Workshop on econometric analysis using Stata, July 26-31, 2010, Enugu, Nigeria

Limited dependent econometric models

By

Araar Abdelkrim

Introduction

- Limited dependent variables have a restricted range, such as the wage or salary income of non-self-employed individuals, which runs from 0 to the highest level recorded.

- Limited dependent variables cannot be modeled by linear regression. These models require more computational effort to fit and are harder to interpret.

Truncated samples

- Some LDVS are generated by truncated processes.
- For truncation, the sample is drawn from a subset of the population so that only certain values are included in the sample.
- We lack observations on both the response variable and explanatory variables.
- For instance, we might have a sample of individuals who have a high school diploma, some college experience, or one or more college degrees.

Truncated samples

- The sample has been generated by interviewing those who completed high school. This is a truncated sample, relative to the population, in that it excludes all individuals who have not completed high school.

- The excluded individuals are not likely to have the same characteristics as those in our sample. For instance, we might expect average or median income of dropouts to be lower than that of graduates.

Truncated samples

- The effect of truncating the distribution of a random variable are:
  - The expected value or mean of the truncated random variable moves away from the truncation point;
  - the variance is reduced.

- We cannot use a sample from this truncated population to make inferences about the entire population without correcting for those excluded individuals' not being randomly selected from the population at large.

- Although it might appear that we could use these A truncated data to make inferences about the subpopulation, we cannot even do that.

Truncated random variable and moments

If $x \sim N[\mu, \sigma^2]$ *and* a *is a constant, then*

$$E[x \mid \text{truncation}] = \mu + \sigma\lambda(\alpha),$$

$$\text{Var}[x \mid \text{truncation}] = \sigma^2[1 - \delta(\alpha)],$$

where $\alpha = (a - \mu)/\sigma$, $\phi(\alpha)$ *is the standard normal density and*

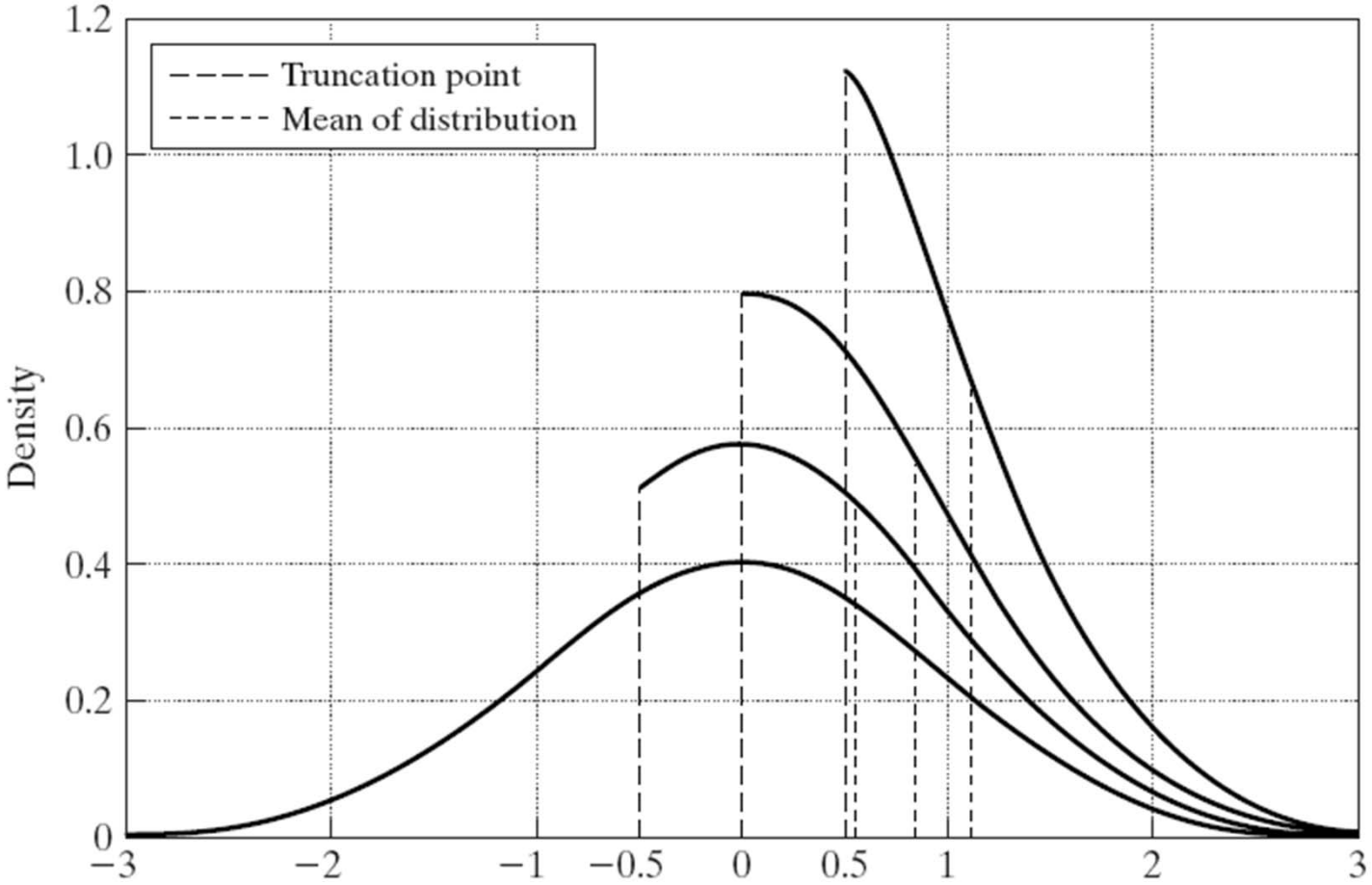
$$\lambda(\alpha) = \phi(\alpha)/[1 - \mathbf{\Phi}(\alpha)] \quad \text{if truncation is } x > a,$$

