frisch elasticity is larger, the standard error is also larger.[12]

[10]Table 2 also reports the mean and median of the estimated OLS standard error across the 10,000 artificial datasets. These are again 0.015. And the variation across samples of this standard error estimate is trivially small. So we have an even stronger result: The estimated standard error in each individual sample is a good guide to the actual variability of the OLS estimates across all samples.

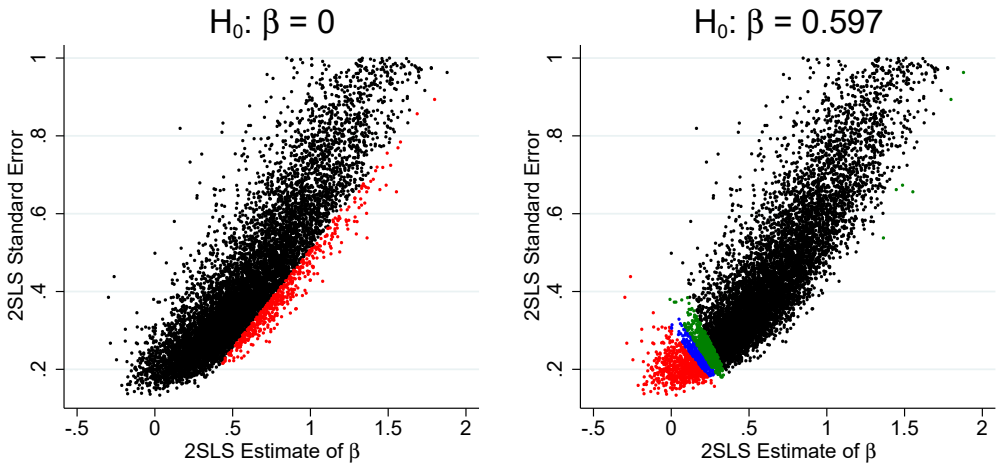
[11]In fact, in the single instrument case the mean and variance of the 2SLS estimator does not exist, which means that if we did many more than 10,000 runs the estimates wouldn't converge to anything in particular. This means the standard deviation of the 2SLS standard error cannot be bootstrapped.

[12]This pattern arises for the following reason: As we discussed in Section III, in the original NLS sample the partial correlation between the ASVAB score and wage growth is 0.04. When we look across our 10,000 subsamples, the correlation fluctuates around that value due to sampling variation. Two things happen in samples where that correlation is relatively high:

First, the 2SLS standard error estimate is smaller: The stronger is the correlation between the instrument and the endogenous variable, the smaller is the 2SLS standard error.

Second, the 2SLS estimate is more shifted in the direction of the OLS bias (which is negative). This is because, as we discussed in Section III, if the predictable part of wage growth is small, then a high

FIGURE 1. STANDARD ERROR OF $\hat{\beta}_{2SLS}$ PLOTTED AGAINST $\hat{\beta}_{2SLS}$ ITSELF



Note: Runs with standard error > 1 *are not shown. In the left panel, the red dots indicate* $H_0 : \beta = 0$ *is rejected at the 5% level, while in the right panel red dots indicate* $H_0 : \beta = 0.597$ *is rejected at the 5% level. Blue and green indicate 10% and 20%.*

Because of this pattern, large positive 2SLS estimates of the Frisch elasticity will have large standard errors. This mechanical relationship makes it very hard for a 2SLS t-test to detect a true positive Frisch elasticity.

Recall that our 10,000 simulated data sets are constructed in such a way that the true value of the Frisch elasticity in these data sets is 0.597. This means that if the 2SLS t-test is reliable it should have two properties: First, if we run 5% t-tests of the hypothesis that the true Frisch elasticity is zero we should reject that false hypothesis at a high rate. This means the test has good "power." Second, if we run 5% t-tests of the true hypothesis that the Frisch elasticity is equal to 0.597 (the true value) we should reject that hypothesis only 5% of the time (i.e., not very often). This means the test has correct "size." Furthermore, those rejections should be evenly split between cases where the estimated Frisch elasticity is above and below the true value.

In the left panel of Figure 1 we shade in red the cases where the 2SLS t-test rejects the false null hypothesis that the true Frisch elasticity is equal to zero. In other words, the cases where the absolute value of the ratio of the estimate to the standard error exceeds the 5% critical level of 1.96. Notice how the red shaded

correlation between the instrument and the endogenous variable is not really a good thing. In samples where that correlation rises above 0.04, the instrument is picking up some of the endogenous part of wage growth that arises due to measurement error and surprise wage growth. This in turn means the 2SLS estimate will be shifted in the direction of the OLS bias (negative).

