5A: Censored and truncated data

© A. Colin Cameron
U. of Calif. - Davis

.

OeNB Summer School 2010
Microeconometrics
Oesterreichische Nationalbank (OeNB), Vienna, Austria

Based on
A. Colin Cameron and Pravin K. Trivedi,
Microeconometrics: Methods and Applications (MMA), ch.16.
Microeconometrics using Stata (MUS), ch.16.
Data examples are from MUS.

Aug 30 - Sep 3, 2010

1. Introduction

- Censored Data:
  - For part of the range of y we observe only that y is in that range, rather than observing the exact value of y.
    - ★ e.g. Expenditures or hours worked bunched at 0 (censored from below).
    - ★ e.g. Annual income top-coded at $75,000 (censored from above).
- Truncated data:
  - For part of range of y we do not observe y at all.
    - ★ e.g. Those with expenditures of $0 are not observed.
    - ★ e.g. Sample excludes those with annual income $>$ $75,000 per year.

- Censored and truncated regression models
  - considerably more difficult conceptually than many other models.
  - sample is not reflective of the population (selection on y)
    - whereas more common selection on $\mathbf{x}$ (exogenous stratification) okay
  - standard solutions rely on strong distributional assumptions.
- Focus on Tobit models
  - linear models with normal errors that are censored or truncated
- Issues carry over to censoring for other types of data
  - censored counts, censored durations, ...
- And also building block for more general selection models
  - sample selection model
  - Roy model

Outline

2. Tobit Example with Simulated Data

- Latent variable y^*, generated by model

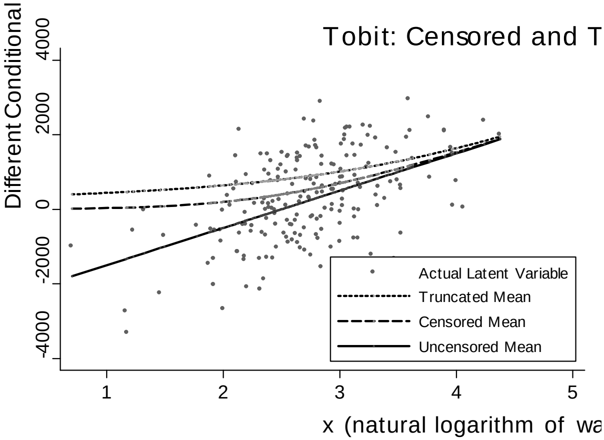
$$\begin{aligned} y_i^* &= -2500 + 1000x_i + \varepsilon_i, \quad i = 1, ..., 250, \\ \varepsilon_i &\sim \mathcal{N}[0, 1000^2], \end{aligned}$$

  - and $x_i \sim \mathcal{N}[2.75, 0.6^2]$.
  - e.g. y : annual hours worked and x : log hourly wage $(w_i \sim [18.7, 12.3^2])$.
- Complication: y^* is not fully observed.
- Censored Tobit model: We observe y_i where

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

  - Here if desired hours are negative people do not work and $y = 0$.
- Truncated Tobit: We observe only $y_i = y_i^*$ if $y_i^* > 0$.

- Scatterplot & true regression curves (derived later) for three samples:
  - truncated (top), censored (middle) and completely observed (bottom).



- Censored and truncated data the model is now nonlinear
  - and linear model will be flatter line than true line $(\widehat{\beta} \simeq 0.5\beta)$.

3. Tobit Model Definition

- Latent dependent variable y^* follows regular linear regression

$$\begin{aligned} y^* &= \mathbf{x}'\boldsymbol{\beta} + \varepsilon \\ \varepsilon &\sim \mathcal{N}[0, \sigma^2] \end{aligned}$$

  - But this latent variable is only partially observed.

