Lecture 6

Limited Variables 5 of 5: Censored & Truncated Models & Correcting for Sample Selection

Michael Curran

Trinity College Dublin

JS Econometrics

Lecture 6 Outline

Censored & Truncated Models

- Censored & Truncated Regression Models

Sample Selection

- Correcting for Sample Selection

Summary & References

- Summary & References

Censored & Truncated Regression Models

Introduction

Data observability taken into account.

Distinction between lumpiness in an outcome variable and problems of data censoring can be confusing; Wooldridge has Tobit only for corner solution outcomes but literature on Tobit models usually treats another situation within same framework: response variable censored above or below some threshold.

Survey design and institutional constraints.

Solve data censoring by applying a **censored regression model** – problem solved by censored regression model is one of missing data on response variable y but where we have info about the variable when it is missing, *viz.* whether it is above or below a known threshold. A **truncated regression model** arises when we exclude on the basis of y a subset of the population in our sampling scheme – we've not got a random sample from the population but we know the rule that was used to include units in the sample. This rule is determined by whether y is above or below a certain threshold.

Censored & Truncated Regression Models

Censored Regression Models

Focus on **censored normal regression model**. y follows CLM and letting i emphasise a random draw from the population:

$$y_i = \beta_0 + \mathbf{x}_i\boldsymbol{\beta} + u_i \quad u_i|\mathbf{x}_i, c_i \sim N(0, \sigma^2) \tag{1}$$

$$w_i = \min(y_i, c_i) \tag{2}$$

Only observe y_i if it is less than a censoring value c_i: censoring from above (*right censoring*); censoring from below (*left censoring*) is handled similarly. Example of right data censoring is **top coding** – we know its value only up to a certain threshold; for responses greater than the threshold, we only know the variable is at least as large as the threshold. E.g. family wealth in some surveys is top coded: can respond 'more than \$500, 000', in this e.g. c_i censoring threshold is same for all i; in many cases, c_i changes with individual/family characteristics.

Censored & Truncated Regression Models

Censored Regression Models

Censoring induces problems: OLS regression using only uncensored observations (i.e. only those with $y_i < c_i$) produces inconsistent estimates of β_j. OLS regression of w_i on $\mathbf{x}_i$ using all observation doesn't consistently estimate β_j unless there is no censoring – similar to Tobit but problem is much different. In Tobit, we are modeling economic behaviour, which often yields zero outcomes but the Tobit model is supposed to reflect this; with censored regression, we have a data collection problem since for some reason the data are censored.

Under assumptions (1) and (2), we can estimate $\boldsymbol{\beta}$ and σ^2 by ML given a random sample $(\mathbf{x}_i, w_i)$. Need density of w_i given $(\mathbf{x}_i, c_i)$. For uncensored observations $w_i = y_i$ and the density of w_i is the same as y_i. For censored observations, we need:

$$P(w_i = c_i|\mathbf{x}_i) = P(y_i \geq c_i|\mathbf{x}_i) = P(u_i \geq c_i - \mathbf{x}_i\boldsymbol{\beta}) = 1 - \Phi[\frac{(c_i - \mathbf{x_i}\boldsymbol{\beta})}{\sigma}]$$

Censored & Truncated Regression Models

Censored Regression Models

Combine these two parts to obtain density of w_i given $\mathbf{x}_i$ and c_i:

$$f\left(w|\mathbf{x}_i, c_i\right) = \begin{cases} 1 - \Phi\left[\frac{(c_i - \mathbf{x}_i\boldsymbol{\beta})}{\sigma}\right] & w = c_i \\ \frac{1}{\sigma}\phi\left[\frac{(w - \mathbf{x}_i\boldsymbol{\beta})}{\sigma}\right] & w < c_i \end{cases}$$

Log-likelihood for observation i is obtained by taking the natural log of the density for each i. Can maximise the sum of these across i wrt β_j and σ to get MLEs. Interpret β_j just as in linear reg under random sampling – much different to Tobit where expectations of interest are NL functions of the β_j.

