MECT Microeconometrics
Blundell Lecture 2
Censored Data Models

Richard Blundell
http://www.ucl.ac.uk/~uctp39a/

University College London

February-March 2016

Censored Data Models

- **Censored and truncated data**

Examples:

earnings

hours of work (mroz.dta is a ‘typical’ data set to play with)

top coding of wealth

expenditure on cars (this was James Tobin’s original example which became know as Tobin’s Probit model or the **Tobit** model.)

- **Typical definitions:**

Censored data *includes* the censoring points

Truncated data *excludes* the censoring points

- A mixture of discrete and continuous processes. In general we should model the process of censoring or truncation as a separate discrete mechanism, i.e. the 'selectivity' model.
- To begin with we have a model in which the two processes are generated from the same underlying continuous latent variable model e.g. corner solution models in economics.

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

with

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* > 0 \\ 0 & \text{otherwise} \end{cases}$$

or

$$y_i = \begin{cases} y_i^* & \text{if } u_i > -x_i\beta \\ 0 & \text{otherwise} \end{cases}$$

- Sometimes also define D_i

$$D_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{otherwise} \end{cases}$$

The general specification for the censored regression model is

$$\begin{aligned} y_i^* &= x_i\beta + u_i \\ y_i &= \max\{0, y_i^*\} \end{aligned}$$

where y^* is the unobservable underlying process (similar to what was used with discrete choice models) and y is the data observation.

- When u are normally distributed - $u|x \sim \mathcal{N}(0, \sigma^2)$ - the model is the *Tobit* model.

- Note that

$$P(y > 0|x) = P(u > -x'\beta|x) = \Phi\left(\frac{x'\beta}{\sigma}\right)$$

■ Consider the moments of the truncated normal.

▶ Assume $w \backsim \mathcal{N}(0, \sigma)$. Then $w|w > c$ where c is an arbitrary constant, is a truncated normal.

▶ The density function for the truncated normal is:

$$
\begin{aligned}
f(w|w > c) &= \frac{f(w)}{1 - F(c)} \\
&= \frac{\sigma\phi\left(\frac{w}{\sigma}\right)}{1 - \Phi\left(\frac{c}{\sigma}\right)}
\end{aligned}
$$

where f is the density function of w and F is the cumulative density function of w.

- We can now write

$$
\begin{aligned}
E(w|w>c) &= \int_c^{\infty} wf(w|w>c)dw \\
&= \sigma \frac{\phi\left(\frac{c}{\sigma}\right)}{1-\Phi\left(\frac{c}{\sigma}\right)}
\end{aligned}
$$

  - Applying this result to the regression model yields

$$
E(y|x, y>0) = x'\beta + E(u|u>-x'\beta) = x'\beta + \sigma \frac{\phi\left(\frac{x'\beta}{\sigma}\right)}{\Phi\left(\frac{x'\beta}{\sigma}\right)}
$$

- Note that $\phi(w)/\Phi(w)$ is the Inverse Mills Ratio, usually written $\lambda(w)$.
- Also note that, contrary to the discrete choice models, the variance of the residual plays a central role here: it determines the size of the partial effects.

OLS Bias ▶ Truncated Data:

▶ Suppose one uses only the positive observations to estimate the model and the unobservables are normally distributed. Then, we have seen that,

$$E(y|x, y > 0) = x'\beta + \sigma\lambda\left(\frac{x'\beta}{\sigma}\right)$$

where the last term is $E(u|x, u > -x'\beta)$, which is generally non-zero.

▶ A model of the form:

$$y = x'\beta + \sigma\lambda + v$$

would have $E(v|x, y > 0) = 0$.

▶ This implies the inconsistency of OLS: omitted variable problem. Thus, the resulting error term will be correlated with x.