Putting these two facts together, it means that 2SLS estimates that are most shifted in the direction of the OLS bias (negative) appear to be more precise. This is exactly the pattern we see in Figure 1.

area is very small. In fact, the false null hypothesis is only rejected 5.1% of the time. This is an abysmally low level of power. In fact, if the hypothesis were true, we would expect a well behaved 5% level t-test to reject it 5% of the time, and this is scarcely better than that!

In the right panel of Figure 1 we shade in red the cases where the 2SLS t-test rejects the null hypothesis that the true Frisch elasticity is equal to the true value of 0.597. The test rejects the null hypothesis 6.6% of the time. This is not so bad when viewed in isolation, as it is not too far from the correct rate of 5%. But more importantly, the rate of rejecting the true hypothesis that the Frisch equals 0.597 is actually ***greater*** than the rate of rejecting the false hypothesis that the Frisch equals 0. This is truly awful behavior for a statistical test.

Another notable aspect of the right panel of Figure 1 is that cases where we reject the null of Frisch = 0.597 are not evenly split between cases where the estimate is above and below the true value. In fact, all the rejections occur when the estimated Frisch elasticity is very small (near zero). This is a direct consequence of the positive covariance between 2SLS estimates and their standard errors. As large positive estimates of the Frisch elasticity have large standard errors, there is very little chance of concluding a large positive estimate is significant.

C. Anderson-Rubin Test Results

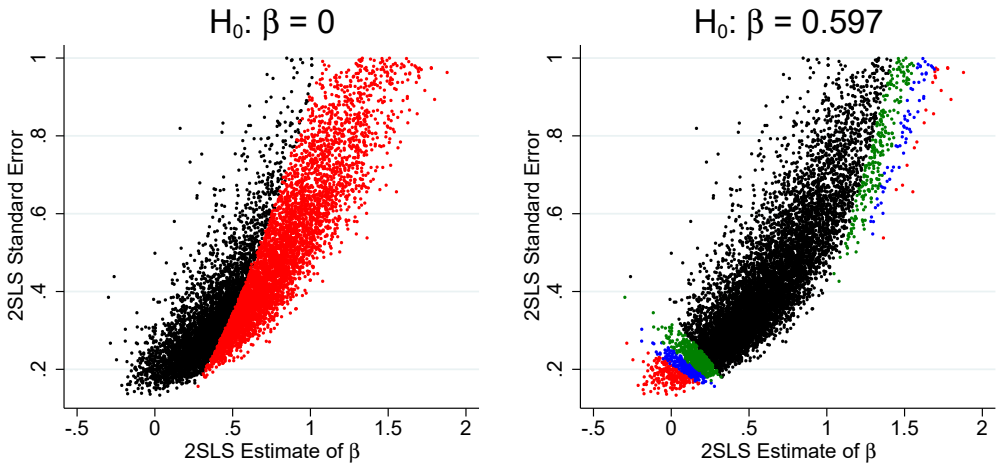
Figure 2 reports the same results for the AR test. The contrast with the t-test could not be more dramatic. The figure again plots the 2SLS standard errors against the 2SLS estimates, just as in Figure 1. But now we plot in red the cases where the AR test rejects the null hypothesis. In the left panel we see that the red region is quite large. The AR test rejects the false null that the Frisch is equal to zero 56.5% of the time. This is a good level of power that is more than ten times greater than the t-test.

The right panel shows the rate of rejecting the true null that the Frisch = 0.597. The AR test rejects 4.9% of the time, which is almost exactly equal to the correct 5% rate. Furthermore, we plot in blue and green the cases where 10% and 20% AR tests reject. These rates are 10% and 19.5%, so again almost perfect. This illustrates how the AR test is "robust" to weak instruments, meaning that is has correct size (rejection rates) even if instruments are weak or borderline.

The only limitation of the AR test is that it doesn't quite generate symmetric rejections when the estimates are above and below the true value. For example, of the 4.9% rejections in the 5% test, 3.6% occur when the estimate is below 0.597 and 1.3% occur when it is above. This is because, like the t-test, the AR test tends to attribute greater precision to estimates shifted in the direction of the OLS bias. But this problem is much less severe for the AR test than the t-test.