$$\lambda(\alpha) = -\phi(\alpha)/\mathbf{\Phi}(\alpha) \quad \text{if truncation is } x < a,$$

and

$$\delta(\alpha) = \lambda(\alpha)[\lambda(\alpha) - \alpha].$$

Truncated random variable and moments



Truncated samples

Assume that $y = x_i\beta + u_i$ is observed if only it exceed τ.

Let:

$$\alpha_i = (\tau - x_i\beta) / \sigma_u \qquad (1.01)$$

and

$$\lambda(\alpha_i) = \phi(\alpha_i) / \left(1 - \Phi(\alpha_i)\right) \qquad (1.02)$$

where

$\phi(\alpha_i)$ is the density function

$\Phi(\alpha_i)$ is the cumulative density function

$\lambda(\alpha_i)$ is the Inverse Mills Ratio

Truncated samples and the appropriate econometric model

Standard manipulation of normally distributed random variables shows that:

$$E\left[y_i > \tau, x_i\right] = x_i\beta + \sigma_u\lambda(\alpha_i) + u_i \qquad (1.03)$$

- The above equation implies that a simple OLS regression of y on x suffers from the exclusion of the term $\lambda(\alpha_i)$.
- This regression is mis-specified, and the effect of that mis-specification will differ across observations, with a heteroskedastic error term whose variance depends on xi.
- To deal with these problems, we include the IMR as an additional regressor, so we can use a truncated sample to make consistent inferences about the subpopulation.

Truncated samples and the appropriate econometric model

The marginal effects in this model in the subpopulation can be obtained by writing

$$\partial E\left[y_i > \tau, x_i\right] / \partial x_i = \beta(1-\delta_i) \qquad (1.04)$$

$(1-\delta_i)$	**is the truncated variance which is between 0 and 1**
$\beta(1-\delta_i)$	**is the marginal effect at subpopulation level**
β	**is the marginal effect at population level**

Truncated samples and estimations with Stata

- If we assume that the regression errors in the population are normally distributed, we can estimate an equation for a truncated sample with the Stata command truncreg.

