# Sample Selection

Walter Sosa-Escudero

Econ 507. Econometric Analysis. Spring 2012

April 23, 2012

# Preliminaries 1) Truncated normal distribution

$X \sim f(x)$, $X|X < a$: $X$ *truncated* in $a$. Then

$$f(x|X < a) = \frac{f(x)}{Pr(X < a)}$$

If $X \sim N(\mu, \sigma^2)$, and recalling that

$$Pr(X < a) = Pr\Big(1/\sigma(X - \mu) < 1/\sigma(a - \mu)\Big) = Pr(z < \alpha),$$

$\alpha \equiv (a - \mu)/\sigma, \quad z \equiv (x - \mu)/\sigma.$

$$\begin{aligned} f(x|X < a) &= \frac{\frac{1}{\sigma}\frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]}{\Phi(\alpha)} \\ &= \frac{(1/\sigma)\phi(z)}{\Phi(\alpha)} \end{aligned}$$

Result (no proof): if $X \sim N(\mu, \sigma^2)$, then:

$$E(X|X < a) = \mu - \sigma \frac{\phi(\alpha)}{\Phi(\alpha)}$$

- Truncated to the right: expected value moves to the left (general).
- How much? Depends on $\alpha$ and $\sigma^2$
- $\lambda(z) \equiv \phi(z)/\Phi(z)$ is known as the inverse Mills ratio

# Preliminaries 2) Latent variables and probits

Recall the probit model:

$$Pr(y = 1|x) = \Phi(x'\beta)$$

$\beta$ can be consistently estimated by MLE based on a random sample $(y_i, x_i)$, $i = 1, \ldots, n$.

Consider the regression model

$$y^* = x'\beta^* + u, \qquad u \sim N(0, \sigma^2)$$

$y^*$ not directly observable (a *latent* variable) but, instead, we observe $y = 1[y^* > 0]$.

Which parameters of the regression models can be estimated consistently with this information?

Note that:

$$\begin{aligned} P(y=1|x) &= P(y^* > 0|x) \\ &= P(u > -x'\beta^*|x) \\ &= P(u < x'\beta^*|x) \\ &= P(u/\sigma < x'\beta^*/\sigma \mid x) \\ &= \Phi(x'\beta) \end{aligned}$$

This is a *probit* model with $\beta \equiv \beta^*/\sigma$.

Based on $(y_i, x_i)$, we can estimate $\beta$ consistenly by MLE, even when we cannot estimate $\beta^*$ and $\sigma^2$ separately.

- $\sigma^2$ and $\beta^*$ are not identified with the sample $(y_i, x_i)$. For example $\beta^* = 10$ and $\sigma^2 = 2$ are observationally equivalent to the case $\beta^* = 5$ y $\sigma^2 = 1$.
- $\beta \equiv \beta^*/\sigma$ is identified.

# Sample selectivity

Consider the regression model

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

$s_i$ is a *selectivity* variable: $s_i = 1$ observed, 0 if not.

- We can think that there is a 'super sample' of size $N$ of $y_i^*, x_i, s_i$ and that we observe the 'sub sample' $y_i^*, x_i$, only when $s_i = 1$.
- Example: female labor productivity

# OLS under selectivity

With a random sample $(y_i^*, x_i)$, consistency relies on:

$$E(u_i|x_i) = 0$$

which implies $E(y_i|x_i) = x_i'\beta$.

Now we do not have a random sample, but one conditioned on $s_i = 1$. Taking conditional expectations:

$$E(y_i|x_i, s_i = 1) = x_i'\beta + E(u_i|x_i, s_i = 1)$$

OLS based on the selected sample will be inconsistent, unless $E(u_i|x_i, s_i = 1) = 0$.

- Not every selectivity mechanism makes OLS inconsistent.
- If $u$ is independent of $x$, OLS still consistent (why?).
- If $s = g(x)$, OLS still consistent.
- *Examples:* wages and education. Males with odd SSN. Males with formal education?

# An estimable model under selectivity

Consider the following equations:

$$\begin{cases} y_{1i} & = & x'_{1i}\,\beta_1 + u_{1i} & \text{(regression)} \\ y_{2i^*} & = & x'_{2i}\,\beta_2 + u_{2i} & \text{(selectivity)} \end{cases}$$

Let $y_{2i} = 1[y^*_{2i} > 0]$.

*Example:* $y_{1i} =$ wage, regression equantion determines wages based on person's characteristics $(x_{1i})$. $y_{2i} =$ net utility of work, $x_{2i}$ are observed determinantes of utility.

