Lecture 9
Models for Censored and Truncated Data – Truncated Regression and Sample Selection

1

Censored and Truncated Data: Definitions

• Y is **censored** when we observe X for all observations, but we only know the true value of Y for a restricted range of observations. Values of Y in a certain range are reported as a single value or there is significant clustering around a value, say 0.

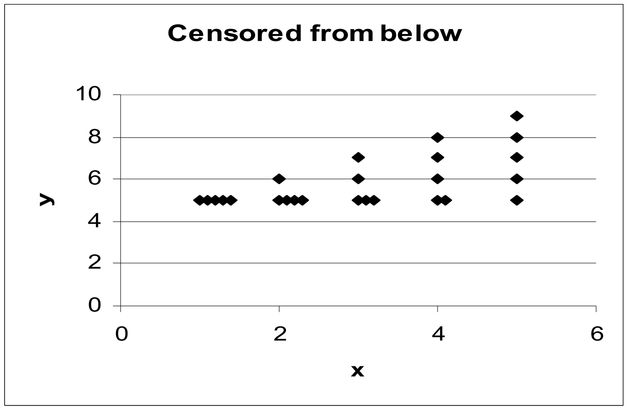
- If $Y = k$ or $Y > k$ for all $Y \Rightarrow Y$ is *censored from below* or *left-censored.*

- If $Y = k$ or $Y < k$ for all $Y \Rightarrow Y$ is *censored from above* or *right-censored.*

We usually think of an uncensored Y, Y^*, the true value of Y when the censoring mechanism is not applied. We typically have all the observations for $\{Y,X\}$, but not $\{Y^*, X\}$.

• Y is **truncated** when we only observe X for observations where Y would not be censored. We do not have a full sample for $\{Y, X\}$, we exclude observations based on characteristics of Y.

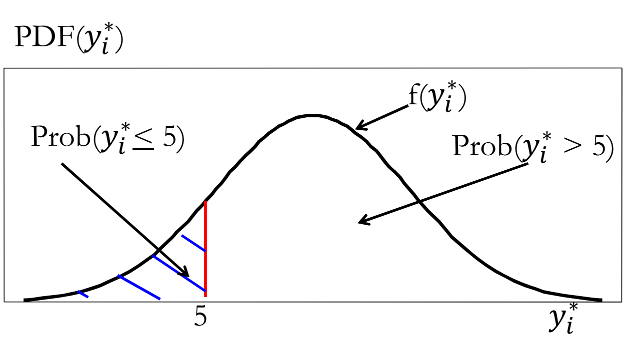
Censored from below: Example



• If $Y \leq 5$, we do not know its exact value.

Example: A Central Bank intervenes if the exchange rate, Y, hits the band's lower limit. If $Y \leq \bar{E} \Rightarrow Y = \bar{E}$.

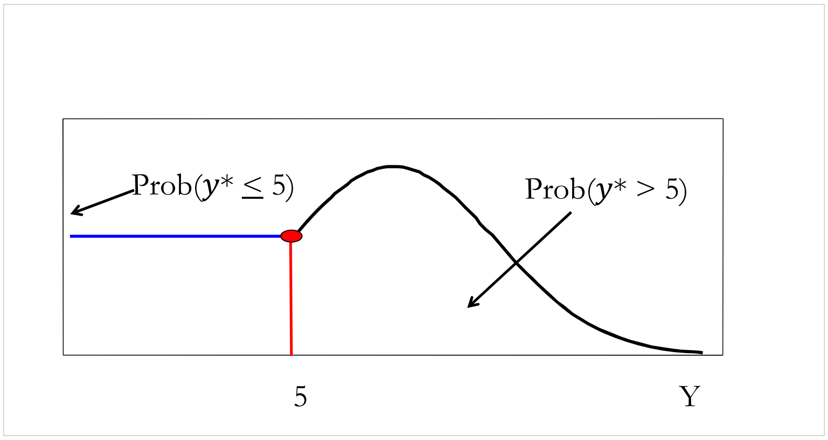
Censored from below: Example



• The pdf of the observable variable, y, is a mixture of discrete (prob. mass at Y=5) and continuous (Prob[Y^*>5]) distributions.

4

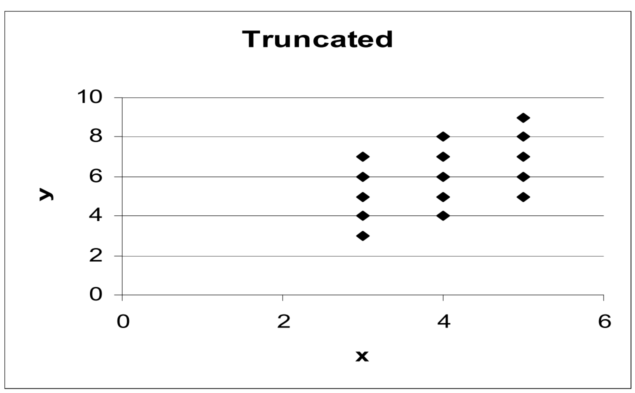
Censored from below: Example



• Under censoring we assign the full probability in the censored region to the censoring point, 5.

5

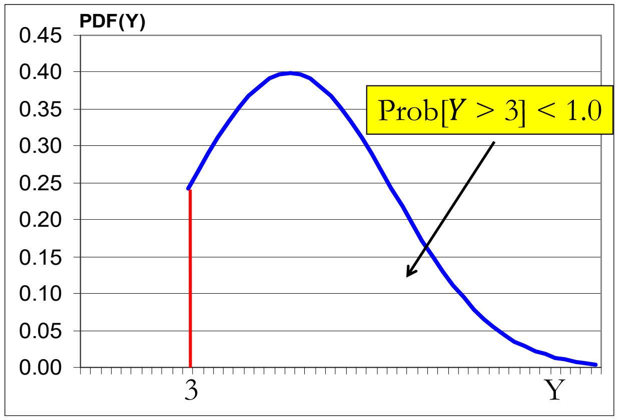
Truncated Data: Example



• If $Y < 3$, the value of X (or Y) is unknown. *(Truncation from below.)*

Example: If a family's income is below certain level, we have no information about the family's characteristics.

