

# Cross-Section Regression Estimates of Labor Supply Elasticities:

## Procedures and Problems

Beginning in the 1970's labor economists tried to learn about how households' labor supply decisions respond to offered wages by running cross-section regressions in survey samples of (usually) thousands of individuals. The idea was that different people in a population face different prices ( $w$ ) for their labor and are endowed with different amounts of non-labor income ( $G$ ). By asking whether people who earn higher hourly wages work more hours, economists hoped to learn whether a "representative" household would respond to a permanently higher wage by working more or less (i.e. to estimate the sign and magnitude of the uncompensated labor supply elasticity).

This exercise generally had two main goals (the first of which was probably most prevalent):

1. *Estimating a parameter* (the size of the labor supply response to wages) that is useful to know for policy purposes, whether or not any particular theory of labor supply behavior—including the classic, static model—is correct. For example, labor supply elasticities tell us how much (or whether) tax cuts will stimulate extra work effort, and help us estimate who "really" bears the burden of payroll taxes.

2. *Testing the predictions* of the static labor supply model. Some practitioners (e.g. Ashenfelter and Heckman) used the estimated responses to  $G$  and  $w$  in the data to compute a pure substitution effect, and checked to see if it had the predicted sign.

Before going into the details, it is worth reflecting on the notion of estimating a price elasticity (in this case the price is the wage) from pure cross-section data. If the price in question were, say, the price of home heating oil and the sample was all the households in a city, this would be a pretty hopeless exercise. The reason is that, at any given point in time, competition between firms tends to ensure that all households pay the same price. There would be no price variation in your sample from which you could estimate how households behave when they face different prices.

In the labor supply case, different persons do face different prices for their labor in a cross-section of individuals, for various reasons such as different family backgrounds, intelligence, motivation, and past education decisions. So we *can* ask whether, at any point in time, high-wage people work more or fewer hours. But you should also immediately be suspect of this idea: high-wage people may work more hours than low-wage people because they differ from low-wage people in many ways other than just their wage, and controlling for all these other differences may not be that easy. So it is unclear we can really isolate a causal effect of wages on labor supply this way. We will come back to this problem under the heading of "omitted variables bias" below.

First, let's describe what a typical economist did. They had a sample of persons, all of whom did some work for pay in a certain reference period (say the last year). Thus, we observed, for each person:

- total hours worked ( $h$ )
- the hourly wage rate ( $w$ ); often calculated by dividing annual earnings by total hours worked
- some measure of nonlabor income ( $G$ ), generally including investment income, sometimes including government transfers, and in many cases including their spouse's earnings.
- a vector of demographic characteristics,  $X$ , such as age, race, education, presence of kids, etc. Generally people ran separate regressions for men and women, rendering a gender control unnecessary.

They then ran the following regression:

$$h_i = a + bw_i + cG_i + dX_i + e_i \quad (1)$$

where  $i$  indexes individuals, and  $h$ ,  $G$ , and  $w$  are measured in logs. This is done so that the estimated coefficients  $b$  and  $c$  directly take the form of elasticities, and are therefore independent of the units in which wages, income and labor supply are measured.

What results did they typically get? For both men and women, estimated income elasticities of labor supply ( $c$ ) from this approach are typically negative, though fairly small in magnitude (.2 or so). For women, wage elasticities of labor supply were positive and quite large (about 1); for men they were essentially zero. Estimated substitution effects were generally positive (in the sense that hours respond positively to wages if utility is held fixed), supporting the static labor supply model.

If the above estimates represent true, average behavioral responses that would occur in the U.S. labor force when offered wages are altered, a number of important implications would follow. For one, raising income tax rates paid by men is predicted not to reduce the hours they will work at all: so much for deleterious incentive effects of taxes! The same is *not* true for women though. But are such conclusions about the likely effects of policy changes really warranted from this evidence? Before jumping to conclusions, it might be wise to consider some possible econometric problems with the above procedure, as I do below. Please note: not only are these problems relevant to the estimation of labor supply responses, they are common problems that arise repeatedly in applied microeconomic work; it just happens that we are “meeting” them first in the context of labor supply because of the way this course is organized. So, while the illustrative context is labor supply, the lessons learned should be much broader.

**1. Functional form.** Static labor supply suggests the possibility (indeed the likelihood) of a nonmonotonic (“backward-bending”) labor supply curve. Yet equation 1 assumes the labor supply curve is linear (in logs). If, so, a zero “overall” effect could be masking an inverted-U-shaped relationship. How to fix this? Add a higher-order polynomial in  $w$  to the equation. This usually doesn’t make much difference.