- The truncreg option ll(#) indicates that values of the response variable less than or equal to # are truncated. Similarly, the upper truncation can be handled with the ul(#) option

Truncated samples and estimations with Stata

Example

Consider a sample of married women from the laborsub dataset whose:

- hours of work (whrs) are truncated from below at zero.

The other variables of interest are:

- the number of preschool children (k16);
- The number of school-aged children (k618);
- The age (wa);
- The number of years of education (we).

Truncated samples and estimations with Stata

```
. use data/laborsub, clear

. sum
```

Variable	Obs	Mean	Std. Dev.	Min	Max
lfp	250	.6	.4908807	0	1
whrs	250	799.84	915.6035	0	4950
k16	250	.236	.5112234	0	3
k618	250	1.364	1.370774	0	8
wa	250	42.92	8.426483	30	60
we	250	12.352	2.164912	5	17

```
. count if whrs == 0
  100
```

Truncated samples and estimations with Stata

To illustrate the consequences of ignoring truncation, we fit a model of hours worked with OLS, including only working women.

```
. regress whrs kl6 k618 wa we if whrs>0
```

Source	SS	df	MS
Model	7326995.15	4	1831748.79
Residual	94793104.2	145	653745.546
Total	102120099	149	685369.794

Number of obs	= 150
F(4, 145)	= 2.80
Prob > F	= 0.0281
R-squared	= 0.0717
Adj R-squared	= 0.0461
Root MSE	= 808.55

whrs	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
kl6	-421.4822	167.9734	-2.51	0.013	-753.4748	-89.48953
k618	-104.4571	54.18616	-1.93	0.056	-211.5538	2.639668
wa	-4.784917	9.690502	-0.49	0.622	-23.9378	14.36797
we	9.353195	31.23793	0.30	0.765	-52.38731	71.0937
_cons	1629.817	615.1301	2.65	0.009	414.0371	2845.597

Truncated samples and estimations with Stata

now refit the model with truncreg, taking into account that 100 of the 250 observations have zero recorded whrs.

```
. truncreg whrs kl6 k618 wa we, ll(0) nolog
(note: 100 obs. truncated)

Truncated regression
Limit:   lower =          0                      Number of obs =     150
         upper =       +inf                      Wald chi2(4)  =   10.05
Log likelihood = -1200.9157                      Prob > chi2   =  0.0395
```

whrs	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
kl6	-803.0042	321.3614	-2.50	0.012	-1432.861	-173.1474
k618	-172.875	88.72898	-1.95	0.051	-346.7806	1.030578
wa	-8.821123	14.36848	-0.61	0.539	-36.98283	19.34059
we	16.52873	46.50375	0.36	0.722	-74.61695	107.6744
_cons	1586.26	912.355	1.74	0.082	-201.9233	3374.442
/sigma	983.7262	94.44303	10.42	0.000	798.6213	1168.831

Truncated samples and estimations with Stata

- Whether truncated regression is more appropriate than the ordinary least-squares estimation depends on the purpose of that estimation.

- If we are interested in the mean of wife's working hours conditional on the subsample of market laborers, least-squares estimation is appropriate.

- However if we are interested in the mean of wife's working hours regardless of market or nonmarket labor status, least-squares estimates could be seriously misleading.

Truncated samples and estimations with Stata

- Some of the attenuated coefficient estimates from regress are no more than half as large as their counterparts from truncreg.

- The parameter sigma -cons, comparable to Root MSE in the OLS regression, is considerably larger in the truncated regression, reflecting its downward bias in a truncated sample.

- We can use the coefficient estimates and marginal effects from truncreg to make inferences about the entire population.

Censored dependant variables

- Censoring is another common mechanism that restricts the range of dependent variables. Censoring occurs when a response variable is set to an arbitrary value when the variable is beyond the censoring point.

- In the truncated case, we observe neither the dependent nor the explanatory variables for individuals whose y_i lies in the truncation region.

- In contrast, when the data are censored we do not observe the value of the dependent variable for individuals whose y_i is beyond the censoring point, but we do observe the values of the explanatory variables.