Truncated Data: Example



• Under data censoring, the censored distribution is a combination of a pmf plus a pdf. They add up to 1. We have a different situation under truncation. To create a pdf for Y we will use a conditional pdf.

Truncated regression

• Truncated regression is different from censored regression in the following way:

Censored regressions: The dependent variable may be censored, but you can include the censored observations in the regression

Truncated regressions: A subset of observations are dropped, thus, only the truncated data are available for the regression.

• Q: Why do we have truncation?

(1) ***Truncation by survey design***: Studies of poverty. By survey's design, families whose incomes are greater than that threshold are dropped from the sample.

(2) **Incidental Truncation**: Wage offer married women. Only those who are working have wage information. It is the people's decision, not the survey's design, that determines the sample selection.

8

Truncation and OLS

Q: What happens when we apply OLS to a truncated data?

- Suppose that you consider the following regression:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \boldsymbol{x}_i'\beta + \varepsilon_i,$$

- We have a random sample of size N. All CLM assumptions are satisfied. (The most important assumption is (**A2**) E$[\varepsilon_i|\boldsymbol{x}_i] = 0$.)

- Instead of using all the N observations, we use a subsample. Then, run OLS using this sub-sample (truncated sample) only.

• Q: Under what conditions, does **sample selection** matter to OLS?

(A) OLS is Unbiased

(A-1) Sample selection is randomly done.

(A-2) Sample selection is determined solely by the value of ***x*-variable**. For example, suppose that x is age. Then if you select sample if age is greater than 20 years old, this OLS is unbiased.

9

Truncation and OLS

(B) OLS is Biased

(B-1) Sample selection is determined by the value of ***y*-variable**.

Example: We are studying the determinants of hedging, y. We select the sample if y is greater than certain threshold. Then this OLS is biased.

(B-2) Sample selection is correlated with ε_i.

Example: We run a wage regression $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where ε_i contains unobserved ability. If sample is selected based on the unobserved ability, this OLS is biased.

- In practice, this situation happens when the selection is based on the survey participant's decision. Since the decision to participate is likely to be based on unobserved factors which are contained in ε_i, the selection is likely to be correlated with ε_i.

10

Truncation and OLS: When does (A2) hold?

• Consider the previous regression:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- All CLM assumptions are satisfied.

- Instead of using all the N observations, we use a subsample. Let s_i be a selection indicator: If $s_i = 1$, then person i is included in the regression. If $s_i = 0$, then person i is dropped from the data.

• If we run OLS using the selected subsample, we use only the observation with $s_i = 1$. That is, we run the following regression:

$$s_i y_i = \beta_0 s_i + \beta_1 s_i x_i + s_i \varepsilon_i$$

Now, $s_i x_i$ is the explanatory variable, and $u_i = s_i \varepsilon_i$ is the error term.

• OLS is unbiased if $E[u_i = s_i \varepsilon_i | s_i x_i] = 0$.

⇒ under what conditions is this new (**A2**) satisfied?

11

Truncation and OLS: When does (A2) hold?

Q: When does $E[u_i = s_i \varepsilon_i | s_i x_i]$ hold?

It is sufficient to check: $E[u_i | s_i x_i]$. (If this is zero, then new (**A2**) is also zero.)

• $E[u_i | s_i x_i] = s_i E[\varepsilon_i | s_i x_i]$ - s_i is in the conditional set.

• It is sufficient to check the condition which ensures $E[\varepsilon_i | x_i, s_i] = 0$.

• CASES:

(A-1) Sample selection is done randomly.

s_i is independent of ε_i and x_i ⇒ $E[\varepsilon_i | x_i, s_i] = E[\varepsilon_i | x_i]$

Since CLM assumptions are satisfied ⇒ we have $E[\varepsilon_i | x_i] = 0$.

⇒ OLS is unbiased.

12

Truncation and OLS: When does (A2) hold?

(A-2) Sample is selected based solely on the value of x-variable.

Example: We study trading in stocks, y_i. One of the dependent variables, x_i, is wealth, and we select person i if wealth is greater than 50K. Then,

$s_i = 1$ if $x_i \geq$ 50K,

$s_i = 0$ if $x_i <$ 50K.

-Now, s_i is a deterministic function of x_i.

• Since s_i is a deterministic function of x_i, $s_i(x_i)$, it drops out from the conditioning set. Then,

$$\text{E}[\varepsilon_i | x_i, s_i] = \text{E}[\varepsilon_i | x_i, s_i(x_i)]$$

$$= \text{E}[\varepsilon_i | x_i] = 0$$ - CLM assumptions satisfied.

⇒ OLS is unbiased.

13

Truncation and OLS: When does (A2) hold?

(B-1) *Sample selection is based on the value of y-variable.*

Example: We study determinants of wealth, y. We select individuals whose wealth is smaller than 150K. Then, $s_i = 1$ if $y_i <$150K.

- Now, s_i depends on y_i (and ε_i). It cannot be dropped out from the conditioning set like we did before. Then,

$$\text{E}[\varepsilon_i | x_i, \ s_i] \neq \text{E}[\varepsilon_i | x_i] = 0.$$

• For example, $\text{E}[\varepsilon_i | x_i, s_i] = \text{E}[\varepsilon_i | x_i, s_i(x_i)]$

$$\text{E}[\varepsilon_i | x_i, s_i = 1] = \text{E}[\varepsilon_i | x_i, y_i \leq 150\text{K}]$$

$$= \text{E}[\varepsilon_i | x_i, \beta_0 + \beta_1 x_i + \varepsilon_i \leq 150\text{K}]$$

$$= [\varepsilon_i | x_i, \varepsilon_i \leq 150\text{K} - (\beta_0 + \beta_1 x_i)]$$

$$\neq \text{E}[\varepsilon_i | x_i] = 0$$ ⇒ OLS is biased.

14

Truncation and OLS: When does (A2) hold?

(B-2) *Sample selection is correlated with* u_i.
The inclusion of a person in the sample depends on the person's decision, not the surveyor's decision. This type of truncation is called the *incidental truncation*. The bias that arises from this type of sample selection is called the **Sample Selection Bias**.

