Models for Incomplete Observations: Censoring, Truncation and Selection

Matteo Paradisi
(EIEF)

Applied Micro - Lecture 12
Incomplete Observations

- Today we study models where the dependent variable is not completely observed.
- We study two main cases:
  - **censoring**: $y$ is censored at some point of the distribution.
  - **truncation**: $y$ is set to missing above some point in the distribution.
A variable can be either top or bottom coded

- **Top coded**
  \[
  y = \begin{cases} 
  a & \text{if } y^* > a \\
  y^* & \text{if } y^* \leq a 
  \end{cases}
  \]

- **Bottom coded**
  \[
  y = \begin{cases} 
  b & \text{if } y^* < b \\
  y^* & \text{if } y^* \geq b 
  \end{cases}
  \]
Censored data can arise for two main reasons.

▶ First, data **artificially** top or bottom coded
  - e.g. wages above some level (ceiling on social security contributions)
  - sometimes censoring imposed to prevent identification

▶ Second, data **arise naturally** from the problem under consideration
  - e.g. charity donations, people decide not to donate and the distribution shows a mass point at zero
  - in natural censoring, the uncensored variable does not exist, true variable is already censored
Truncated Data

- Similar to censoring, but replaced with missing
- Hence, we have
  \[ y = \begin{cases} 
  y^* & \text{if } a < y^* < b \\
  \text{otherwise} 
  \end{cases} \]
- Sometimes truncation due to fact that X are missing
Implications of Censoring in OLS

Let’s consider the model

\[ y^* = X\beta + u \]

Suppose that \( y^* \) is the complete variable

Assume the model satisfies

\[ E(u) = 0 \]
\[ E(X'u) = 0 \]

However, we do not observe \( y^* \)
Implications of Censoring in OLS

- The conditional mean or regression function of the OLS is

\[ E(y^* | X) = X\beta \]

- If we run OLS on censored variable we assume that conditional mean is linear

- Consider some censoring

\[ y = \begin{cases} 
  y^* & \text{if } y^* > 0 \\
  0 & \text{if } y^* \leq 0 
\end{cases} \]
Implications of Censoring in OLS

- The conditional mean can be decomposed as

\[ E(y|X) = Pr(y = 0|X) \times 0 + Pr(y > 0|X) E(y|X, y > 0) \]

\[ = Pr(y > 0|X) E(y|X, y > 0) \]

\[ = Pr(u > -\beta X) [X\beta + E(u|u > -X\beta)] \]

- this is not linear!

- We can also rewrite it as

\[ E(y|X) = X\beta + [Pr(u > -\beta X) E(u|u > -\beta X) - (1 - Pr(u > -\beta X)) X\beta] \]

- Hence, estimation of OLS with censored variable is essentially an OLS with omitted variable!

- Notice that the omitted term is correlated with X
Implications of Truncation in OLS

- Now, consider truncated data

\[ y = \begin{cases} y^* & \text{if } y^* > 0 \\ 0 & \text{if } y^* \leq 0 \end{cases} \]

- Here the conditional mean is

\[
E(y|X) = E(y^*|X, y^* > 0) = E(X\beta + u|X, X\beta + u > 0) = X\beta + E(u|X, u > -X\beta)
\]

- We have an omitted variable problem
Dealing with Censored Data: Tobit Model

- We now introduce the **Tobit model** to solve the OLS bias.
- As we have seen before when censoring at 0

\[
E(y|X) = \Pr(u > -\beta X) [X\beta + E(u|u > -X\beta)]
\]

- **Tobit assumptions:**
  1. \(E(u) = 0\)
  2. \(E(X'u) = 0\)
  3. \(u \sim N(0, \sigma^2)\)
Dealing with Censored Data: Tobit Model

- The distributional assumption allows to derive the density of $y \mid X$
- Then we apply maximum likelihood

The likelihood contribution of censored observations is

$$Pr(y_i = 0 \mid X_i) = 1 - \Phi \left( \frac{X_i \beta}{\sigma} \right)$$
Dealing with Censored Data: Tobit Model

The likelihood contribution of non-censored observations ($y_i > 0$) is

$$f(y_i|X, y_i > 0) = f(y_i^*|X, y_i^* > 0)$$

We need to find an expression for $f$

Consider the cdf of $f$

$$F(c|y^* > 0) = Pr(y^* < c|y^* > 0) = \frac{Pr(y^* < c, y^* > 0)}{Pr(y^* > 0)} = \frac{Pr(0 < y^* < c)}{Pr(y^* > 0)} = \frac{F(c) - F(0)}{1 - F(0)}$$
Dealing with Censored Data: Tobit Model