Assumptions:

1. $(y_{2i}, x_{2i})$ observed for everyone.
2. $(y_{1i}, x_{1i})$ observed only if $y_{2i} = 1$ (selected sample).
3. $(u_{1i}, u_{2i})$ are independent of $x_{2i}$ and have zero mean.
4. $u_{2i} \sim N(0, \sigma_2^2)$.
5. $E(u_{1i}|u_{2i}) = \gamma\, u_2$. No-observables can be related.

# Selectivity bias

Here $s_i \equiv y_{2i}$.

$$
\begin{aligned}
E(y_{1i}|x_{1i}, y_{2i}=1) &= x'_{1i}\beta_1 + E(u_{1i} \mid x_{1i}, y_{2i}=1) \\
&= x'_{1i}\beta_1 + E\big[E(u_{1i}|u_{2i}) \mid x_{1i}, y_{2i}=1\big] \\
&= x'_{1i}\beta_1 + E\big[\gamma u_{2i} \mid x_{1i}, y_{2i}=1\big] \\
&= x'_{1i}\beta_1 + \gamma E\big[u_{2i} \mid x_{1i}, y^*_{2i}>0\big] \\
&= x'_{1i}\beta_1 + \gamma E\big[u_{2i} \mid x_{1i}\ ,\ u_{2i} < x'_{2i}\beta_2\big] \\
&= x'_{1i}\beta_1 - \gamma\sigma_2\ \lambda(x'_{2i}\beta_2/\sigma_2) \\
&= x'_{1i}\beta_1 - \gamma\sigma_2\ z_i \neq x'_{1i}\beta_1
\end{aligned}
$$

with $z_i \equiv \lambda(x'_{2i}\beta_2/\sigma_2)$. OLS with the selected sample is inconsistent.

$$E(y_{1i}|x_{1i}, y_{2i} = 1) = x'_{1i}\beta_1 - \gamma\sigma_2 \; z_i \neq x'_{1i}\beta_1$$

- Inconsistency: omission of $z_i$. Heckman (1979): selectivity bias as misspecification.
- Inconsistency due to the correlation between $u_{1i}$ y $u_{2i}$, , that is $\gamma \neq 0$.

# Heckman's two-step estimator

Define $u_{1i}^* \equiv y_{1i} - x_{1i}'\beta_1 - \gamma^* z_i$, $\gamma^* \equiv -\gamma\sigma_2$. Solving:

$$y_{1i} = x_{1i}'\beta + \gamma^* z + u_{1i}^*$$

where, by construction $E(u_{1i}^* | x_{1i}, y_{2i} = 1) = 0$.

- $x_{1i}$, $z_i$ observable when $y_{2i} = 1$: OLS of $y_{1i}$ on $x_{1i}$ and $z_i$ using the selected sample gives consistent esimates of $\beta_1$ and $\gamma^*$.
- Problem: $z_i \equiv \lambda(x_{2i}'\beta_2/\sigma_2)$ is NOT observable, it depends on $\beta_2$ and $\sigma_2$.

Note that $u_{2i} \sim N(0, \sigma_2^2)$, hence:

$$P(y_{2i} = 1) = P(y_{2i}^* > 0) = P(u_{2i}/\sigma_2 < x_{2i}'\beta_2/\sigma_2) = \Phi(x_{2i}'\delta)$$

- $P(y_{2i} = 1)$ is a probit model with unknown coefficients $\delta$.
- $x_{2i}$ and $y_{2i}$ are observed for everybody: $\delta$ can be estimated using probit.
- Important: we cannot identify $\beta_{2i}$ and $\sigma_2$ separately but $\delta = \beta_{2i}/\sigma_{2i}$.

This suggests the following two-stage method:

- *Stage 1:* Obtain estimates $\hat{\delta}$ based on the probit model $P(y_{2i} = 1) = \Phi(x'_{2i}\delta)$ using the full sample. Predict $z_i$ using $\hat{z}_i = \lambda(x'_{2i}\hat{\delta})$.
- *Stage 2:* Regress $y_{2i}$ on $x_{1i}$ and $\hat{z}_i$ using the selected sample. This produces consistent estimates of $\beta_1$ and $\gamma^*$.

# Practical remarks

- The method is consistent and asymptotically normal (method of moments estimator). Standard inference works fine.
- Careful with asympototic variance. The second stage is heteroskedastic. Requires correction. See Greene (Ch. 20).
- A test of $H_0 : \gamma = 0$ may be used to check sample selectivity. Under $H_0$, the regression model with the selected sample is homoskedastic, test can be based in a model without taking care of heteroskedasticity.
- Classic issue: low power when $x_1$ is very similar to $x_2$.
- MLE? Requires bivariate normality. Complicated likelihood, rarely used.