Econometrics II
Censoring & Truncation

Måns Söderbom

May 5, 2011

1 Censored and Truncated Models

Recall that a corner solution is an actual economic outcome, e.g. zero expenditure on health by a household in a given period. In this section we discuss briefly two close cousins of the corner solution model, namely the censored regression model and the truncated regression model.

The good news is that the econometric techniques used for censored and truncated dependent variables are very similar to what we have already studied.

1.1 Censored regression models

In contrast to corner solutions, censoring is essentially a **data problem**. Censoring occurs, for example, if whenever y exceeds some upper threshold c the

actual value of y gets **recorded** as equal to c, rather than the true value. Of course, censoring may also occur at the lower end of the dependent variable. **Top coding** in income surveys is the most common example of censoring, however. Such surveys are sometimes designed so that that people with incomes higher than some upper threshold, say $\$500,000$, are allowed to respond "more than $\$500,000$". In contrast, for people with incomes lower than $\$500,000$ the actual income gets recorded. If we want to run a regression explaining income based on such data, we clearly need to deal with the top coding. A reasonable way of writing down the model might be

$$y^* = \boldsymbol{x}\boldsymbol{\beta} + u,$$

$$y = \min\left(y^*, c\right),$$

where y^* is **actual** income (which is not fully observed due to the censoring), u is a normally distributed and homoskedastic residual, and y is measured income,

which in this example is bounded above at $c = \$500,000$ due to the censoring produced by the design of the survey.

You now see that the censored regression is very similar to the corner solution model. In fact, if $c = 0$ and this is a lower bound, the econometric model for corner solution models and censored regressions coincide: in both cases we would have the tobit model. If the threshold c is not zero and/or represents an upper rather than a lower bound on what is observed, then we still use tobit but with a simple (and uninteresting) adjustment of the log likelihood.

The only substantive difference between censored regressions models and corner solution models lies in the **interpretation of the results**. Suppose we have two models:

- Model 1: the dependent variable is a corner solution variable, with the corner at zero

- Model 2: the dependent variable is censored below at zero.

We could use exactly the same econometric estimator for both models, i.e. the tobit model. In the corner solution model we are probably mainly interested in how the expected value of the observed dependent variable varies with the explanatory variable(s). This means we should look at $E(y|\boldsymbol{x}, y > 0)$ or $E(y|\boldsymbol{x})$, and we have seen in the previous section how to obtain the relevant marginal effects. However, for the censored regression model we are mostly interested in learning how the expected value of the **unobserved and censored** variable y^* varies with the explanatory variable(s), i.e. $E(y^*|\boldsymbol{x})$:

$$E(y^*|\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta},$$

and so the partial effect of x_j is simply β_j.

1.1.1 Duration Data

One field in which censored regression models are very common is in the econometric analysis of **duration data.** Duration is the time that elapses between the 'beginning' and the 'end' of some specified state. The most common example is unemployment duration, where the 'beginning' is the day the individual becomes unemployed and the 'end' is when the same individual gets a new job. Other examples are the duration of wars, duration of marriages, time between first and second child, the lifetimes of firms, the length of stay in graduate school, time to adoption of new technologies, length of financial crises etc etc.

Data on durations are often censored, either to the right (common) or to the left (not so common) or both (even less common). Right censoring means that we don't know from the data when a certain duration ended; left censoring

means that we don't know when it began. I will not cover duration data as part of this course, but you can find an old lecture introducing duration data models on my web page.

1.2 Truncated regression models

A truncated regression model is similar to a censored regression model, but there is one important difference:

- If the dependent variable is truncated we do not observe **any** information about a certain segment in the population.

- In other words, we do not have a representative (random) sample from the population. This can happen if a survey targets a sub-group of the population. For instance when surveying firms in developing countries, the World Bank often excludes firms with less than 10 employees. Clearly if we are modelling employment based on such data we need to recognize the fact that firms with less than 10 employees are not covered in our dataset.

- Alternatively, it could be that we target poor individuals, and so exclude everyone with an income higher than some upper threshold c.