These results make it very obvious that in our empirical application the AR test provides a far more reliable guide to the significance of the estimate of the Frisch elasticity than does the t-test. Yet, in the vast literature on estimating the Frisch elasticity, we are not aware of any work that has used the AR test. The

FIGURE 2. STANDARD ERROR OF $\hat{\beta}_{2SLS}$ PLOTTED AGAINST $\hat{\beta}_{2SLS}$ ITSELF (AR TEST)



Note: Runs with standard error > 1 *are not shown. In the left panel, red dots indicate* $H_0 : \beta = 0$ *rejected at the 5% level using the AR test. In the right panel red dots indicate* $H_0 : \beta = 0.597$ *rejected at the 5% level using the AR test. Blue and green indicate 10% and 20%.*

consequence is that prior work that relied on 2SLS t-tests will have tended to obtain insignificant results even if the true Frisch elasticity is well above zero.

VII. Conclusion

The magnitude of the Frisch labor supply elasticity – how work hours respond to predictable wage changes – lies at the center of many economic policy debates, as the pure substitution effect measured by the Frisch is a vital input into tax policy. For example, higher values of the Frisch imply lower optimal tax rates on labor income. Because of its importance, there is a large literature estimating the Frisch elasticity using instrumental variable methods. But this literature has been plagued by weak instrument problems due to the fact that it is hard to find instruments that strongly predict wage growth. Hence the value of the Frisch remains a topic of intense debate.

Here we revisit that debate. Using the ASVAB ability test as an instrument for wage growth, we estimate a large Frisch elasticity of 0.597 for young men using data from the NLSY97. But, as is typical of this literature, the 2SLS standard error is 0.403, implying our estimate of the Frisch elasticity is very imprecise. Based on this, we can't even reject the hypothesis that it is zero at conventional levels – a result that is typical of many prior papers.

Importantly, the first stage F-statistic for our ASVAB instrument is 10.12, which is right on the borderline for whether weak instrument problems are a

concern. This is again typical of prior work on estimating the Frisch elasticity.

Econometric theory strongly suggests that if weak instruments are a concern, 2SLS t-tests are unreliable, and the Anderson-Rubin (AR) test should be used instead. The AR test is robust to weak instruments and it is efficient. When we implement the AR test, we find our estimate of the Frisch elasticity is significantly greater than zero at the 5% level. In fact, it is significant at the 3.5% level.

These contradictory results led us to conduct an experiment to evaluate the reliability of the t-test vs. the AR test. In our data environment we find the AR test has correct size and ten times the power of the t-test. In fact, the power of the t-test is so poor that a 5% level test is more likely to reject a hypothesis that the Frisch equals its true value than a false hypothesis that it equals zero.

Given the clear theoretical guidance, along with empirical results like these, it is difficult to understand why applied researchers have not widely adopted the AR test in preference to 2SLS t-tests.[13] When we use the appropriate inferential procedure (the AR test) we conclude the Frisch elasticity is fairly large (.597) and highly significant for young males in the NLSY97.

Our estimate of the Frisch elasticity for young men (.597) is quite consistent with values of 1.0 or more often used to calibrate macro models, once one considers the accumulating evidence that the Frisch elasticity increases substantially with age (see, e.g., Borella, De Nardi and Yang (2019), Erosa, Fuster and Kambourov (2016), French (2005) and Keane (2021)), as well as the clear evidence that it is greater for women than men (see Keane (2011)).

Of course, there have been numerous attempts to reconcile low 2SLS estimates of the Frisch elasticity using micro data with the large values often found in macro calibrations. The various approaches are detailed in Keane and Rogerson (2012, 2015). These reconciliations fall into two broad categories: One set of explanations, exemplified by Imai and Keane (2004), Low (2005) and Domeij and Floden (2006) takes issue with the specification of equation (1), arguing that more general models of labor supply (e.g., models that account for human capital or liquidity constraints) imply that estimation of this equation will give downward biased estimates of the Frisch elasticity. The other set of explanations, exemplified by Chang and Kim (2006) and Rogerson and Wallenius (2009), argue that, once one accounts for the participation margin of labor supply and aggregation issues, it is possible for the macro level Frisch elasticity to be large even if the micro level elasticity is small. More recently, Gottlieb, Onken and Valladares-Esteban (2021) have shown how a large macro level Frisch elasticity can be reconciled with modest reactions to tax holidays due to a combination of income and equilibrium

[13]We conjecture there are a few potential explanations: First, applied researchers may not be aware of how serious the problems with 2SLS t-tests are, or how superior the AR test can be. We hope the analysis presented here will help to shed light on these issues. Second, applied researchers may think AR tests are difficult to implement. That is obviously not true, but the econometric theory literature on AR tests presents them at such a high level of generality that it is indeed difficult for applied researchers to penetrate. Finally, applied researchers may be wedded to t-tests simply because they are so familiar. But we hope that inertia may be overcome so that empirical practice can be improved.

effects. These arguments are complementary to our argument here.