Censored & Truncated Regression Models

Censored Regression Models

- **Duration analysis**: a *duration* is a variable measuring time before a certain event occurs (e.g. days before felon is released – may never happen for some felons or may happen after so long that we must censor the duration to analyse data).
- Logarithm of dependent variable (censoring threshold) as in (2), so parameters can be interpreted as percentage change.
- Alternatives exist for measuring effects of each explanatory variable on duration rather than only expected duration.
- If any assumptions of censored normal regression model are violated (e.g. heterogeneity / non-Normality), MLEs are generally inconsistent. So censoring is potentially very costly as OLS using uncensored sample requires neither Normality nor homogeneity for consistency. There are methods that do not require us to assume a distribution, but they are more advanced.

Censored & Truncated Regression Models

Truncated Regression Models

Difference to censored regression model: don't observe any information about a certain segment of the population. The **truncated normal regression model** begins with an underlying population model that satisfies CLM assumptions:

$$y = \beta_0 + \mathbf{x}\boldsymbol{\beta} + u \quad u|\mathbf{x} \sim N(0, \sigma^2) \tag{3}$$

Under (3), given random sample from population, OLS is most efficient estimation procedure. Problem: don't observe a random sample from population – assumption MLR.2 is violated. A random draw $(\mathbf{x}_i, y_i)$ is only observed if $y_i \leq c_i$, where c_i is truncated threshold that can depend on exogenous variables, in particular $\mathbf{x}_i$ so if $\{(\mathbf{x}_i, y_i) : i = 1, \ldots, n\}$ is our *observed* sample, then y_i is necessarily less than or equal to c_i. **Important:** this differs from the censored regression model since there we observe $\mathbf{x}_i$ for any randomly drawn observation from the population; in truncated model, we only observe $\mathbf{x}_i$ if $y_i \leq c_i$.

Censored & Truncated Regression Models

Truncated Regression Models

Estimation (β_j and σ)

$$g(y|\mathbf{x}_i, c_i) = \frac{f(y|\mathbf{x}_i\boldsymbol{\beta}, \sigma^2)}{F(c_i|\mathbf{x}_i\boldsymbol{\beta}, \sigma^2)} \quad y \leq c_i \tag{4}$$

$f() \sim N(\beta_0 + \mathbf{x}_i\boldsymbol{\beta}, \sigma^2)$ and F is a Normal cdf with same mean and variance evaluated at c_i. Renormalise density by dividing by area under $f(\cdot|\mathbf{x}_i\boldsymbol{\beta}, \sigma^2)$ that is to the left of c_i. Take log of (4), sum across i and maximise result wrt β_j and σ^2, get MLEs, leading to consistent, approximately Normal estimators; inference is standard.

Censored & Truncated Regression Models

Truncated Regression Models

- Can analyse data from example 17.4 as truncated sample if drop all observations whenever it is censored.
- However, we'd never analyse duration data (or top-coded data) in this way since it eliminates useful info: fact that we know lower bound for 893 durations along with explanatory variables is useful info and censored regression uses this info while truncated regression doesn't.
- Note: OLS applied to a sample truncated from above generally gives est biased towards 0.
- Like censored regression, if homogeneity and Normality assumptions in (3) are violated, truncated Normal MLE is biased and inconsistent.
- Methods not requiring these assumptions are available (see 'Advanced' Wooldridge).

Lecture 6 Outline

Censored & Truncated Models

Censored & Truncated Regression Models

Sample Selection

Correcting for Sample Selection

Summary & References

Summary & References

Sample Selection Corrections

Truncated regression is a special case of a general problem: **nonrandom sample selection**. Survey design is not the only cause of this – respondents fail to answer some questions, which leads to missing data for dependent / independent variables; since we can't use these observations, we wonder if dropping them leads to bias in our estimators.

Another general e.g.: **incidental truncation** – not observe y because of the outcome of another variable, e.g. *wage offer function* estimations: how factors affect wage you could earn, but only observe wage offer for those in workforce. Since working may be correlated with unobservables that affect wage offer, using only working people might produce biased estimates of parameters in wage offer equation. Nonrandom sample selection may arise in panel data, e.g. with two years of data, due to attrition, some people leave sample – particular problem in policy analysis, where attrition may be related to effectiveness of a program.

Conditions for consistency of OLS on selected sample

Truncated Tobit: endogenous sample selection and OLS biased and inconsistent. If sample determined solely by an exogenous explanatory variable, we've exog sample selection. Cases between extremes are less clear. Population model:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u \quad E(u|x_1, x_2, \ldots, x_k) = 0 \tag{5}$$

Population model for for a *random draw*:

$$y_i = \mathbf{x}_i \boldsymbol{\beta} + u_i \tag{6}$$

Let n be size of *random sample* from population. If observe y_i and each x_{ij} for all i, then use OLS.

