Tobit and Selection Models

Manuel Arellano
CEMFI

January 2014

Censored Regression

Illustration 1: Top-coding in wages

- Suppose Y (log wages) are subject to "top coding" (as with social security records):

$$Y = \begin{cases} Y^* \text{ if } Y^* \leq c \\ c \text{ if } Y^* > c \end{cases}$$

- Suppose we are interested in $E(Y^*)$. Effectively it is not identified but if we assume $Y^* \sim \mathcal{N}(\mu, \sigma^2)$, then μ can be determined from the distribution of Y.
- The density of Y is of the form

$$f(r) = \begin{cases} \frac{1}{\sigma}\phi\left(\frac{r-\mu}{\sigma}\right) \text{ if } r < c \\ \Pr(Y^* \geq c) = 1 - \Phi\left(\frac{r-\mu}{\sigma}\right) \text{ if } r \geq c \end{cases}$$

- The log-likelihood function of the sample $\{y_1, ..., y_N\}$ is

$$\mathcal{L}\left(\mu, \sigma^2\right) = \prod_{y_i < c} \frac{1}{\sigma}\phi\left(\frac{y_i - \mu}{\sigma}\right) \prod_{y_i = c}\left[1 - \Phi\left(\frac{c-\mu}{\sigma}\right)\right].$$

- Usually, we shall be interested in a regression version of this model:

$$Y^* \mid X = x \sim \mathcal{N}\left(x'\beta, \sigma^2\right),$$

  in which case the likelihood takes the form

$$\mathcal{L}\left(\beta, \sigma^2\right) = \prod_{y_i < c} \frac{1}{\sigma}\phi\left(\frac{y_i - x_i'\beta}{\sigma}\right) \prod_{y_i = c}\left[1 - \Phi\left(\frac{c - x'\beta}{\sigma}\right)\right].$$

Means of censored normal variables

- Consider the following right-censored variable:

$$Y = \begin{cases} Y^* \text{ if } Y^* \leq c \\ c \text{ if } Y^* > c \end{cases}$$

with $Y^* \sim \mathcal{N}\left(\mu, \sigma^2\right)$. Therefore,

$$E(Y) = E(Y^* \mid Y^* \leq c) \Pr(Y^* \leq c) + c \Pr(Y^* > c)$$

- Letting $Y^* = \mu + \sigma\varepsilon$ with $\varepsilon \sim \mathcal{N}(0,1)$

$$\Pr(Y^* \leq c) = \Phi\left(\frac{c-\mu}{\sigma}\right)$$

$$E(Y^* \mid Y^* \leq c) = \mu + \sigma E\left(\varepsilon \mid \varepsilon \leq \frac{c-\mu}{\sigma}\right) = \mu - \sigma\lambda\left(\frac{c-\mu}{\sigma}\right).$$

- Note that

$$E(\varepsilon \mid \varepsilon \leq r) = \int_{-\infty}^{r} e \frac{\phi(e)}{\Phi(r)} de = -\frac{1}{\Phi(r)} \int_{-\infty}^{r} \phi'(e)\, de = -\frac{\phi(r)}{\Phi(r)} = -\lambda(r)$$

and

$$E(\varepsilon \mid \varepsilon > r) = \int_{r}^{\infty} e \frac{\phi(e)}{\Phi(-r)} de = -\frac{1}{\Phi(-r)} \int_{r}^{\infty} \phi'(e)\, de = -\frac{-\phi(r)}{\Phi(-r)} = \lambda(-r).$$

Illustration 2: Censoring at zero (Tobit model)

- Tobin (1958) considered the following model for expenditure on durables

$$\begin{aligned} Y &= \max\left(X'\beta + U, 0\right) \\ U &\mid X \sim \mathcal{N}\left(0, \sigma^2\right). \end{aligned}$$

- This is similar to the first example, but now we have left-censoring at zero.
- However, the nature of the application is very different because there is no physical censoring (the variable Y^* is just a model's construct).
- We are interested in the model as a way of capturing a particular form of nonlinearity in the relationship between X and Y.
- In a utility based model, the variable Y^* might be interpreted as a notional demand before non-negativity is imposed.
- With censoring at zero we have

$$Y = \begin{cases} Y^* \text{ if } Y^* > 0 \\ 0 \text{ if } Y^* \leq 0 \end{cases}$$

$$E(Y) = E(Y^* \mid Y^* > 0) \Pr(Y^* > 0)$$

$$\Pr(Y^* > 0) = \Pr\left(\varepsilon > -\frac{\mu}{\sigma}\right) = \Phi\left(\frac{\mu}{\sigma}\right)$$

$$E(Y^* \mid Y^* > 0) = \mu + \sigma E\left(\varepsilon \mid \varepsilon > -\frac{\mu}{\sigma}\right) = \mu + \sigma\lambda\left(\frac{\mu}{\sigma}\right).$$