- The standard truncated regression model is written

$$y = \boldsymbol{x}\boldsymbol{\beta} + u,$$

  where the residual u is assumed normally distributed, homoskedastic and uncorrelated with $\boldsymbol{x}$ (the latter assumption can be relaxed if we have instruments). Suppose that all observations for which $y_i > c$ are excluded from the sample. Our objective is to estimate the parameter $\boldsymbol{\beta}$.

- See example in appendix, Section 5.

It is clear from the example in the appendix that ignoring the truncation leads to substantial downward bias in the estimate of $\boldsymbol{\beta}$. Fortunately, we can correct this bias fairly easily, by using the normality assumption in combination with the information about the threshold. The density of y, conditional on $\boldsymbol{x}$ and y observed, takes a familiar form:

$$f\left(y|\boldsymbol{x};\boldsymbol{\beta},\boldsymbol{\gamma}\right)=\left[\frac{\phi\left(\left(y-\boldsymbol{x\beta}\right)/\sigma\right)/\sigma}{\boldsymbol{\Phi}\left(\boldsymbol{x\beta}/\sigma\right)}\right],$$

and the individual log likelihood contribution is

$$\ln L_i=\ln\left[\phi\left(\left(y_i-\boldsymbol{x}_i\boldsymbol{\beta}\right)/\sigma\right)/\sigma\right]-\ln\boldsymbol{\Phi}\left(\boldsymbol{x}_i\boldsymbol{\beta}/\sigma\right)$$

The conditional expected value of y is also of a familiar form:

$$E\left(y|y>0,\boldsymbol{x}\right)=\boldsymbol{x\beta}+\sigma_u\lambda\left(\boldsymbol{x\beta}/\sigma_u\right)$$

In Stata we can implement this model using the **truncreg** command (see appendix).

5. Illustration: The truncated regression model

Consider a simple simulation, obtained by the following Stata code:

```
clear
set seed 2355
set obs 500

ge u=invnorm(uniform())

ge x=2*uniform()

/* true population model: y = -1 + 1*x + u /

ge y=-1+x+u

/* no truncation */
reg y x
predict yh_ols_nt

/* truncation of y at 0.8*/
reg y x if y<.8
predict yh_ols_t

/* truncated regression corrects for the trunctation. ul(.) indicates the upper limit */

truncreg y x, ul(0.8)
```

Consider three different regressions based on these artificial data:

i) OLS using the full sample of 500 observations (i.e. no truncation)

```
. reg y x

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  1,   498) =  156.47
       Model |  139.883218     1  139.883218           Prob > F      =  0.0000
    Residual |  445.219899   498  .894015862           R-squared     =  0.2391
-------------+------------------------------           Adj R-squared =  0.2375
       Total |  585.103118   499  1.17255134           Root MSE      =  .94552

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .8940591   .0714753    12.51   0.000     .7536288    1.034489
       _cons |  -.9019037   .0834538   -10.81   0.000    -1.065869   -.7379389
------------------------------------------------------------------------------
```

ii) OLS using the truncated sample of 380 observations

```
. reg y x if y<.8

      Source |       SS       df       MS              Number of obs =     380
-------------+------------------------------           F(  1,   378) =   47.00
       Model |   28.616886     1   28.616886           Prob > F      =  0.0000
    Residual |  230.164146   378  .608899857           R-squared     =  0.1106
-------------+------------------------------           Adj R-squared =  0.1082
       Total |  258.781032   379  .682799556           Root MSE      =  .78032

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .4811388    .070183     6.86   0.000     .3431407    .6191369
       _cons |  -.8577185   .0732374   -11.71   0.000    -1.001722   -.7137147
------------------------------------------------------------------------------
```

Notice coefficient on x is much lower than the true value of one. It is clearly significantly different from one, indicating significant bias.