► \( f \) is just the derivative of the cdf

\[
f (c | X, y^* > 0) = \frac{\partial F (c | y^* > 0)}{\partial c} = \frac{\partial}{\partial c} \left[ F(c) - F(0) \right] \frac{1}{1 - F(0)} = \frac{f (c)}{1 - F (0)}
\]

► Under the distributional assumptions

\[
f (c) = \frac{1}{\sigma} \phi \left( \frac{c - X\beta}{\sigma} \right) \quad \text{and} \quad 1 - F (0) = \Phi \left( \frac{X\beta}{\sigma} \right)
\]
Dealing with Censored Data: Tobit Model

- $f(c)$ is the density of a variable that integrates to 1 in $(0, +\infty)$
- We must weight this density for the share of obs above 0
- Hence

$$\Pr(y > 0|X) = \Pr(X\beta + u > 0|X) = \Pr(u > -X\beta|X)$$

$$= 1 - \Phi(-X\beta/\sigma) = \Phi(X\beta/\sigma)$$

- We have

$$f(y_i|X_i, y_i > 0) = \Phi(X_i\beta/\sigma) f(y_i|X_i, y_i^* > 0)$$

$$= \frac{1}{\sigma} \phi \left( \frac{y_i - X_i\beta}{\sigma} \right)$$
Tobit Model: Maximum Likelihood

- The individual contribution to the log-likelihood is

\[ \ell (\beta, \sigma) = 1(y_i = 0) \ln [1 - \Phi (X_i \beta / \sigma)] + 1(y_i > 0) \ln \left[ \frac{1}{\sigma} \phi \left( \frac{y_i - X_i \beta}{\sigma} \right) \right] \]

- The log-likelihood therefore is

\[ L(\beta, \sigma) = \sum_{i=1}^{N} \left\{ 1(y_i = 0) \ln [1 - \Phi (X_i \beta / \sigma)] + 1(y_i > 0) \ln \left[ \frac{1}{\sigma} \phi \left( \frac{y_i - X_i \beta}{\sigma} \right) \right] \right\} \]

- The maximization delivers estimates of \((\beta, \sigma)\)
Truncated Data Models

- Using a similar procedure, we can write a likelihood function for truncated data
- Let’s keep the assumption that $u \sim N(0, \sigma^2)$
- Take the model truncated below 0

$$y = \begin{cases} 
  y^* & \text{if } y^* > 0 \\
  \text{.} & \text{otherwise}
\end{cases}$$
Truncated Data Models

- We know that the density of the model is

\[
    f(y|X) = f(y^*|X, y^* > 0) = \frac{f(y)}{1 - F(0)}
\]

\[
    = \frac{1}{\sigma} \phi \left( \frac{y - X\beta}{\sigma} \right)
    \Phi \left( \frac{X\beta}{\sigma} \right)
\]

- The log-likelihood contribution is

\[
    \ell_i(\beta, \sigma) = -\ln \sigma + \ln \phi \left( \frac{y_i - X_i\beta}{\sigma} \right) - \ln \Phi \left( \frac{X_i\beta}{\sigma} \right)
\]

- Total log-likelihood is

\[
    L(\beta, \sigma) = -N \ln \sigma + \sum_{i=1}^{N} \left\{ \ln \phi \left( \frac{y_i - X_i\beta}{\sigma} \right) - \ln \Phi \left( \frac{X_i\beta}{\sigma} \right) \right\}
\]
Comments on Censoring and Truncation

- Censoring is 'better' than truncation
- Censored data contain more information about the true underlying distribution
- Censored observations are available (i.e. the X’s are observable)
- Truncated observations are not available
Comments on Censoring and Truncation

- Think about the marginal effects
- The type of marginal effects of main interest depends on the specific analysis
- If interested in effects on $y^*$, then $E(y^*|X) = X\beta$ and $\beta$s are already the marginal effects we need
- If interested in effects on $y$

  Censoring: $E(y|X) = \Pr(u > -X\beta) [X\beta + E(u|u > -X\beta)]$
  Truncation: $E(y|X) = X\beta + E(u|u > -X\beta)$

- When truncation or censoring is “natural” consequence of data structure, we want marginal effect on $y$
- When it arises because of some artifact, then we probably want marginal effect on $y^*$
Marginal Effects

- To write the marginal effects, we must write $E(u|u > -X\beta)$
- Use the normality assumption on $u$ distribution
- Rule with normal distributions

$$E(z|z > c) = \mu + \sigma \frac{\varphi \left( \frac{c-\mu}{\sigma} \right)}{1 - \Phi \left( \frac{c-\mu}{\sigma} \right)}$$

