SAMPLE SELECTION AND MISSING DATA

Econometric Analysis of Cross Section and Panel Data, 2e
MIT Press
Jeffrey M. Wooldridge

1. Introduction
2. When Can Sample Selection Be Ignored?
3. Selection on the Response Variable: Truncated Regression
4. Incidental Truncation: A Probit Selection Equation
5. Incidental Truncation: A Tobit Selection Equation

1. INTRODUCTION

- Now turn to the problem of using only a subset of a random sample obtained from a well-defined population (presumably, the one of interest).
- Obvious but important point: There cannot be an issue of nonrandom sample selection if a random sample has been obtained from a given population. The population is not immutable. We can choose a population of interest from a bigger population.

- For example, if we are interested in the effect of a job training program on a population of men with poor labor market histories, we can define the population based on observed past labor market outcomes, such as unemployment status or labor earnings. If we can collect a random sample from the defined population, we just apply standard methods.
- Sample selection becomes an issue when the sample we can obtain are not representative of the population of interest.

- As an example, suppose we are interested in a wealth equation,

$$wealth = \beta_0 + \beta_1 plan + \beta_2 educ + \beta_3 age + \beta_4 income + u$$

which describes the population of all families in the United States (where *educ* and *age* are for the self-described "household head"). If we assume that u has zero mean and is uncorrelated with each explanatory variable, we would use OLS if we have a random sample from the population.

- Suppose, though, that only people less than 65 years old were sampled. What if we use OLS on the **selected sample**?

- As we will see, OLS on the nonrandom sample nevertheless consistently estimates the β_j provided

$$E(u|plan, educ, age, income) = 0.$$

- Zero correlation is not enough! Must have the conditional mean correctly specified. This falls under "exogenous sampling."

- Next suppose that only families with wealth greater than zero are included in the sample. Now, the data are selected on the basies of the response variable, wealth. As we will see, using standard methods (including OLS) on such as sample leads to biased and inconsistent estimators of the β_j, even under the zero conditional mean assumption.

• A different setup is when sample selection is not a deterministic function of either they x_j or y, but it may be related to them. This includes the problem of missing data, where data are missing on one ore more elements of $(\mathbf{x}, y)$ for some units drawn randomly from the population.

• Another example is when y is observed only when a certain event is true. A leading example is when y is $\log(wage^o)$, the log of the "wage offer" – the hourly wage someone could get paid if in the work force. We observe $wage^o$ only if the person decides to enter the work force.

- Generally called the problem of **incidental truncation**.
- The hallmark of the incidental truncation problem is the notion of "self-selection." For example, we only observe the wage offer if the person "self-selects" into the workforce.
- Whether someone chooses to report, say, their annual income has a self-selection component.

2. WHEN CAN SAMPLE SELECTION BE IGNORED?

Linear Model

- Assume there is a population represented by the random vector $(\mathbf{x}, y, \mathbf{z})$, where $\mathbf{x}$ is a $1 \times K$ vector of explanatory variables, y is the scalar response variable, and $\mathbf{z}$ is a $1 \times L$ vector of instrumental variables.

- Standard linear model with (possibly) endogenous explanatory variables:

$$
\begin{aligned}
y &= \beta_1 + \beta_2 x_2 + \ldots + \beta_K x_K + u = \mathbf{x}\boldsymbol{\beta} + u \\
E(\mathbf{z}' u) &= \mathbf{0},
\end{aligned}
$$

with $x_1 \equiv 1$ (so z_1 is almost certainly equal to unity, too).

- Given a random sample from the population, we can, under a rank condition, use 2SLS to consistently estimate $\boldsymbol{\beta}$.
- Unfortunately, the rank condition (essentially *rank* $E(\mathbf{z}'\mathbf{x}) = K$) and $E(\mathbf{z}'u) = \mathbf{0}$ are not usually enough to consistently estimate $\boldsymbol{\beta}$ with a selected sample.
- A leading special case is $\mathbf{z} = \mathbf{x}$, in which case the explanatory variables are assumed to be uncorrelated with the error.

- Analysis is simplified by thinking of drawing units randomly from the population, but now the random draw for unit i, $(\mathbf{x}_i, y_i, \mathbf{z}_i)$, is supplemented by drawing a **selection indicator**, s_i. By definition, $s_i = 1$ if unit i is used in the estimation, and $s_i = 0$ if we do not use random draw i.
- Therefore, our "data" consists of $\{(\mathbf{x}_i, y_i, \mathbf{z}_i, s_i) : i = 1, \ldots, N\}$, where the value of s_i determines whether we observe all of $(\mathbf{x}_i, y_i, \mathbf{z}_i)$.
- Because identification is properly studied in the population, let s denote a random variable with the distribution of s_i for all i. In other words, $(\mathbf{x}, y, \mathbf{z}, s)$ now represents the population.

- To determine the properties of any estimation procedure using the selected sample, we need to know about the distribution of s and its dependence on $(\mathbf{x}, y, \mathbf{z})$.
- Consider the algebraically simpler case of just identification (in the population!), that is, $L = K$. Let $\{(\mathbf{x}_i, y_i, \mathbf{z}_i, s_i) : i = 1, \ldots, N\}$ be a random sample from the population.

- The IV estimator using the selected sample can be written as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{IV} &= \left(N^{-1}\sum_{i=1}^{N} s_i \mathbf{z}_i'\mathbf{x}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N} s_i \mathbf{z}_i' y_i\right) \\ &= \boldsymbol{\beta} + \left(N^{-1}\sum_{i=1}^{N} s_i \mathbf{z}_i'\mathbf{x}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N} s_i \mathbf{z}_i' u_i\right).\end{aligned}$$

- In the statistics literature, often called the “complete case” estimator.
- By the law of larger numbers for random samples,

$$plim_{N\to\infty}(\hat{\boldsymbol{\beta}}_{IV}) = \boldsymbol{\beta} + [E(s\mathbf{z}'\mathbf{x})]^{-1}E(s\mathbf{z}'u).$$

- Weak assumptions sufficient for consistency are

$$rank\, E(\mathbf{z}'\mathbf{x}|s = 1) = K$$

and

$$E(s\mathbf{z}'u) = \mathbf{0},$$

• For the general 2SLS case, the conditions are only slightly more complicated. Regularity conditions, such as finite second moments, are assumed to hold. Then the conditions are

$$\begin{aligned} E(s\mathbf{z}'u) &= \mathbf{0} \\ rank\, E(\mathbf{z}'\mathbf{z}|s = 1) &= L \\ rank\, E(\mathbf{z}'\mathbf{x}|s = 1) &= K \end{aligned}$$

• These ensure also that the 2SLS estimator using the selected sample is $\sqrt{N}$ –asymptotically normal.

• Practically, for the rank condition to hold on the subpopulation, we should have it holding in the population and then the subpopulation not being "to small."

• More interesting is $E(s\mathbf{z}'u) = \mathbf{0}$. Holds is when s is independent of $(\mathbf{z}, u)$ along with zero correlation n the population:

$$E(\mathbf{z}'u) = \mathbf{0}.$$

Why? If s is independent of $(\mathbf{z}, u)$ then

$$E(s\mathbf{z}'u) = E(s)E(\mathbf{z}'u) = \rho \cdot \mathbf{0} = \mathbf{0}$$

where $\rho = E(s)$ is the (unconditional) probability that a randomly draw observation is kept.

- In statistics, if s is independent of $(\mathbf{x}, y, \mathbf{z})$, the data are said to be **missing completely at random**.
- Another sufficient condition is

$$E(u|\mathbf{z}, s) = E(u|\mathbf{z}) = 0,$$

where the second equality would be a strengthening of the exogeneity requirement on the instruments. The first equality rules out correlation between s and u.

- Sufficient for this latter condition is $E(u|\mathbf{z}) = 0$ and s is a deterministic function of $\mathbf{z}$, say $s = h(\mathbf{z})$. Then $E(u|\mathbf{z}, s) = E(u|\mathbf{z}, h(\mathbf{z})) = E(u|\mathbf{z}).$ This is the case of **exogenous sampling**.

• With $\mathbf{z} = \mathbf{x}$, a sufficient condition is

$$E(y|\mathbf{x}, s) = E(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta},$$

which means s can be an abitrary function of the exogenous variables. The rank condition is that $E(\mathbf{x}'\mathbf{x}|s = 1)$ has rank K.

• Generally, though, linear projections are not consistently estimated using a selected sample when s is a function of $\mathbf{x}$. In other words, even with exogenous sampling we must use a conditional mean assumption in the underlying population.

- If $y = \mathbf{x}\boldsymbol{\beta} + u$, $E(\mathbf{x}'u) = \mathbf{0}$, and s is independent of $(\mathbf{x}, y)$, then OLS using $s_i = 1$ is consistent for $\boldsymbol{\beta}$.
- The cases with $\mathbf{x}$ exogenous and with instruments are very important for sample selection corrections. If we can obtain an equation where the selection indicator is a function of the explanatory variables (or instruments), we can apply OLS or 2SLS to that equation for consistent estimation.

- Application of previous results. Suppose the population model is

$$y = \mathbf{x}\boldsymbol{\beta} + u$$
$$E(u|\mathbf{x}) = 0$$

and s is correlated with u. But suppose s is a determinstic function of $(\mathbf{x}, v)$ for a variable v. Further, suppose (u, v) is independent of $\mathbf{x}$. Then

$$E(y|\mathbf{x}, v) = \mathbf{x}\boldsymbol{\beta} + E(u|\mathbf{x}, v) = \mathbf{x}\boldsymbol{\beta} + E(u|v)$$

where the last equality follows by the independence assumption.

