Advanced Econometrics

Chapter 15: Censored Regression Models

In Choi

Sogang University

Useful references:

This note heavily depends on Wooldridge (2002).

- Amemiya, T. (1985). Advanced econometrics. Harvard university press.
- Cameron, A.C., and P.K. Trivedi (2005). Microeconometrics: Methods and Applications. Cambridge University Press: New York.
- Maddala, G. S. (1986). Limited-dependent and qualitative variables in econometrics (No. 3). Cambridge university press.
- Wooldridge, J. (2002). *Econometric Analysis of Cross Section and Panel Data*, The MIT Press.

Introduction

- Model

$$
\begin{aligned}
y_i^* &= x_i'\beta + u_i, \;\; u_i \mid x_i \sim N(0, \sigma^2) \\
y_i &= \max(0, y_i^*).
\end{aligned}
$$

- This model is called the *standard censored Tobit model* or *type I Tobit model*.

Example

y_i : health expenditure
Some individuals' health expenditures are zero.

Example

y_i : expenditure on cars
Some people in the sample do not buy cars .

Expected values

- We are interested in $E(y \mid x, y > 0)$ and $E(y \mid x)$.
- Since the function $g(z) = \max(0, z)$ is convex, it follows from the conditional Jensen's inequality[1]

$$E(y \mid x) \geq \max(0, x'\beta)$$

if $E(y^* \mid x) = x'\beta$.

[1] For a random variable X, if $g(x)$ is a convex function

$$Eg(X) \geq g(EX).$$

Expected values

- Assume that u is independent of x and has a normal distribution. Then,

$$\begin{aligned} E(y \mid x) &= p(y=0 \mid x) \cdot 0 + P(y>0 \mid x) E(y \mid x, y>0) \\ &= P(y>0 \mid x) E(y \mid x, y>0). \end{aligned}$$

- (i)

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

Expected values

- Assume that u is independent of x and has a normal distribution. Then,

$$\begin{aligned} E(y \mid x) &= p(y=0 \mid x)\cdot 0 + P(y>0 \mid x)E(y \mid x, y>0) \quad (1) \\ &= P(y>0 \mid x)E(y \mid x, y>0). \end{aligned}$$

- (i)

$$\begin{aligned} P(y > 0 \mid x) &= P(y^* > 0 \mid x) = P(u > -x'\beta \mid x) \\ &= P(u/\sigma > -x'\beta/\sigma \mid x) = \Phi(x'\beta/\sigma) \quad (2) \end{aligned}$$

Expected values

- (ii)

$$\begin{aligned} E(y \mid x, y>0) &= x'\beta + E(u \mid u > -x'\beta) \\ &= x'\beta + E(u/\sigma \mid u/\sigma > -x'\beta/\sigma) \\ &= x'\beta + \sigma \left[\frac{\phi(x'\beta/\sigma)}{\Phi(x'\beta/\sigma)} \right].^2 \end{aligned} \tag{3}$$

- The quantity $\lambda(c) = \phi(c)/\Phi(c)$ is called the inverse Mills ratio.

(ii) $1 - \Phi(-c) = \Phi(c)$.

- If x_j is a continuous explanatory variable, it can be shown that

$$\frac{\partial E(y \mid x, y > 0)}{\partial x_j} = \beta_j \left\{1 - \lambda(x'\beta/\sigma)\left[x'\beta/\sigma + \lambda(x'\beta/\sigma)\right]\right\}. \quad (4)$$

  This implies that the partial effect of x_j on $E(y \mid x, y > 0)$ is not entirely determined by β_j. Because the term in the bracket is strictly between 0 and 1, the sign of β_j is the same as the sign of the partial effect of x_j. [3]

- Elasticities can be calculated by using equations (3) and (4). For example,

$$\frac{\partial E(y \mid x, y > 0)}{\partial x_1} \cdot \frac{x_1}{E(y \mid x, y > 0)}$$

[3] If $z \sim N(0, 1)$,

$$Var(z \mid z > -c) = 1 - \lambda(c)[c + \lambda(c)].$$

Expected values

- We have from (1), (2) and (3)

$$E(y \mid x) = \Phi(x'\beta/\sigma)\left\{x'\beta + \sigma\left[\frac{\phi(x'\beta/\sigma)}{\Phi(x'\beta/\sigma)}\right]\right\}.$$

Using this, we obtain

$$\frac{\partial E(y \mid x)}{\partial x_j} = \Phi(x'\beta/\sigma)\beta_j.$$

Inconsistency of OLS

- Consider using the subsample with strictly positive y_i. Since

$$E(y \mid x, y > 0) = x'\beta + \sigma\lambda(x'\beta/\sigma),$$

we may write

$$\begin{aligned} y_i &= x'\beta + \sigma\lambda(x'\beta/\sigma) + e_i; \\ E(e_i &\mid x_i, y_i > 0) = 0. \end{aligned}$$

Thus, running OLS using $\{y_i, x_i\}$ encounters the problem of omitted variable bias.