Example: Dividend payments model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

Since it is a company's decision to pay dividends –i.e., to participate–, this sample selection is likely to be based on some unobservable factors which are contained in ε_i. Like in **(B-1)**, s_i cannot be dropped out from the conditioning set:

$$\text{E}[\varepsilon_i | x_i, s_i] \neq \text{E}[\varepsilon_i | x_i] = 0 \qquad \Rightarrow \text{OLS is biased.}$$

15

Truncation and OLS: When does (A2) hold?

• CASE (A-2) can be more complicated, when the selection rule based on the ***x*-variable** may be correlated with ε_i.

Example: x is IQ. A survey participant responds if IQ > v.
Now, the sample selection is based on x-variable *and* a random error v.

Q: If we run OLS using only the truncated data, will it cause a bias?
Two cases:
- (1) If v is independent of ε, then it does not cause a bias.
- (2) If v is correlated with ε, then this is the same case as **(B-2)**. Then, OLS will be biased.

16

Estimation with Truncated Data.

• CASES

- Under cases (A-1) and (A-2), OLS is appropriate.
- Under case **(B-1)**, we use **Truncated regression**.
- Under case **(B-2)** –i.e., incidental truncation-, we use the ***Heckman Sample Selection Correction*** method. This is also called the **Heckit model**.

17

Truncated Regression

• Data truncation is **(B-1)**: the truncation is based on the ***y*-variable**.

• We have the following regression satisfies all CLM assumptions:

$$y_i = \boldsymbol{x}_i'\beta + \varepsilon_i, \qquad \varepsilon_i \sim \text{N}(0, \sigma^2)$$

- We sample only if $y_i < c_i$

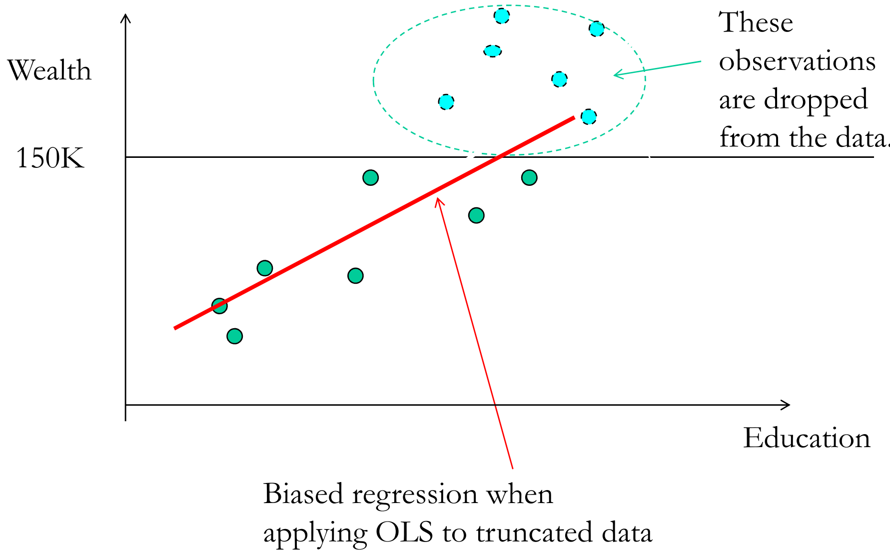
$\Rightarrow$ Observations are dropped if $y_i \geq c_i$ by design.

- We know the exact value of c_i for each person.

• We know that OLS on the truncated data will be biased. The model that produces unbiased estimate is based on ML Estimation.

18

Truncated Regression



Truncated Regression: Conditional Distribution

• Given the normality assumption for ε_i, ML is easy to apply.

- For each, $\varepsilon_i = y_i - \boldsymbol{x}_i'\beta$, the likelihood contribution is $f(\varepsilon_i)$.

- But, we select sample only if $y_i < c_i$

⇒ we have to use the density function of ε_i conditional on $y_i < c_i$:

$$f(\varepsilon_i|\, y_i < c_i) = f(\varepsilon_i|\, \varepsilon_i < c_i - \boldsymbol{x}_i'\beta) = \frac{f(\varepsilon_i)}{P(\varepsilon_i < c_i - \boldsymbol{x}_i'\beta)}$$

$$= \frac{f(\varepsilon_i)}{P\left(\frac{\varepsilon_i}{\sigma} < \frac{c_i - \boldsymbol{x}_i'\beta}{\sigma}\right)} = \frac{f(\varepsilon_i)}{\Phi\left(\frac{c_i - \boldsymbol{x}_i'\beta}{\sigma}\right)}$$

$$= \frac{\frac{1}{\sigma}\phi\left(\frac{\varepsilon_i}{\sigma}\right)}{\Phi\left(\frac{c_i - \boldsymbol{x}_i'\beta}{\sigma}\right)}$$

Truncated Normal (Again)

• Moments:

Let $y^* \sim$ N(μ^*, σ²) and $\alpha = \frac{(c - \mu^*)}{\sigma}$.

- **First moment**:

E[y^* | $y > c$] = μ^* + σ λ(α) <= This is the truncated regression.

⇒ If μ >0 and the truncation is from below –i.e., λ(α)>0–, the mean of the truncated variable is greater than the original mean

Note: For the standard normal distribution λ(α) is the mean of the truncated distribution.

- **Second moment**:

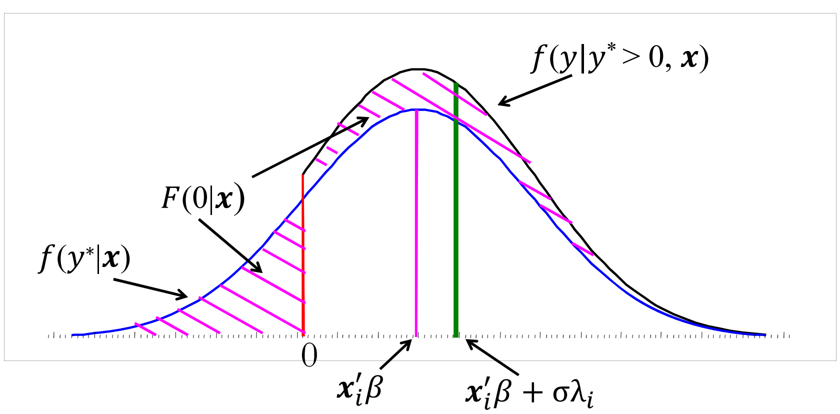
- Var[y^* | $y > c$] = σ²[1 − $\delta(\alpha)$] where $\delta(\alpha)$= λ(α) [λ(α) − α]