- Hence

$$E(u|u > -X\beta) = \sigma \frac{\varphi \left( \frac{-X\beta}{\sigma} \right)}{\Phi \left( \frac{X\beta}{\sigma} \right)} = \sigma \cdot \lambda \left( \frac{X\beta}{\sigma} \right)$$

- where $\lambda \left( \frac{X\beta}{\sigma} \right) = \frac{\varphi}{\Phi}$ is called inverse Mills ratio
Marginal Effects

- Using this result, we have

\[
\text{Censoring: } E(y|X) = \Phi \left( \frac{X\beta}{\sigma} \right) X\beta + \sigma \varphi \left( \frac{X\beta}{\sigma} \right)
\]

\[
\text{Truncation: } E(y|X) = X\beta + \sigma \cdot \lambda \left( \frac{X\beta}{\sigma} \right)
\]

- Marginal effects can be easily computed with this formulas
Sample Selection: Heckman Model

- In many cases the sample is not a random draw from the population of interest
- In many applications this is not the case
- Consider the model

\[ y = \beta_0 + \beta_1 x_1 + \ldots + \beta_K x_K + u \]

- where \( E(u|X) = 0 \)
Sample Selection: Heckman Model

- Suppose some info is missing
- we can run the model only on a selected set of $N$
- Indicator equal to 1 for those observations

$$s_i = \begin{cases} 
1 & \text{if } \{y_i, X_i\} \text{ exists} \\
0 & \text{if } \{y_i, X_i\} \text{ does not exist or is incomplete} 
\end{cases}$$
Let’s write the OLS estimator for this model

\[ \hat{\beta}_{OLS} = \left[ \sum_{i=1}^{N} s_i X'_i X_i \right]^{-1} \left[ \sum_{i=1}^{N} s_i X'_i y_i \right] \]

\[ = \beta + \left[ \sum_{i=1}^{N} s_i X'_i X_i \right]^{-1} \left[ \sum_{i=1}^{N} s_i X'_i u_i \right] \]

This estimator is consistent only if \( E(sX'u) = 0 \), which is true if \( E(u|s) = 0 \)

Hence, \( u \) must be independent of the selection process
Random Selection

- Example: suppose that $s \sim \text{Bernoulli}(p)$
- $p$ determines which fraction of the data we select
- you might do this to reduce the computational power needed
- or, data provider might give you only a random sample
- In this case, $E(u|s) = 0$
Deterministic Selection

- Suppose that selection is based on deterministic rule $g(x)$
- e.g. selection is based on age, gender, region, etc.
- Since $E(u|X) = 0$, and $s$ is a function of $X$, then $E(u|s) = 0$
- Important: $X$s that determine selection do not have to be in the dataset
Selection Based on Dependent Variable

- Truncated data arise from sample selection
- Selection based on \( y \)
- Hence \( s \) is
  \[
  s_i = \begin{cases} 
  1 & \text{if } a_1 < y < a_2 \\
  0 & \text{otherwise}
  \end{cases}
  \]
- Obviously, this selection is not exogenous
- Indeed, \( E(u|y) \) cannot be equal to 0 since \( y \) is itself a function of \( u \)
Endogenous Selection

- **Endogenous selection** arises whenever \( E(u|s) \neq 0 \)
- e.g. survey data where people asked about income,
- people at the tails of the distribution refuse to answer.
- We only observe income data for those who actually answered the question.
Endogenous Selection: Motivating Example

Motivating example in the literature: wages and labor market participation

- Individuals heterogeneous in productivity and preference for work
- More productive will receive higher offers
- $w^0_i$: wage offer received by $i$
- Workers with higher preferences for work have lower reservation wages
- $w^r_i$: reservation wage for $i$, lowest $w$ he/she would accept
Endogenous Selection: Motivating Example

- Define \( w^0_i \) and \( w^r_i \) as

\[
\begin{align*}
  w^0_i &= X_{i1} \beta_1 + u_{i1} \\
  w^r_i &= X_{i2} \beta_2 + u_{i2}
\end{align*}
\]

- Assume that \( E(u_{i1} | X_{i1}) = 0 \) and \( E(u_{i2} | X_{i2}) = 0 \)

- We want to estimate \( \beta_1 \), but people work only if wage offer high enough

\[
\begin{align*}
  w^0_i \geq w^r_i &\implies i \text{ works} \\
  w^0_i < w^r_i &\implies i \text{ is inactive/unemployed}
\end{align*}
\]
Endogenous Selection: Motivating Example