- Suppose v also has zero mean, and

$$E(u|v) = \gamma v.$$

Then

$$E(y|\mathbf{x}, v) = \mathbf{x}\boldsymbol{\beta} + \gamma v.$$

Now, because s is a function of $(\mathbf{x}, v)$, we can use OLS of y_i on $\mathbf{x}_i, v_i$ using the selected sample ($s_i = 1$) to consistently estimate $\boldsymbol{\beta}$ and γ. Notice that all variables, include v_i, only need to be observed when $s_i = 1$.

• In effect, controlling for v in the regression on the selected sample solves the sample selection problem. We will use this result, and the IV version of it, later.

• In practice, v depends on unknown parameters that have to be estimated in a first stage.

• Notice that we could assume, say, $E(u|v) = \gamma_1 v + \gamma_2(v^2 - \sigma_v^2)$, where $\sigma_v^2 = E(v^2)$, and use a very similar approach.

Nonlinear Models

- Suppose we know, for a parametric function $m(\cdot,\cdot)$,

$$E(y|\mathbf{x}) = m(\mathbf{x},\boldsymbol{\beta}_o),$$

and suppose that selection is exogenous in the sense that

$$E(y|\mathbf{x},s) = E(y|\mathbf{x}).$$

- The NLS estimator on the selected sample solves

$$\min_{\boldsymbol{\beta}} N^{-1}\sum_{i=1}^{N} s_i[y_i - m(\mathbf{x},\boldsymbol{\beta})]^2.$$

- The expected value of the objective function is

$$E\{s \cdot [y - m(\mathbf{x}, \boldsymbol{\beta})]^2\} = E(s \cdot E\{[y - m(\mathbf{x}, \boldsymbol{\beta})]^2 | \mathbf{x}, s\})$$

and the conditional expectation can be expanded as

$$\begin{aligned} E\{[y - m(\mathbf{x}, \boldsymbol{\beta})]^2 | \mathbf{x}, s\} = & \; E(u^2 | \mathbf{x}, s) + [m(\mathbf{x}, \boldsymbol{\beta}_o) - m(\mathbf{x}, \boldsymbol{\beta})]^2 \\ & + 2\{[m(\mathbf{x}, \boldsymbol{\beta}_o) - m(\mathbf{x}, \boldsymbol{\beta})]\} E(u | \mathbf{x}, s) \end{aligned}$$

where $u \equiv y - m(\mathbf{x}, \boldsymbol{\beta}_o)$. Give the exogenous selection condition, $E(u | \mathbf{x}, s) = 0$ so that last term is zero.

• The first term $E(u^2|\mathbf{x}, s)$ does not depend on $\boldsymbol{\beta}$ and the the second term is minimized at $\boldsymbol{\beta} = \boldsymbol{\beta}_o$ (not usually uniquely for give $\mathbf{x}$). The unconditional expectation of the objective function is

$$E\{s \cdot [y - m(\mathbf{x}, \boldsymbol{\beta})]^2\} = E[s \cdot E(u^2|\mathbf{x}, s)] + E\{s \cdot [m(\mathbf{x}, \boldsymbol{\beta}_o) - m(\mathbf{x}, \boldsymbol{\beta})]^2\}$$

and we need to assume the second part is uniquely minimized at $\boldsymbol{\beta} = \boldsymbol{\beta}_o$, that is, the the selected subpopulation.

- Argument generally fails if s is correlated with y even after controlling for $\mathbf{x}$.
- MLE is similar. The log likelihood in the selected sample is

$$\sum_{i=1}^{N} s_i \ell(\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\theta}).$$

If selection is exogenous in the sense that

$$D(\mathbf{y}_i | \mathbf{x}_i, s_i) = D(\mathbf{y}_i | \mathbf{x}_i)$$

then the population value, $\boldsymbol{\theta}_o$, also maximizes the expected value of the selected log likelihood:

$$\begin{aligned} E[s \cdot \ell(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta})] &= E\{s \cdot E[\ell(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta})|\mathbf{x}, s]\} \\ &= E\{s \cdot E[\ell(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta})|\mathbf{x}]\} \end{aligned}$$

Because $E[\ell(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta}_o)|\mathbf{x}] \geq E[\ell(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta})|\mathbf{x}]$ for all $\boldsymbol{\theta}$ – the key result for consistency of conditional MLE – and $s \geq 1$, it follows that

$$E[s \cdot \ell(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta}_o)] \geq E[s \cdot \ell(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta})], \text{ all } \boldsymbol{\theta}.$$

- Uniqueness of $\boldsymbol{\theta}_o$ as the maximizer must be established using the structure of the problem (including the distribution of $\mathbf{x}$ and the nature of selection).

- Conditions for other methods, such as GMM, are similar. But zero conditional mean assumptions of errors given exogenous variables play a key role. Zero correlation orthogonality conditions in the population are not enough even to consistently estimate the parameters on the selected sample even when selection depends on exogenous variables.

3. Selection on the Response Variable: Truncated Regression

- Now consider the case where the rule for observing a data point depends in a known, deterministic way on the response variable. Start with the premise we are interested in $D(y|\mathbf{x})$ in a given population.
- For simplicity, assume y has a continuous distribution. Let $(\mathbf{x}_i, y_i)$ denote a random draw from the population, but where we only observe (or, at least, we only use) the data point if $s_i = 1$.
- Assume the rule is that, for known constants a_1 and a_2,

$$s_i = 1[a_1 < y_i < a_2].$$

- Allow for the cases $a_1 = -\infty$ and $a_2 = +\infty$.
- While the analysis can be made much more general, assume we are primarily interested in

$$E(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}.$$

- But now using OLS on the selected sample, because selection is a function of y_i, results in an inconsistent estimator of $\boldsymbol{\beta}$.
- In a parametric context, assume that the population conditional density is $f(y|\mathbf{x};\boldsymbol{\beta},\boldsymbol{\gamma})$.

• The density conditional on $s = 1$ is

$$p(y|\mathbf{x}, s = 1) = \frac{f(y|\mathbf{x};\boldsymbol{\beta},\boldsymbol{\gamma})}{P(a_1 < y < a_2|\mathbf{x})} = \frac{f(y|\mathbf{x};\boldsymbol{\beta},\boldsymbol{\gamma})}{F(a_2|\mathbf{x};\boldsymbol{\beta},\boldsymbol{\gamma}) - F(a_1|\mathbf{x};\boldsymbol{\beta},\boldsymbol{\gamma})}$$

where $F(\cdot|\mathbf{x};\boldsymbol{\beta},\boldsymbol{\gamma})$ is the cdf of $f(\cdot|\mathbf{x};\boldsymbol{\beta},\boldsymbol{\gamma})$.

• Now that we have the density for the subpopulation with $s = 1$, we can use MLE. The log-likelihood function for a sample of size N from the subpopulation with $a_1 < y_i < a_2$ is

$$\sum_{i=1}^{N}\{\log[f(y_i|\mathbf{x}_i;\boldsymbol{\beta},\boldsymbol{\gamma})] - \log[F(a_2|\mathbf{x}_i;\boldsymbol{\beta},\boldsymbol{\gamma}) - F(a_1|\mathbf{x}_i;\boldsymbol{\beta},\boldsymbol{\gamma})]\}$$

- When $D(y|\mathbf{x}) = Normal(\mathbf{x}\boldsymbol{\beta}, \sigma^2)$, this is often called the "truncated Tobit mode," but a better name is the **truncated normal regression model**.
- As with censoring, truncated the sample is costly. We are interested in $E(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$ in the entire population, but because of the truncated sampling, we specify all of $D(y|\mathbf{x})$.
- Differs from the censored normal regression model in that we observe no information on units not in the subpopulation with $a_1 < y < a_2$. In the censored case, we have a random sample of units, which means we observe $\mathbf{x}_i$, and we can use that in estimation.

• For simplicity, consider the case $a_1 = -\infty$. It is useful to reintroduce the selection indicator s_i and let N be the number of random draws from the population. The likelihood in the truncated case is

$$\prod_{i=1}^{N} \left\{ \frac{\sigma^{-1}\phi[(y_i - \mathbf{x}_i\boldsymbol{\beta})/\sigma]}{\Phi[(a_2 - \mathbf{x}_i\boldsymbol{\beta})/\sigma]} \right\}^{s_i},$$

which emphasizes that we completely drop all units with $s_i = 0$.