Censored dependant variables

- A solution to the problem with censoring at 0 was first proposed by Tobin (1958) as the censored regression model; it became known as "Tobin's probit" or the tobit model The model can be expressed in terms of a latent variable:

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

$$y_i = \begin{cases} 0 & if \quad y_i^* \leq 0 \\ y_i^* & if \quad y_i^* > 0 \end{cases} \qquad (1.05)$$

- The model combines aspects of the binomial Probit for the distinction of y = 0 versus y_i > 0 and the regression model for E[y_i | y_i>0, x_i].

Censored dependant variables

- Of course, we could collapse all positive observations on yi and treat this as a binomial probit (or logit) estimation problem, but doing so would discard the information on the dollar amounts spent by purchasers.

- Likewise, we could throw away the y_i = 0 observations, but we would then be left with a truncated distribution, with the various problems that creates.

- To take account of all the information in yi properly, we must fit the model with the **tobit** estimation method, which uses maximum likelihood to combine the probit and regression components of the log-likelihood function.

$$\ln L = \sum_{y_i>0} -\frac{1}{2}\left[\log(2\pi) + \ln\sigma^2 + \frac{(y_i - \mathbf{x}_i'\beta)^2}{\sigma^2}\right] + \sum_{y_i=0} \ln\left[1 - \Phi\left(\frac{\mathbf{x}_i'\beta}{\sigma}\right)\right]$$

Censored models and estimations with Stata

- We can define tobit models with a threshold other than zero. We can specify censoring from below at any point on the y scale with the ll (#) option for left censoring.

- Similarly, the standard tobit formulation may use an upper threshold (censoring from above, or right censoring) using the ul(#) option to specify the upper limit. Stata's

- t o b i t command also supports the two-limit tobit model where observations on y are censored from both left and right by specifying both the ll (#) and ul(#) options.

Censored models and estimations with Stata

The marginal effect with censored models is:

$$\frac{\partial E[y \mid x]}{\partial x_j} = \beta_j \times Pr(ll < y < ul) \qquad (1.06)$$

A change in x_j has two effects: It affects the conditional mean of y∗_i in the positive part of the distribution, and it affects the probability that the observation will fall in that part of the distribution.

Censored models and estimations with Stata

Example

The following example uses a modified version of the womenwk dataset, which contains information on 2,000 women, 657 of which are not recorded as wage earners. The indicator variable work is set to zero for the nonworking and to one for those reporting positive wages:

```
. use data/womenwk, replace

. summarize work age married children education
```

Variable	Obs	Mean	Std. Dev.	Min	Max
work	2000	.6715	.4697852	0	1
age	2000	36.208	8.28656	20	59
married	2000	.6705	.4701492	0	1
children	2000	1.6445	1.398963	0	5
education	2000	13.084	3.045912	10	20

Censored models and estimations with Stata

We generate the log of the wage (lw) for working women and set lwf equal to lw for working women and zero for nonworking women. We first fit the model with OLS, ignoring the censored nature of the response variable:

```
. regress lwf age married children education
```

Source	SS	df	MS
Model	937.873188	4	234.468297
Residual	3485.34135	1995	1.74703827
Total	4423.21454	1999	2.21271363

Number of obs	= 2000
F(4, 1995)	= 134.21
Prob > F	= 0.0000
R-squared	= 0.2120
Adj R-squared	= 0.2105
Root MSE	= 1.3218

lwf	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0363624	.003862	9.42	0.000	.0287885	.0439362
married	.3188214	.0690834	4.62	0.000	.1833381	.4543046
children	.3305009	.0213143	15.51	0.000	.2887004	.3723015
education	.0843345	.0102295	8.24	0.000	.0642729	.1043961
_cons	-1.077738	.1703218	-6.33	0.000	-1.411765	-.7437105