- Censored regression (from below at 0): we observe

$$y = \begin{cases} y^* & \text{if } y^* > 0 \\ 0 & \text{if } y^* \leq 0. \end{cases}$$

- Truncated regression (from below at 0): we observe only

$$y = y^* \quad \text{if } y^* > 0.$$

Tobit Model Overview

- Consistency of the MLE (and other results such as prediction and marginal effects) requires
  - constant censoring point (here 0)
  - model errors to be normal
  - model errors to be homoskedastic.
- The Tobit model is therefore often too restrictive in practice.
- But it is the key building block for other models so analyze in detail.
- We consider
  - Censored and truncated means
  - ML Estimation
  - Prediction
  - Marginal effects

4. Tobit Model: Truncated Mean

- Truncated mean: We observe y only when $y > 0$.
- The truncated conditional mean (suppressing conditioning on $\mathbf{x}$) is

$$
\begin{array}{ll}
\mathrm{E}[y^*|y^* > 0] & \\
= \mathrm{E}\left[\mathbf{x}'\boldsymbol{\beta} + \varepsilon | \mathbf{x}'\boldsymbol{\beta} + \varepsilon > 0\right] & \text{as } y^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon \\
= \mathbf{x}'\boldsymbol{\beta} + \mathrm{E}\left[\varepsilon | \varepsilon > -\mathbf{x}'\boldsymbol{\beta}\right] & \text{as } \mathbf{x} \text{ and } \varepsilon \text{ independent} \\
= \mathbf{x}'\boldsymbol{\beta} + \sigma \mathrm{E}\left[\frac{\varepsilon}{\sigma} | \frac{\varepsilon}{\sigma} > \frac{-\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right] & \text{transform to } \varepsilon/\sigma \sim \mathcal{N}[0,1] \\
= \mathbf{x}'\boldsymbol{\beta} + \sigma\lambda\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right) & \text{using next slide: key result for } \mathcal{N}[0,1].
\end{array}
$$

  - where $\lambda(z) = \phi(z)/\Phi(z)$ is called the inverse Mills ratio.
- The regression function is not just $\mathbf{x}'\boldsymbol{\beta}$ (and is nonlinear).
  - OLS of y on $\mathbf{x}$ is inconsistent for $\boldsymbol{\beta}$
  - Need NLS or MLE for consistent estimates.

- Proof: Truncated mean $\mathsf{E}[z|z>c]$ for the standard normal
  - key result used in the previous slide
  - consider $z \sim \mathcal{N}[0,1]$, with density $\phi(z)$ and c.d.f. $\Phi(z)$.
  - conditional density of $z|z>c$ is $\phi(z)/(1-\Phi(c))$.
  - truncated conditional mean is

$$
\begin{aligned}
\mathsf{E}[z|z>c] &= \int_c^\infty z\,(\phi(z)/(1-\Phi(c)))\;dz \\
&= \int_c^\infty z\frac{1}{\sqrt{2\pi}}\exp(-\tfrac{1}{2}z^2)\;dz \Big/ (1-\Phi(c)) \\
&= \left[-\frac{1}{\sqrt{2\pi}}\exp(-\tfrac{1}{2}z^2)\right]_c^\infty \Big/ (1-\Phi(c)) \\
&= \frac{\phi(c)}{1-\Phi(c)} \\
&= \frac{\phi(-c)}{\Phi(-c)} \\
&= \lambda(-c),\ \text{where } \lambda(c)=\phi(c)/\Phi(c).
\end{aligned}
$$

Tobit Model: Censored Mean

- Censored mean: We observe $y = 0$ if $y^* < 0$ and $y = y^*$otherwise.
- The censored conditional mean (suppressing conditioning on $\mathbf{x}$) is

$$\begin{aligned} \mathrm{E}[y] &= \mathrm{E}_{y^*}[\mathrm{E}[y|y^*]] \\ &= \Pr[y^* \leq 0] \times 0 + \Pr[y^* > 0] \times \mathrm{E}[y^*|y^* > 0] \\ &= 0 + \Phi(\mathbf{x}'\boldsymbol{\beta}/\sigma)\left\{\mathbf{x}'\boldsymbol{\beta} + \sigma\frac{\phi\left(\mathbf{x}'\boldsymbol{\beta}/\sigma\right)}{\Phi\left(\mathbf{x}'\boldsymbol{\beta}/\sigma\right)}\right\} \\ \mathrm{E}[y|\mathbf{x}] &= \Phi(\mathbf{x}'\boldsymbol{\beta}/\sigma)\mathbf{x}'\boldsymbol{\beta} + \sigma\phi\left(\mathbf{x}'\boldsymbol{\beta}/\sigma\right), \end{aligned}$$

  using earlier result for the truncated mean $\mathrm{E}[y^*|y^* > 0]$.
- This conditional mean is again nonlinear.
  - OLS of y on $\mathbf{x}$ is inconsistent for $\boldsymbol{\beta}$
  - Need NLS or MLE for consistent estimates.