- In the censored case, the likelihood is

$$\prod_{i=1}^{N}\{\sigma^{-1}\phi[(y_i - \mathbf{x}_i\boldsymbol{\beta})/\sigma]\}^{s_i}\{\Phi[(a_2 - \mathbf{x}_i\boldsymbol{\beta})/\sigma]\}^{1-s_i}$$

$$= \prod_{i=1}^{N}\left\{\frac{\sigma^{-1}\phi[(y_i - \mathbf{x}_i\boldsymbol{\beta})/\sigma]}{\Phi[(a_2 - \mathbf{x}_i\boldsymbol{\beta})/\sigma]}\right\}^{s_i}\{\Phi[(a_2 - \mathbf{x}_i\boldsymbol{\beta})/\sigma]\}^{s_i}\{\Phi[(a_2 - \mathbf{x}_i\boldsymbol{\beta})/\sigma]\}^{1-s_i}$$

- If we take the log of each likelihood and focus on observation i, we can write

$$
\begin{aligned}
\ell_i^{censored}(\boldsymbol{\theta}) &= s_i \log\{\sigma^{-1}\phi[(y_i - \mathbf{x}_i\boldsymbol{\beta})/\sigma]\} \\
&\quad + (1 - s_i) \log\{1 - \Phi[(a_2 - \mathbf{x}_i\boldsymbol{\beta})/\sigma]\} \\
&= s_i(\log\{\sigma^{-1}\phi[(y_i - \mathbf{x}_i\boldsymbol{\beta})/\sigma]\} \ - \log\{\Phi[(a_2 - \mathbf{x}_i\boldsymbol{\beta})/\sigma]\}) \\
&\quad + s_i \log\{\Phi[(a_2 - \mathbf{x}_i\boldsymbol{\beta})/\sigma]\} \ + (1 - s_i) \log\{1 - \Phi[(a_2 - \mathbf{x}_i\boldsymbol{\beta})/\sigma]\} \\
&= \ell_i^{truncated}(\boldsymbol{\theta}) \\
&\quad + s_i \log\{\Phi[(a_2 - \mathbf{x}_i\boldsymbol{\beta})/\sigma]\} \ + (1 - s_i) \log\{1 - \Phi[(a_2 - \mathbf{x}_i\boldsymbol{\beta})/\sigma]\}
\end{aligned}
$$

- $\ell_i^{censored}(\boldsymbol{\theta})$ uses additional information in the form of the model for the binary selection indicator, s_i (y_i uncensored or not), which depends on the parameters $\boldsymbol{\beta}$ and σ. (Remember, we are not specifying a separate model for s_i; it is implied by the underlying classical linear model and the right censoring.) We can use this information in the censored case because we observe $\mathbf{x}_i$ even when $s_i = 0$. In the truncated case, we do not observe this information.
- The same can be shown in the general case with other forms of censoring and other distributions.
- If you have a choice, you should use censored regression, not truncated regression.

• In Stata. Suppose we only observe a unit if $y < 50$:

```
truncreg y x1 ... xK, ul(50)
```

• Again, we interpret the results as if we had run a linear regression using a random sample from the entire population. This is much different from applying Tobit to a corner solution.

• Easy to extend to case where limits change with i, so (a_{i1}, a_{i2}). Must assume

$$D(y_i|\mathbf{x}_i, a_{i1}, a_{i2}) = D(y_i|\mathbf{x}_i),$$

which is always true if a_{i1} and a_{i2} are deterministic functions of $\mathbf{x}_i$.

• The Hausman and Wise (1974) analysis of data from a negative income tax experiment has this form. Eligibility depended on family size in addition to income (where $y = income$).

• The log likelihood just adds an i subscript on the truncation points:

$$\sum_{i=1}^{N}\{\log[f(y_i|\mathbf{x}_i;\boldsymbol{\beta},\boldsymbol{\gamma})] - \log[F(a_{i2}|\mathbf{x}_i;\boldsymbol{\beta},\boldsymbol{\gamma}) - F(a_{i1}|\mathbf{x}_i;\boldsymbol{\beta},\boldsymbol{\gamma})]\}$$

and the general Stata command is

```
truncreg y x1 ... xK, ll(lower) ul(upper)
```

where “lower” and “upper” are variables defined in the data set (and should be nonmissing, otherwise those observations will be dropped).

EXAMPLE: Truncating the Wealth Distribution

```
. truncreg nettfac inc incsq age agesq male e401k, ul(50)
(note: 224 obs. truncated)

Truncated regression
Limit:   lower =       -inf                             Number of obs =    751
         upper =         50                             Wald chi2(6)  =  57.16
Log likelihood = -3351.6879                             Prob > chi2   = 0.0000

------------------------------------------------------------------------------
     nettfac |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inc |   .6447338   .1412821     4.56   0.000      .367826    .9216416
       incsq |  -.0034965   .0010597    -3.30   0.001    -.0055735   -.0014194
         age |   .1806256   .7896731     0.23   0.819    -1.367105    1.728356
       agesq |   .0032957   .0090657     0.36   0.716    -.0144727    .0210641
        male |   .1300546   3.379858     0.04   0.969    -6.494346    6.754455
       e401k |    4.09938   2.224616     1.84   0.065    -.2607873    8.459548
       _cons |  -24.23088   16.15679    -1.50   0.134    -55.89761    7.435844
-------------+----------------------------------------------------------------
      /sigma |   25.12179   .9167748    27.40   0.000     23.32494    26.91863
------------------------------------------------------------------------------
```

```
. truncreg nettfac inc incsq age agesq male e401k if ~cens, ul(50)
(note: 0 obs. truncated)

Truncated regression
Limit:   lower =       -inf                             Number of obs =    751
         upper =         50                             Wald chi2(6)  =  57.16
Log likelihood = -3351.6879                             Prob > chi2   = 0.0000

------------------------------------------------------------------------------
     nettfac |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inc |   .6447338   .1412821     4.56   0.000      .367826    .9216416
       incsq |  -.0034965   .0010597    -3.30   0.001    -.0055735   -.0014194
         age |   .1806256   .7896731     0.23   0.819    -1.367105    1.728356
       agesq |   .0032957   .0090657     0.36   0.716    -.0144727    .0210641
        male |   .1300546   3.379858     0.04   0.969    -6.494346    6.754455
       e401k |    4.09938   2.224616     1.84   0.065    -.2607873    8.459548
       _cons |  -24.23088   16.15679    -1.50   0.134    -55.89761    7.435844
-------------+----------------------------------------------------------------
      /sigma |   25.12179   .9167748    27.40   0.000     23.32494    26.91863
------------------------------------------------------------------------------
```

```
. * If underlying CLM is correct, truncated and censored regression should
. * give similar answers, with censored more efficient.

. cnreg nettfac inc incsq age agesq male e401k, cen(cens)

Censored-normal regression                        Number of obs   =        975
                                                  LR chi2(6)      =     301.64
                                                  Prob > chi2     =     0.0000
Log likelihood = -3774.6932                       Pseudo R2       =     0.0384

------------------------------------------------------------------------------
     nettfac |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inc |   .7225285   .1192285     6.06   0.000     .4885527    .9565043
       incsq |  -.0018362   .0008255    -2.22   0.026    -.0034562   -.0002162
         age |  -.1480192   .7230439    -0.20   0.838    -1.566932    1.270893
       agesq |   .0122743   .0081677     1.50   0.133    -.0037542    .0283028
        male |  -2.032747   3.123538    -0.65   0.515    -8.162425    4.096931
       e401k |   7.496106    2.00374     3.74   0.000     3.563936    11.42828
       _cons |  -31.34548   15.02683    -2.09   0.037    -60.83437   -1.856601
-------------+----------------------------------------------------------------
      /sigma |   28.67045   .7756753                      27.14825    30.19264
------------------------------------------------------------------------------
  Observation summary:         0  left-censored observations
                             751     uncensored observations
                             224 right-censored observations
```

4. Incidental Truncation: A Probit Selection Equation

Exogenous Explanatory Variables

- Motivation: Interested in estimating $E(wage^o|\mathbf{x})$, where $wage^o$ is the wage offer. But need to recognize that if we randomly sample adults, some will not be working, so $wage^o$ is unoberved.

• Simple utility maximization approach (with w^o the wage offer) to choosing weekly hours:

$$\max_h \; util_i(w_i^o h + a_i, h) \;\; \text{subject to } 0 \leq h \leq 168$$

Assume can rule out a solution at $h_i = 168$. Can show hat if $ds_i(0)/dh \leq 0$, where $s_i(h) = util_i(w_i^o h + a_i, h)$, then $h_i = 0$ is the optimum.

- Equivalent to

$$w_i^o \leq -mu_i^h(a_i, 0)/mu_i^q(a_i, 0)$$

where $mu_i^h(\bullet, \bullet)$ is the marginal disutility of working and $mu_i^q(\bullet, \bullet)$ is the marginal utility of income.

- Can think of the right hand side as the reservation wage, w_i^r.
- Assume the person works only if

$$w_i^o > w_i^r$$

(where we can ignore ties under continuity).

• This looks like censoring the wage offer from below, but there is a key difference: we do not observe w_i^r. Called **incidental truncation**. (Perhaps "incidental censoring" would be a better name, as we can generally draw a random sample from the population of working-age adults, and then observe other attributes.)

• Model the wage offer and reservation wages as

$$\begin{aligned} w_i^o &= \exp(\mathbf{x}_{i1}\boldsymbol{\beta}_1 + u_{i1}) \\ w_i^r &= \exp(\mathbf{x}_{i2}\boldsymbol{\beta}_2 + \gamma_2 a_i + u_{i2}) \end{aligned}$$

- We observe w_i^o if $\log(w_i^o) - \log(w_i^r) > 0$ or

$$\mathbf{x}_{i1}\boldsymbol{\beta}_1 + u_{i1} - \mathbf{x}_{i2}\boldsymbol{\beta}_2 - \gamma_2 a_i - u_{i2} > 0$$

or

$$\mathbf{x}_i\boldsymbol{\delta}_2 + v_{i2} > 0,$$

where $\mathbf{x}_i$ includes all nonredundant elements of $\mathbf{x}_{i1}$ and $\mathbf{x}_{i2}$ ans also a_i, nonwage income.

- Having a_i (at least) affect the reservation wage, and therefore the labor force participation decision, but having no affect on the wage offer, is very important for identification.

• In the population, we can write the Gronau-Heckman model as

$$\log(wage^o) = \mathbf{x}_1\boldsymbol{\beta}_1 + u_1$$
$$inlf = 1[\mathbf{x}\boldsymbol{\delta}_2 + v_2 > 0]$$

where *inlf* is equal to unity if a person is in the labor force. We observe $wage^o$, and therefore $\log(wage^o)$, only if $inlf = 1$.

• We have some interest in estimating the factors that affect *inlf*, but we are primarily interested in the wage offer equation.

- Notation for the general population model

$$
\begin{aligned}
y_1 &= \mathbf{x}_1\boldsymbol{\beta}_1 + u_1 \\
y_2 &= 1[\mathbf{x}\boldsymbol{\delta}_2 + v_2 > 0]
\end{aligned}
$$

where y_1 is the response that is only partially observed, and now y_2 is the selection indicator.

