ECON 2007: Quant Econ and Econometrics
Censored, Truncated and Count Data

Dr. Áureo de Paula

Department of Economics
University College London

Censoring

The censored regression model focusses on a classical linear regression model where outcome observations are censored. For a certain range of values of the outcome, the econometrician does not know the exact value but only that the variable is within a certain interval.

Usually censoring occurs because of survey limitations such as top coding, cost considerations or attrition.

Censored regression models are mathematically very similar to the Tobit model. Whereas there were no observability issues there, here data on the censoring region are “incomplete”.

Censoring

Adopting the notation in the Tobit model, the model is

$$y^* = \beta^\top \mathbf{x} + u \quad u|\mathbf{x}, c \sim \mathcal{N}(0, \sigma^2)$$
$$y = \min(c, y^*)$$

where now c is possibly random and instead of a max operator, we focus on the min.

The max case is analogous and relates to censoring from below, i.e. left-censoring, instead of from above, i.e. right-censoring.

Censoring

As in the Tobit case,

$$\begin{aligned} Pr(y = c|\mathbf{x}) &= Pr(y^* \geq 0|\mathbf{x}) = Pr(u \geq c - \beta^\top \mathbf{x}|\mathbf{x}) \\ &= 1 - \Phi[(c - \beta^\top \mathbf{x})/\sigma] \end{aligned}$$

And the density of y conditional on $\mathbf{x} = \mathbf{x}_i$ and $c = c_i$ is given by

$$(1/\sigma)\phi[(y - \beta^\top \mathbf{x}_i)/\sigma], y < c_i \qquad 1 - \Phi[(c_i - \beta^\top \mathbf{x}_i)/\sigma], y = c_i$$

which is then used to construct a MLE much as in the Tobit case.

Censoring is typical in duration analysis (i.e., time-to-event).

Notice that in this case the marginal effects are given by β (as opposed to Tobit)!

Censoring

For example, Costa and Kahn (2003) use data on soldiers in the American Civil War to study how individual and community level variables affected group loyalty as measured by time until desertion, arrest or AWOL.

They use a more sophisticated statistical model (e.g., competing risks hazard model), but we could also analyze the data using a regression model where duration is censored.

Censoring

VARIABLE MEANS FOR ALL MEN, FOR DESERTED, ARRESTED, AND AWOL COMBINED AND FOR DESERTED, ARRESTED, AND AWOL SEPARATELY

	Combined	Std dev	All outcomes	Deserted	Arrested	AWOL
Days from muster until			237.181	190.644	385.175	356.181

because morale varies over time, because men can become more committed soldiers, and because of censoring—some men may have died, been discharged, changed company, become prisoners of war, or be missing in action before they could desert. We treat these men as censored in our estimation strategy. When we

Source: *Costa and Kahn, (2003): "Cowards and Heroes: Group Loyalty in the American Civil War", QJE, V.118(2)*

Truncation

In a truncated regression model, certain observations are not selected into the sample. This can arise, for instance, in surveys where for cost considerations only a subset of the population is targeted.

The model relies on a regression model:

$$y = \beta^\top \mathbf{x} + u, \quad E(u|\mathbf{x}) = 0. \tag{1}$$

Using a random sample with n observations, we could simply use OLS and obtain an unbiased estimator for β.

Truncation

Let's assume instead that certain observations are fully observed whereas others are not observed at all. Mark the selection *into* the observed sample by the indicator variable s_i. This variable is $= 1$ if unit i is observed and $= 0$ otherwise.

Instead of estimating equation (1), we instead focus on

$$s_i y_i = \beta^\top s_i \mathbf{x}_i + s_i u_i \tag{2}$$

When $s_i = 1$, we have (1) for the random draw i. Otherwise, when $s_i = 0$, we obtain $0 = 0 + 0$ which is vacuous and adds nothing to the estimation. In the end, running OLS on (2) is equivalent to running OLS *only* on those observations selected out of the n initial draws.

Truncation

When and how does truncation affect the estimation properties?

For consistency remember that we require that

$$E(su) = 0 \qquad E[(sx_j)(su)] = E[sx_j u] = 0.$$

These are implied by the stronger condition:

$$E(su|s\mathbf{x}) = 0$$

which would also imply that OLS is unbiased.