Censored Data:

- Now suppose we use all observations, both positive and zero.
- Under normality of the residual, we obtain,

$$E(y|x) = \Phi\left(\frac{x'\beta}{\sigma}\right)x'\beta + \sigma\phi\left(\frac{x'\beta}{\sigma}\right)$$

- Thus, once again the OLS estimates will be biased and inconsistent.

The Maximum Likelihood Estimator

- Let $\{(y_i, x_i), i = 1, ..., N\}$ be a random sample of data on the model. The contribution to the likelihood of a zero observation is determined by,

$$P(y_i = 0|x_i) = 1 - \Phi\left(\frac{x_i'\beta}{\sigma}\right)$$

The contribution to the likelihood of a non-zero observation is determined by,

$$f(y_i|x_i) = \frac{1}{\sigma}\phi\left(\frac{y_i - x_i'\beta}{\sigma}\right)$$

which is not invariant to σ.

Thus, the overall contribution of observation i to the loglikelihood function is,

$$\begin{aligned} \ln l_i(x_i; \beta, \sigma) &= \mathbf{1}(y_i = 0) \ln\left[1 - \Phi\left(\frac{x_i'\beta}{\sigma}\right)\right] \\ +\mathbf{1}(y_i &= 1) \ln\left[\frac{1}{\sigma}\phi\left(\frac{y_i - x_i'\beta}{\sigma}\right)\right] \end{aligned}$$

and the sample loglikelihood is,

$$\ln \mathcal{L}_N(\beta, \sigma) = \sum_{i=1}^{N} \left\{ \begin{array}{c} (1 - D_i) \ln \left[1 - \Phi\left(\frac{x_i'\beta}{\sigma} \right) \right] \\ + D_i \left[\ln \phi\left(\frac{y_i - x_i'\beta}{\sigma} \right) - \ln \sigma \right] \end{array} \right\}$$

where D equals one when $y^* > 0$ and equals zero otherwise.

- Notice that both β and σ are separately identified. Moreover, if $D = 1$ for all i, the ML and the OLS estimators will be the same.
- FOC

$$\frac{\partial \ln \mathcal{L}}{\partial \beta} = \sum_{i=1}^{N} \frac{1}{\sigma^2} \left\{ D_i (y_i - x_i'\beta) x_i - (1 - D_i) \frac{\sigma \phi\left(\frac{x_i'\beta}{\sigma} \right)}{1 - \Phi\left(\frac{x_i'\beta}{\sigma} \right)} x_i \right\}$$

$$\frac{\partial \ln \mathcal{L}}{\partial \sigma^2} = \sum_{i=1}^{N} \left\{ (1 - D_i) \frac{x_i \beta \phi\left(\frac{x_i'\beta}{\sigma} \right)}{2\sigma^2 \left[1 - \Phi\left(\frac{x_i'\beta}{\sigma} \right) \right]} + D_i \left[\frac{(y_i - x_i'\beta)^2}{2\sigma^4} - \frac{1}{2\sigma^2} \right] \right\}$$

Or write as:

$$
\begin{aligned}
(1)\ \frac{\partial \ln \mathcal{L}}{\partial \beta} &= -\sum_{i\in 0} \frac{1}{\sigma^2} \frac{\sigma \phi_i}{1-\Phi_i} x_i + \frac{1}{\sigma^2} \sum_{i\in +} (y_i - x_i'\beta) x_i \\
(2)\ \frac{\partial \ln \mathcal{L}}{\partial \sigma^2} &= \frac{1}{2\sigma^2} \sum_{i\in 0} \frac{x_i \beta \phi_i}{1-\Phi_i} + \frac{1}{2\sigma^4} \sum_{i\in +} (y_i - x_i'\beta)^2 - \frac{N_+}{2\sigma^2}
\end{aligned}
$$

note that $\frac{\beta'}{2\sigma^2}$ x (1) + (2) $\rightarrow$

$$
\widehat{\sigma}^2 = \frac{1}{N_+} \sum_{i\in +} (y_i - x_i'\beta)^2
$$

that is the positive observations only contribute to the estimation of σ.

