How To Correct for Sampling Biases

James J. Heckman
University of Chicago

Econ 312, Spring 2021

THE UNIVERSITY OF CHICAGO

References:

- Amemiya, Ch. 10
- Different types of sampling
  - a. random sampling
  - b. censored sampling
  - c. truncated sampling
  - d. other non-random (exogenous stratified, choice-based)

Standard Tobit Model (Tobin, 1958) "Type I Tobit"

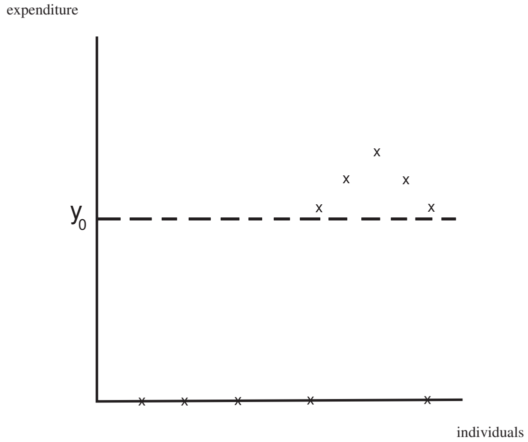
$$y_i^* = x_i\beta + u_i$$

- Observe, i.e.,

$$\begin{aligned} y_i &= y_i^* \quad \text{if } y_i^* \geq y_0 \text{ or } y_i = 1\left(y_i^* \geq y_0\right) y_i^* \\ y_i &= 0 \quad \text{if } y_i^* < y_0 \\ y_i &= 1 \quad (y_i^* < y_0) y_i^* \end{aligned}$$

- Tobin's example-expenditure on a durable good only observed if good is purchased

Figure 1



Note: Censored observations might have bought the good if price had been lower.

- Estimator. Assume $y_i^* / x_i \sim N\left(x_i \beta, \sigma_u^2\right)$

Density of Latent Variables

$$g(y^*) = f(y_i^*|y_i^* < y_0)\Pr(y_i^* < y_0) + f(y_i^*|y_i \geq y_0) \cdot \Pr(y_i^* \geq y_0)$$

$$\Pr(y_i^* < y_0) = \Pr(x_i\beta + u_i < y_0) = \Pr\left(\frac{u_i}{\sigma_u} < \frac{y_{0-}x_i\beta}{\sigma_u}\right) = \Phi\left(\frac{y_{0-}x_i\beta}{\sigma_u}\right)$$

$$f(y_i^*|y_i^* \geq y_0) = \frac{\frac{1}{\sigma_u}\phi\left(\frac{y_i^* - x_i\beta}{\sigma_u}\right)}{1 - \Phi\left(\frac{y_0 - x_i\beta}{\sigma_u}\right)}$$

- **Question: Why?**

$$\begin{aligned} &\Pr(y^* = y_i^*|y_0 \leq y^*) \\ =&\Pr(x\beta + u = y_i^*|y_0 \leq x\beta + u) \\ &\Pr\left(\frac{u}{\sigma_u} = \frac{y_i^* - x\beta}{\sigma_u}|\frac{u}{\sigma_u} \geq \frac{y_0 - x\beta}{\sigma_u}\right) \end{aligned}$$

- Note that likelihood can be written as:

$$\mathcal{L} = \underbrace{\Pi_0 \Phi\left(\frac{y_0 - x_i\beta}{\sigma_u}\right)\Pi_1\left(1 - \Phi\left(\frac{y_0 - x_i\beta}{\sigma_u}\right)\right)}_{\substack{\text{This part you would set} \\ \text{with just a simple probit}}} \Pi_1 \underbrace{\frac{\frac{1}{\sigma_u}\phi\left(\frac{y_i^* - x_i\beta}{\sigma_u}\right)}{\left\{1 - \Phi\left(\frac{y_0 - x_i\beta}{\sigma_u}\right)\right\}}}_{\text{Additional information}}$$

- You could estimate β up to scale using only the information on whether $y_i \gtreqless y_0$, but will get more efficient estimate using additional information.