Truncation

Now we can lay out a few scenarios:

- If the selection rule s depends only on the explanatory variables $\mathbf{x}$, then sx_j depends only on $\mathbf{x}$ and there is nothing in $s\mathbf{x}$ that is not known beyond $\mathbf{x}$.Consequently $E(u|s\mathbf{x}) = 0$ since $E(u|\mathbf{x}) = 0$ by (1). Then, $E(su|s\mathbf{x}) = sE(u|s\mathbf{x}) = 0$ and the estimator is unbiased and consistent.

- If the selection is completly independent of $(\mathbf{x}, u)$, then $E(sx_j u) = E(s)E(x_j u) = 0$ and OLS is consistent. It can also be shown that OLS is unbiased.

Truncation

- If the selection depends on the explanatory variables and randomness that is independent of u, again one obtain unbiasedness since $E(u|\mathbf{x}, s) = E(u|\mathbf{x})$. This follows because, conditional on $\mathbf{x}$, s is independent of u.

- If the selection rule relies on the regressand y, OLS will typically be inconsistent. For example, let $s = 1$ if $y \leq c$ where c is a random variable and $s = 0$ otherwise. Then

$$s = 1 \text{ if and only if } u \leq c - \beta^\top \mathbf{x}$$

  Since s depends on u, they will not be uncorrelated even as we condition on $\mathbf{x}$. In this case, typically $E(sx_j u) \neq 0$ and OLS will be inconsistent.

Truncation

Consider for example a linear regression model where:

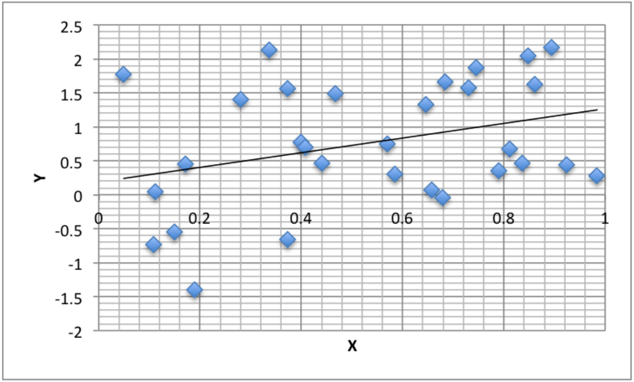
$$y = \beta_0 + \beta_1 \mathbf{x} + u$$

with $\beta_0 = 0$ and $\beta_1 = 1$.

Assuming distributions for $\mathbf{x}$ and u, we can simulate a sample for the model above and examine the regression line obtained in the sample.

Truncation

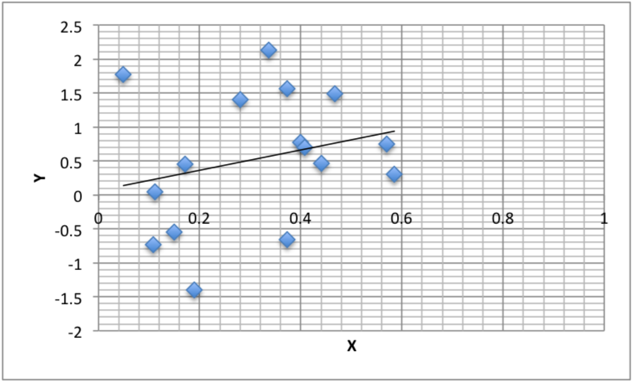
Take a sample with 30 observations:



The slope is close (though not quite equal) to $\beta_1 = 1$. In fact, $\hat{\beta}_1 = 1.08$.