⇒ Truncation reduces variance! This result is general, it applies to upper or lower truncation given that 0 ≤ $\delta(\alpha)$ ≤ 1

Truncated Normal (Again)

Model: $y_i^* = \boldsymbol{x}_i'\beta + \varepsilon_i$

Observed Data: $y_i = y_i^* \,|\, y_i^* > 0$



• Truncated (from below, $y_i^* > 0$) regression model:

$$\text{E}[y_i|\, y_i^* > 0, \boldsymbol{x}_i] = \boldsymbol{x}_i'\beta + \sigma\lambda_i > \text{E}[y_i|\boldsymbol{x}_i]$$

Truncated Regression: ML Estimation

• The likelihood contribution for i^{th} observation is given by

$$L_i(\boldsymbol{\beta}, \sigma) = \frac{\frac{1}{\sigma}\phi(\frac{y_i - x_i'\beta}{\sigma})}{\Phi\left(\frac{c_i - x_i'\beta}{\sigma}\right)}$$

ln(joint density of N values of y_i^*)

• The likelihood function is given by (with $c_i = 0$):

$$Log\ L(\beta, \sigma) = \sum_{i=1}^{N} \log L_i = -\frac{N}{2}[\log(2\pi) + \log(\sigma^2)] - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\varepsilon_i^2$$

$$-\sum_{i=1}^{N}\log\left[\Phi(\frac{x_i'\beta}{\sigma})\right]$$

log(joint probability of $y_i^* > 0$)

• The values of $(\boldsymbol{\beta}, \sigma)$ that maximizes Log L are the ML estimators of the *Truncated Regression.*

23

The partial effects

• The estimated parameters β_k measures the effect of x_k on y for participating individual. Thus,

$$\frac{\delta E[y_i | y_i > 0, \boldsymbol{x_i}'\boldsymbol{\beta}]}{\delta x} = \beta_k + \sigma\,\frac{\delta\lambda(\boldsymbol{x}_i'\boldsymbol{\beta})}{\delta x} = \beta_k + \sigma\,\frac{\delta\lambda(\boldsymbol{x}_i'\boldsymbol{\beta})}{\delta x} =$$
$$= \beta_k * (1 - d_i)$$

with $d_i = \lambda(\boldsymbol{x}_i'\boldsymbol{\beta}) * [\lambda(\boldsymbol{x}_i'\boldsymbol{\beta}) + \boldsymbol{x}_i'\boldsymbol{\beta}]$.

24

Truncated Regression: MLE – Example

• DATA: From a survey of family income in Japan (JPSC_familyinc.dta). The data is originally not truncated.

Model: $y_i = \beta_0 + \beta_1 x_i + \upsilon_i$

y_i = family income in JPY 10,000

x_i: husband's education

• Three cases:

EX1. Use all observations to estimate model

EX2. Truncate sample from above (y_i < 800). Then run the OLS using on the truncated sample.

EXe. Run the truncated regression model for the data truncated from above.


```
. reg familyinc huseduc

      Source |       SS       df       MS              Number of obs =    7695
-------------+------------------------------           F(  1,  7693) =  924.22
       Model |  38305900.9     1  38305900.9           Prob > F      =  0.0000
    Residual |   318850122  7693  41446.7856           R-squared     =  0.1073
-------------+------------------------------           Adj R-squared =  0.1071
       Total |   357156023  7694  46420.0705           Root MSE      =  203.58

------------------------------------------------------------------------------
   familyinc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     huseduc |   32.93413   1.083325    30.40   0.000     30.81052    35.05775
       _cons |    143.895   15.09181     9.53   0.000     114.3109     173.479
------------------------------------------------------------------------------
```

OLS using all the observations, unbiased estimated β_1= 32.93413.

```
. reg familyinc huseduc if familyinc<800

      Source |       SS       df       MS              Number of obs =    6274
-------------+------------------------------           F(  1,  6272) =  602.70
       Model |  11593241.1     1  11593241.1           Prob > F      =  0.0000
    Residual |   120645494  6272  19235.5699           R-squared     =  0.0877
-------------+------------------------------           Adj R-squared =  0.0875
       Total |   132238735  6273  21080.621            Root MSE      =  138.69

------------------------------------------------------------------------------
   familyinc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     huseduc |   20.27929   .8260432    24.55   0.000     18.65996    21.89861
       _cons |   244.5233   11.33218    21.58   0.000     222.3084    266.7383
------------------------------------------------------------------------------
```

OLS on truncated sample.

The parameter on husband's education is biased towards zero.

```
. truncreg familyinc huseduc, ul(800)
(note: 1421 obs. truncated)

Fitting full model:

Iteration 0:   log likelihood = -39676.782
Iteration 1:   log likelihood = -39618.757
Iteration 2:   log likelihood = -39618.629
Iteration 3:   log likelihood = -39618.629

Truncated regression
Limit:   lower =       -inf                    Number of obs =    6274
         upper =        800                    Wald chi2(1)  =  569.90
Log likelihood = -39618.629                    Prob > chi2   =  0.0000
```

familyinc	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
huseduc	24.50276	1.0264	23.87	0.000	22.49105	26.51446
_cons	203.6856	13.75721	14.81	0.000	176.7219	230.6492
/sigma	153.1291	1.805717	84.80	0.000	149.59	156.6683

Truncated regression model on the truncated sample

Note: Bias seems to be corrected, but not perfect in this example.

27

Sample Selection Bias Correction Model

• The most common case of truncation is **(B-2)**: Incidental truncation.

• This data truncation usually occurs because sample selection is determined by the people's decision, not the surveyor's decision.

• Back to the wage regression example. If person i has chosen to participate (work), person i has *self-selected into the sample.* If person i has decided not to participate, person i has self-selected out of the sample.

• The bias caused by this type of truncation is called *sample selection bias.*

• This model involves two decisions: (1) participation and (2) amount. It is a generalization of the Tobit Model.

28

Tobit Model: Type II

• Different ways of thinking about how the latent variable and the observed variable interact produce different Tobit Models.