Figure 3 illustrates the problem of truncation.

iii) Truncated regression which corrects for the truncation

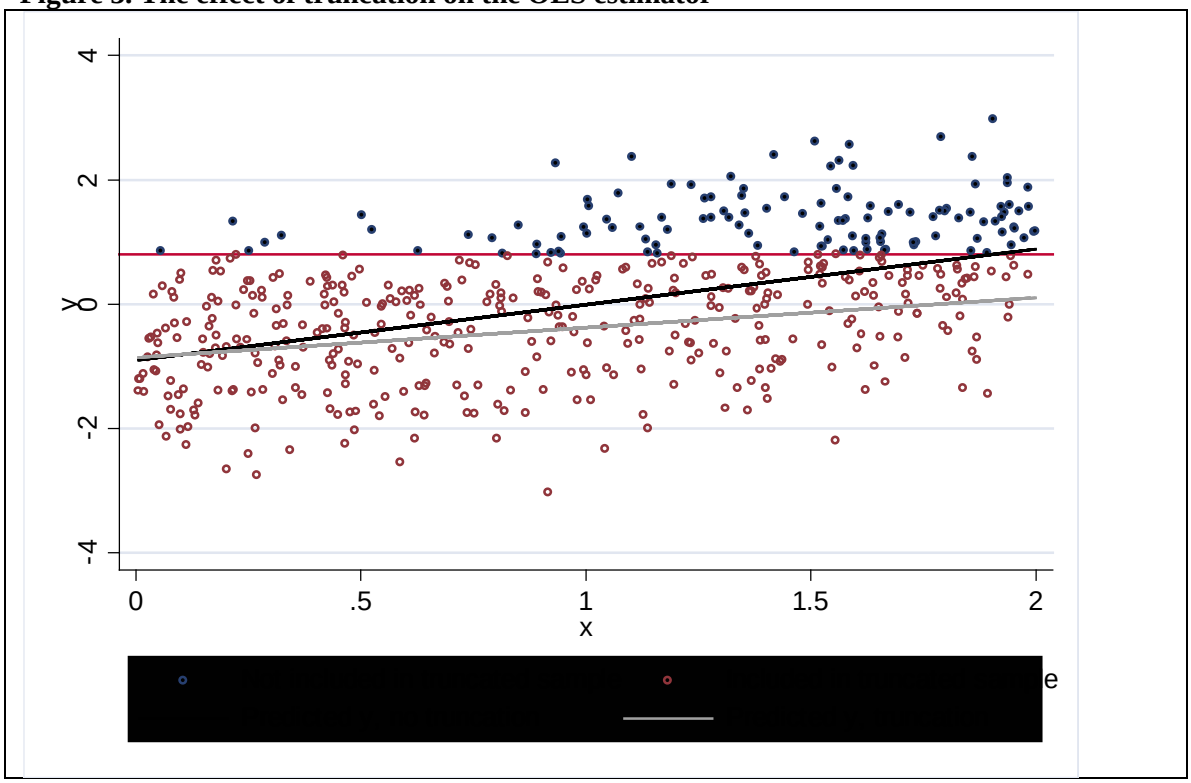
```
. truncreg y x, ul(0.8)
(note: 120 obs. truncated)

Truncated regression
Limit:   lower =       -inf                          Number of obs =    380
         upper =         .8                          Wald chi2(1)  =  37.41
Log likelihood = -398.51329                          Prob > chi2   = 0.0000

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
eq1          |
           x |   .8506762   .1390748     6.12   0.000     .5780947    1.123258
       _cons |  -.7836381   .1214471    -6.45   0.000     -1.02167   -.5456061
-------------+----------------------------------------------------------------
sigma        |
       _cons |   1.019341    .067624    15.07   0.000     .8868003    1.151882
------------------------------------------------------------------------------
```

Coefficient increases as a result and is similar to the OLS estimate in (i) and not significantly different from the true value of 1.

Figure 3. The effect of truncation on the OLS estimator



Note: The predications have been generated from the OLS estimates shown in (i) and (ii) above.

References

Bigsten, Arne, Paul Collier, Stefan Dercon, Marcel Fafchamps, Bernard Gauthier, Jan Willem Gunning, Remco Oostendorp, Catherine Pattillo, Måns Söderbom, and Francis Teal (2005). "Adjustment Costs, Irreversibility and Investment Patterns in African Manufacturing," *The B.E. Journals in Economic Analysis & Policy: Contributions to Economic Analysis & Policy* 4:1, Article 12, pp. 1-27.

Söderbom, Måns, and Francis Teal (2004). "Size and Efficiency in African Manufacturing Firms: Evidence from Firm-Level Panel Data," *Journal of Development Economics* 73, pp. 369-394.