  * if you know y_0, you can estimate σ_u.

Truncated Version of Type I Tobit

Observe $y_i = y_i^*$ if $y_i^* > o$

$$\left(\begin{array}{c} \text{observe nothing for truncated observations} \\ \text{example: only observe wages for workers} \end{array} \right)$$

$$\begin{aligned} \text{Likelihood: } \mathcal{L} &= \Pi_1 \frac{\frac{1}{\sigma_u} \phi \left(\frac{y_i^* - x_i \beta}{\sigma_u} \right)}{\Phi \left(\frac{x_i \beta}{\sigma_u} \right)} \\ \Pr\left(y_i^* > 0\right) &= \Pr\left(x\beta + u > 0\right) \\ &= \Pr\left(\frac{u}{\sigma_u} > \frac{-x\beta}{\sigma_u} \right) \\ &= \Pr\left(u < \frac{x\beta}{\sigma_u} \right) \end{aligned}$$

Different Ways of Estimating Tobit

a. if censored, could obtain estimates of $\frac{\beta}{\sigma_u}$ by simple probit
b. run OLS on observations for which y_i^* is observed

$$E\left(y_i | x_i\beta + u_i \geq 0\right) = x_i\beta + \sigma_u E\left(\frac{u_i}{\sigma_u} | \frac{u_i}{\sigma_u} > \frac{-x\beta}{\sigma_u}\right) \qquad (y_0 = 0)$$

- where $E\left(y_i | x_i\beta + u_i \geq 0\right)$ is the conditional mean for truncated normal r.v and

$$\sigma_u E\left(\frac{u_i}{\sigma_u} | \frac{u_i}{\sigma_u} > \frac{-x\beta}{\sigma_u}\right) \longrightarrow \lambda\left(\frac{x_i\beta}{\sigma_u}\right) = \frac{\phi\left(\frac{-x\beta}{\sigma_u}\right)}{\Phi\left(\frac{\pi_i\beta}{\sigma_u}\right)}$$

- $\lambda\left(\frac{x_i\beta}{\sigma_u}\right)$ known as "Mill's ratio" ; bias due to censoring, can be viewed as an omitted variables problem

Heckman Two-Step procedure

- Step 1: estimate $\frac{\beta}{\sigma_u}$ by probit
- Step 2:

$$\text{form } \hat{\lambda}\left(\frac{x_i\hat{\beta}}{\sigma}\right)$$

regress

$$\begin{aligned} y_i &= x_i\beta + \sigma\,\hat{\lambda}\left(\frac{x_i\beta}{\sigma}\right) + v + \varepsilon \\ v &= \sigma\left\{\lambda\left(\frac{x_i\beta}{\sigma}\right) - \hat{\lambda}\left(\frac{x_i\beta}{\sigma}\right)\right\} \\ \varepsilon &= u_i - E\left(u_i \middle| u_i > x_i\beta\right) \end{aligned}$$

THE UNIVERSITY OF CHICAGO

- Note: errors $(v + \varepsilon)$ will be heteroskedatic;
- need to account for fact that λ is estimated (Durbin problem)
- Two ways of doing this:
  a. Delta method
  b. GMM (Newey, Economic Letters, 1984)
  c. Suppose you run OLS using all the data

$$
\begin{aligned}
E(y_i) =& \Pr(y_i^* \leq 0) \cdot 0 + \Pr(y_i^* > 0) \left[x_i\beta + \sigma_u E\left(\frac{u_i}{\sigma_u} | \frac{u_i}{\sigma_u} > \frac{-x_i\beta}{\sigma} \right) \right] \\
=& \Phi\left(\frac{x_i\beta}{\sigma}\right) \left[x_i\beta + \sigma_u \lambda \left(\frac{x_i\beta}{\sigma} \right) \right]
\end{aligned}
$$

- Could estimate model by replacing Φ with $\hat{\phi}$ and λ with $\hat{\lambda}$.
- For both (b) and (c), errors are heteroskedatic, meaning that you could use weights to improve efficiency.
- Also need to adjust for estimated regressor.
  (d) Estimate model by Tobit maximum likelihood directly.