Truncation

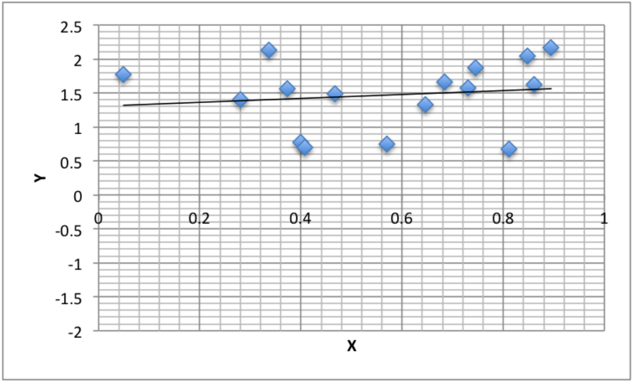
Consider now only the observations where $\mathbf{x} \leq 0.6$:



The slope is (still!) close (though not quite equal) to $\beta_1 = 1$. Here $\hat{\beta}_1 = 1.01$.

Truncation

Consider now only the observations where $y \geq 0.5$:



In this case, the slope is much flatter than before: $\hat{\beta}_1 = 0.27$!

Truncation

To address the last scenario, we expand the model to:

$$y = \beta^\top \mathbf{x} + u, \quad u|\mathbf{x}, c \sim \mathcal{N}(0, \sigma^2).$$

Remember that the selection rule is that a random draw $(\mathbf{x}_i, y_i)$ is observed only if $y_i \leq c_i$.

(In contrast, in a censored model the realizations of $\mathbf{x}_i$ would be known for $y_i > c_i$.)

Truncation

The density of y given $\mathbf{x} = \mathbf{x}_i$ and $c = c_i$ is then given by

$$\frac{f(y|\mathbf{x}_i, \beta, \sigma)}{F(c_i|\mathbf{x}_i, \beta, \sigma)} \qquad \text{if } y < c_i$$

where $f(y|\mathbf{x}, \beta, \sigma)$ is the normal density with mean $\beta^\top \mathbf{x}$ and variance σ^2 and $F(y|\mathbf{x}_i, \beta, \sigma)$ is the corresponding cumulative distribution function.

This is then used to compute the MLE for the truncated regression.

Truncation

An important type of truncation in Economics is (what Wooldridge calls "incidental truncation"), more commonly referred to simply as sample selection.

We only observe y for a subset of the population and the selection rule depends indirectly on the outcome.

The canonical example relates wages $y = \log(\text{wage})$ to variables such as experience. This variable is nevertheless only observed for those who *choose* to participate in the labor force, a decision that may depend on variables such as non-labor income.

Truncation

The model here is

$$
\begin{aligned}
y &= \beta^\top \mathbf{x} + u, \quad E(u|\mathbf{x}, \mathbf{z}) = 0 & (3) \\
s &= 1[\gamma^\top \mathbf{z} + \nu \geq 0] & (4)
\end{aligned}
$$

where $s = 1$ if we observe y and zero otherwise. The variables $\mathbf{x}$ and $\mathbf{z}$ are *always* observed. We will assume that $\mathbf{x}$ is a strict sub-vector of $\mathbf{z}$ and $\mathbf{z}$ is independent of (u, ν).

Taking the expectation of (3) conditional on $\mathbf{z}$ and ν we have

$$E[y|\mathbf{z}, \nu] = \beta^\top \mathbf{x} + E[u|\mathbf{z}, \nu] = \beta^\top \mathbf{x} + E[u|\nu]$$

Truncation

It follows that, if (u, ν) are jointly normal with mean zero, $E[u|\nu] = \rho\nu$ for some constant ρ. So,

$$E[y|\mathbf{z}, \nu] = \beta^\top \mathbf{x} + \rho\nu.$$

We do not observe ν, but know whether $s = 1$. Using the formula above, it can be shown that

$$E[y|\mathbf{z}, s = 1] = \beta^\top \mathbf{x} + \rho E[\nu|s = 1] = \beta^\top \mathbf{x} + \rho\lambda(\gamma^\top \mathbf{z}).$$

The last equality derives from normality of ν and (4) in a similar way as in our analysis of censored regressions.

Truncation

The parameter ρ will be zero when u and ν are independent. In this case, the selection is based on explanatory variables and a random component that is independent of u. As we saw before, OLS on the truncated sample will be consistent.