- **Assumptions**: (a) $(\mathbf{x}, y_2)$ are always observed, y_1 is observed only when $y_2 = 1$; (b) (u_1, v_2) is independent of $\mathbf{x}$ with zero mean; (c) $v_2 \sim Normal(0,1)$; (d) $E(u_1|v_2) = \gamma_1 v_2$.
- So, we can think of a random draw $(\mathbf{x}_i, y_{i1}, y_{i2})$ from the population, but we only observe y_{i1} if $y_{i2} = 1$.

- This is sometimes called the **Type II Tobit model**, but it is important to recognized it as a sample selection model. Not surprisingly, it has some statistical similarities with the "selection model" for corner solutions we discussed previously. But it does *not* make sense to set $y_1 = 0$, say, just because we do not observe it. (In the wage offer example, it means we set $wage^o = 1$ whenever we do not observe it.)
- Contrast the sample selection setup with a hurdle model for a corner solution. If, say, y_1 is charitable contributions, and we define $y_2 = 1[y_1 > 0]$, then of course it makes sense that $y_1 = 0$ when $y_2 = 0$; it holds by definition.

• Joint normality of (u_1, v_2) is not necessary for a two-step estimation method, but it is often imposed for a (partial) MLE analysis.

• Because v_2 is independent of $\mathbf{x}$ and standard normal, y_2 follows a probit: $P(y_2 = 1|\mathbf{x}) = \Phi(\mathbf{x}\boldsymbol{\delta}_2)$.

• Because $(\mathbf{x}, y_2)$ is assumed to always be observed, $\boldsymbol{\delta}_2$ is identified, and so we can treat it as known for the purposes of deriving an estimating equation for $\boldsymbol{\beta}_1$.

• How can we obtain an estimating equation for $\boldsymbol{\beta}_1$? Under the previous assumptions,

$$
\begin{aligned}
E(y_1|\mathbf{x}, v_2) &= \mathbf{x}_1\boldsymbol{\beta}_1 + E(u_1|\mathbf{x}, v_2) \\
&= \mathbf{x}_1\boldsymbol{\beta}_1 + E(u_1|v_2) = \mathbf{x}_1\boldsymbol{\beta}_1 + \gamma_1 v_2.
\end{aligned}
$$

• If we could observe (or, in effect, estimate) v_2, we could solve the selection problem by adding v_2 as a regressor and using OLS on the selected sample.

• But we only observe $y_2 = 1[\mathbf{x}\boldsymbol{\delta}_2 + v_2 > 0]$. So we need to obtain $E(y_1|\mathbf{x}, y_2)$. But $(\mathbf{x}, y_2)$ is a function of $(\mathbf{x}, v_2)$, so we can apply iterated expectations:

$$E(y_1|\mathbf{x}, y_2) = E[E(y_1|\mathbf{x}, v_2)|\mathbf{x}, y_2] = \mathbf{x}_1\boldsymbol{\beta}_1 + \gamma_1 E(v_2|\mathbf{x}, y_2).$$

• When $y_2 = 1[\mathbf{x}\boldsymbol{\delta}_2 + v_2 > 0]$ and $v_2|\mathbf{x} \sim Normal(0,1)$, $E(v_2|\mathbf{x}, y_2)$ has a well-known form: it is the inverse Mills ratio. (Actually, its form depends on whether $y_2 = 1$ or $y_2 = 0$, and we only need the $y_2 = 1$ expression here.)

• For completeness (and because it is useful later for treatment effect estimation),

$$E(v_2|\mathbf{x},y_2) = y_2\lambda(\mathbf{x}\boldsymbol{\delta}_2) - (1-y_2)\lambda(-\mathbf{x}\boldsymbol{\delta}_2) \equiv r(y_2,\mathbf{x}\boldsymbol{\delta}_2)$$

where

$$\lambda(\cdot) = \frac{\phi(\cdot)}{\Phi(\cdot)}$$

is the IMR. The function $r(y_2,\mathbf{x}\boldsymbol{\delta}_2)$ is sometimes called a **generalized residual**. Note that $E[r(y_2,\mathbf{x}\boldsymbol{\delta}_2)|\mathbf{x}] = 0$ necessarily follows by iterated expectations because $E(v_2|\mathbf{x}) = 0$, but it can also be shown directly.

- Therefore, on the selected sample we have

$$E(y_1|\mathbf{x}, y_2 = 1) = \mathbf{x}_1\boldsymbol{\beta}_1 + \gamma_1\lambda(\mathbf{x}\boldsymbol{\delta}_2)$$

- If we just regress y_{i1} on $\mathbf{x}_{i1}$ using the $y_{i2} = 1$ sample, then, in effect, we omit the variable $\lambda(\mathbf{x}_i\boldsymbol{\delta}_2)$ from the regression. (It is *possible* that, in the subpopulation with $y_2 = 1$, $\lambda(\mathbf{x}\boldsymbol{\delta}_2)$ is uncorrelated with $\mathbf{x}_1$, in which case OLS would be consistent for the slopes in $\boldsymbol{\beta}_1$. But this would be a fluke and cannot be relied on.)
- The equation for $E(y_1|\mathbf{x}, y_2 = 1)$ is properly viewed as an estimating equation, not a model that we begin with!

• The expression for $E(y_1|\mathbf{x}, y_2 = 1)$ suggests a simple two-step estimation method. (i) Estimate probit of y_{i2} on $\mathbf{x}_i$ using all of the data, $i = 1, \ldots, N$, to obtain $\hat{\boldsymbol{\delta}}_2$ and

$$\hat{\lambda}_{i2} = \lambda(\mathbf{x}_i \hat{\boldsymbol{\delta}}_2).$$

(ii) Run OLS of y_{i1} on $\mathbf{x}_{i1}$, $\hat{\lambda}_{i2}$, $i = 1, \ldots, N_1$ where the data have been ordered so that $y_{i2} = 1$ for $i = 1, \ldots, N_1$.

• This has been called the **Heckit method** after Heckman (1976).

Comments

• When we write $y_1 = \mathbf{x}_1\boldsymbol{\beta}_1 + u_1$ and $y_2 = 1[\mathbf{x}\boldsymbol{\delta}_2 + v_2 > 0]$, we call the first equation the "regression equation" and the second the "selection equation."

• We are using this procedure to solve a missing data problem, or a sample selection problem. Thus, we are interested in estimating $\boldsymbol{\beta}_1$. In the case of two-part models, the partial effects we want are much more complicated.

• Should adjust our standard errors and inference for two-step estimation. Many packages, including Stata, make the adjustment routinely. Bootstrapping is also valid.

• If $\gamma_1 = 0$, it turns out no adjustment to the asymptotic variance of $(\hat{\boldsymbol{\beta}}_1, \hat{\gamma}_1)$ is necessary. In particular, under the null $H_0 : \gamma_1 = 0$ – which means there is no sample selection problem – we can ignore estimation of $\boldsymbol{\delta}_2$. So, we can use the usual OLS t statistic on $\hat{\lambda}_{i2}$ or the heteroskedastic-robust version.

• Technically, the procedure goes through with $\mathbf{x}_1 = \mathbf{x}$, that is, without an exclusion restriction. But then identification of $\boldsymbol{\beta}_1$ is possible only because $\lambda(\cdot)$ is a nonlinear function.

• Generally, should be hesitant to achieve identification "off of a nonlinearity." Cannot really tell if $\lambda(\mathbf{x}_i\hat{\boldsymbol{\delta}}_2)$ is statistically significant because selection is an issue or the functional form $E(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}_1$ is misspecified (in the population).

- If we write $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$, we are assuming $E(y_1|\mathbf{x})$ (the population regression) does not depend on $\mathbf{x}_2$. The only reason $E(y_1|\mathbf{x}, y_2 = 1)$ depends on $\mathbf{x}_2$ is because $\mathbf{x}_2$ predicts selection and selection is correlated with u_1.
- Often, over the range of $\mathbf{x}_i\hat{\boldsymbol{\delta}}_2$ in the data, $\lambda(\cdot)$ is pretty close to linear. Very high collinearity is usually present unless $\mathbf{x}_i$ contains something not in $\mathbf{x}_{i1}$ that is useful for predicting selection.

- If we allowed $\lambda(\cdot)$ to be replaced by an unknown function, say

$$E(y|\mathbf{x}, y_2 = 1) = \mathbf{x}\boldsymbol{\beta}_1 + \gamma_1 h(\mathbf{x}\boldsymbol{\delta}_2),$$

as in semiparametric approaches, then $\boldsymbol{\beta}_1$ would not be identified: we would have to allow $h(\cdot)$ to be arbitrarily close to a linear function. We say that $\boldsymbol{\beta}_1$ is "nonparametrically unidentified" without an exclusion restriction.

- There exist semiparametric methods that allow $h(\cdot)$ to be a generally smooth function.

- Bottom line: the Heckit approach is not believable unless one has at least one exclusion restriction in the regression equation. And, if we write $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$, so that

$$P(y_2 = 1|\mathbf{x}) = \boldsymbol{\Phi}(\mathbf{x}_1\boldsymbol{\delta}_{21} + \mathbf{x}_2\boldsymbol{\delta}_{22}),$$

then we must be able to reject $H_0 : \boldsymbol{\delta}_{22} = \mathbf{0}$ at some low significance level. (Just like with instrumental variables.) What we cannot generally test is whether excluding $\mathbf{x}_2$ from the regression equation is appropriate (just like with IV).