Conditions for consistency of OLS on selected sample

Now assume either y_i or some of the independent variables are not observed for some i. For at least some observations, observe full set of variables. Define a *selection indicator* s_i for each i by $s_i = 1$ if we observe all of $(y_i, \mathbf{x}_i)$ and $s_i = 0$ otherwise, so $s_i = 1$ implies use observation in our analysis and $s_i = 0$ means the observation will not be used.

Interest: statistical properties of OLS estimates using the **selected sample**, i.e. using observations for which $s_i = 1$, so use fewer than n obs (n_i). Easy to get conditions where OLS consistent and unbiased. Rather than estimate (6), we can only estimate:

$$s_i y_i = s_i \mathbf{x}_i \boldsymbol{\beta} + s_i u_i \tag{7}$$

When $s_i = 1$, we've (6); when $s_i = 0$, we've $0 = 0 + 0$ and this tells us nothing about $\boldsymbol{\beta}$. Regressing $s_i y_i$ on $s_i \mathbf{x_i}$ for $i = 1, 2, \ldots, n$ is same as regressing y_i on $\mathbf{x}_i$ using observations when $s_i = 1$ so can learn about consistency of $\hat{\beta}_j$ by studying (7) on a random sample.

Conditions for consistency of OLS on selected sample

OLS estimates from (7) are consistent if error term has zero mean and is uncorrelated with each explanatory variable. In population, 0 mean assumption is $E(su) = 0$ and zero correlation assumption:

$$E[(sx_j)(su)] = E(sx_j u) = 0 \tag{8}$$

where s, x_j and u are random variables representing the population and we used the fact that $s^2 = s$ since s is binary. Condition (8) is different from what we need if we observe all variables for a random sample $E(x_j u) = 0$ so in population we need u to be uncorrelated with sx_j. Key cond for unbiasedness: $E(su|sx_1, \ldots, sx_k) = 0$, which (as usual) is a stronger assumption than that needed for consistency.

Conditions for consistency of OLS on selected sample

If s is a function only of the explanatory variables, then sx_j is just a function of $x_1, x_2, \ldots, x_k$; by conditional mean assumption (5), sx_j is also uncorrelated with u. Actually since $E(u|x_1, \ldots, x_k) = 0$, $E(su|sx_1, \ldots, sx_k) = sE(u|sx_1, \ldots, sx_k) = 0$. This is the case of **exogenous sample selection**, where $s_i = 1$ is determined entirely by $x_{i1}, \ldots, x_{ik}$. If sample selection is entirely random in the sense that s_i is *independent* of $(\mathbf{x_i}, u_i)$, then $\because E(x_j u) = 0$ under (5), $E(sx_j u) = E(s)E(x_j u) = 0$. So, beginning with a random sample and randomly dropping observations, OLS is still consistent and is unbiased in this case if there is no perfect multicollinearity in the selected sample. If s depends on explanatory variables and extra random terms that are independent of $\mathbf{x}$ and u, OLS is consistent and unbiased. Conditional on explanatory variables, $s \perp\!\!\!\perp u$ so $E(u|x_1, \ldots, x_k, s) = E(u|x_1, \ldots, x_k)$ and the last is 0 by assumption on population model. If we add homogeneity assumption $E(u^2|\mathbf{x}, s) = E(u^2) = \sigma^2$, then usual OLS SE and test statistics are valid.

Conditions for consistency of OLS on selected sample

When is OLS on selected sample inconsistent? When truncated from above: $s_i = 1$ if $y_i \leq c_i$ or equivalently $s_i = 1$ if $u_i \leq c_i - \mathbf{x}_i\boldsymbol{\beta}$. Since s_i depends directly on u_i, s_i and u_i won't be uncorrelated even conditional on $\mathbf{x}_i$. This is why OLS on selected sample doesn't consistently estimate β_j. There are less obvious ways that s and u can be correlated – consider them in next subsection. Results on consistency of OLS extend to IV (lectures 7-8). If IVs are denoted z_h in the population, the key condition for consistency of 2SLS is $E(sz_hu) = 0$, which holds if $E(u|\mathbf{z}, s) = 0$. Thus, if selection is determined entirely by exogenous variables $\mathbf{z}$, or if s depends on other factors that are independent of u and $\mathbf{z}$, then 2SLS on selected sample is generally consistent. Need to assume explanatory and IV are appropriately correlated in selected part of population. When selection is entirely a function of exogenous variables, MLE of a nonlinear model (e.g. logit/probit) produces consistent, asymptotically Normal estimates and the usual SE and test statistics are valid.