**2. Parameter heterogeneity.** The model in equation (1) assumes that everyone’s  $b$  and  $c$  are the same. But this seems unlikely. There are three ways to deal with this, the first of which is to do nothing. In that case it can still be shown that the estimated values of  $b$  and  $c$  are weighted averages of the heterogeneous response parameters in the sample. The second is to estimate separate regressions for groups you believe might have very different parameters, such as women and men. The third is to introduce interaction terms between the  $X$ s on the one hand, and  $w$  and  $G$  on the other into the regression.

**3. Measurement error.** A well known result in basic regression analysis is that measurement error in the *dependent* variable, (in our case,  $h$ ), is not a cause of bias or inconsistency in the parameter estimates, so if this is the only variable that is imperfectly measured in the survey data, it is not a cause for concern (subject to an important proviso below).

But what if the hourly wage rate (a right-hand-side variable) is measured with error? If this measurement error is of the “classic” garden-variety white noise, our estimate of  $b$  is *biased towards zero*. Perhaps this explains the low estimated wage elasticities of men in the data?

The important proviso regarding measurement error in  $h$  has been dubbed “division bias” (see Borjas, Journal of Human Resources, 1979). It stems from the fact that, in practice, hourly wage rates in many cases are not measured directly. Instead they are calculated by dividing survey information on total earnings ( $E$ ) by hours worked ( $h$ ). In this case, measurement error in  $h$  will lead to bias in OLS estimates of  $b$  in (1).

Why? If measured hours,  $\hat{h}_i = h_i + u_i$ , where  $u_i$  is independent white noise, measured hourly wages are now  $\hat{w}_i = E_i / (h_i + u_i)$ . As a result there is a “built-in” negative correlation between measured wages and hours worked. Thus  $b$  is *biased downwards*. Think of it this way. Suppose the true  $b$  was zero: there is no causal effect of wages on hours at all. But whenever hours are randomly underestimated by the imperfect measurement technology, the calculated wage rate will be randomly overestimated (because  $h$  appears in the denominator). Thus our regression will, incorrectly “tell” us that labor supply falls with  $w$ , i.e. that  $b$  is negative.

Both the above sources of measurement-error bias can be addressed using instrumental-variables techniques, or (of course) by improving data quality.

**4. Omitted-variables bias.** Viewed in the context of labor supply theory, the reason for including the  $X$  variables in the labor supply equation is to control for factors, such as the presence of small children in the household, that might affect tastes for work. If these factors happen to be correlated with  $w$  or  $G$  in our sample, failing to include them in the regression could result in biased estimates of  $b$  or  $c$ . For example, suppose educated people (perhaps because their jobs tend to be more pleasant) tend to work more hours than less-educated people, even when offered exactly the same wage. In this case, failing to include education in the list of  $X$  variables would result in an overestimate of  $b$ , mistakenly attributing to the wage what is really an effect of higher education on the type of job occupied.

More seriously, it seems likely that there are other determinants of labor supply that could potentially be correlated with  $w$  or  $G$ , which we will never be able to measure well enough to include in the list of  $X$  variables. For one example, consider an unobserved taste for work you might label “drive” or “ambition”. If, over time, people with high values of this unobserved trait tend to win promotions at a faster rate than others (or they have accumulated more human capital because by worked more hours in the past), then in a cross-section of people at any point in time they will tend both to work longer hours ( $h$ ) and to have higher wages ( $w$ ). But this does not mean that the higher wages these people can command are causing them to work harder. Instead both long hours and high wages are driven by a third, unobserved factor. Thus,  **$b$  is biased upwards**. At the same time, such people (because they had higher incomes in the past as well as today, and therefore are likely to have more assets on average) will likely have higher nonwage incomes,  $G$ . This **biases  $c$  (which is negative) upwards, i.e. towards zero**.

One possible treatment for this kind of omitted variable bias is to find an instrumental variable that affects  $w$  but is arguably uncorrelated with tastes for work. In my opinion, no one has yet proposed such a variable in cross-section data. Another is to introduce individual fixed effects. This requires us to observe the same person at two different times, facing different wages. This, however, introduces dynamic considerations which are best postponed to our analysis of dynamic labor supply models.