• Sometimes one sees exclusion restrictions in the selection probit. This is not usually advised. Now let $\mathbf{x}_1$ and $\mathbf{x}_2$ both be subsets of $\mathbf{x}$, which generally overlap but where $\mathbf{x}_1$ is not a subset of $\mathbf{x}_2$. If we use

$$y_2 = 1[\mathbf{x}_2\boldsymbol{\delta}_2 + v_2 > 0]$$

then we are assuming

$$P(y_2 = 1|\mathbf{x}) = P(y_2 = 1|\mathbf{x}_2).$$

• But, as in the Gronau-Heckman example, the selection equation is usually a reduced form. (So, nonlabor income appears in the selection equation, as do all other characteristics that affect the wage offer or the reservation wage.)

- Exclusion restrictions are not needed in the probit selection equation. So, if it makes a difference for estimating $\boldsymbol{\beta}_1$, one must always include all of $\mathbf{x}$ in the selection equation. Making exclusion restrictions in the selection equation is tantatmount to making exclusion restrictions in a reduced form. In special cases, this might be warranted, but it is less robust than allowing an unrestricted reduced form. (Think 2SLS estimation of a single equation versus 3SLS of two equations where the second is a restricted reduced form.)
- Better to treat missing explanatory variables as endogenous, provided we have extra instrumental variables.

• If we assume (u_1, v_2) is bivariate normal, then we can apply partial MLE. It is "partial" because we can only use y_{i1} when $y_{i2} = 1$. See text for log likelihood function. The MLE is more efficient if joint normality holds, and the standard errors are readily available.

• But the two-step method does have some robustness because it only uses

$$E(u_1|\mathbf{x}, v_2) = E(u_1|v_2) = \gamma_1 v_2$$

$$v_2|\mathbf{x} \sim Normal(0, 1)$$

- Can pretty easily relax the linear conditional mean:

$$E(u_1|v_2) = \gamma_1 v_2 + \psi_1(v_2^2 - 1).$$

- Can show

$$E(v_2^2 - 1|\mathbf{x}, y_2 = 1) = -\lambda(\mathbf{x}\boldsymbol{\delta}_2)\mathbf{x}\boldsymbol{\delta}_2$$

so

$$E(y_1|\mathbf{x}, y_2 = 1) = \mathbf{x}_1\boldsymbol{\beta}_1 + \gamma_1\lambda(\mathbf{x}\boldsymbol{\delta}_2) - \psi_1\lambda(\mathbf{x}\boldsymbol{\delta}_2)\mathbf{x}\boldsymbol{\delta}_2$$

- The estimating equation has changed, but the underlying population model, $E(y_1|\mathbf{x}) = \mathbf{x}_1\boldsymbol{\beta}_1$, has not!
- Two step procedure. Start with probit, as usual, and then regression

$$y_{i1} \text{ on } \mathbf{x}_{i1}, \hat{\lambda}_{i2}, \hat{\lambda}_{i2} \cdot (\mathbf{x}_i\hat{\boldsymbol{\delta}}_2),\ i = 1,\ldots,N_1.$$

- Bootstrapping very attractive here for standard errors and inference.
- MLE would be much more cumbersome.

EXAMPLE: Wage Offer for Married Women

```
. use mroz

. des lwage inlf nwifeinc

              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
lwage           float  %9.0g                  log(wage)
inlf            byte   %9.0g                  =1 if in lab frce, 1975
nwifeinc        float  %9.0g                  (faminc - wage*hours)/1000
. sum lwage inlf educ kidslt6 nwifeinc

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       lwage |       428    1.190173    .7231978  -2.054164   3.218876
        inlf |       753    .5683931    .4956295          0          1
        educ |       753    12.28685    2.280246          5         17
     kidslt6 |       753    .2377158     .523959          0          3
    nwifeinc |       753    20.12896     11.6348  -.0290575         96
```

```
. reg lwage educ exper expersq

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  3,   424) =   26.29
       Model |  35.0222967     3  11.6740989           Prob > F      =  0.0000
    Residual |  188.305144   424  .444115906           R-squared     =  0.1568
-------------+------------------------------           Adj R-squared =  0.1509
       Total |  223.327441   427  .523015084           Root MSE      =  .66642

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1074896   .0141465     7.60   0.000     .0796837    .1352956
       exper |   .0415665   .0131752     3.15   0.002     .0156697    .0674633
     expersq |  -.0008112   .0003932    -2.06   0.040    -.0015841   -.0000382
       _cons |  -.5220406   .1986321    -2.63   0.009    -.9124667   -.1316144
------------------------------------------------------------------------------
```

```
. heckman lwage educ exper expersq, select(inlf = educ exper expersq nwifeinc
          age kidslt6 kidsge6) twostep

Heckman selection model -- two-step estimates   Number of obs      =       753
(regression model with sample selection)        Censored obs       =       325
                                                Uncensored obs     =       428

                                                Wald chi2(6)       =    180.10
                                                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lwage        |
        educ |   .1090655    .015523     7.03   0.000     .0786411      .13949
       exper |   .0438873   .0162611     2.70   0.007     .0120163    .0757584
     expersq |  -.0008591   .0004389    -1.96   0.050    -.0017194    1.15e-06
       _cons |  -.5781032   .3050062    -1.90   0.058    -1.175904     .019698
-------------+----------------------------------------------------------------
```

inlf						
educ	.1309047	.0252542	5.18	0.000	.0814074	.180402
exper	.1233476	.0187164	6.59	0.000	.0866641	.1600311
expersq	-.0018871	.0006	-3.15	0.002	-.003063	-.0007111
nwifeinc	-.0120237	.0048398	-2.48	0.013	-.0215096	-.0025378
age	-.0528527	.0084772	-6.23	0.000	-.0694678	-.0362376
kidslt6	-.8683285	.1185223	-7.33	0.000	-1.100628	-.636029
kidsge6	.036005	.0434768	0.83	0.408	-.049208	.1212179
_cons	.2700768	.508593	0.53	0.595	-.7267473	1.266901
mills						
lambda	.0322619	.1336246	0.24	0.809	-.2296376	.2941613
rho	0.04861					
sigma	.66362875					
lambda	.03226186	.1336246				

```
. heckman lwage educ exper expersq, select(inlf =educ exper expersq nwifeinc age
          kidslt6 kidsge6)

Heckman selection model                                Number of obs      =       753
(regression model with sample selection)               Censored obs       =       325
                                                       Uncensored obs     =       428

                                                       Wald chi2(3)       =     59.67
Log likelihood = -832.8851                             Prob > chi2        =    0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lwage        |
        educ |   .1083502   .0148607     7.29   0.000     .0792238    .1374767
       exper |   .0428369   .0148785     2.88   0.004     .0136755    .0719983
     expersq |  -.0008374   .0004175    -2.01   0.045    -.0016556   -.0000192
       _cons |  -.5526973   .2603784    -2.12   0.034     -1.06303   -.0423651
-------------+----------------------------------------------------------------
```

```
inlf         |
        educ |   .1313415   .0253823     5.17   0.000     .0815931    .1810899
       exper |   .1232818   .0187242     6.58   0.000     .0865831    .1599806
     expersq |  -.0018863   .0006004    -3.14   0.002     -.003063   -.0007095
    nwifeinc |  -.0121321   .0048767    -2.49   0.013    -.0216903    -.002574
         age |  -.0528287   .0084792    -6.23   0.000    -.0694476   -.0362098
     kidslt6 |  -.8673988   .1186509    -7.31   0.000     -1.09995   -.6348472
     kidsge6 |   .0358723   .0434753     0.83   0.409    -.0493377    .1210824
       _cons |   .2664491   .5089578     0.52   0.601    -.7310898    1.263988
-------------+----------------------------------------------------------------
     /athrho |    .026614    .147182     0.18   0.857    -.2618573    .3150854
    /lnsigma |  -.4103809   .0342291   -11.99   0.000    -.4774687   -.3432931
-------------+----------------------------------------------------------------
         rho |   .0266078   .1470778                     -.2560319    .3050564
       sigma |   .6633975   .0227075                      .6203517    .7094303
      lambda |   .0176515   .0976057                     -.1736521    .2089552
------------------------------------------------------------------------------
LR test of indep. eqns. (rho = 0):   chi2(1) =     0.03   Prob > chi2 = 0.8577
------------------------------------------------------------------------------
```

• Olsen (1980) proposed an alternative two-step estimator that enforces discipline by requiring an exclusion restriction. It can be derived by assuming, in $y_2 = 1[\mathbf{x}\boldsymbol{\delta}_2 + v_2 > 0]$, that v_2 has a $Uniform[-c,c]$ distribution for any contant $c > 0$ (rather than standard normal). Then $y_2 = 1[-v_2 \leq \mathbf{x}\boldsymbol{\delta}_2]$ and $e_2 = -v_2$ also has a $Uniform[-c,c]$ distribution. For concreteness, choose $c = 1/2$. Then

$$P(y_2 = 1|\mathbf{x}) = P(e_2 \leq \mathbf{x}\boldsymbol{\delta}_2|\mathbf{x}) = \mathbf{x}\boldsymbol{\delta}_2 + 1/2 \equiv \mathbf{x}\boldsymbol{\pi}_2$$

where $\boldsymbol{\pi}_2$ is $\boldsymbol{\delta}_2$ but with 1/2 added to the intercept.