Heckman's generalized selection model

- Consider the model

$$\begin{aligned} y^* &= x'\beta + \sigma u \\ d &= 1\left(z'\gamma + v \geq 0\right) \end{aligned}$$

$$\begin{pmatrix} u \\ v \end{pmatrix} \mid z \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

  so that

$$v \mid z, u \sim \mathcal{N}\left(\rho u, 1-\rho^2\right) \qquad \text{or} \qquad \Pr\left(v \leq r \mid z, u\right) = \Phi\left(\frac{r-\rho u}{\sqrt{1-\rho^2}}\right).$$

- In Heckman's original model, y^* denotes female log market wage and d is an indicator of participation in the labor force.
- The index $\{z'\gamma + v\}$ is a reduced form of the difference between market wage and reservation wage.

Joint likelihood function

- The joint likelihood is:

$$L = \sum_{d=1} \ln \{p(d=1, y^* \mid z)\} + \sum_{d=0} \ln \Pr(d=0 \mid z)$$

we have

$$p(d=1, y^* \mid z) = \Pr(d=1 \mid z, y^*) f(y^* \mid z)$$

$$f(y^* \mid z) = \frac{1}{\sigma}\phi\left(\frac{y^* - x'\beta}{\sigma}\right)$$

$$\Pr(d=1 \mid z, y^*) = 1 - \Pr\left(v \leq -z'\gamma \mid z, u\right) = 1 - \Phi\left(\frac{-z'\gamma - \rho u}{\sqrt{1-\rho^2}}\right) = \Phi\left(\frac{z'\gamma + \rho u}{\sqrt{1-\rho^2}}\right)$$

- Thus

$$L(\gamma, \beta, \sigma) = \sum_{d=1}\left\{\ln\left[\frac{1}{\sigma}\phi(u)\right] + \ln\Phi\left(\frac{z'\gamma + \rho u}{\sqrt{1-\rho^2}}\right)\right\} + \sum_{d=0}\ln\left[1 - \Phi\left(z'\gamma\right)\right]$$

where

$$u = \frac{y^* - x'\beta}{\sigma}.$$

- Note that if $\rho = 0$ this log likelihood boils down to the sum a Gaussian linear regression log likelihood and a probit log likelihood.

Density of y^* conditioned on $d = 1$

- From the previous result we know that

$$p\left(d = 1, y^* \mid z\right) = \frac{1}{\sigma}\phi\left(\frac{y^* - x'\beta}{\sigma}\right)\Phi\left(\frac{z'\gamma + \rho u}{\sqrt{1-\rho^2}}\right).$$

- Alternatively, to obtain it we could factorize as follows

$$p\left(d = 1, y^* \mid z\right) = \Pr\left(d = 1 \mid z\right) f\left(y^* \mid z, d = 1\right) = \Phi\left(z'\gamma\right) f\left(y^* \mid z, d = 1\right).$$

- From the previous expression we know that

$$f\left(y^* \mid z, d = 1\right) = \frac{p\left(d = 1, y^* \mid z\right)}{\Phi\left(z'\gamma\right)} = \frac{1}{\Phi\left(z'\gamma\right)}\Phi\left(\frac{z'\gamma + \rho u}{\sqrt{1-\rho^2}}\right)\frac{1}{\sigma}\phi\left(u\right).$$

- Note that if $\rho = 0$ we have $f\left(y^* \mid z, d = 1\right) = f\left(y^* \mid z\right) = \sigma^{-1}\phi\left(u\right)$.

Two-step method

- Then mean of $f\left(y^* \mid z, d=1\right)$ is given by

$$\begin{aligned} E\left(y^* \mid z, d=1\right) &= x'\beta + \sigma E\left(u \mid z'\gamma + v \geq 0\right) \\ &= x'\beta + \sigma\rho E\left(v \mid v \geq -z'\gamma\right) = x'\beta + \sigma\rho\lambda\left(z'\gamma\right) \end{aligned}$$

- Form $w_i = \left(x_i', \widehat{\lambda}_i\right)'$, where $\widehat{\lambda}_i = \lambda\left(z_i'\widehat{\gamma}\right)$ and $\widehat{\gamma}$ is the probit estimate.
- Then do the OLS regression of y on x and $\widehat{\lambda}$ in the subsample with $d=1$ to get consistent estimates of β and $\sigma_{uv}\ (=\sigma\rho)$:

$$\begin{pmatrix} \widehat{\beta} \\ \widehat{\sigma}_{uv} \end{pmatrix} = \left(\sum_{d_i=1} w_i w_i'\right)^{-1} \sum_{d_i=1} w_i y_i .$$

Nonparametric identification: The fundamental role of exclusion restrictions

- The role of exclusion restrictions for identification in a selection model is paramount.
- In applications there is a marked contrast in credibility between estimates that rely exclusively on the nonlinearity and those that use exclusion restrictions.
- The model of interest is

$$\begin{aligned} Y &= g_0(X) + U \\ D &= 1(p(X,Z) - V > 0) \end{aligned}$$

where (U, V) are independent of (X, Z) and V is uniform in the $(0,1)$ interval.