**5. Sample Selection Bias.** I mentioned at the start of this discussion that regressions like this are essentially always run on a sample of people who did at least *some* work during the sample period. Thus those persons who choose not to work at all (or are involuntarily unemployed – another issue that warrants discussion; we’ll address this in the dynamic labor supply section) are simply excluded from the sample.

The reason for this is neither a good theoretical one nor absent-mindedness; it is purely practical. Labor supply theory says (entirely correctly) that people choose *whether* or not to work based on what wage they would earn *if* they worked. Generally, however, survey designers only ask respondents their wage or earnings if they actually worked in the reference period. I am not aware of any survey that asks people who voluntarily choose not to work how much they think they *would* command per hour if they did, and I’m not sure such data would be very reliable if it were collected. So, because one of the regressors ( $w$ ) is not observed for nonworkers, these people *can’t* be included in labor supply regressions (at least not in any obvious way: see below for attempted solutions).

Unfortunately, the fact that we have a “self-selected sample” also causes bias in the estimated coefficients,  $b$  and  $c$ . The easiest way to see this is to jointly model the decision *to* work and the number of work hours chosen using the latent variable  $h^*$ , which we can think of as “desired hours”. Unlike actual hours, which can only be zero or positive, desired hours can be any real number. Further, let:

$$h_i^* = a + bw_i + cG_i + dX_i + e_i; \quad (2)$$

$$h_i = h_i^* \text{ if } h_i^* \geq 0; \quad h_i = 0 \text{ if } h_i^* < 0.$$

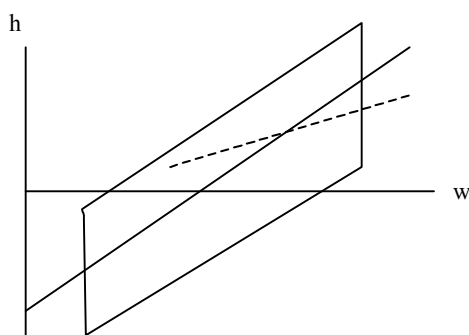
If we want to know the causal effect of  $w$  on a randomly-selected individual's labor supply, we need to know the parameters of (2).

Because only workers are in the sample, all observations in our data must satisfy the condition:

$$h_i^* > 0, \text{ or } e_i > -a - bw_i - cG_i - dX_i.$$

Inclusion in the sample thus depends on the realized value of the dependent variable,  $h$ , and therefore also on the realization of  $e$ . This automatically induces a correlation between the regressors ( $w$ ,  $G$  and  $X$ ) and the error term,  $e$ , again causing bias in OLS estimates of (2).

The bias is easily and intuitively signed in the case where there is only one regressor (say  $w$ ) using the following figure:



In the figure, the solid upward-sloping line represents the true regression line  $h_i^* = a + bw_i + e_i$  that we want to estimate (the figure assumes for the sake of argument that the true  $b > 0$ ). The parallelogram surrounding this line represents a homoscedastic scattering of observed data around the true line. If we observed all values of  $h^*$  (including the negative ones) this would be estimated without bias by OLS. However our sample contains only the data points that lie above the horizontal line defined by  $h = 0$ . Among the observed data points, note that the expectation of the error term,  $e$ , falls as  $w$  increases. (Intuitively, the only way a very low-wage person would ever get into our sample would be if he/she had very high unobserved tastes for work,  $e$ ; among high-wage people our sample is more representative). Thus  $\text{Cov}(w, e) < 0$ , with the consequence that  **$b$  is biased downwards**. Diagrammatically, fitting an OLS line through the scatter of points with  $h > 0$  gives the dashed line, whose slope is less than  $b$ .

Intuitively, the dashed line in the above figure (the line that OLS estimates) is not without interest or meaning. Every regression, no matter how poorly specified, yields an unbiased estimate of *some* conditional expectation, and in this case this is *the average work hours of the subset of the sample who would voluntarily choose to work at a given wage,  $w$* . This average increases with  $w$  at a rate that is smaller than  $b$  because, as  $w$  rises, more and more individuals with low tastes for work (low  $e$ 's) are induced to join the sample of workers. This draws average hours in the subsample of workers down. Thus the slope that is estimated by OLS is a composite of two effects: a causal effect of wages on hours,  $b$ , and a composition effect whereby wages change the mix of people who choose to work.