5. Tobit Model: Censored MLE

- Density varies according to whether $y > 0$ or $y = 0$.
- Positives: for $y > 0$ we observe $y \sim \mathcal{N}[\mathbf{x}'\boldsymbol{\beta}, \sigma^2]$.

$$\begin{aligned} f(y) &= f^*(y) \\ &= \left(1/\sqrt{2\pi\sigma^2}\right) \times \exp\left(-(y - \mathbf{x}'\boldsymbol{\beta})^2/2\sigma^2\right) \\ &= \frac{1}{\sigma}\,\phi\left(\frac{y - \mathbf{x}'\boldsymbol{\beta}}{\sigma}\right) \text{ where } \phi(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}. \end{aligned}$$

- Zeroes: for $y = 0$ we observe only that $y^* \leq 0$.

$$\begin{aligned} f(0) &= \Pr[y = 0] = \Pr[y^* \leq 0] \\ &= \Pr[\mathbf{x}'\boldsymbol{\beta} + \varepsilon \leq 0] \\ &= \Pr[\varepsilon/\sigma \leq -\mathbf{x}'\boldsymbol{\beta}/\sigma] = \Phi\left(\frac{-\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right), \end{aligned}$$

- Now combine positives and zeroes.
- Introduce indicator:

$$d = \begin{cases} 1 & \text{if } y > 0 \\ 0 & \text{if } y = 0. \end{cases}$$

- Censored Tobit density:

$$f(y) = \left[\frac{1}{\sigma} \phi \left(\frac{y - \mathbf{x}'\boldsymbol{\beta}}{\sigma} \right) \right]^{d} \times \left[\Phi \left(\frac{-\mathbf{x}'\boldsymbol{\beta}}{\sigma} \right) \right]^{1-d} .$$

- Log-likelihood function for censored Tobit:

$$\ln L(\boldsymbol{\beta}, \sigma^2) = \sum_{i=1}^{N} \left\{ d_i \ln \frac{1}{\sigma} \phi \left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma} \right) + (1 - d_i) \ln \Phi \left(-\mathbf{x}_i'\boldsymbol{\beta}/\sigma \right) \right\} .$$

- MLE maximizes this with respect to $\boldsymbol{\beta}$ and σ^2.

Truncated MLE

- For left truncated at 0 the density is

$$\begin{aligned} f(y) &= f^*(y^*|y^*>0) \\ &= f^*(y)/\Pr[y^*>0] \\ &= \left(1/\sqrt{2\pi\sigma^2}\right)\times\exp\left(-(y-\mathbf{x}'\boldsymbol{\beta})^2/2\sigma^2\right)/\Phi\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right) \\ &= \frac{1}{\sigma}\phi\left(\frac{y-\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right)/\Phi\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right). \end{aligned}$$

- Log-likelihood function for truncated Tobit:

$$\ln L(\boldsymbol{\beta},\sigma^2)=\sum_{i=1}^{N}\left\{\ln\frac{1}{\sigma}\phi\left(\frac{y_i-\mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)-\ln\Phi\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right\}.$$

6. Tobit MLE: Data Example

- Data from 2001 Medical Expenditure Panel Survey (MUS chapter 16).
  - ambexp (ambulatory expenditure = physician and hospital outpatient).
  - dambexp (=1 if ambexp>0 and =0 if ambexp=0).
  - Regressors: age (in tens of years), female, educ (years of completed schooling), blhisp (=1 if black or hispanic) , totchr (number of chronic conditions), and ins (=1 if PPO or HMO health insurance).