- Thus,

$$E(U \mid X, Z, D = 1) = E[U \mid V < p(X,Z)] = \lambda_0[p(X,Z)]$$
$$E(Y \mid X, Z) = g_0(X)$$

(i.e. enforcing the exclusion restriction), but we observe

$$\begin{aligned} E(Y \mid X, Z, D = 1) &= \mu(X,Z) = g_0(X) + \lambda_0[p(X,Z)] \\ E(D \mid X, Z) &= p(X,Z). \end{aligned}$$

- The question is whether $g_0(.)$ and $\lambda_0(.)$ can be identified from knowledge of $\mu(X,Z)$ and $p(X,Z)$.

- Let us consider first the case where X and Z are continuous. Suppose there is an alternative solution (g^*, λ^*). Then

$$g_0(X) - g^*(X) + \lambda_0(p) - \lambda^*(p) = 0.$$

Differentiating

$$\begin{aligned} \frac{\partial(\lambda_0 - \lambda^*)}{\partial p} \frac{\partial p}{\partial Z} &= 0 \\ \frac{\partial(g_0 - g^*)}{\partial X} + \frac{\partial(\lambda_0 - \lambda^*)}{\partial p} \frac{\partial p}{\partial X} &= 0. \end{aligned}$$

- Under the assumption that $\partial p/\partial Z \neq 0$ (instrument relevance), we have

$$\frac{\partial(\lambda_0 - \lambda^*)}{\partial p} = 0, \qquad \frac{\partial(g_0 - g^*)}{\partial X} = 0$$

so that $\lambda_0 - \lambda^*$ and $g_0 - g^*$ are constant (i.e. $g_0(X)$ is identified up to an unknown constant).
- This is the identification result in Das, Newey, and Vella (2003).
- $E(Y \mid X)$ is identified up to a constant, provided we have a continuous instrument.
- Identification of the constant requires units for which the probability of selection is arbitrarily close to one ("identification at infinity").
- Unfortunately, the constants are important for identifying average treatment effects.

Z discrete

- With binary Z, functional form assumptions play a more fundamental role in securing identification than in the case of an exclusion restriction of a continuous variable.
- Suppose X is continuous but Z is a dummy variable. In general $g_0(X)$ is not identified. To see this, consider

$$\begin{aligned} \mu(X,1) &= g_0(X) + \lambda_0[p(X,1)] \\ \mu(X,0) &= g_0(X) + \lambda_0[p(X,0)], \end{aligned}$$

so that we identify the difference

$$\nu(X) = \lambda_0[p(X,1)] - \lambda_0[p(X,0)],$$

but this does not suffice to determine λ_0 up to a constant.

- Take as an example the case where $p(X,Z)$ is a simple logit or probit model:

$$p(X,Z) = F(\beta X + \gamma Z),$$

then letting $h_0(.) = \lambda_0[F(.)]$,

$$\nu(X) = h_0(\beta X + \gamma) - h_0(\beta X).$$

- Suppose the existence of another solution h^*. We should have

$$h_0(\beta X + \gamma) - h^*(\beta X + \gamma) = h_0(\beta X) - h^*(\beta X),$$

which is satisfied by a multiplicity of periodic functions.

X and Z discrete

- If X is also discrete, there is clearly lack of identification.
- For example, suppose X and Z are dummy variables:

$$\begin{aligned}
\mu(0,0) &= g_0(0) + \lambda_0[p(0,0)] \\
\mu(0,1) &= g_0(0) + \lambda_0[p(0,1)] \\
\mu(1,0) &= g_0(1) + \lambda_0[p(1,0)] \\
\mu(1,1) &= g_0(1) + \lambda_0[p(1,1)].
\end{aligned}$$

- Since $\lambda_0(.)$ is unknown $g_0(1) - g_0(0)$ is not identified.
- Only $\lambda_0[p(1,1)] - \lambda_0[p(1,0)]$ and $\lambda_0[p(0,1)] - \lambda_0[p(0,0)]$ are identified.