**IN SUM**, because of this long list of potential biases, cross-section OLS estimates may or may not provide good, policy-relevant estimates of the causal effects of wages and nonlabor income on individuals' labor supplies. As a result, a great deal of attention has been devoted to finding good "solutions" to the above econometric problems in the labor supply case, as well as in other cases. More recently, attention has also been devoted to alternative methods of estimating labor supply elasticities. Before we discuss these, however, we will spend some more time on possible treatments for *sample selection bias*, because of the central role it plays in so many economic contexts.

**How to fix sample selection bias (i.e. how to estimate the parameters of (2))?**

**One solution that *won't* work:**

Use another regression, say:

$$w_i = \alpha + \beta Z_i + v_i \quad (3)$$

to *impute* wages to nonparticipants based on characteristics,  $Z$ , such as age, education, etc. Then include these people in the labor supply regression with their imputed wages.

There are two problems with this:

1. The wage regression, (3) can only be estimated on the sample of workers. If decisions to work depend at all on the wage, it *must* therefore also be subject to sample selection bias. (more intuitively, it isn't clear the sample of workers is representative enough of nonworkers that we can use wages earned by the former to estimate wages that would be earned by the latter, should they choose to work).
- 2.. Even if we could somehow solve problem 1 above, the problem remains that we don't observe  $h^*$  for nonworkers. If we assign them  $h = 0$  instead of  $h^*$  we'll have a bunch of data lined up on the horizontal axis in the above figure, and  $b$  will still be biased towards zero.

**A solution that *will* work** (under certain assumptions) is Heckman's approach. (the classical selection model, or "Heckit" regression).

## Sample Selection Bias, The Classical Selection Model, and “Heckit” Regression

### 1. Setting up the model

In a large number of cases (of which labor supply is only one), regressions are estimated on a sample where the probability of selection into the sample is correlated with the realized value of the dependent variable. A quarter century ago, Jim Heckman introduced a way of thinking about the biases that arise in this situation that has come to be called the “classical selection model”. He also proposed a solution that has come to be called “Heckit” regression. In 2000, he was awarded the [Nobel Prize in Economics](#) for this and related work.

In what follows I describe the model and technique as it applies to estimating labor supply and wage equations when the decision to work (i.e. to be in the sample in the first place) can, of course, depend on the wage. The treatment is close to that in Heckman’s July 1974 *Econometrica* article.

Let desired hours in a cross section of persons be given by:

$$h_i^* = \delta_0 + \delta_1 w_i + \delta_2 Z_i + \varepsilon_i, \quad (1)$$

where  $Z$  includes non-labor income ( $G$ ) as well as taste variables such as number of kids, etc. Think of  $\varepsilon_i$  as unobserved “tastes for work” (an unobserved, person-specific factor that makes person  $i$  work more or fewer hours than other observationally-identical persons). We will refer to (1) as the *structural labor supply equation*; it represents the behavioral response of an individual’s labor supply decision to his/her economic environment and one of our ultimate goals is to estimate  $\delta_i$ .

Suppose the market wage that person  $i$  can command is given by:

$$w_i = \beta_0 + \beta_1 X_i + \mu_i \quad (2)$$

where  $X$  includes productivity and human capital variables such as age, education, etc. In practice there may be considerable overlap between the variables in  $X$  and  $Z$ . It may be helpful to think of  $\mu_i$  as “unobserved (wage-earning) ability” here. We will refer to (2) as the *structural wage equation*.

In the above situation we already know that OLS estimates of either (1) or (2) on the sample of workers only will be biased (in the case of (1) because the sample includes only those with positive hours; in the case of (2) because the sample includes only those with wages above their reservation wage). So how can we formalize the nature and size of these biases, and obtain unbiased estimates of the  $\delta$ ’s and  $\beta$ ’s<sup>1</sup>?

We begin by substituting (2) into (1), yielding:

$$h_i^* = \delta_0 + \delta_1[\beta_0 + \beta_1 X_i + \mu_i] + \delta_2 Z_i + \varepsilon_i$$

or, 
$$h_i^* = [\delta_0 + \delta_1 \beta_0] + \delta_1 \beta_1 X_i + \delta_2 Z_i + [\varepsilon_i + \delta_1 \mu_i]$$

or just: 
$$h_i^* = \alpha_0 + \alpha_1 X_i + \alpha_2 Z_i + \eta_i \quad (3)$$

where  $\alpha_0 \equiv \delta_0 + \delta_1 \beta_0$ ;  $\alpha_1 \equiv \delta_1 \beta_1$ ;  $\alpha_2 \equiv \delta_2$ ;  $\eta_i \equiv \varepsilon_i + \delta_1 \mu_i$ . We’ll refer to equation (3) as the *reduced form hours equation*.