Variations on Standard Tobit Model

$$
\begin{aligned}
y_{1i}^* &= x_{1i}\beta + u_{1i} \\
y_{2i}^* &= x_{2i}\beta + u_{2i} \\
y_{2i} &= y_{2i}^* \quad \text{if} \quad y_{1i}^* \geq 0 \\
&= 0 \quad \text{else}
\end{aligned}
$$

- Example
  - y_{2i} student test scores
  - y_{1i}^* index representing parents propensity to enroll students in school
  - Test scores only observed for population enrolled

$$
\begin{aligned}
\mathcal{L} =& \Pi_1 \left[\Pr\left(y_{1i}^* > 0\right) f\left(y_{2i} | y_{1i}^* > 0\right)\right] \Pi_0 \left[\Pr\left(y_{1i}^* \leq 0\right)\right] \\
f\left(y_{2i}^* | y_{1i}^* \geq 0\right) =& \frac{\int_0^\infty f\left(y_{1i}^*, y_{2i}^*\right) dy_{1i}^*}{\int_0^\infty f\left(y_{1i}^*\right) dy_{1i}^*} \\
=& \frac{f\left(y_{2i}\right) \int_0^\infty f\left(y_{1i}^* | y_{2i}^*\right) dy_{1i}^*}{\int_0^\infty f\left(y_{1i}^*\right) dy_{1i}^*} \\
=& \frac{1}{\sigma^2} \phi\left(\frac{y_{2i}^* - x_{2i}\beta_2}{\sigma^2}\right) \cdot \frac{\int_0^\infty f\left(y_{1i}^* | y_{2i}^*\right) dy_{1i}^*}{\Pr\left(y_{1i}^* > 0\right)}
\end{aligned}
$$

$$
\begin{aligned}
y_{1i} &\sim N\left(x_{1i}\beta_1, \sigma^2\right) \\
y_{2i} &\sim N\left(x_{2i}\beta_2, \right)
\end{aligned}
$$

THE UNIVERSITY OF CHICAGO

$$y_{1i}^{*} \mid y_{2i}^{*} \sim N\left(x_{1i}\beta_1 + \frac{\sigma_{12}}{\sigma_2^2}\left(y_{2i} - x_{2i}\beta_2\right), \sigma_1^2 - \frac{\sigma_{12}}{\sigma_2^2}\right)$$

$$E\left(y_{1i}^{*} \mid u_{2i} = y_{2i}^{*} - x_{2i}\beta\right) = x_{1i}\beta_1 + E\left(u_{1i} \mid u_{2i} = y_{2i}^{*} - x_{2i}\beta\right)$$

THE UNIVERSITY OF CHICAGO

Estimation by MLE

$$
\begin{aligned}
L &= \Pi_0 \left[1 - \Phi\left(\frac{x_{1i}\beta}{\sigma_1}\right)\right] \Pi_1 \frac{1}{\sigma_2} \cdot \phi\left(\frac{y_{2i}^* - x_{2i}\beta_2}{\sigma_2}\right) \\
&\cdot \left\{1 - \Phi\left(\frac{-\left\{x_{1i}\beta_1 + \frac{\sigma_{12}}{\sigma_2^2}\left(y_{2i} - x_{2i}\beta_2\right)\right\}}{\sigma^x}\right)\right\}
\end{aligned}
$$