- Further, the distribution of e_2 conditional on $e_2 \leq \mathbf{x}\boldsymbol{\delta}_2$ is *Uniform*$[-1/2, \mathbf{x}\boldsymbol{\delta}_2]$, and so, using the usual formula for the expected value of a uniform random variable,

$$E(e_2|\mathbf{x}, e_2 \leq \mathbf{x}\boldsymbol{\delta}_2) = E(e_2|\mathbf{x}, y_2 = 1) = (\mathbf{x}\boldsymbol{\delta}_2 - 1/2)/2 = (\mathbf{x}\boldsymbol{\pi}_2 - 1)/2.$$

- As before, make a linearity assumption relating u_1 and e_2:

$$E(u_1|e_2) = \gamma_1 e_2.$$

- Then

$$E(y_1|\mathbf{x}, y_2 = 1) = \mathbf{x}_1\boldsymbol{\beta}_1 + (\gamma_1/2)(\mathbf{x}\boldsymbol{\pi}_2 - 1)$$

- Two-step method is now clear. (1) Estimate a linear probability model by OLS, regressing y_{i2} on $\mathbf{x}_i$, using all of the data, to get the fitted values, $\hat{y}_{i2} = \mathbf{x}_i\hat{\boldsymbol{\pi}}_2$. (As always, $\mathbf{x}_i$ should include a constant.) (2) Using the selected sample, run the regression y_{i1} on $\mathbf{x}_{i1}$, $(\hat{y}_{i2} - 1)$.
- The test for the null of no sample selection bias is the t statistic on $(\hat{y}_{i2} - 1)$.
- Standard errors should account for the two-step estimation.

- Unlike with Heckman's approach, one cannot apply Olsen's method unless $\mathbf{x}_{i1}$ is a strict subset of $\mathbf{x}_i$. That is because $\hat{y}_{i2} - 1 = \mathbf{x}_i\hat{\boldsymbol{\pi}}_2 - 1$ is a linear combination of $\mathbf{x}_i$.
- Might carry this idea further. Model $E(u_1|e_2)$ as a polynomial in e_2, imposing the restriction $E(u_1) = 0$. Or just add polynomials in $\mathbf{x}_i\hat{\boldsymbol{\pi}}_2$, for example,

$$y_{i1} \text{ on } \mathbf{x}_{i1},\ \hat{y}_{i2}, \hat{y}_{i2}^2, \hat{y}_{i2}^3 \text{ using } y_{i2} = 1.$$

- The intercept in $y_1 = \mathbf{x}_1\boldsymbol{\beta}_1 + u_1$ is generally unidentified using this approach. Not very important for sample selection, but is for "self-selection" and treatment effects later on.

• Can use the same trick with probit fitted probabilities because $\Phi(\bullet)$ is a strictly monotonic function. In particular, note that we can always write the IMR as a function of $\Phi(\bullet)$:

$$\lambda(z) = h(\Phi(z))$$

where

$$h(a) = \lambda(\Phi^{-1}(a)).$$

So, approximate $h(\bullet)$ by polynomials. Then

$$E(y_1|\mathbf{x}, y_2 = 1) \approx \mathbf{x}_1\boldsymbol{\beta}_1 + \alpha_0 + \alpha_1\Phi(\mathbf{x}\boldsymbol{\delta}_2) + \alpha_2[\Phi(\mathbf{x}\boldsymbol{\delta}_2)]^2 + \ldots + \alpha_q[\Phi(\mathbf{x}\boldsymbol{\delta}_2)]^q.$$

• As before, lose identification of the intercept because α_0 gets absorbed in the intercept.

Endogenous Explanatory Variables

- Let y_1 be the response variable, as before. Let y_2 be the endogenous explanatory variable. (Easy to extend to a vector.) Now, y_3 is the binary selection indicator.

- Think of the model and selection mechanism as follows:

$$\begin{aligned} y_1 &= \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1 \\ y_2 &= \mathbf{z}_2\boldsymbol{\delta}_2 + v_2 \\ y_3 &= 1[\mathbf{z}\boldsymbol{\delta}_3 + v_3 > 0] \end{aligned}$$

- If we are careful, we only need the equation for y_2 to be a linear projection, so that y_2 can be any kind of variable (discrete, continuous, some combination).
- As in the case of standard 2SLS applied to random sampling contexts, the equation for y_2 is a reduced form, and so $\mathbf{z}_1$ should be a subset of $\mathbf{z}_2$.
- Even if y_2 is always observed, get some robustness by acting as if it is not.
- For reasons we will see, $\mathbf{z}$ should include all elements of $\mathbf{z}_2$, and at least one more element.

• **Assumptions**: (a) $(\mathbf{z}, y_3)$ is always observed, (y_1, y_2) is observed when $y_3 = 1$; (b) (u_1, v_3) is independent of $\mathbf{z}$; (c) $v_3 \sim Normal(0,1)$; (d) $E(u_1|v_3) = \gamma_1 v_3$; (e) $E(\mathbf{z}' v_2) = \mathbf{0}$ and $\boldsymbol{\delta}_{22} \neq 0$, where $\mathbf{z}_2\boldsymbol{\delta}_2 = \mathbf{z}_1\boldsymbol{\delta}_{21} + \mathbf{z}_{22}\boldsymbol{\delta}_{22}$.

• Notice that (e) assumes $\mathbf{z}_2 = (\mathbf{z}_1, \mathbf{z}_{22})$, that is, $\mathbf{z}_1$ is contained in $\mathbf{z}_2$. As we will see, (e) is technically not quite right. But the point is that identification should hold in the population or there is no hope of its holding in the selected subpopulation.

- Estimating equation: in the population, write $g(\mathbf{z}, y_3) \equiv E(u_1|\mathbf{z}, y_3)$ so that

$$
\begin{aligned}
y_1 &= \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + g(\mathbf{z}, y_3) + e_1 \\
E(e_1|\mathbf{z}, y_3) &= 0.
\end{aligned}
$$

- $\mathbf{z}_1$ and $g(\mathbf{z}, y_3)$ are, by construction, exogenous in this equation, but y_2 is generally endogenous. So, we will have to apply IV.

• From earlier results on applying IV to a selected sample, IV applied to the $y_{i3} = 1$ subsample consistently estimates the parameters. We only need $g(\mathbf{z}, y_3)$, which can act as its own instrument, when $y_3 = 1$:

$$g(\mathbf{z}, 1) = \gamma_1 \lambda(\mathbf{z}\boldsymbol{\delta}_3).$$

• Two-step procedure: (i) Probit of y_{i3} on $\mathbf{z}_i$ (*all* exogenous variables) using the full sample. Obtain $\hat{\lambda}_{i3} = \lambda(\mathbf{z}_i\hat{\boldsymbol{\delta}}_3)$.

(ii) Apply 2SLS to

$$y_{i1} = \mathbf{z}_{i1}\boldsymbol{\delta}_1 + \alpha_{i1}y_{i2} + \gamma_1\hat{\lambda}_{i3} + error_i$$

using instruments $(\mathbf{z}_{i2}, \hat{\lambda}_{i3})$.

• The first-stage regression in the 2SLS estimation makes it clear that $\mathbf{z}_{22}$ actually needs to appear in the linear projection of y_2 on $\mathbf{z}_1, \mathbf{z}_{22}, \lambda(\mathbf{z}\boldsymbol{\lambda}_3)$ in the subpopulation with $y_3 = 1$. Can test this (account for two-step estimation).

• Simple test for $H_0 : \gamma_1 = 0$ (no selection bias) without taking a stand on endogeneity of y_2: use the usual 2SLS or heteroskedasticity-robust t statistic on $\hat{\lambda}_{i3}$.

Comments

• Practically speaking, $\mathbf{z}$ should have at least two elements not in $\mathbf{z}_1$. It is helpful to force oneself to include one at least more element in $\mathbf{z}_2$ not in $\mathbf{z}_1$, and then one more element in $\mathbf{z}$ not in $\mathbf{z}_2$. The idea is that we need something to predict y_2 (in the absense of sample selection) and something else to predict selection, y_3.

• In the wage offer equation, we might use parents' education as IVs for *educ*, and then other income and number of children as variables largely predicting workforce participation. The selection equation should include *all* such variables.

- Because only fitted values are used for 2SLS, one can use as IVs $(\mathbf{z}_i, \hat{\lambda}_{i3})$ rather than $(\mathbf{z}_{i2}, \hat{\lambda}_{i3})$. We must include in the instruments at least all exogenous elements in the estimating equation – $(\mathbf{z}_{i1}, \lambda_{i3})$ – and then some additional instruments for y_{i2}.
- The first stage regression using $(\mathbf{z}_i, \hat{\lambda}_{i3})$ likely will suffer from multicollinearity but we only use the fitted values as IVs for y_{i2}.

- Even if y_2 is exogenous in the population model, we usually need an IV for it if it is sometimes missing. In effect, the missingness of y_2 when $y_3 = 0$ can cause it to be endogenous in the subpopulation.
- A different, less robust approach is possible. Suppose y_2 is always observed. Then can estimate $\boldsymbol{\delta}_2$ from the OLS regression y_{i2} on $\mathbf{z}_{i2}$ using all of the observations.
- We can write

$$\begin{aligned} y_1 &= \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1(\mathbf{z}_2\boldsymbol{\delta}_2) + \alpha_1 v_2 + u_1 \\ &\equiv \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1(\mathbf{z}_2\boldsymbol{\delta}_2) + v_1. \end{aligned}$$

Because we can insert $\hat{\boldsymbol{\delta}}_2$ for $\boldsymbol{\delta}_2$, we might just apply the usual Heckit to this equation.

• Why is this less robust than the previous method? Because it requires something like v_1 independent of $\mathbf{z}$, which essentially means v_2 should be independent of $\mathbf{z}$. This severely restricts the nature of y_2 because $y_2 = \mathbf{z}_2\boldsymbol{\delta}_2 + v_2$ where v_2 is independent of $\mathbf{z}$ effectively rules out discreteness in y_2.