---

<sup>1</sup> Although our original motivation for this exercise was to get an estimate of  $\delta_1$ , obtaining unbiased estimates of the  $\beta$ ’s, i.e. of the effects of factors such as education, experience, race, union status, or participation in a job training program on *wages* in (2) is also a major area of research in labor economics. Because we can only measure wages for workers, sample selection can be an important source of bias in these contexts as well.

As a final step in setting up the problem, note that given our assumptions individual  $i$  will work a positive number of hours iff:

$$\text{Work iff: } h_i^* > 0; \text{ i.e. } \eta_i > -\alpha_0 - \alpha_1 X_i - \alpha_2 Z_i \quad (4).$$

Just to keep track of the intuition, note that conditional on observables ( $X$  and  $Z$ ) either high unobserved tastes for work ( $\varepsilon$ ) or (provided  $\delta_1 > 0$ ) high unobserved wage-earning ability ( $\mu_i$ ) tend to put people into the sample of workers.

Next, to greatly simplify matters, assume that the two underlying error terms ( $\varepsilon_i$  and  $\mu_i$ ) follow a joint normal distribution (lots of work has been done since to generalize this assumption). Note that (a) it therefore follows that the “composite” error term  $\eta_i$  is distributed as a joint normal with  $\varepsilon_i$  and  $\mu_i$ ; and (b) we have *not* assumed that  $\varepsilon_i$  and  $\mu_i$  are independent. In fact, it seems plausible that work decisions and wages could have a common unobserved component; indeed one probably wouldn’t have much confidence in an estimation strategy that required them to be independent.

## 2. Characterizing the bias in OLS estimates of (1) and (2).

Let’s start with the labor supply equation, (1). OLS estimates of (1) will be unbiased if the expectation of the error term equals zero, regardless of the values taken by the regressors, ( $X$ ). Is this the case? Let’s work it out, recalling that an observation is in the sample iff (4) is satisfied for that observation.

$$\begin{aligned} E(\varepsilon_i | h_i > 0) &= E(\varepsilon_i | \eta_i > -\alpha_0 - \alpha_1 X_i - \alpha_2 Z_i) \\ &= \frac{\sigma_{\varepsilon\eta}}{\sigma_\eta^2} E(\eta_i | \eta_i > -\alpha_0 - \alpha_1 X_i - \alpha_2 Z_i) \\ &= \frac{\sigma_{\varepsilon\eta}}{\sigma_\eta^2} \cdot \frac{\phi\left(\frac{-\alpha_0 - \alpha_1 X_i - \alpha_2 Z_i}{\sigma_\eta}\right)}{1 - \Phi\left(\frac{-\alpha_0 - \alpha_1 X_i - \alpha_2 Z_i}{\sigma_\eta}\right)} \\ &\equiv \theta_1 \lambda_i \end{aligned} \quad (5)$$

where  $\Phi$  is the standard normal cdf and  $\phi$  is the standard normal density function. The second and third lines of this derivation follow from properties of the normal distribution.

How should we understand the two terms in equation (5)? The first term,  $\theta_1$ , is a parameter that does not vary across observations. It is the coefficient from a regression of  $\eta_i$  on  $\varepsilon_i$ ; therefore of  $\varepsilon_i + \delta_1 \mu_i$  on  $\varepsilon_i$ . Unless  $\delta_1$  (the true labor supply elasticity) is zero or negative, or there is a strong negative correlation between underlying tastes for work,  $\varepsilon_i$  and wage-earning ability,  $\mu_i$ , this will be positive. In words, conditioning on observables, people who are more likely to make it into our sample –i.e. have a high  $\eta_i$  -- will on average have a higher residual in the labor supply equation,  $\varepsilon_i$ ).

The second term in (5),  $\lambda_i$ , has an  $i$  subscript and therefore varies across observations. Mathematically, it is the ratio of the normal density to one minus the normal cdf (both evaluated at the same point, which in turn depends on  $X$  and  $Z$ ); this ratio is sometimes called the *inverse Mills ratio*. For the normal distribution –**and this is a property of the normal distribution only**—this ratio gives the *mean of a truncated normal distribution*. Specifically, the standard normal distribution has the following property: If  $x$  is a standard normal variate,  $E(x | x > a) = \phi(a)/(1 - \Phi(a))$ .