Censored models and estimations with Stata

Refitting the model as a tobit and indicating that lwf is left censored at zero with the ll () option yields

```
. tobit lwf age married children education, ll(0)

Tobit regression                                  Number of obs     =       2000
                                                  LR chi2(4)        =     461.85
                                                  Prob > chi2       =     0.0000
Log likelihood = -3349.9685                       Pseudo R2         =     0.0645

------------------------------------------------------------------------------
         lwf |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .052157   .0057457     9.08   0.000     .0408888    .0634252
     married |   .4841801   .1035188     4.68   0.000     .2811639    .6871964
    children |   .4860021   .0317054    15.33   0.000     .4238229    .5481812
   education |   .1149492   .0150913     7.62   0.000     .0853529    .1445454
       _cons |  -2.807696   .2632565   -10.67   0.000    -3.323982   -2.291409
-------------+----------------------------------------------------------------
       /sigma |  1.872811    .040014                      1.794337    1.951285
------------------------------------------------------------------------------
  Obs. summary:        657  left-censored observations at lwf<=0
                      1343     uncensored observations
                         0 right-censored observations
```

Censored models and estimations with Stata

- The tobit estimates of lwf show positive, significant effects for age, marital status, the number of children, and the number of years of education.

- We expect each of these factors to increase the probability that a woman will work as well as increase her wage conditional on employment status.

Censored models and estimations with Stata

Following tobit estimation, we first generate the marginal effects of each explanatory variable on the probability that an individual will have a positive log(wage) by using the pr(a, b) option of predict.

```
. mfx compute, predict (pr (0, . ) )

Marginal effects after tobit
      y  = Pr(lwf>0) (predict, pr (0, . ))
         =  .81920975
```

variable	dy/dx	Std. Err.	z	P>|z|	[95% C.I.]		X
age	.0073278	.00083	8.84	0.000	.005703	.008952	36.208
married*	.0706994	.01576	4.48	0.000	.039803	.101596	.6705
children	.0682813	.00479	14.26	0.000	.058899	.077663	1.6445
educat~n	.0161499	.00216	7.48	0.000	.011918	.020382	13.084

(*) dy/dx is for discrete change of dummy variable from 0 to 1

Censored models and estimations with Stata

We then calculate the marginal effect of each explanatory variable on the expected log wage, given that the individual has not been censored (i.e., was working).

```
. mfx compute, predict(e(0,.))

Marginal effects after tobit
      y  = E(lwf|lwf>0) (predict, e(0,.))
         =  2.3102021
```

variable	dy/dx	Std. Err.	z	P>|z|	[95% C.I.]		X
age	.0314922	.00347	9.08	0.000	.024695	.03829	36.208
married*	.2861047	.05982	4.78	0.000	.168855	.403354	.6705
children	.2934463	.01908	15.38	0.000	.256041	.330852	1.6445
educat~n	.0694059	.00912	7.61	0.000	.051531	.087281	13.084

(*) dy/dx is for discrete change of dummy variable from 0 to 1

Incidental truncation and sample-selection models

For incidental truncation, the sample is representative of the entire population, but the observations on the dependent variable are truncated according to a rule whose errors are correlated with the errors from the equation of interest. We do not observe y because of the outcome of some other variable, which generates the selection indicator, s. Formally let

$$s_i y_i = s_i x_i \beta + s_i u_i \qquad (1.07)$$

- If, s_i is set randomly, the OLS is BLUE estimator;
- If si is set by a rule, such as s_i = 1 if y_i < c, we have a truncated model.

Incidental truncation and sample-selection models

Incidental truncation means that we observe y_i based not on its value but rather on the observed outcome of another variable. For instance, we observe hourly wage when an individual participates in the labor force. We can imagine fitting a binomial probit or logit model that predicts the individual's probability of participation. In this circumstance, s_i is set to zero or one based on the factors underlying that decision:

$$y_i = x_i\beta + u_i$$
$$s_i = I\left[z_i\gamma + v_i\right] \qquad (1.08)$$

- z contains all x but must also contain more factors that do not appear in x;

Incidental truncation and sample-selection models

Incidental truncation arises when there is a nonzero correlation between u and v. If both these processes are normally distributed with zero means, the conditional expectation E[u|v] = ρv, where ρ is the correlation of u and v.

$$E\left[y \mid z, s=1\right] = x\beta + \rho\lambda(z\gamma) \qquad (1.09)$$

- If ρ ≠ 0, OLS estimates from the incidentally truncated sample will not consistently estimate ρ unless the IMR term is included.