Our argument is new in that we criticise the micro-econometric literature on its own terms: Suppose the (strong) assumptions necessary for 2SLS estimation of equation (1) to deliver consistent estimates of the Frisch elasticity do hold. Even then, we show that the econometric methods that have been used to draw inferences from those estimates are inherently biased against finding that the Frisch elasticity is both large and significant. We hope our straightforward econometric argument will prove convincing to economists who have not been convinced by the more subtle theoretical arguments based on more complex labor supply models or aggregation issues.

REFERENCES

Altonji, J.G. 1986. "Intertemporal substitution in labor supply: Evidence from micro data." *Journal of Political Economy*, 94(3, Part 2): S176–S215.

Anderson, T.W., and H. Rubin. 1949. "Estimation of the parameters of a single equation in a complete system of stochastic equations." *Annals of Mathematical statistics*, 20(1): 46–63.

Andrews, I., J. Stock, and L. Sun. 2019. "Weak instruments in instrumental variables regression: Theory and practice." *Annual Review of Economics*, 11: 727–753.

Borella, M., M. De Nardi, and F. Yang. 2019. "Are marriage-related taxes and Social Security benefits holding back female labor supply?" National Bureau of Economic Research.

Bound, J., D. Jaeger, and R. Baker. 1995. "Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak." *Journal of the American Statistical Association*, 90(430): 443–450.

Chang, Y., and S. Kim. 2006. "From individual to aggregate labor supply: A quantitative analysis based on a heterogeneous agent macroeconomy." *International Economic Review*, 47(1): 1–27.

Conesa, J.C., S. Kitao, and D. Krueger. 2009. "Taxing capital? Not a bad idea after all!" *American Economic Review*, 99(1): 25–48.

Domeij, D, and M. Floden. 2006. "The labor-supply elasticity and borrowing constraints: Why estimates are biased." *Review of Economic dynamics*, 9(2): 242–262.

Erosa, A., L. Fuster, and G. Kambourov. 2016. "Towards a micro-founded theory of aggregate labour supply." *The Review of Economic Studies*, 83(3): 1001–1039.

French, E. 2005. "The effects of health, wealth, and wages on labour supply and retirement behaviour." *The Review of Economic Studies*, 72(2): 395–427.

Gottlieb, C., J. Onken, and A. Valladares-Esteban. 2021. "On the Measurement of the Elasticity of Labour." Working Paper, Swiss Institute of Empirical Economic Research.

Imai, S., and M.P. Keane. 2004. "Intertemporal labor supply and human capital accumulation." *International Economic Review*, 45(2): 601–641.

Keane, M.P. 2011. "Labor supply and taxes: A survey." *Journal of Economic Literature*, 49(4): 961–1075.

Keane, M.P. 2021. "Recent Research on Labor Supply: Implications for Tax and Transfer Policy." *Labour Economics*, 102026.

Keane, M.P., and R. Rogerson. 2012. "Micro and macro labor supply elasticities: A reassessment of conventional wisdom." *Journal of Economic Literature*, 50(2): 464–76.

Keane, M.P., and R. Rogerson. 2015. "Reconciling micro and macro labor supply elasticities: A structural perspective." *Annu. Rev. Econ.*, 7(1): 89–117.

Low, H.W. 2005. "Self-insurance in a life-cycle model of labour supply and savings." *Review of Economic dynamics*, 8(4): 945–975.

MaCurdy, T.E. 1981. "An empirical model of labor supply in a life-cycle setting." *Journal of political Economy*, 89(6): 1059–1085.

Moreira, M.J. 2009. "Tests with correct size when instruments can be arbitrarily weak." *Journal of Econometrics*, 152(2): 131–140.

Prescott, E.C. 2006. "Nobel lecture: The transformation of macroeconomic policy and research." *Journal of Political Economy*, 114(2): 203–235.

Rogerson, R., and J. Wallenius. 2009. "Micro and macro elasticities in a life cycle model with taxes." *Journal of Economic theory*, 144(6): 2277–2292.

Staiger, D., and J. Stock. 1997. "Instrumental variables regression with weak instruments." *Econometrica*, 65(3): 557–586.

Stock, J., and M. Watson. 2015. *Introduction to econometrics (3rd global ed.).* Pearson Education.

Stock, J., and M. Yogo. 2005. "Testing for weak instruments in linear IV regression." *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg*, 80(4.2): 1.