Now that we have an expression for the expectation of the error term in the structural labor supply equation (1) we can write:

$$\varepsilon_i = E(\varepsilon_i | h_i > 0) + \varepsilon'_i = \theta_1 \lambda_i, \text{ where } E(\varepsilon'_i) = 0.$$

In a sample of participants, we can therefore write (1) as:

$$h_i^* = \delta_0 + \delta_1 w_i + \delta_2 Z_i + \theta_1 \lambda_i + \varepsilon'_i \quad (1')$$

Call this the *augmented labor supply equation*. It demonstrates that we can decompose the error term in a selected sample into a part that potentially depends on the values of the regressors ( $X$  and  $Z$ ) and a part that doesn't. It also tells us that, *if we had data on  $\lambda_i$*  and included it in the above regression, we could estimate (1) by OLS and not encounter any bias. In what follows, we'll discuss how to get *consistent estimates* of  $\lambda_i$ , which for some purposes are just as good. Thus, one can think of *sample selection bias* as [a specific type of] *omitted variables bias*, which happens to be the title of another one of Heckman's famous articles on this topic.

Before moving on to how we "solve" the sample selection bias problem for the labor supply equation (1), let's discuss how the above reasoning applies to the market wage equation (2). Following the same train of reasoning,

$$\begin{aligned} E(\mu_i | h_i > 0) &= E(\mu_i | \eta_i > -\alpha_0 - \alpha_1 X_i - \alpha_2 Z_i) \\ &= \frac{\sigma_{\mu\eta}}{\sigma_\eta^2} E(\eta | \eta > -\alpha_0 - \alpha_1 X_i - \alpha_2 Z_i) \\ &= \frac{\sigma_{\mu\eta}}{\sigma_\eta^2} \cdot \frac{\phi\left(\frac{-\alpha_0 - \alpha_1 X_i - \alpha_2 Z_i}{\sigma_\eta}\right)}{1 - \Phi\left(\frac{-\alpha_0 - \alpha_1 X_i - \alpha_2 Z_i}{\sigma_\eta}\right)} \\ &\equiv \theta_2 \lambda_i \end{aligned} \quad (6)$$

Note that the  $\lambda_i$  in (6) is exactly the same  $\lambda_i$  that appeared in equation (5). The parameter  $\theta_2$  is the coefficient from a regression of  $\eta_i$  on  $\varepsilon_i$ ; therefore of  $\varepsilon_i + \delta_1 \mu_i$  on  $\mu_i$ . As before, unless  $\delta_1$  (the true labor supply elasticity) is zero or negative, or there is a strong negative correlation between  $\varepsilon_i$  and  $\mu_i$ , this will be positive (on average, conditioning on observables, people who are more likely to make it into our sample -- i.e. have a high  $\eta_i$  -- will have a higher residual in the wage equation,  $\mu_i$ ).

Equation (6) allows us to write an *augmented wage equation*:

$$w_i = \beta_0 + \beta_1 X_i + \theta_2 \lambda_i + \mu'_i, \text{ where } E(\mu'_i) = 0. \quad (2')$$

Thus, data on  $\lambda_i$  would allow us to eliminate the bias in wage equations fitted to the sample of workers only.

### 3. Obtaining unbiased estimates of the structural hours and wage equations .

When (as we have assumed) all our error terms follow a joint normal distribution, the reduced form hours equation (3) defines a *probit* equation where the dependent variable is the dichotomous decision of whether to work or not (i.e. whether to be in the sample for which we can estimate our wage and hours equations). Note that all the variables in this probit (the  $X$ 's,  $Z$ 's and *whether* a person works) are observed for both workers and nonworkers. Thus we can estimate the parameters of this equation consistently. In particular (recall that the variance term in a probit is not identified) we can get consistent estimates of  $\alpha_0/\sigma_\eta$

,  $\alpha_1/\sigma_{\eta}$  and  $\alpha_2/\sigma_{\eta}$ . Combined with data on the  $X$ 's and  $Z$ 's, these estimates allow us to calculate an estimated  $\lambda_i$  for each observation in our data. (We can of course calculate it both for the workers and the nonworkers but we'll only actually make use of the values calculated for the workers).

Now that we have consistent estimates of  $\lambda_i$  we can include them as regressors in a labor supply equation estimated on the sample of participants only. Once we do so, the expectation of the error term in that equation is identically zero, so it can be estimated consistently via OLS. We can do the same thing in the wage equation. This procedure is known as *Heckman's 2-step estimator*. When we implement this, we will as a matter of course get estimates of the  $\theta$  parameters ( $\theta_1$  in the case where the second stage is an hours equation;  $\theta_2$  in the case where the second stage is an hours equation). These in turn provide some information about the covariance between the underlying error terms  $\varepsilon_i$  and  $\mu_i$ .