Inconsistency of OLS

- Consider using the whole sample with strictly positive y_i. Since

$$E(y \mid x) = \Phi(x'\beta/\sigma)x'\beta + \sigma\phi(x'\beta/\sigma),$$

running OLS using $\{y_i, x_i\}$ yields an inconsistent estimator.

Estimation and inference

- Let $\{(x_i, y_i) : i = 1, ..., N\}$ be a random sample following the censored Tobit model. We need to derive the density of y_i given x_i for estimation and inference. First,

$$f(0 \mid x_i) = 1 - P(y > 0 \mid x) = 1 - \Phi(x'\beta/\sigma).$$

Second, since $P(y_i \leq y \mid x_i) = P(y_i^* \leq y \mid x_i)$,

$$f(y \mid x_i) = f^*(y \mid x_i) \text{ for } y > 0,$$

where $f^*(\cdot)$ denotes the density of y^* given x_i. By assumption

$$f^*(y \mid x_i) = \frac{1}{\sigma}\phi\left[\left(y - x_i'\beta\right)/\sigma\right], \ -\infty < y < \infty.$$

Estimation and inference

- The density of y_i given x_i is written as

$$f(y \mid x_i) = \left[1 - \Phi(x'\beta/\sigma)\right]^{1\{y=0\}} \left\{ \frac{1}{\sigma} \phi \left[\left(y - x_i'\beta \right) / \sigma \right] \right\}^{1\{y>0\}}.$$

- Let $\theta = (\beta' \ \sigma^2)'$. The conditional log likelihood function for the i-th observation is

$$\begin{aligned} l_i(\theta) &= 1\{y_i = 0\} \log \left[1 - \Phi(x'\beta/\sigma)\right] \\ +1\{y_i &> 0\} \log \left\{ \phi \left[\left(y - x_i'\beta \right) / \sigma \right] - \frac{1}{2} \log \left(\sigma^2 \right) \right\}. \end{aligned}$$

- The MLE of θ is obtained by maximizing $\sum_{i=1}^{N} l_i(\theta)$. The standard theory of MLE applies to the MLE for the censored Tobit model.

Endogenous explanatory variables

- Consider the model

$$\begin{aligned} y_1 &= \max(0, z_1'\delta_1 + \alpha_1 y_2 + u_1) \\ y_2 &= z_1'\delta_{21} + z_2'\delta_{22} + v_2, \end{aligned}$$

where (u_1, v_2) are zero-mean normally distributed, independent of z. If u_1 and v_2 are correlated, y_2 is endogenous.

- Similar to Rivers and Vuong's (1988; JOE) procedure.
- Under joint normality of (u_1, v_2), we can write

$$u_1 = \theta_1 v_2 + e_1,$$

where $\theta_1 = \eta_1/\tau_2^2$, $\eta_1 = Cov(v_2, u_1)$, $\tau_2^2 = Var(v_2)$, and e_1 is independent of z and v_2. We also have $E(e_1) = 0$ and $Var(e_1) = Var(u_1) - \eta_1^2/\tau_2^2 = \tau_1^2$.

- Now write
$$y_1 = \max(0, z_1'\delta_1 + \alpha_1 y_2 + \theta_1 v_2 + e_1).$$
- Since v_2 is not observed, we replace it with its OLS estimate obtained by regressing y_2 on z.
- Then, perform the standard Tobit procedure. This gives consistent estimators of the parameters.
- When $\theta_1 \neq 0$, however, the second-stage Tobit standard errors and test statistics are not asymptotically valid.

Endogenous explanatory variables

Maximum likelihood approach

- The joint distribution of (y_1, y_2) conditional on z is

$$f(y_1, y_2) = f(y_1 \mid y_2, z) f(y_2 \mid z).$$

- $y_2 \mid z \sim N(z\delta_2, \tau_2^2)$
- $y_1 \mid y_2, z \sim N(z_1'\delta_1 + \alpha_1 y_2 + (\eta_1/\tau_2^2)(y_2 - z'\delta_2), \tau_1^2)$.
- We can construct the likelihood function using these and perform the maximum likelihood procedure.