• Suppose $y_2 = \mathit{benefits}^o/\mathit{wage}^o$. This is zero for some job offers. It is unlikely we can write $y_2 = \mathbf{z}_2\boldsymbol{\delta}_2 + v_2$ with v_2 independent of $\mathbf{z}$.

• It is more robust to leave y_2 in the equation, add $\hat{\lambda}_{i3}$, and then use IV (probably 2SLS) on the resulting equation.

- In addition, the first approach outlined applies easily to more complicated models, such as when y_2^2 or interactions enter. We need to simply specify instruments for these. Plugging fitted values into the nonlinear function, as always, leads to trouble (even if we did not have a sample selection problem).

• For example, suppose the structural equation is

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \eta_1 y_2^2 + y_2\mathbf{z}_1\boldsymbol{\psi}_1 + u_1,$$

and (y_1, y_2) are observed when $y_3 = 1$.

• Before we estimate the selection equation, it makes sense to decide what the IVs would be if we did not have a selection problem. Suppose they are $[\mathbf{z}_2, \mathbf{g}(\mathbf{z}_2)]$ where $\mathbf{g}(\mathbf{z}_2)$ consists of nonlinear functions of $\mathbf{z}_2$, such as squares and cross products.

• Then, use probit of y_{i3} on $\mathbf{z}_i, \mathbf{g}(\mathbf{z}_{i2})$ to get the IMRs, $\hat{\lambda}_{i3}$. Then, on the selected sample, use IV (2SLS) on

$$y_{i1} = \mathbf{z}_{i1}\boldsymbol{\delta}_1 + \alpha_1 y_{i2} + \eta_1 y_{i2}^2 + y_2 \mathbf{z}_{i1}\boldsymbol{\psi}_1 + \gamma_1 \hat{\lambda}_{i3} + error_{i1},$$

using instruments $[\mathbf{z}_{i2}, \mathbf{g}(\mathbf{z}_{i2}), \hat{\lambda}_{i3}]$.

• Of course, it is possible that $\mathbf{g}(\mathbf{z}_{i2})$ is not needed in the selection probit – that is a functional form issue – or some other nonlinear functions of $\mathbf{z}_2$ should be used. But a safe approach is to use the same functions in both places.

EXAMPLE: Education Endogenous in the Wage Offer Equation

```
. probit inlf educ exper expersq nwifeinc age kidslt6 kidsge6 motheduc fatheduc

Probit regression                                 Number of obs   =        753
                                                  LR chi2(9)      =     227.43
                                                  Prob > chi2     =     0.0000
Log likelihood =  -401.1592                       Pseudo R2       =     0.2209

------------------------------------------------------------------------------
        inlf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1260833   .0279019     4.52   0.000     .0713965    .1807701
       exper |    .123625    .018729     6.60   0.000     .0869167    .1603332
     expersq |  -.0018905   .0006005    -3.15   0.002    -.0030674   -.0007136
    nwifeinc |  -.0120713   .0048593    -2.48   0.013    -.0215953   -.0025472
         age |  -.0520759   .0086085    -6.05   0.000    -.0689483   -.0352035
     kidslt6 |  -.8663033   .1185224    -7.31   0.000    -1.098603   -.6340038
     kidsge6 |   .0371177   .0436089     0.85   0.395    -.0483541    .1225896
    motheduc |   .0099308   .0191914     0.52   0.605    -.0276837    .0475452
    fatheduc |  -.0018494   .0181487    -0.10   0.919    -.0374201    .0337214
       _cons |    .217918    .519246     0.42   0.675    -.7997853    1.235621
------------------------------------------------------------------------------

. predict xd3h, xb

. gen phi3 = normalden(xd3h)

. gen PHI3 = normal(xd3h)

. gen lambda3 = phi3/PHI3
```

```
. ivreg lwage exper expersq lambda3 (educ = nwifeinc age kidslt6 kidsge6
              motheduc fatheduc)

Instrumental variables (2SLS) regression

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  4,   423) =   16.15
       Model |   35.018841     4  8.75471025           Prob > F      =  0.0000
    Residual |    188.3086   423  .445173995           R-squared     =  0.1568
-------------+------------------------------           Adj R-squared =  0.1488
       Total |  223.327441   427  .523015084           Root MSE      =  .66721

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1044079   .0175683     5.94   0.000     .0698759      .13894
       exper |   .0435482   .0164173     2.65   0.008     .0112785    .0758178
     expersq |  -.0008552    .000442    -1.93   0.054    -.0017241    .0000136
     lambda3 |   .0241612    .136629     0.18   0.860     -.244395    .2927175
       _cons |  -.5113313   .3331186    -1.53   0.126    -1.166105    .1434426
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   exper expersq lambda3 nwifeinc age kidslt6 kidsge6 motheduc
               fatheduc
------------------------------------------------------------------------------

. * Virtually no evidence of sample selection.
```

```
. * Estimated effect ignoring sample selection is similar:

. ivreg lwage exper expersq (educ = nwifeinc age kidslt6 kidsge6 motheduc fatheduc)

Instrumental variables (2SLS) regression

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  3,   424) =   11.20
       Model |  34.6262515     3  11.5420838           Prob > F      =  0.0000
    Residual |  188.701189   424  .445049975           R-squared     =  0.1550
-------------+------------------------------           Adj R-squared =  0.1491
       Total |  223.327441   427  .523015084           Root MSE      =  .66712

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0941307   .0266148     3.54   0.000     .0418172    .1464441
       exper |   .0423212   .0132503     3.19   0.002     .0162766    .0683657
     expersq |  -.0008366    .000396    -2.11   0.035     -.001615   -.0000583
       _cons |  -.3567989   .3423923    -1.04   0.298    -1.029796    .3161987
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   exper expersq nwifeinc age kidslt6 kidsge6 motheduc fatheduc
------------------------------------------------------------------------------
```

```
. reg educ exper expersq nwifeinc age kidslt6 kidsge6 motheduc fatheduc lambda3
      if inlf

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  9,   418) =  216.68
       Model |  1836.5383     9  204.059811           Prob > F      =  0.0000
    Residual | 393.657965   418  .941765465           R-squared     =  0.8235
-------------+------------------------------           Adj R-squared =  0.8197
       Total | 2230.19626   427  5.22294206           Root MSE      =  .97045

------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |  -.7855876   .0304343   -25.81   0.000    -.8454109   -.7257644
     expersq |    .012524   .0006945    18.03   0.000     .0111588    .0138892
    nwifeinc |   .0904503   .0047679    18.97   0.000     .0810784    .0998223
         age |   .3109581   .0120253    25.86   0.000     .2873205    .3345957
     kidslt6 |   5.666308   .1891528    29.96   0.000     5.294499    6.038117
     kidsge6 |  -.2643391   .0406855    -6.50   0.000    -.3443127   -.1843655
    motheduc |  -.0307841    .018222    -1.69   0.092    -.0666022    .0050341
    fatheduc |   .0573622   .0165472     3.47   0.001     .0248361    .0898883
     lambda3 |  -12.00563   .3367136   -35.66   0.000    -12.66749   -11.34376
       _cons |   10.89935   .4296475    25.37   0.000     10.05482    11.74389
------------------------------------------------------------------------------
```

Binary Response with Sample Selection

• The selection problem with a binary response can be solved by partial MLE. Write

$$
\begin{aligned}
y_1 &= 1[\mathbf{x}_1\boldsymbol{\beta}_1 + u_1 > 0] \\
y_2 &= 1[\mathbf{x}\boldsymbol{\delta}_2 + v_2 > 0]
\end{aligned}
$$

where $(\mathbf{x}, y_2)$ is always observed, $\mathbf{x}_1 \subset \mathbf{x}$, and y_1 is observed when $y_2 = 1$.

• Assume that (u_1, v_2) is independent of $\mathbf{x}$ with a bivariate normal distribution, where the variance of each is unity and $Corr(u_1, v_2) = \rho_1$.

• Similar to probit with an endogenous binary explanatory variable. Note that y_2 does not appear in the equation for y_1 (it cannot, and it makes no sense in most sample selection contexts).

• Estimate by partial MLE. Not believable without an exclusion restriction in $\mathbf{x}_1$, even though parameters are technically identified.

• Remember, we interpret the estimates as if we had been able to use a random sample to estimate

$$P(y_1 = 1|\mathbf{x}) = P(y_1 = 1|\mathbf{x}_1) = \Phi(\mathbf{x}_1\boldsymbol{\beta}_1)$$

directly.

• In Stata, suppose *healthins* is a binary variable indicating whether health insurance is included as part of a job offer. We only observe this variable if the person is in the workforce.

```
heckprob healthins educ exper expersq,
select(inlf = educ exper expersq otherinc)
```

• Important: Some simple strategies for "correcting" for sample selection cannot be justified. It is tempting to estimate the selection equation by probit and then plug the estimated inverse Mills ratio into the second stage probit, using only the observations with $y_{i2} = 1$. There is no way to justify this as a sample selection correction.

- Inserting the IMR into the second stage probit is a legitimate test of the null hypothesis of no selection bias. Can show this by finding $E(y_1|\mathbf{x}, y_2 = 1)$, as in the case of a probit model with a binary endogenous variable. Under the null $\rho_1 = 0$, the mean function is probit, so we will just do probit on the selected sample in obtaining a score-type test.
- Let $m(\mathbf{x}, \boldsymbol{\beta}_1, \rho_1; \boldsymbol{\delta}_2)$ be the mean function. Can show

$$\nabla_{\boldsymbol{\beta}_1} m(\mathbf{x}, \boldsymbol{\beta}_1, 0; \boldsymbol{\delta}_2) = \phi(\mathbf{x}_1\boldsymbol{\beta}_1)\mathbf{x}_1$$
$$\nabla_{\rho_1} m(\mathbf{x}, \boldsymbol{\beta}_1, 0; \boldsymbol{\delta}_2) = \phi(\mathbf{x}_1\boldsymbol{\beta}_1)\lambda(\mathbf{x}\boldsymbol{\delta}_2)$$

where $\lambda(\cdot)$ is the IMR.