• The Type I Tobit Model presents a simple relation:

- $y_i = 0$ if $y_i^* = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i \leq 0$
  $= y_i^* = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i,$ if $y_i^* = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i > 0$

The effect of the X's on the probability that an observation is censored and the effect on the conditional mean of the non-censored observations are the same: $\boldsymbol{\beta}$.

• The Type II Tobit Model presents a more complex relation:

- $y_i = 0$ if $y_i^* = \boldsymbol{x}_i'\alpha + \varepsilon_{1,i} \leq 0$, $\varepsilon_{1,i} \sim N(0, 1)$
  $= \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_{2,i}$ if $y_i^* = \boldsymbol{x}_i'\alpha + \varepsilon_{2,i} > 0$, $\varepsilon_{2,i} \sim N(0, \sigma_2^2)$

Now, we have different effects of the $\boldsymbol{x}$'s.

29

Tobit Model: Type II

• The Type II Tobit Model:

- $y_i = 0$ if $y_i^* = \boldsymbol{x}_i'\boldsymbol{\alpha} + \varepsilon_{1,i} \leq 0$, $\varepsilon_{1,i} \sim N(0, \sigma_1^2=1)$
  $= \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_{2,i}$ if $y_i^* = \boldsymbol{x}_i'\boldsymbol{\alpha} + \varepsilon_{2,i} > 0$, $\varepsilon_{2,i} \sim N(0, \sigma_2^2)$

- A more flexible model. $\boldsymbol{x}$ can have an effect on the decision to participate (Probit part) and a different effect on the amount decision (truncated regression).

• Type I is a special case: $\varepsilon_{2,i} = \varepsilon_{1,i}$ and $\boldsymbol{\alpha} = \boldsymbol{\beta}$.

Example: Age affects the decision to donate to charity. But it can have a different effect on the amount donated. We may find that age has a positive effect on the decision to donate, but given a positive donation, younger individuals donate more than older individuals.

30

Tobit Model: Type II

• The model assumes a **bivariate normal distribution** for ($\varepsilon_{1,i}$, $\varepsilon_{2,i}$); with covariance given by $\sigma_{12}(= \rho\ \sigma_1 \sigma_2)$.

- Conditional expectation:

$$E[y_i \mid y_i > 0, \boldsymbol{x}_i] = \boldsymbol{x}_i{}'\boldsymbol{\beta} + \sigma_{12}\ \lambda(\boldsymbol{x}_i{}'\boldsymbol{\alpha}) \qquad (\sigma_{12}(= \rho\ \sigma_2))$$

- Unconditional Expectation

$$E[y_i \mid \boldsymbol{x}_i] = \text{Prob}(y_i{>}0 \mid \boldsymbol{x}_i) * E[y_i \mid y_i{>}0, \boldsymbol{x}_i] + \text{Prob}(y_i{=}0 \mid \boldsymbol{x}_i) * 0$$
$$= \text{Prob}(y_i > 0 \mid \boldsymbol{x}_i) * E[y_i \mid y_i > 0, \boldsymbol{x}_i]$$
$$= \Phi(\boldsymbol{x}_i{}'\boldsymbol{\alpha}) * [\boldsymbol{x}_i{}'\boldsymbol{\beta} + \sigma_{12}\ \lambda(\boldsymbol{x}_i{}'\boldsymbol{\alpha})]$$

Note: This model is known as the Heckman selection model, or the Type II Tobit model (Amemiya), or the probit selection model (Wooldridge). 31

Tobit Model: Type II – Sample selection

• Now, we generalize the model presented, making the decision to participate dependent on a different variable, z. Then,

- $y_i = 0$ if $y_i^* = z_i{}'\boldsymbol{\alpha} + \varepsilon_{1,i} \leq 0$, $\varepsilon_{1,i} \sim N(0, \sigma_2{}^2{=}1)$
 $= \boldsymbol{x}_i{}'\boldsymbol{\beta} + \varepsilon_{2,i}$ if $y_i^* = z_i{}'\boldsymbol{\alpha} + \varepsilon_{2,i} > 0$, $\varepsilon_{2,i} \sim N(0, \sigma_2{}^2)$

• This model is called the **Sample selection model**, due to Heckman.

Example (from Heckman (*Econometrica*, 1979): Structural Labor model:

• Labor Supply equation:

$$h_i^* = \delta_0 + \delta_1\ w_i + \boldsymbol{Z}_i{}'\boldsymbol{\delta}_2 + \varepsilon_i \qquad (1)$$

- h_i^*: desired hours by i^{th} person (latent variable)
- w_i: wage that could be earned
- $\boldsymbol{Z}_i$: non-labor income, taste variables (married, kids, etc.)
- ε_i (error term): unobserved taste for work. 32

Tobit Model: Type II – Sample selection

Example (from Heckman) (continuation)
• Market wage equation (equation of interest):

$$w_i = \boldsymbol{x}_i{}'\boldsymbol{\beta} + u_i \qquad (2)$$

- $\boldsymbol{x}_i$: productivity, age, education, previous experience, etc.
- u_i (error term): unobserved wage earning ability.
- u_i & ε_i are assumed to follow a bivariate distribution (usually, a normal)

We observe w_i for only those who work –i.e., $h_i^* > 0$.

Goal: Estimation of wage offer equation for people of working age

Q: The sample is non longer random. How can we estimate (2) if we only observe w_i (wages) for those who work?

• Problem: Selection bias. Non-participation is rarely random

33

Tobit Model: Type II – Selection Bias

• Selection bias: Non-participation is rarely random
- Not distributed equally across subgroups
- Agents decide to participate or not –i.e., self-select into a group.

Q: Can we test for selection bias?

34

Tobit Model: Type II – Terminology

• Terminology:

- **Selection equation**:

$y_i^* = \boldsymbol{z}_i{}'\boldsymbol{\alpha} + \varepsilon_{1,i}$ (often, a latent variable equation, say market wage vs. value of home production)

- **Selection Rule**:

- $y_i = 0$ if $y_i^* \leq 0$ $\Rightarrow D_i = 0,$
  $= \boldsymbol{x}_i{}'\boldsymbol{\beta} + \varepsilon_{2,i}$ if $y_i^* > 0$ $\Rightarrow D_i = 1,$

- **Outcome equation:**

$y_i = \boldsymbol{x}_i{}'\boldsymbol{\beta} + \varepsilon_{2,i}$ (the primary equation of interest)

• To derive moments, we need to make an assumption about the distribution of the errors, $\varepsilon_{1,i}$ & $\varepsilon_{2,i}$. In the Heckman model, we assume a bivariate normal joint distribution.