- In the data we **only observe the wage for those who work**

- Hence

\[
s_i = 1 \left( w_i^0 \geq w_i^r \right) \\
= 1 \left( X_{i1} \beta_1 + u_{i1} \geq X_{i2} \beta_2 + u_{i2} \right) \\
= 1 \left( Z_i \delta + v_i \geq 0 \right)
\]

- where \( Z_i = (X_{i1}, X_{i2}) \), \( \delta = (\beta_1, \beta_2)' \) and \( v_i = u_{i1} - u_{i2} \)

- The model is

\[
w_i^0 = X_{i1} \beta_1 + u_{i1} \\
s_i = 1 \left( Z_i \delta + v_i \geq 0 \right)
\]

- Selection is endogenous since \( v_i \) depends on \( u_{i1} \)
Solving the problem: Heckman Selection

- Let’s study a model to solve the selection problem
- This model will only work if we have some data on obs that were not selected
- Take a general model with main equation and selection equation
  \[ y_i = X_i \beta + u_i \]
  \[ s_i = 1 (Z_i \delta + v_i \geq 0) \]

- Assume: \((s_i, Z_i)\) always observed for all N
- \((y_i, X_i)\) are observed only if \(s_i = 1\)
- \(E(u|X, Z) = E(v|X, Z) = 0\)
- \(v \sim N(0, 1)\) (can be relaxed to have \(N(0, \sigma^2)\))
- \(E(u|v) = \gamma v\): imposes a linear structure to conditional mean
Heckman Selection

- Take the conditional mean

\[
E(y|X, s = 1) = X\beta + E(u|X, s = 1) = X\beta + E(u|X, v > -Z\delta)
\]

- Using the assumptions \( u = \gamma v + \zeta \) where \( \zeta \) is non-systematic with zero mean

\[
E(y|X, s = 1) = X\beta + E(u|X, v > -Z\delta) = X\beta + E(\gamma v + \zeta|X, v > -Z\delta) = X\beta + \gamma E(v|X, v > -Z\delta)
\]
Heckman Selection

- Now, let’s exploit the assumption on v’s distribution

\[
E(y|X, s = 1) = X\beta + \gamma E(v|X, v > -Z\delta)
\]

\[
= X\beta + \gamma \frac{\phi(-Z\delta)}{1 - \Phi(-Z\delta)}
\]

\[
= X\beta + \gamma \frac{\phi(Z\delta)}{\Phi(Z\delta)}
\]

\[
= X\beta + \gamma \cdot \lambda(Z\delta)
\]

- where \( \lambda(Z\delta) \) is the inverse Mills ratio
- The true conditional mean includes a second term \( \gamma \cdot \lambda(Z\delta) \)
- Excluding this term we introduce a bias (X and Z most likely overlap)
Heckman Selection

\[ E (y | X, s = 1) = X \beta + \gamma \cdot \lambda (Z \delta) \]

- Heckman: let’s include the omitted variable and estimate \( \gamma \)
- However, we must first estimate \( \delta \)
- Recover the \( \delta \) from a probit of \( s_i \) on \( Z_i \)

\[ Pr (s = 1 | Z) = \Phi (Z \delta) \]
Heckman Selection

\[ \Pr(s = 1|Z) = \Phi(Z\delta) \]

- With consistent estimates of \( \delta \) called \( \hat{\delta} \) we have
  \[ \hat{\lambda}_i = \lambda (Z_i\hat{\delta}) \]

- Then use it in regression
  \[ y_i = X_i\beta + \gamma\hat{\lambda}_i + u_i \]

- Standard errors are more complicated since \( \hat{\lambda} \) comes from a separate estimate

- Notice: estimating \( \gamma \) you can test endogeneity of selection
Heckman Selection: Additional Comments

- Consider the relationship between X and Z
- May be completely separated or completely identical
- If completely separated omitting \( \lambda (Z\delta) \) does not generate OVB
  - OLS on selected sample gives consistent estimates (we still have exogeneity)
  - unless \( E[\lambda (Z\delta)] = 0 \) the constant will be inconsistent
Heckman Selection: Additional Comments

- If completely identical: $X = Z$

- Problem of multicollinearity: Mills ratio approximately linear

$$E(y|X) \approx X\beta + a + bZ\delta = X(\beta + b\delta) + a$$

- So that cannot estimate $\beta$ consistently

- Hence, when $X = Z$ identification will only be guaranteed by non-linearity of Mills ratio

- In general, it is better to have $Z = X + Z_1$ so that there are "excluded variables", but all $X$ appear in selection equation

- This is very much like with instrumental variables

- Without $Z_1$ identification with instrumental variables would be impossible