Incidental Truncation

Start with population model (5). However, assume always observe explanatory variables $\mathbf{x_j}$. Problem: only observe y for subset of population. Rule determining whether we observe y does *not* depend directly on outcome of y. E.g. $y = \log\left(wage^o\right)$ where $wage^o$ is wage offer or hourly wage individual could receive in labour market. Can only observe if working: wage offer is assumed to be observed wage so truncation of wage offer is is *incidental* because it depends on another variable, *viz* labour force participation. NB: we'd generally observe all other info about an individual, e.g. education, prior experience, gender, marital status, etc. Usual approach to incidental truncation is to add an explicit selection equation to population model of interest:

$$y = \mathbf{x}\boldsymbol{\beta} + u \quad E(u|\mathbf{x}) = 0 \tag{9}$$
$$s = 1[\mathbf{z}\gamma + \nu \geq 0] \tag{10}$$

where $s = 1$ if observe y and 0 otherwise. Assume elements of $\mathbf{x}$ and $\mathbf{z}$ are always observed.

Incidental Truncation

- Interest: (9) and estimate $\boldsymbol{\beta}$ by OLS given a random sample.
- Selection equation (10) depends on observed variables z_h and an unobserved error ν.
- Standard assumption is $\mathbf{z}$ is exogenous in (9):

$$E(u|\mathbf{x}, \mathbf{z}) = 0$$

- NB: let $\mathbf{x}$ be a strict subset of $\mathbf{z}$: any x_j is also an element of $\mathbf{z}$ and we've some elements of $\mathbf{z}$ that are not also in $\mathbf{x}$.
- Assume $\nu \perp\!\!\!\perp \mathbf{z}$ (thus $\perp\!\!\!\perp$ $\mathbf{x}$) and $\nu \sim N(0, 1)$.

Incidental Truncation

Correlation between u and ν generally causes a sample selection problem: assume $(u, v) \perp\!\!\!\perp \mathbf{z}$, using $\mathbf{x}$ strict subset of $\mathbf{z}$ take:

$$E(y|\mathbf{z},\nu) = \mathbf{x}\boldsymbol{\beta} + E(u|\mathbf{z},\nu) \overset{(u,\nu)\perp\!\!\!\perp\mathbf{z}}{=} \mathbf{x}\boldsymbol{\beta} + E(u|\nu)$$

If (u,ν) jointly Normal (0 mean), then $E(u|\nu) = \rho\nu$

$$E(y|\mathbf{z},\nu) = \mathbf{x}\boldsymbol{\beta} + \rho\nu$$

Don't observe ν, but we can use this equation to compute $E(y|\mathbf{z},s)$ and then specialise this to $s = 1$ to get

$$E(y|\mathbf{z},s) = \mathbf{x}\boldsymbol{\beta} + \rho E(\nu|\mathbf{z},s)$$

Since s and ν are related by (10) and $\nu \sim N(0,1)$, we can show that $E(\nu|\mathbf{z},s)$ is simply inverse Mills ratio $\lambda(\mathbf{z}\gamma)$ when $s = 1$, so:

$$E(y|\mathbf{z},s=1) = \mathbf{x}\boldsymbol{\beta} + \rho\lambda(\mathbf{z}\gamma)$$

Want to estimate $\boldsymbol{\beta}$ and this shows we can using only the selected sample as long as we include the term $\lambda(\mathbf{z}\gamma)$ as an additional regressor.