- Using data on y_{2i} for which $y_{1i} > 0$

$$
\begin{aligned}
E\left(y_{2i} \mid y_{1i} > 0\right) &= x_{2i}\beta + E\left(u_{2i} \mid x_i\beta + u_{1i} > 0\right) \\
&= x_{2i}\beta + \sigma_2 E\left(\frac{u_{2i}}{\sigma_2} \mid \frac{u_{1i}}{\sigma_1} > \frac{-x_{1i}\beta_1}{\sigma_1}\right) \\
&= x_{2i}\beta + \cancel{\sigma}_2 \frac{\sigma_{12}}{\sigma_1 \cancel{\sigma}_2} E\left(\frac{u_{1i}}{\sigma_1} \mid \frac{u_{1i}}{\sigma_1} > \frac{-x_{1i}\beta_1}{\sigma_1}\right) \\
&= x_{2i}\beta_2 + \frac{\sigma_{12}}{\sigma_1}\lambda\left(\frac{-x_i\beta}{\sigma}\right)
\end{aligned}
$$

Example: Female labor supply model

$$\max \quad u(L, x)$$

$$\text{s.t.} \quad x = wH + v \qquad H = 1 - L$$

where H : hours worked

v : asset income

w given

$P_x = 1$

L : time spent at home for child care

$$\frac{\frac{\partial u}{\partial L}}{\frac{\partial u}{\partial x}} = w \qquad \text{when } L < 1$$

$$\text{reservation wage} = MRS \mid_{H=0} = w_R$$

Example: Female labor supply model

- We don't observe w_R directly.

$$\begin{aligned} \text{Model} \qquad w^0 &= x\beta + u \quad \text{(wage person would earn if they worked)} \\ w^R &= z\gamma + v \\ w_i &= w_i^0 \quad \text{if} \quad w_i^R < w_i^0 \\ &= 0 \qquad \text{else} \end{aligned}$$

- Fits within previous Tobit framework if we set

$$\begin{aligned} y_{1i}^* &= x\beta - z\gamma + u - v = w^0 - w^R \\ y_{2i} &= w_i \end{aligned}$$

Incorporate choice of H

$$
\begin{aligned}
w^0 &= x_{2i}\beta_2 + u_{2i} \qquad \text{given} \\
MRS &= \frac{\frac{\partial u}{\partial L}}{\frac{\partial u}{\partial x}} = \gamma H_i + z_i'\alpha + v_i
\end{aligned}
$$

(Assume functional form for utility function that yields this)

$$
\begin{aligned}
w^{r}\left(H_{i}=0\right) &= z_{i}'\alpha + v_{i} \\
\text{work if} \qquad w^{0} &= x_{2i}\beta_{2} + u_{2i} > z_{i}\alpha + v_{i} \\
\text{if work, then} \qquad w_{i}^{0} &= MRS \Longrightarrow x_{2i}\beta_{2} + u_{2i} = \alpha H_{i} + z_{i}\alpha + v_{i} \\
&\Longrightarrow H_{i} = \frac{x_{2i}\beta_{2-} z_{i}'\alpha + u_{2i} - v_{i}}{\gamma} \\
&= x_{1i}\beta_{1} + u_{1i} \\
\text{where} \qquad x_{1i}\beta_{1} &= \left(x_{2i}\beta_{2} - z_{i}\alpha\right)\gamma^{-1} \\
u_{1i} &= u_{2i} - v_{i}
\end{aligned}
$$

Type 3 Tobit Model

$$y_{1i}^* = x_{1i}\beta_1 + u_{1i} \longleftarrow \text{ hours}$$

$$y_{2i}^* = x_{2i}\beta_1 + u_{2i} \longleftarrow \text{wage}$$

$$\begin{aligned} y_{1i} &= y_{1i}^* \qquad \text{if } \; y_{1i}^* > 0 \\ &= 0 \qquad \textit{if } \; y_{1i}^* \leq 0 \end{aligned}$$

$$\begin{aligned} y_{2i} &= y_{2i}^* \qquad \text{if } \; y_{1i}^* > 0 \\ &= 0 \qquad \text{if } \; y_{1i}^* \leq 0 \end{aligned}$$

$$\text{Here } H_i = H_i^* \quad \text{if} \quad H_i^* > 0$$
$$= 0 \quad \text{if} \quad H_i^* \leq 0$$

$$w_i = w_i^0 \quad \text{if} \quad H_i^* > 0$$
$$= 0 \quad \text{if} \quad H_i^* \leq 0$$