- Also if we define $m_i \equiv E(y_i^*|y_i)$ then we can write (1) as

$$\frac{\partial \ln \mathcal{L}}{\partial \beta} = c \sum_{i=1}^{N} x_i (m_i - x_i' \beta)$$

or

$$\sum_{i=1}^{N} x_i m_i = \sum_{i=1}^{N} x_i x_i' \beta$$

which defines an *EM* algorithm for the Tobit model. Note also that

$$m_i = \begin{cases} y^* & \text{if } y_i^* > 0 \\ x_i'\beta - \sigma \frac{\phi_i}{1-\Phi_i} & \text{otherwise} \end{cases}$$

again replacing y^* with its best guess, given y, when it is unobserved.

- Using the Theorems 1 and 2 from Lecture 6, MLE of β and σ^2 is consistent and asymptotically normally distributed.

- Exercise: Derive the asymptotic covariance matrix from the expected values of the 2nd partial derivatives of $\ln\mathcal{L}$.

- Note is has the general form

$$-\begin{bmatrix} E\frac{\partial^2 \ln \mathcal{L}}{\partial \beta^2} & E\frac{\partial^2 \ln \mathcal{L}}{\partial \beta \partial \sigma^2} \\ . & E\frac{\partial^2 \ln \mathcal{L}}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^N a_i x_i x_i' & \sum_i^N b_i x_i \\ . & \sum_{i=1}^N c_i \end{bmatrix}$$

LM or Score Test

- Let the log likelihood be written

$$\ln \mathcal{L}(\theta_1, \theta_2)$$

where θ_1 is the set of parameters that are unrestricted under the null hypothesis and θ_2 are k_2 restricted parameters under H_0.

$$\begin{aligned} H_0 &: \theta_2 = 0 \\ H_1 &: \theta_2 \neq 0 \end{aligned}$$

- e.g.

$$y_i^* = x_{1i}'\beta_1 + x_{2i}'\beta_2 + u_i \text{ with } u_i \sim N(0, \sigma^2).$$

where $\theta_1 = (\beta_1', \sigma^2)'$ and $\theta_2 = \beta_2$.

$$\frac{\partial \ln \mathcal{L}(\theta_1, \theta_2)}{\partial \theta} = \sum \frac{\partial \ln l_i(\theta_1, \theta_2)}{\partial \theta}$$

or

$$S(\theta) = \sum S_i(\theta)$$

- Let $\widehat{\theta}$ be the MLE under H_0. Then

$$\frac{1}{\sqrt{N}} S(\widehat{\theta}) \sim^a N(0, H)$$

therefore

$$\frac{1}{N} S(\widehat{\theta})' H^{-1} S(\widehat{\theta}) \sim^a \chi^2_{(k_2)}$$

In the Tobit model consider the case of $H_0 : \beta_2 = 0$

$$\frac{\partial \ln \mathcal{L}}{\partial \beta_2} = \frac{1}{\sigma^2} \sum_i D_i (y_i - x_i' \beta) x_{2i} - \frac{1}{\sigma^2} \sum_i (1 - D_i) \frac{\sigma_i \phi_i}{1 - \Phi_i} x_{2i}$$

$$\frac{\partial \ln \mathcal{L}}{\partial \beta_2} = \frac{1}{\sigma^2} \sum_i e_i^{(1)} x_{2i}$$

where

$$e_i^{(1)} = D_i (y_i - x_i' \beta) + (1 - D_i)(-\frac{\sigma_i \phi_i}{1 - \Phi_i})$$

is known as the first order **'generalised' residual**, which reduces to $u_i = y_i - x_i' \beta$ in the general linear model case.

This kind of **Score or LM test** can be extended to specification tests for heteroskedasticity and for non-normality. Notice that is estimation under the alternative is avoided, at least in terms of the test statistic. If H_0 is rejected then estimation under H_a is unavoidable.