Incidental Truncation

- If $\rho = 0$, $\lambda(\mathbf{z}\gamma)$ doesn't appear and OLS of y on $\mathbf{x}$ using selected sample consistently estimates $\boldsymbol{\beta}$. Else, we've omitted a variable $\lambda(\mathbf{z}\gamma)$, which is generally corr with $\mathbf{x}$.
- When does $\rho = 0$? When u and ν are uncorrelated.
- Since γ is unknown, we can't evaluate $\lambda(\mathbf{z_i}\gamma)$ for each i, but from assumptions, s given z follows probit:

$$P(s = 1|\mathbf{z}) = \Phi(\mathbf{z}\gamma)$$

- Estimate γ by probit of s_i on $\mathbf{z_i}$ using *entire* sample.
- In second step, we can estimate $\boldsymbol{\beta}$.
- This is called the **Heckit method** – Heckman 1976, Nobel in 2000 for this (sample selection correction).

Incidental Truncation

Heckit method – sample selection correction

1. Using all n observations, estimate probit of s_i on $\mathbf{z_i}$ and obtain estimates $\hat{\gamma}_h$. Compute inverse Mills ratio $\hat{\lambda}_i = \lambda(\mathbf{z}_i\hat{\boldsymbol{\gamma}})$ for each i. (Actually, only need these for i with $s_i = 1$.)
2. Using selected sample, i.e. observations for which $s_i = 1$ (say n_1 of them), run the regression of y_i on $\mathbf{x}_i$, $\hat{\lambda}_i$.

The $\hat{\beta}_j$ are consistent and approximately Normally distributed. Simple test of selection bias is available from regressing y_i on $\mathbf{x}_i$, $\hat{\lambda}_i$: use usual t statistic on $\hat{\lambda}_i$ as a test of $H_0 : \rho = 0$. Under H_0, there's no sample selection problem. $\rho \neq 0 \implies$ OLS SE from this regression aren't exactly correct since they don't account for estimation of γ, which uses same observations in this regression and more.

Incidental Truncation

Implications of **x** being a strict subset of **z**

1. Any element appearing as explanatory variable in (9) should also be an explanatory variable in selection equation. Rare cases: sense to exclude elements from selection equation, but including all elements of **x** in **z** is not very costly; extending them can lead to inconsistency if they're incorrectly excluded.
2. At least one element of **z** is not also in **x** so need a variable that affects selection but does *not* have a partial effect on y. Not absolutely necessary to apply the procedure (can carry out two steps when $\mathbf{z} = \mathbf{x}$) but results usually less than convincing unless have an *exclusion restriction* in (9). Reason: inverse Mills ratio is a nonlinear function of **z** but it's often well approx by a linear function. If $\mathbf{z} = \mathbf{x}$, $\hat{\lambda}_i$ can be highly correlated with elements of $\mathbf{x}_i$, leading to very high SE for $\hat{\beta}_j$. Intuitively, if we've no variable that affects selection but not y, it's really hard (almost impossible) to distinguish sample selection from misspecified functional form in (9).

Incidental Truncation

Implications of **x** being a strict subset of **z**

Alt to 2-step estimation: full MLE. More complicated since need joint dist of y and s. What makes sense is to test for sample selection using the 2-step procedure and if find no evidence of sample selection, then no reason to continue; else either use 2-step estimation or estimate regression and selection equations jointly by MLE. This approach has adv of using more info but is less widely applicable.

Many more topics re-sample selection. One: models with endogenous explanatory variables *in addition to* possible sample selection bias. Single endogenous explanatory variable:

$$y_1 = \alpha_1 y_2 + \mathbf{z}_1 \boldsymbol{\beta}_1 + u_1$$

where y_i is obs when $s = 1$ and y_2 may only be obs along with y_1. Example?

Instrumental Variables and Two Stage Least Squares.

Lecture 6 Outline

Censored & Truncated Models

- Censored & Truncated Regression Models

Sample Selection

- Correcting for Sample Selection

Summary & References

- Summary & References

Summary

- With censored regression, we have info when y missing (e.g. top coding).
  - Requires MLE but complicated if assumptions of censored normal regression model are violated.
  - Application: duration analysis (time before event occurs).
- With truncated regression (non-random sample), we only know inclusion rule – not even know how many variables are missing – only observe $(\mathbf{x}_i, y_i)$ if $y_i \leq c_i$.
  - Requires MLE but complicated if assumptions of truncated normal regression model are violated.
- Nonrandom sample selection (truncated regression is a special case) and incidental truncation motivate sample selection corrections (e.g. Heckit method).

References

- Censored & Truncated Regression Models: Wooldridge 17.4.
- Sample Selection Correction: Wooldridge 17.5.