35

Tobit Model: Type II – Expectations

• **Expectations**: Under incidental truncation with a bivariate normal distribution we have:

- Conditional expectation (when is y_i observed) :

$$E[y_i \,|\, y_i > 0, \boldsymbol{x}_i] = \boldsymbol{x}_i{}'\boldsymbol{\beta} + \sigma_{12}\, \lambda(\frac{\boldsymbol{z}_i'\boldsymbol{\alpha}}{\sigma_1})$$

- Unconditional Expectation:

$$E[y_i \,|\, \boldsymbol{x}_i] = \Phi(\frac{\boldsymbol{z}_i'\boldsymbol{\alpha}}{\sigma_1}) * [\boldsymbol{x}_i{}'\boldsymbol{\beta} + \sigma_{12}\, \lambda(\frac{\boldsymbol{z}_i'\boldsymbol{\alpha}}{\sigma_1})]$$

Note: The results look very similar to the results obtained under truncation, but now we have a different variable, $\boldsymbol{z}_i$, determining truncation.

• Again, OLS estimation on the observed part produces a biased and inconsistent estimator. The size of the bias depends on σ_{12} (or ρ).

36

Tobit Model: Type II – Conditional Expectation

• From the conditional expectation:

$$E[y_i \mid y_i > 0, \boldsymbol{x}_i] = \boldsymbol{x}_i{'}\boldsymbol{\beta} + \rho\, \sigma_1 \sigma_2\, \lambda(\frac{\boldsymbol{z}_i'\boldsymbol{\alpha}}{\sigma_1}) \qquad (\sigma_1 = 1)$$

• Above we see that applying OLS to observed sample will produce biased (and inconsistent) estimators. This is called *sample selection bias* (an omitted variable problem). It depends on σ_{12} (or ρ) and $\boldsymbol{z}$.

• But regressing y on $\boldsymbol{x}$ and λ on the sub-sample with $y_i^* > 0$ produces consistent estimates (though SE need correction). But, we need an estimator for λ. This idea is the basis of Heckman's two-step estimation.

• Estimation
- ML –complicated, but efficient
- Two-step –easier, but not efficient. Not the usual standard errors

37

Tobit Model: Type II – Partial Effects

• Marginal effects of changes in exogenous variables have two components:
- Direct effect on mean of y_i, β_i via (2)
- If a variable affects the Prob[$y_i^* > 0$], then it will affect y_i via (1).

• Marginal effect if regressor appears in both z_i and x_i:

$$\frac{\delta E[y_i \mid y_i > 0, x_i{'}\boldsymbol{\beta}]}{\delta x} = \beta_k - \alpha_k * \rho\sigma_2 * \{\lambda(\boldsymbol{z}_i'\boldsymbol{\alpha})^2 - [\,\underbrace{(-\frac{z_i'\alpha}{\sigma_1})\,\lambda(\boldsymbol{z}_i'\boldsymbol{\alpha})}_{\text{between 0 and 1}}]\}$$

• Suppose $\rho > 0$ and $E(y_i)$ is greater when $y_i^* > 0$ and given that the last term above is between 0 and 1, then, the additional term reduces the marginal effects (it controls for increased mean due to probability impacts). That is, β_k overstates partial effects.

Note: If $\rho = 0$, the partial effect is exactly given by β_k.

38

Review: Conditional Bivariate Normal

• To derive the likelihood function for the Sample selection model, we will use results from the conditional distribution of two bivariate normal RVs.

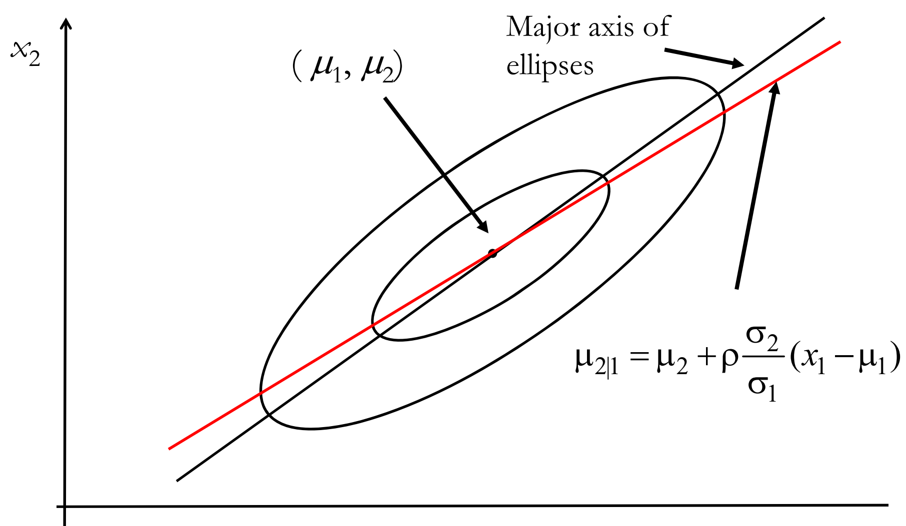
• Recall the definition of conditional distributions for continuous RVs:

$$f_{1|2}\left(x_1|x_2\right)=\frac{f\left(x_1,x_2\right)}{f_2\left(x_2\right)} \quad \text{and} \quad f_{2|1}\left(x_2|x_1\right)=\frac{f\left(x_1,x_2\right)}{f_1\left(x_1\right)}$$

• In the case of the bivariate normal distribution the conditional distribution of x_i given x_j is Normal with mean and standard deviation (using the standard notation):

$$\mu_{i|j}=\mu_i+\rho\frac{\sigma_i}{\sigma_j}\left(x_j-\mu_j\right) \quad \text{and} \quad \sigma_{i|j}=\sigma_i\sqrt{1-\rho^2}$$

Review: Conditional Bivariate Normal



Note: $\mu_{2|1}=\mu_2+(\frac{\sigma_{12}}{\sigma_1\sigma_2})\frac{\sigma_2}{\sigma_1}(x_1-\mu_1)=\mu_2+\frac{\sigma_{12}}{\sigma_1^{\ 2}}(x_1-\mu_1)$

Tobit Model: Type II – ML Estimation

• The model assumes a bivariate normal distribution for $(\varepsilon_{1,i}; \varepsilon_{2,i})$, with covariance given by $\sigma_{12} (= \rho\, \sigma_1 \sigma_2)$. We use a participation dummy variable: $D_i = 0$ (No), $D_i = 1$ (Yes).