Before leaving this topic, note the generality of this technique. Whenever we are running a regression on a sample where there is a *possible* (or likely) correlation between the realization of the dependent variable and the likelihood of being in the sample, one can (in principle) correct for sample selection bias by (1) estimating a *reduced-form probit* in a larger data set where the dependent variable is inclusion in the subsample of interest; then (2) estimating the regression in the selected sample with an extra regressor,  $\lambda_i$ . According to the reasoning above, including this extra regressor should eliminate any inconsistency due to nonrandom selection into our sample.

#### 4. Estimation Techniques

First, a note on efficiency. Two-stage estimates of Heckman's model, as described above, are inefficient for two reasons. First, even though  $E(\varepsilon'_i) = 0$ , it is straightforward to show that, given the structure of our model,  $Var(\varepsilon'_i)$  is not independent of the regressors in (1'). The same is true for (2'). Thus the error term, given the assumptions of our model, *must* be heteroscedastic. Using this information, in principle, should improve efficiency. Second, the very nature of the two-step procedure does not allow for sample information about the "second-stage" outcome (in this case either hours or wages) to be used in estimating the parameters of the first stage (the reduced-form probit). Together, these considerations argue in favor of estimating the entire model described in the previous sections via Full-Information-Maximum-Likelihood (FIML). Indeed, since the model assumes joint normality to begin with, no additional assumptions are required for the FIML model than are already made in the two-step model.

Heckman's technique has been widely used and abused since its introduction. Both the two-step and FIML versions are now widely available in commercial statistics packages. It is fair to say, however, that practitioners have become widely disillusioned with the method, in part because—especially when estimated via FIML—parameter estimates can be highly unstable and sensitive to the precise list of  $X$  and  $Z$  variables that are used. Understanding why this occurs is closely related to the issue of *identification* in the classical selection model, an issue to which we turn next.

#### 5. Identification in the Classical Selection Model.

To see why identification can be marginal at best in the classical selection model, I will now make two relatively minor modifications to the classic selection model. Under these modifications, the parameters of the model are not identified (i.e. it is theoretically *impossible* to obtain consistent estimates of (1) and (2), i.e. to disentangle the effects of causation and selection). Understanding why this occurs helps us understand why, in practice, Heckit regressions yield such unstable results, and why they are so often misused or misinterpreted.

The first modification (actually this isn't a modification at all; it's just a special case of the model) is to set  $X = Z$ . In other words, assume that exactly the same set of regressors enters the wage and hours equations. Since the two vectors are the same, we'll just use  $X$  to denote both in what follows. The second

modification is to change our joint normality assumption in a very particular way, as follows (this is done not for realism but to make a point):

Recall that in the joint normal case,  $\lambda_i$  can be expressed as  $f(I_i)$ , where  $f$  is a monotonically increasing, nonlinear function, and the “participation index”,  $I_i = \alpha_0 + \alpha_1 X_i + \alpha_2 Z_i$  is a linear function of observables (see for example equation 5). When  $X=Z$ , we can therefore write  $\lambda_i = f(\alpha_0 + \alpha_1 X_i)$ . Now suppose (for the sake of argument) that the joint distribution of unobservables (rather than being normal) was such that  $f$  was linear, i.e. suppose we can write  $f(I_i) = a + b I_i$ ; therefore  $\lambda_i = a + b\alpha_0 + b\alpha_1 X_i$ .<sup>2</sup>

In what follows I will suppose (purely for illustrative purposes) that our objective is to get an unbiased estimate of the structural wage equation (2). (We could just as easily have focused on (1) instead). To see the implications of these new assumptions, substitute our new definition of  $\lambda_i$  into the augmented wage equation (2'), yielding:

$$w_i = \beta_0 + \beta_1 X_i + \theta_2 \lambda_i + \mu'_i = \beta_0 + \beta_1 X_i + \theta_2 [a + b\alpha_0 + b\alpha_1 X_i] + \mu'_i \quad (7)$$

Now, none of the parameters ( $\beta$ ) of the structural wage equation are identified. To see this, imagine you calculated your  $\lambda$ 's using the linear formula  $\lambda_i = a + b\alpha_0 + b\alpha_1 X_i$ , then asked the computer to run the regression in (7). It would be unable to do so, because one of your explanatory variables,  $\lambda$ , is a linear combination of all the others ( $X$ ). In this situation, all we can hope to do is to regress  $w$  on our full set of  $X$ 's, a regression that can be written:

$$w_i = \beta_0 + \theta_2 [a + b\alpha_0] + [\beta_1 + \theta b\alpha_1] X_i + \mu'_i$$

The estimated effects of  $X$  on  $w$  represent a combination of causal ( $\beta_1$ ) and sample composition ( $\theta b\alpha_1$ ) effects that cannot be disentangled.