- Consider the normal distribution

$$f(u_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{u_i^2}{\sigma^2}\right)$$

can be written in terms of log scores

$$\frac{\partial \ln f(u_i)}{\partial u_i} = -\frac{u_i}{\sigma^2}.$$

- A popular generalisation (**Pearson family** of distributions) is

$$\frac{\partial \ln f(u_i)}{\partial u_i} = \frac{-u_i + c_1}{\sigma_i^2 - c_1 u_i + c_2 u_i^2}$$

where skedastsic function $\sigma_i^2 = h(\gamma_0 + \gamma_1' z_i)$, z_i observable determinants of heteroskedasticity.

$c_1 \neq 0 \rightarrow$ skewness
$c_2 \neq 0 \rightarrow$ kurtosis
$c_1 = c_2 = 0 \rightarrow$ Normal
$\gamma_1 = 0 \rightarrow$ homoskedastic

- Can write out the loglikelihood with the Pearson family and take derivatives with respect to the c and γ parameters to find the LM or Score test. e.g.

$$\frac{\partial \ln \mathcal{L}}{\partial \gamma_1} = \alpha \sum_i e_i^{(2)} z_i$$

where $e_i^{(2)}$ is the second order generalised residual.
- Also

$$\frac{\partial \ln \mathcal{L}}{\partial c_2} = \frac{1}{4\sigma^2} \sum_i D_i (u_i^4 - \int_{-x_i'\beta}^{\infty} t^4 f dt)$$

which is the 4th order generalised residual.

Semiparametric Estimators:

What if normality is rejected or not a credible prior assumption anyway?

Suppose we just assume symmetry:
We can write the model as

$$\begin{aligned} y_i^* &= x_i'\beta + u_i, \text{ or} \\ y_i &= x_i'\beta + u_i^*, \text{ where} \\ u_i^* &= \max\left\{u_i, -x_i'\beta\right\} \end{aligned}$$

We can define the new residuals

$$u_i^{**} = min\left\{u_i^*, x_i'\beta\right\}$$

where the $x_i'\beta$ reflects 'upper' trimming. Drop observations where $x_i'\beta \leqslant 0$ as no symmetric trimming is possible here.

- Adapt EM algorithm for least squares by replacing y by

$$y_i^* = \min\left\{y_i, 2x_i'\beta\right\}$$

$\rightarrow$ **symmetrically censored least squares**: Applying OLS for all $i : x_i\beta \geqslant 0$ yields consistent and asymptotically normal estimates: the error term now satisfies $E(u^{**}|x) = 0$.

- Requires a symmetric distribution of the error term, u^*, but no normality or homoskedasticity.
- Estimation requires an iterative procedure (EM algorithm)

$$\widehat{\beta} = \left(\sum x_i x_i'\right)^{-1} \sum x_i m_i$$

with

$$m_i = \min\{y_i, 2x_i'\beta\}$$

- Monte-Carlo results.

Censored Least Absolute Deviations

Assume: conditional median of u_i is zero $\longrightarrow$ median of y_i is

$$x_i'\beta.1(x_i'\beta > 0)$$

CLAD minimises the absolute distance of y_i from its median

$$\widehat{\beta}_{CLAD} = \arg\min_{\beta} \sum \left| y_i - x_i'\beta.1(x_i'\beta > 0) \right|$$

- Powell (1984) shows that $\widehat{\beta}_{CLAD}$ is $\sqrt{N}-$ consistent and asymptotically normal.

- Blundell and Powell (2007) develop this idea further for the case of endogenous variables in x. So let's turn to the case of the censored regression model with endogenous regressors.

Endogenous Variables

As in the previous lecture we can consider the following (triangular) model

$$y_{1i}^{*} = x_{1i}'\beta + \gamma y_{2i} + u_{1i} \quad (1)$$

$$y_{2i} = z_i'\pi_2 + v_{2i} \quad (2)$$

where in the censored regression case $y_{1i} = y_{1i}^{*}1(y_{1i}^{*} > 0)$. $z_i' = (x_{1i}', x_{2i}')$. The x_{2i}' are the excluded 'instruments' from the equation for y_1. The first equation is a the 'structural' equation of interest and the second equation is the 'reduced form' for y_2.