- Note: Type IV Tobit simply adds

$$y_{3i} = y_{3i}^* \quad \text{if} \quad y_{1i}^* > 0$$
$$= 0 \quad \textit{if} \quad y_{1i}^* \leq 0$$

- Can estimate by
  1. maximum likelihood
  2. Two-step method

$$E\left(w_i^0 \mid H_i > 0\right) = \gamma H_i + z_i\alpha + E\left(v_i \mid H_i > 0\right)$$

Type V Tobit Model of Heckman (1978)

$$
\begin{aligned}
y_{1j}^{*} &= \gamma y_{2i} + x_{1i}\beta + \delta_2 w_i + u_{1i} \\
y_{2i} &= \gamma_2 y_{1i}^{*} + x_{2i}\beta_2 + \delta_2 w_i + u_{2i}
\end{aligned}
$$

- Analysis of an antidiscrimination law on average income of African Americans in ith state.
- Observe x_{1i}, x_{2i}, y_{2i} and w_i

$$
\begin{aligned}
w_i &= 1 \quad \text{if } y_{1i}^{*} > 0 \\
w_i &= 0 \quad \text{if } y_{1i}^{*} \leq 0
\end{aligned}
$$

- y_{2i} = average income of African Americans in the state
- y_{1i}^{*} = unobservable sentiment towards African Americans
- w_i = if law is in effect

- Adoption of Law is endogenous
- Require restriction $\gamma\delta_2 + \delta_1 = 0$ so that we can solve for y_{1j}^* as a function that does not depend on w_i.
- This class of models known as "dummy endogenous variable" models.

Coherency Problem
(Suppose Restriction Does Not Bind?)

- See notes on "Dummy Endogenous Variables in simultaneous equations."

References:

- Heckman (AER, 1990) "Varieties of Selection Bias"
- Heckman (1980), "Addendum to Sample Selection Bias as Specification Error"
- Heckman and Robb (1985, 1986)

$$
\begin{aligned}
y_1^* &= x\beta + u \\
y_2^* &= z\gamma + v \\
y_1 &= y_1^* \quad \text{if } y_2^* > 0
\end{aligned}
$$

Relaxing Parametric Assumptions in the Selection Model

$$\begin{aligned} E\left(y_1^* \mid \text{observed}\right) = & x\beta + E\left(u \mid x, z\gamma + u > 0\right) \\ & + \left[u - E\left(u \mid x, z\gamma + u > 0\right)\right] \\ & \frac{\int_{-\infty}^{\infty} \int_{-\infty}^{-z\gamma} uf\left(u, v \mid x, z\right) dvdu}{\int_{-\infty}^{\infty} \int_{-\infty}^{-z\gamma} f\left(uv \mid x, z\right) dvdu} \end{aligned}$$

- Note:
  $\Pr\left(y_2^* > 0 \mid z\right) = \Pr\left(z\gamma + u > 0 \mid z\right) = P\left(Z\right) = 1 - F_v(-z\gamma)$

$$\Rightarrow \quad F_v(-z\gamma) = 1 - P(Z)$$
$$\Rightarrow \quad -z\gamma = F_v^{-1}(1 - P(Z)) \quad \text{if } F_v$$

- Can replace $-z\gamma$ in integrals in integrals by $F_v^{-1}(1 - P(Z))$ if in addition $f(u, v \mid x, z) = f(u, v \mid z\gamma)$ (index sufficiency)
- Then

$$E(y_1^* \mid y_2 > 0) = x\beta + g(P(z)) + \varepsilon \text{ where } g(P(Z))$$

is bias or "control function."

- Semiparametric selection model-Approximate bias function by Taylor series in $P(z\gamma)$, truncated power series.