Incidental truncation and sample-selection models

The IMR term includes the unknown population parameters y, which may be fitted by a binomial probit model $\Pr(s = 1 | z) = \Phi(z\gamma)$ from the entire sample.

With estimates of y, we can compute the IMR term for eachobservation for which y_i is observed (s_i = 1) and fit the model.

This two-step procedure, based on the work of Heckman (1976), is often termed the Heckit model. Instead, we can use a full maximum-likelihood procedure to jointly estimate all parameters.

Incidental truncation and sample-selection models and Stata

- Stata's heckman command fits the full maximum-likelihood version of the Heckit model with the following syntax:

heckman depvar [indepvars] [i f] [in] , select (varlist2)

- We should code the depvar as missing (.) for those observations that are not selected.

- The heckman command can also generate the two-step estimator of the selection model (Heckman 1979) if we specify the twostep option.

Incidental truncation and sample-selection models

- In a wage equation, the number of preschool children in the family is likely to influence whether a woman participates in the labor force but might be omitted from the wage determination equation: it appears in z but not x.

- We assume also that marital status affects selection (whether a woman is observed in the labor force) but does not enter the log(wage) equation. All factors in both the log(wage) and selection equations are significant.

Incidental truncation and the estimation with Stata

```
. heckman lw education age children, select(age married children education) nolog

Heckman selection model                         Number of obs      =      2000
(regression model with sample selection)        Censored obs       =       657
                                                Uncensored obs     =      1343

                                                Wald chi2(3)       =    454.78
Log likelihood = -1052.857                      Prob > chi2        =    0.0000
```

lw	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lw						
education	.0397189	.0024525	16.20	0.000	.0349121	.0445256
age	.0075872	.0009748	7.78	0.000	.0056767	.0094977
children	-.0180477	.0064544	-2.80	0.005	-.0306981	-.0053973
_cons	2.305499	.0653024	35.30	0.000	2.177509	2.43349
select						
age	.0350233	.0042344	8.27	0.000	.0267241	.0433225
married	.4547724	.0735876	6.18	0.000	.3105434	.5990014
children	.4538372	.0288398	15.74	0.000	.3973122	.5103621
education	.0565136	.0110025	5.14	0.000	.0349492	.0780781
_cons	-2.478055	.1927823	-12.85	0.000	-2.855901	-2.100208
/athrho	.3377674	.1152251	2.93	0.003	.1119304	.5636045
/lnsigma	-1.375543	.0246873	-55.72	0.000	-1.423929	-1.327156
rho	.3254828	.1030183			.1114653	.5106469
sigma	.2527024	.0062385			.2407662	.2652304
lambda	.0822503	.0273475			.0286501	.1358505

LR test of indep. eqns. (rho = 0): chi2(1) = 5.53 Prob > chi2 = 0.0187

Incidental truncation and the estimation with Stata

By using the selection model, we have relaxed the assumption that the factors determining participation and the wage are identical and of the same sign. The effect of more children increases the probability of selection (participation) but decreases the predicted wage, conditional on participation.

The likelihood-ratio test for $\rho = 0$ rejects its null, so that estimation of the logtwage) equation without taking selection into account would yield inconsistent results.

The output produces an estimate of /athrho, the hyperbolic arctangent of ρ. That parameter is entered in the log-likelihood function to enforce the constraint that $-1 < \rho < 1$. The point and interval estimates of ρ are derived from the inverse transformation.

Incidental truncation and sample-selection models

```
cap drop s
gen s=lw!=.
qui probit s age married children education
predict xb, xb
cap drop lambda
gen lambda = normalden(xb)/normal(xb)

heckman lw education age children, select (age married children education) twostep

regress lw education age children lambda
```