If $\rho \neq 0$, OLS will not be consistent: the inverse Mills' ratio would be an ommited variable. We cannot immediately include that variable though, since it depends on γ, which is an unknown parameter. It can nevertheless be estimated on the *whole* sample since

$$Pr(s = 1|\mathbf{z}) = \Phi(\gamma^\top \mathbf{z}).$$

Truncation

Once γ is estimated we can plug in $\lambda(\hat{\gamma}^\top \mathbf{z}_i)$ as an additional explanatory variable for each observation in the truncated sample and run OLS.

This procedure, originally due to Heckman (1976), turns out to provide a consistent estimator for β.

Truncation

Some caveats apply.

1. Standard errors need to be corrected to account for the first step estimation of γ.

2. Strictly speaking we can have $\mathbf{x} = \mathbf{z}$ as $\lambda(\cdot)$ is a nonlinear function. Depending on the range of $\mathbf{x}$ though, $\lambda(\cdot)$ behaves very much like a linear function and multicollinearity issues may arise. If there are excluded variables appearing in the selection equation, this is less of an issue.

Truncation

```
Probit regression                                 Number of obs   =        753
                                                  LR chi2(7)      =     227.14
                                                  Prob > chi2     =     0.0000
Log likelihood = -401.30219                       Pseudo R2       =     0.2206

------------------------------------------------------------------------------
        inlf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nwifeinc |  -.0120237   .0048398    -2.48   0.013    -.0215096   -.0025378
        educ |   .1309047   .0252542     5.18   0.000     .0814074     .180402
       exper |   .1233476   .0187164     6.59   0.000     .0866641    .1600311
     expersq |  -.0018871      .0006    -3.15   0.002     -.003063   -.0007111
         age |  -.0528527   .0084772    -6.23   0.000    -.0694678   -.0362376
     kidslt6 |  -.8683285   .1185223    -7.33   0.000    -1.100628    -.636029
     kidsge6 |    .036005   .0434768     0.83   0.408     -.049208    .1212179
       _cons |   .2700768    .508593     0.53   0.595    -.7267473    1.266901
------------------------------------------------------------------------------
```

Truncation

Source	SS	df	MS
Model	35.0479487	4	8.76198719
Residual	188.279492	423	.445105182
Total	223.327441	427	.523015084

Number of obs	= 428
F(4, 423)	= 19.69
Prob > F	= 0.0000
R-squared	= 0.1569
Adj R-squared	= 0.1490
Root MSE	= .66716

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.1090655	.0156096	6.99	0.000	.0783835	.1397476
exper	.0438873	.0163534	2.68	0.008	.0117434	.0760313
expersq	-.0008591	.0004414	-1.95	0.052	-.0017267	8.49e-06
lambda	.0322619	.1343877	0.24	0.810	-.2318889	.2964126
_cons	-.5781032	.306723	-1.88	0.060	-1.180994	.024788

Truncation

```
Heckman selection model -- two-step estimates   Number of obs      =       753
(regression model with sample selection)        Censored obs       =       325
                                                Uncensored obs     =       428

                                                Wald chi2(3)       =     51.53
                                                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lwage        |
        educ |   .1090655    .015523     7.03   0.000     .0786411      .13949
       exper |   .0438873   .0162611     2.70   0.007     .0120163    .0757584
     expersq |  -.0008591   .0004389    -1.96   0.050    -.0017194    1.15e-06
       _cons |  -.5781032   .3050062    -1.90   0.058    -1.175904     .019698
-------------+----------------------------------------------------------------
select       |
    nwifeinc |  -.0120237   .0048398    -2.48   0.013    -.0215096   -.0025378
        educ |   .1309047   .0252542     5.18   0.000     .0814074     .180402
       exper |   .1233476   .0187164     6.59   0.000     .0866641    .1600311
     expersq |  -.0018871      .0006    -3.15   0.002     -.003063   -.0007111
         age |  -.0528527   .0084772    -6.23   0.000    -.0694678   -.0362376
     kidslt6 |  -.8683285   .1185223    -7.33   0.000    -1.100628    -.636029
     kidsge6 |    .036005   .0434768     0.83   0.408     -.049208    .1212179
       _cons |   .2700768    .508593     0.53   0.595    -.7267473    1.266901
-------------+----------------------------------------------------------------
mills        |
      lambda |   .0322619   .1336246     0.24   0.809    -.2296376    .2941613
-------------+----------------------------------------------------------------
         rho |    0.04861
       sigma |  .66362875
------------------------------------------------------------------------------
```

Count Data

A count variable is a random variable that takes on non-negative integer values: $\{0, 1, 2, \ldots\}$.