• The likelihood reflects two contributions:

(1) Observations with $y_i = 0$ –i.e., $y_i^* = \boldsymbol{z}_i'\boldsymbol{\alpha} + \varepsilon_{1,i} \leq 0 \;\Rightarrow D_i = 0$.

- Prob$(D_i = 0\,|\,\boldsymbol{x}_i)$ = P$(y_i^* = \boldsymbol{z}_i'\boldsymbol{\alpha} + \varepsilon_{1,i} \leq 0\,|\,\boldsymbol{x}_i)$ = P$(\varepsilon_{1,i} \leq -\boldsymbol{z}_i'\boldsymbol{\alpha}\,|\,\boldsymbol{x}_i)$

$= 1 - \Phi(\boldsymbol{z}_i'\boldsymbol{\alpha})$

(2) Observations with $y_i > 0$ –i.e., $y_i^* = \boldsymbol{z}_i'\boldsymbol{\alpha} + \varepsilon_{1,i} > 0 \;\Rightarrow D_i = 1$.

- $f(y_i|D_i = 1, \boldsymbol{x}_i, \boldsymbol{z}_i)$ * Prob$(D_i = 1\,|\,\boldsymbol{x}_i, \boldsymbol{z}_i, y_i)$

(2.a) $f(y_i|D_i = 1, \boldsymbol{x}_i) = \dfrac{P(D_i = 1\,|\,\boldsymbol{x}_i, y_i) * f(y_i|\boldsymbol{x}_i)}{P(D_i = 1, \boldsymbol{x}_i)}$ (Bayes' Rule)

where $f(y_i|\boldsymbol{x}_i) = (1/\sigma_2)\, \phi((y_i - \boldsymbol{x}_i'\boldsymbol{\beta})/\sigma_2)$

41

Tobit Model: Type II – ML Estimation

(2.b) $P(y_i|D_i = 1\,|\,\boldsymbol{x}_i, \boldsymbol{z}_i, y_i) = P(\varepsilon_{1,i} > -\boldsymbol{z}_i'\boldsymbol{\alpha}|\boldsymbol{x}_i, y_i)$

$$= P\left[\frac{\varepsilon_{1,i} - (\rho/\sigma_2) * (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})}{\sqrt{\sigma_1^2(1-\rho)^2}} > \frac{-\boldsymbol{z}_i'\boldsymbol{\alpha} - (\rho/\sigma_2) * (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})}{\sqrt{\sigma_1^2(1-\rho)^2}}\right]$$

$$= 1 - \Phi\left(-\frac{\boldsymbol{z}_i'\boldsymbol{\alpha} + (\rho/\sigma_2) * (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})}{\sqrt{\sigma_1^2(1-\rho)^2}}\right)$$

$$= \Phi\left(\frac{\boldsymbol{z}_i'\boldsymbol{\alpha} + (\rho/\sigma_2) * (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})}{\sqrt{\sigma_1^2(1-\rho)^2}}\right)$$

- Moments of the conditional distribution $(y_1|y_2)$ of a normal RV:
  - Mean for RV 1: $\mu_1 + (\sigma_{12}/\sigma_2^2)\,(y_2 - \mu_2) = (\rho/\sigma_2) * (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})$
  - Variance for RV 1: $\sigma_1^2\,(1 - \rho^2) = 1 - \rho^2$ (Recall: $\sigma_1 = 1$)

42

Tobit Model: Type II – ML Estimation

• Now, we can put all the contributions together:

$L(\boldsymbol{\beta}) = \prod_{i,y_i=0} P(y_i = 0) * \prod_{i,y_i>0} \{P(y_i > 0) * f(y_i|\boldsymbol{x}_i, \boldsymbol{z}_i)\}$

• Taking logs:

$\log L(\beta,\boldsymbol{\alpha},\sigma,\rho) = \sum_{i=1}^{N}(1 - D_i) * \ln(1 - \Phi(\boldsymbol{z}_i'\boldsymbol{\alpha})) +$

$$+ \sum_{i=1}^{N} D_i * \ln\{\Phi\left(\frac{\boldsymbol{z}_i'\boldsymbol{\alpha} + (\rho/\sigma_2) * (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})}{\sqrt{\sigma_1{}^2(1-\rho)^2}}\right)\}$$

$$+ \sum_{i=1}^{N} D_i * \ln\{\frac{1}{\sigma_2}\phi(\frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma_2})\}$$

• Complicated likelihood. The algorithm tends to be badly behaved:
 ⇒ Iterative methods do not always converge to the MLE.

Note: If $\rho = 0$ this log likelihood is just the sum a Gaussian linear regression log likelihood and a probit log likelihood.

43

Tobit Model: Type II – Two-step estimator

• It is much easier two use Heckman's two-step (**Heckit**) estimator:

(1) Probit part: Estimate $\boldsymbol{\alpha}$ using ML ⇒ get $\hat{\boldsymbol{\alpha}}$

(2) Truncated regression:
- For each $D_i = 1$ (participation), calculate $\lambda_i = \lambda(\boldsymbol{z}_i'\hat{\boldsymbol{\alpha}})$.
- Regress y_i against $\boldsymbol{x}_i$ & $\lambda(\boldsymbol{z}_i'\hat{\boldsymbol{\alpha}})$ ⇒ get **b** & b_λ ($= \rho\sigma_2$).

• Problems:
- Consistent, but not efficient (relative to MLE)
- Getting Var[**b**] is not easy (we are estimating $\boldsymbol{\alpha}$ too).