What are the practical implications of this? Unless you are very confident that your error terms are joint normally distributed (i.e. that  $f(I_i)$  is NOT linear), **identification of selection effects requires that at least one variable be contained in  $Z$**  (the equation determining selection into the sample for analysis) **that is not in  $X$**  (the “outcome” equation of interest). By analogy to identifying simultaneous equation systems, variables of the latter type are sometimes referred to as *instruments*. If you are running a Heckit model without an instrument, i.e. with the same list of variables in both equations, we say you are achieving *identification via functional form*. In other words, you are relying on an assumption of nonlinearity (of a very particular form) in the function  $f(I)$  to distinguish selection from causation. Perhaps unsurprisingly, this often leads to unstable and unbelievable results.

One last, really important point. Just *mechanically* deciding to exclude one variable from the outcome equation that is included in the participation equation won't solve this problem. For Heckit to work, **you really need to have advance knowledge (or a strong, credible prior belief) that the excluded instrument truly has no causal effect on the outcome of interest (nor can the excluded variable be correlated with unobservable determinants of the outcome ( $\mu_i$ ))**. This is your *identifying assumption*. It **cannot** be tested.<sup>3</sup> To illustrate, here is an example that comes up frequently: One variable that is

<sup>2</sup> Even in the normal case  $f$  is approximately linear over much of its range.

<sup>3</sup> If you have *more than one* instrument there are certain “overidentification tests” that can be done, but if you have only one IT CANNOT BE TESTED. Got that? Once again: IT IS LOGICALLY IMPOSSIBLE TO TEST SUCH A RESTRICTION. You can't regress the outcome on the instrument, look to see if there is a correlation, and if there is none, conclude you have a valid instrument. That's because, as we have already shown, the coefficient in this outcome regression is biased unless you have a valid instrument. If you don't *know* in advance whether your instrument is valid, you can't use the results of a regression (which is unbiased only *if* your instrument is valid) to check *whether* your instrument is valid. I belabor this point because, 30 years after Heckman, so many applied researchers still don't get it.

sometimes used as an instrument in estimating wage equations for married women is the presence of young children (say under the age of 5) in the home. The argument is that children affect one's current taste for working ( $\varepsilon_i$ ) by raising the value of home time, but not one's value to an employer (i.e. it is uncorrelated with  $\mu_i$ ). Given this assumption, Heckit regressions will *interpret* any correlation between wages and the presence of kids in our data as a consequence of sample selection. They do this because we have ruled out, *by assumption*, all other possible ways that kids could affect wages. For example, suppose that (controlling for the other  $X$ 's in the outcome equation) working women with kids earn lower hourly wages than working women without kids in our data. Heckit will *attribute* this to the notion that, when kids are present, a smaller number of women work and that smaller group has systematically different wage-earning ability than the broader sample of women. In fact, Heckit will argue (via the sign of the estimated  $\theta_2$  coefficient) perhaps counterintuitively that when kids arrive, it is disproportionately the *less* "able" (those with lower wage residuals) who tend to stay in the labor market. If you have confidence in your identifying restriction, that's absolutely fine.

But suppose we weren't 100% confident that the presence of kids can't possibly have a causal effect on the wages a woman can command in the labor market, or that the presence of kids is uncorrelated with unobserved determinants of wages. Perhaps caring for kids takes energy that reduces productivity, which in turn reduces the wages firms are willing to pay. Perhaps employers discriminate against women with kids because they expect less job commitment from those women. Perhaps women who are less "able" to command high wages in the labor market choose to have more kids. If any of these are true, then *these processes*, rather than the sample selection effect described above, could explain the observed negative association between kids and wages. The "kids" variable can't be used as an instrument to help us correct for selectivity bias.

Once you see how critically it depends on identifying assumptions, the apparent "magic" via which the Heckit technique solves the problem of sample selection bias doesn't seem so magic after all.