One of the most commons distributions for such variables is the Poisson distribution which postulates the probability mass function:

$$Pr(y) = \frac{\exp(-\lambda)\lambda^y}{y!}, \quad y \in \{0, 1, 2, \ldots\}$$

where $\lambda > 0$ characterizes the distribution and $y! = 1 \times 2 \times \cdots \times y$. It can be shown that

$$E(y) = \lambda \qquad var(y) = \lambda.$$

Count Data

In modelling the dependence of a count variable y on $\mathbf{x}$, the Poisson regression simply assumes that λ is a function of $\mathbf{x}$. Because $\lambda > 0$, it is generally imposed that

$$\lambda = \exp(\beta^\top \mathbf{x})$$

so that

$$Pr(y|\mathbf{x}) = \frac{\exp(-\exp(\beta^\top \mathbf{x}))\exp(\beta^\top \mathbf{x})^y}{y!}, \quad y \in \{0, 1, 2, \dots\} \tag{5}$$

and

$$E(y|\mathbf{x}) = \exp(\beta^\top \mathbf{x}).$$

Count Data

The conditional probability mass function (5) can be used to form the log-likelihood to obtain a MLE:

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} l_i(\beta) = \sum_{i=1}^{n} \{y_i \beta^\top \mathbf{x}_i - \exp(\beta^\top \mathbf{x}_i)\}$$

where terms that do not depend on β are dropped without loss.

Noting that $\partial E(y|\mathbf{x})/\partial x_j = \exp(\beta^\top \mathbf{x})\beta_j$ we can also form APE or PEA estimators for the marginal effect of x_j.

Count Data

Some caveats:

1. The model imposes the restriction $var(y|\mathbf{x}) = E(y|\mathbf{x})$.
2. Even if this is not the case, the estimator for β is still consistent! (In which case we refer to it as **quasi-maximum likelihood estimator**.)
3. In this case, we can allow $var(y|\mathbf{x}) = \sigma^2 E(y|\mathbf{x})$ where sigma is consistently estimated by $(n-k-1)^{-1}\sum_{i=1}^{n}\hat{u}_i^2/\hat{y}_i$ where $\hat{u}_i = y_i - \hat{y}_i$.
4. ...other solutions still allow for robust standard errors.

Count Data

```
Poisson regression                                Number of obs   =       2725
                                                  Wald chi2(8)    =     245.65
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood = -2249.0801                 Pseudo R2       =     0.0790

------------------------------------------------------------------------------
             |               Robust
      narr86 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        pcnv |  -.4052683   .1011874    -4.01   0.000    -.6035919   -.2069447
      avgsen |  -.0236308   .0235749    -1.00   0.316    -.0698367    .0225752
     tottime |   .0243395   .0205298     1.19   0.236    -.0158982    .0645771
     ptime86 |   -.098591   .0223129    -4.42   0.000    -.1423235   -.0548584
      qemp86 |  -.0361079   .0341832    -1.06   0.291    -.1031058      .03089
       inc86 |  -.0081463   .0012306    -6.62   0.000    -.0105582   -.0057345
       black |   .6603471   .0995567     6.63   0.000     .4652195    .8554747
      hispan |   .4995934   .0924134     5.41   0.000     .3184665    .6807204
       _cons |   -.617177    .083212    -7.42   0.000    -.7802696   -.4540844
------------------------------------------------------------------------------
```

Count Data

Variable	Obs	Mean	Std. Dev.	Min	Max
sigma	2725	1.23272	0	1.23272	1.23272

Count Data

Variable	Obs	Mean
t_tottime	2725	.9617503
t_qemp86	2725	-.8568887
t_inc86	2725	-5.370232
t_black	2725	5.380681
t_hispan	2725	4.385482

These slides covered:

Wooldridge 17