- Therefore, a simple variable addition test is to obtain $\hat{\boldsymbol{\delta}}_2$ by probit MLE, and construct the IMRs, $\hat{\lambda}_{i2} = \lambda(\mathbf{x}_i\hat{\boldsymbol{\delta}}_2)$. Next, using the observations for which $y_{i2} = 1$ (that is, for which y_{i1} is observed), run probit of y_{i1} on $\mathbf{x}_{i1}$, $\hat{\lambda}_{i2}$ and use the usual t statistic on $\hat{\lambda}_{i2}$ to test the null hypothesis $H_0 : \rho_1 = 0$.
- Under the null, no need to adjust this t statistic for first-stage estimation.

Exponential Model with Sample Selection

- Start again with an omitted variable formulation, as in the endogenous explanatory variable case:

$$E(y_1|\mathbf{x},c_1) = E(y_1|\mathbf{x}_1,c_1) = \exp(\mathbf{x}_1\boldsymbol{\beta}_1 + c_1).$$

Here, we assume c_1 is independent of $\mathbf{x}$, so it would be harmless to exclude it (because $\mathbf{x}_1$ contains unity) if we could obtain a random sample.

- But again write

$$y_2 = 1[\mathbf{x}\boldsymbol{\delta}_2 + v_2 > 0]$$

and we observe y_1 only if $y_2 = 1$.

- Assume (c_1, v_2) is independent of $\mathbf{x}$ and jointly normally distributed, with $Var(v_2) = 1$. The key expectation is

$$E(y_1|\mathbf{x}, y_2 = 1) = \exp(\tau_1^2/2 + \mathbf{x}_1\boldsymbol{\beta}_1)\{\Phi(\rho_1 + \mathbf{x}\boldsymbol{\delta}_2)/\Phi(\mathbf{x}\boldsymbol{\delta}_2)\}$$

where $\tau_1^2 = Var(c_1)$ and $\rho_1 = Cov(c_1, v_2)$.

- So, estimate $\boldsymbol{\delta}_2$ by probit in the first stage. Then, estimate the above mean function in the second stage. $\tau_1^2/2$ gets absorbed in intercept. This is actually what we want because it appears in the APEs. ρ_1 is estimated along with $\boldsymbol{\beta}_1$. Could use Poisson QMLE. with this more complicated mean function.

- Interpret results as exponential regression on random sample.

- Simple test of sample selection bias is obtained by adding the log of the inverse Mills ratio, $\log[\lambda(\mathbf{x}_i\hat{\boldsymbol{\delta}}_2)]$, to the exponential function, and estimate the resulting "model" by, say, the Poisson QMLE using the selected sample. The robust t statistic for $\log[\lambda(\mathbf{x}_i\hat{\boldsymbol{\delta}}_2)]$ that allows the likelihood to be misspecified is a valid test of the null hypothesis of no selection bias.

- In Stata:

```
probit y2 x1 ... xK
predict xd2hat, xb
gen lamda2h = normalden(xd2hat)/normal(xd2hat)
gen llamda2h = log(lamda2h)
glm y1 x11 ... x1K1 llamda2h, fam(poisson)
```

5. Incidental Truncation: A Tobit Selection Equation

- Occasionally, we observe a partially continuous variable that determines selection. For example, we might observe hours worked, which implies we observe the wage offer if $hours > 0$.
- If *hours* follows a Tobit model, can use that information.
- General model is

$$
\begin{aligned}
y_1 &= \mathbf{x}_1\boldsymbol{\beta}_1 + u_1 \\
y_2 &= \max(0, \mathbf{x}\boldsymbol{\delta}_2 + v_2) \\
s_2 &= 1[y_2 > 0]
\end{aligned}
$$

• Selection is a function of the partially continuous variable y_2.

• Under the same assumptions for the probit selection case, we can derive

$$E(y_1|\mathbf{x}, v_2) = \mathbf{x}_1\boldsymbol{\beta}_1 + \gamma_1 v_2$$

• Therefore, we can apply OLS on the selected sample if we can observe v_2. Now, we effectively can observe v_2 because $v_2 = y_2 - \mathbf{x}\boldsymbol{\delta}_2$ whenever $y_2 > 0$ – just when we need to.

• Two step procedure: (1) Estimate $\boldsymbol{\delta}_2$ by Tobit using the entire sample. Construct $\hat{v}_{i2} = y_{i2} - \mathbf{x}_i\hat{\boldsymbol{\delta}}_2$ when $y_{i2} > 0$. (2) Use OLS on the selected sample of y_{i1} on $\mathbf{x}_{i1}, \hat{v}_{i2}$.

- As usual, correct for two-step estimation. Not needed to test $H_0 : \gamma_1 = 0$ using t statistic on $\hat{v}_{i2}$.
- Unlike in the binary selection case, an exclusion restriction is not needed. That is, we can take $\mathbf{x}_1 = \mathbf{x}$. There is variation in $v_{i2} = y_{i2} - \mathbf{x}_i\boldsymbol{\delta}_2$ that is not a deterministic function of $\mathbf{x}_i$ because y_{i2} has (some) continuous variation.
- Assumes y_2 does not appear in y_1 equation. If

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1$$
$$y_2 = \max(0, \mathbf{z}\boldsymbol{\delta}_2 + v_2)$$

where $s_2 = 1[y_2 > 0]$, above approach works with $\mathbf{x}_1 = (\mathbf{z}_1, y_2)$.

• Including $\hat{v}_2$ simultaneous controls for the endogeneity of y_2 and also the sample selection problem. Now, we do need something appearing in $\mathbf{z}$ (with nonzero coefficient in $\boldsymbol{\delta}_2$) that does not appear in $\mathbf{z}_1$.

• If y_2 (the corner solution) is better described as following, say, a Cragg Hurdle model, than the probit selection approach can be used (because $P(y_2 > 0|\mathbf{x})$ is assumed to follow a probit.

• One can use the previous approach to intentionally select on a corner solution variable to obtain simple estimators. For example, suppose

$$\begin{aligned} y_1 &= 1[\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1 > 0] \\ y_2 &= \max(0, \mathbf{z}\boldsymbol{\delta}_2 + v_2) \end{aligned}$$

and there is *no* selection problem: we observe a random sample on $(y_1, y_2, \mathbf{z})$.

• Assume (u_1, v_2) is independent of $\mathbf{z}$. Under joint normality of (u_1, v_2) with $Var(u_1) = 1$, can use MLE on all the data.

- An alternative is a two-step method that uses only the $y_{i2} > 0$ observations in the second step. But must keep track of parameters.
- Write

$$\begin{aligned} u_1 &= \rho_1 v_2 + e_1 \\ v_2|\mathbf{z} &\sim Normal(0, \tau_2^2) \end{aligned}$$

so that $Var(e_1) = 1 - \rho_1^2 \tau_2^2 \equiv \eta_1^2$.

- Then

$$y_1 = 1[\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1 v_2 + e_1 > 0]$$

and so

$$P(y_1 = 1|\mathbf{z}, v_2) = \Phi(\mathbf{z}_1\boldsymbol{\delta}_{\eta 1} + \alpha_{\eta 1} y_2 + \rho_{\eta 1} v_2) \equiv \Phi(\mathbf{x}_1\boldsymbol{\beta}_{\eta 1} + \rho_{\eta 1} v_2)$$

where $\boldsymbol{\beta}_{\eta 1} = \boldsymbol{\beta}_1/\eta_1$.

- After Tobit to get $\hat{\boldsymbol{\delta}}_2$ and $\hat{\tau}_2^2$, define the Tobit residuals, $\hat{v}_{i2} = y_{i2} - \mathbf{z}_i\hat{\boldsymbol{\delta}}_2$ when $y_{i2} > 0$. Then, probit of y_{i1} on $\mathbf{x}_{i1}, \hat{v}_{i2}$ using only $y_{i2} > 0$ observations to estimate the scaled parameters. Get $\hat{\boldsymbol{\beta}}_{\eta 1}$, $\hat{\rho}_{\eta 1}$.

• Test for endogeneity is immediate. But need to recover $\boldsymbol{\beta}_1 = (\boldsymbol{\delta}_1', \alpha_1)'$ to get APEs, so we need to be able to estimate η_1. But

$$
\begin{aligned}
1 + \rho_{\eta 1}^2 \tau_2^2 &= 1 + (\rho_1^2/\eta_1^2)\tau_2^2 \\
&= 1 + \left(\frac{\rho_1^2}{1 - \rho_1^2 \tau_2^2} \right) \tau_2^2 = \frac{(1 - \rho_1^2 \tau_2^2) + \rho_1^2 \tau_2^2}{1 - \rho_1^2 \tau_2^2} \\
&= \frac{1}{1 - \rho_1^2 \tau_2^2} = 1/\eta_1^2.
\end{aligned}
$$

Therefore,

$$
1/\eta_1 = (1 + \rho_{\eta 1}^2 \tau_2^2)^{1/2}.
$$

- So, after estimating the Tobit and then the probit with $\hat{v}_{i2}$ as a regressor, we estimate the unscaled coefficients as

$$\hat{\boldsymbol{\beta}}_1 = (1 + \hat{\rho}_{\eta 1}^2 \hat{\tau}_2^2)^{1/2} \hat{\boldsymbol{\beta}}_{\eta 1}$$

- The unscaled estimates are too small (although this does not necessarily mean that partial effects would be too small).
- Easy to bootstrap both stages to avoid using the delta method.