- y_2 is endogenous if u_1 and v_2 are correlated. If y_1 was fully observed we could use IV.

Using the othogonal decomposition for u_1

$$u_{1i} = \rho v_{2i} + \epsilon_{1i}$$

where $E(\epsilon_{1i} | v_{2i}) = 0$.

- where y_2 is uncorrelated with u_{1i} **conditional** on the control function v_2.

- As before, under the assumption that u_1 and v_2 are jointly normally distributed, u_2 and ϵ are uncorrelated by definition and ϵ also follows a normal distribution.

Use this to define the **augmented model**

$$\begin{aligned} y_{1i}^* &= x_{1i}'\beta + \gamma y_{2i} + \rho v_{2i} + \epsilon_{1i} \\ y_{2i} &= z_i'\pi_2 + v_{2i} \end{aligned}$$

2-step Estimator:

- **Step 1**: Estimate α by OLS and predict v_2,

$$\widehat{v}_{2i} = y_{2i} - \widehat{\pi}_2' z_i$$

- **Step 2**: use $\widehat{v}_{2i}$ as a 'control function' in the model for y_1^* above and estimate by Tobit or other consistent method.

An Exogeneity test

The null of exogeneity in this model is analogous to

$$H_0 : \rho = 0$$

A test of this null can be performed using a t-test.

- Blundell-Smith (1986, *Econometrica*).
- Specifically for the censored regression model (Tobit model).
- This test follows for the **binary choice** (try this as an exercise) and other related models.

Semiparametric Estimation of the Censored Regression model with Endogenous Variables

We write the structural equation of interest as

$$y_{1i} = \max[0, x_i'\beta_0 + u_{1i}] \tag{3}$$

where $x_i' = (x_{1i}', y_{2i})$.
Now invoke the usual control function conditional independence assumption

$$u_1 \perp x \mid v_2$$

This distributional restriction is equivalent to a restriction that all of the conditional quantiles of u_{1i} given x_i and z_i are functions only of the control variable v_{2i}.

- Such a quantile restriction is useful for models in which the dependent variable is monotonically related to the error term as in the censored model here.

Semiparametric Estimation of the Censored Regression model with Endogenous Variables

Notice, the conditional quantile of the censored dependent variable y_{1i} can be written:

$$\begin{aligned} q_i &= Q_\alpha[y_i \mid x_i, z_i] \equiv q_i(\alpha) \\ &= Q_\alpha[\max\{0, x_i'\beta_0 + u_{1i}\} \mid x_i, z_i] \\ &= \max\{0, x_i'\beta_0 + Q_\alpha[u_{1i} \mid x_i, z_i]\} \\ &= \max\{0, x_i'\beta_0 + \lambda_\alpha(v_{2i})\} \end{aligned}$$

where $\lambda_\alpha(v_{2i}) \equiv Q_\alpha[u_{1i} \mid v_{2i}]$.

- Useful to point out under the exogeneity assumption the control function is constant for all α. The background to some semiparametric estimation methods for the censored regression model under exogeneity (see Powell (1984) and many subsequent papers).

Semiparametric Estimation of the Censored Regression model with Endogenous Variables

- Under the assumption of v_{2i} is known this estimator is a semilinear censored regression model.

- Take the case of two observations with the conditional quantiles of y_1 are positive. The difference in the quantile regression functions is the difference in the regression function plus the difference in the control functions. By restriction attention to pairs of observations with identical control variables v_{2i}, differences in the quantiles only involve differences in the regression function, which then identifies β_0.

- Blundell and Powell (JoE, 2007) develop this idea to form a consistent semiparametric estimator for the censored regression estimator under endogeneity.