• We can get consistent estimators of ρ & σ_2, individually. For each observation, the true conditional variance of the disturbance would be

$$\sigma_i{}^2 = \sigma_2{}^2 (1 - \rho^2 \delta_i) \qquad (\delta(\alpha) = \lambda(\alpha)\,[\lambda(\alpha) - \alpha])$$

where we can estimate

$$\hat{\sigma}_2{}^2 = \frac{e'e}{N} + (\textstyle\sum_{i=1}^{N} \delta_i/N)\, b_\lambda \quad \& \quad \hat{\rho} = \frac{b_\lambda{}^2}{\sigma_2{}^2}.$$

44

Tobit Model: Type II – Two-step estimator

• In theory, we can use the delta method to get SE for ρ & σ_2. But, we have heteroscedasticity and the usual 2-step SE estimation problem.

• Heckman (1979) shows the correct asymptotic covariance matrix for $\boldsymbol{\beta}$ & $\boldsymbol{\beta}_\lambda$ is given by: the following:

$$\text{Est.Asy.Var}[\beta,\beta_\lambda] = \hat{\sigma}_\varepsilon^2 [X_*'X_*]^{-1} \left[X_*' \left(I - \hat{\rho}\hat{\Delta} \right) X_* + Q \right] [X_*'X_*]^{-1}$$

where $\left(I - \hat{\rho}\hat{\Delta} \right)$ is a diagonal matrix with

$(1 - \rho^2 \delta_i)$ on the diagonal

$X_{i*} = [X_i, \lambda_i]$

$Q = \hat{\rho}^{\,2} (\boldsymbol{z}' \hat{\Delta} \boldsymbol{X}_*) Var[\hat{\boldsymbol{\alpha}}] (\boldsymbol{z}' \hat{\Delta} \boldsymbol{X}_*)$

Note: Murphy and Topel (1985) SE for 2-step estimators can be used. 45

Tobit Model: Type II – Identification

• In general, it is difficult to justify different variables for $\boldsymbol{z}_i$ and $\boldsymbol{x}_i$. This is a problem for the estimates. It creates an identification problem.

• Technically, the parameters of the model are identified, even when $\boldsymbol{z}_i = \boldsymbol{x}_i$. But, identification is based on the distributional assumptions.

• Estimates are very sensitive to assumption of bivariate normality - Winship and Mare (1992) and $\boldsymbol{z}_i = \boldsymbol{x}_i$.

• ρ parameter very sensitive in some common applications. Sartori (2003) comes with 95% C.I. for ρ = -.999999 to +0.99255!

• Identification is driven by the non-linearity in the selection equation, through λ_i (and, thus, we need variation in the $\boldsymbol{z}_i$'s too!).

46

Tobit Model: Type II – Identification

• In general, it is difficult to justify different variables for $\boldsymbol{z}_i$ and $\boldsymbol{x}_i$. This is a problem for the estimates. It creates an identification problem.

• We find that when $\boldsymbol{z}_i = \boldsymbol{x}_i$, identification tends to be tenuous unless there are many observations in the tails, where there is substantial nonlinearity in the λ_i. We need exclusion restrictions.

47

Tobit Model: Type II – Testing the model

• Q: Do we have a sample selection problem?
Based on the conditional expectation, a test is very simple. We need to test if there is an omitted variable. That is, we need to test if λ_i belongs in the conditional expectation E$[y_i \,|\, y_i > 0]$.

• Easy test: H_0: $\beta_\lambda = 0$.

We can do this test using the estimator for β_λ, b_λ, from the second step of Heckman's two-step procedure.

• Usual problems with testing.
- The test assumes correct specification. If the selection equation is incorrect, we may be unable to reject H_0.
- Rejection of H_0 does not imply accepting the alternative –i.e., sample selection problem. We may have non-linearities in the data!

48

Tobit Model: Type II – Testing the model

• Rejection of H_0 does not imply accepting the alternative –i.e., sample selection problem. We may have non-linearities in the data!

Identification issue II

We are not sure about the functional form. We may not be comfortable interpreting nonlinearities as evidence for endogeneity of the covariates.

49

Tobit Model: Type II – Application

Estimating Heckit Manually. (note: you will not get the correct standard errors.

```
. ***********************************************
. * Estimating heckit model manually            *
. ***********************************************
. **************************
. * First create selection  *
. * Variable                *
. **************************
.     gen s=0 if wage==.
(428 missing values generated)

.     replace s=1 if wage~=.
(428 real changes made)

. ******************************
. *Next, estimate the probit    *
. *selection equation           *
. ******************************
. probit s educ exper expersq nwifeinc age kidslt6 kidsge6

Iteration 0:   log likelihood =  -514.8732
Iteration 1:   log likelihood = -405.78215
Iteration 2:   log likelihood = -401.32924
Iteration 3:   log likelihood = -401.30219
Iteration 4:   log likelihood = -401.30219

Probit regression                                 Number of obs   =        753
                                                  LR chi2(7)      =     227.14
                                                  Prob > chi2     =     0.0000
Log likelihood = -401.30219                       Pseudo R2       =     0.2206
```

First step:
Probit selection equation

s	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
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

50

Tobit Model: Type II – Application

```
. ******************************
. *Then create inverse lambda   *
. ******************************
.  predict xdelta, xb

.  gen lambda =normalden(xdelta)/normal(xdelta)

. *************************************
. *Finally, estimate the Heckit model *
. *************************************
. reg lwage educ exper expersq lambda

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  4,   423) =   19.69
       Model |  35.0479487     4  8.76198719           Prob > F      =  0.0000
    Residual |  188.279492   423  .445105182           R-squared     =  0.1569
-------------+------------------------------           Adj R-squared =  0.1490
       Total |  223.327441   427  .523015084           Root MSE      =  .66716

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1090655   .0156096     6.99   0.000     .0783835    .1397476
       exper |   .0438873   .0163534     2.68   0.008     .0117434    .0760313
     expersq |  -.0008591   .0004414    -1.95   0.052    -.0017267    8.49e-06
      lambda |   .0322619   .1343877     0.24   0.810    -.2318889    .2964126
       _cons |  -.5781032    .306723    -1.88   0.060    -1.180994     .024788
------------------------------------------------------------------------------
```

Second step: Truncated regression

<u>Note</u>: The standard errors are not correct.

Tobit Model: Type II – Application

```
. heckman lwage educ exper expersq, select(s=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(3)       =     51.53
                                                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
-------------+----------------------------------------------------------------
s            |
        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
------------------------------------------------------------------------------
```

Heckit Model estimated automatically.

Note H_0: ρ =0 cannot be rejected. There is little evidence that sample selection bias is present.