Variable	Obs	Mean	Std. Dev.	Min	Max
ambexp	3328	1386.519	2530.406	0	49960
dambexp	3328	.8419471	.3648454	0	1
age	3328	4.056881	1.121212	2.1	6.4
female	3328	.5084135	.5000043	0	1
educ	3328	13.40565	2.574199	0	17
blhisp	3328	.3085938	.4619824	0	1
totchr	3328	.4831731	.7720426	0	5
ins	3328	.3650841	.4815261	0	1

- 16% of sample are censored (since dambexp has mean 0.84).

Censored MLE

- Stata command `tobit, ll(0)` yields

```
. * Tobit on censored data
. tobit ambexp age female educ blhisp totchr ins, ll(0)

Tobit regression                                  Number of obs   =       3328
                                                  LR chi2(6)      =     694.07
                                                  Prob > chi2     =     0.0000
Log likelihood = -26359.424                       Pseudo R2       =     0.0130

------------------------------------------------------------------------------
      ambexp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   314.1479   42.63358     7.37   0.000     230.5572    397.7387
      female |   684.9918   92.85445     7.38   0.000     502.9341    867.0495
        educ |    70.8656   18.57361     3.82   0.000     34.44873    107.2825
      blhisp |   -530.311   104.2667    -5.09   0.000    -734.7443   -325.8776
      totchr |   1244.578   60.51364    20.57   0.000      1125.93    1363.226
         ins |  -167.4714   96.46068    -1.74   0.083    -356.5998    21.65696
       _cons |  -1882.591   317.4299    -5.93   0.000    -2504.969   -1260.214
-------------+----------------------------------------------------------------
      /sigma |   2575.907   34.79296                      2507.689    2644.125
------------------------------------------------------------------------------
  Obs. summary:        526  left-censored observations at ambexp<=0
                      2802     uncensored observations
                         0 right-censored observations
```

- The OLS coefficients were 250, 374, 33, -310, 1076 and -237.

Marginal Effects

- Question: How do we interpret the coefficients?
- Marginal effects for uncensored mean:

$$\frac{\partial \mathrm{E}[y^*|\mathbf{x}]}{\partial x_j} = \beta_j.$$

  - cannot always interpret this - what is y^*?
- MEs for censored mean (after some algebra):

$$\frac{\partial \mathrm{E}[y|\mathbf{x}]}{\partial x_j} = \Phi(\mathbf{x}'\boldsymbol{\beta}/\sigma)\beta_j.$$

  - this is of more interest
  - the AME approximately equals β_j times the fraction uncensored
  - AME in Stata 11 use `margins, dydx(*) predict(ystar(0,.))`
  - MEM in Stata 10 use `mfx, predict(ystar(0,.))`
  - can decompose into ME at corner and ME at interior

Tobit Marginal Effects: Example for Censored Mean

- Marginal effect for censored mean $E[y|\mathbf{x}]$ evaluated at $\mathbf{x} = \bar{\mathbf{x}}$ (MEM).

```
. * Marginal effects for censored conditional mean evaluated at x = xbar
. mfx compute, predict(ystar(0,.))

Marginal effects after tobit
      y  = E(ambexp*|ambexp>0) (predict, ystar(0,.))
         =  1647.8507
```

variable	dy/dx	Std. Err.	z	P>|z|	[95% C.I.	]	X
age	207.526	28.205	7.36	0.000	152.245	262.807	4.05688
female*	451.6399	61.029	7.40	0.000	332.026	571.254	.508413
educ	46.81378	12.265	3.82	0.000	22.7739	70.8537	13.4056
blhisp*	-342.4803	65.756	-5.21	0.000	-471.361	-213.6	.308594
totchr	822.1678	40.61	20.25	0.000	742.573	901.763	.483173
ins*	-110.0883	63.117	-1.74	0.081	-233.795	13.6185	.365084

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

- The marginal effects are approximately 65% of the estimated coefficients (314, 684, 70, -530, 1244 and -167).
- The OLS coefficients were 250, 374, 33, -310, 1076 and -237.

7. Tobit Extensions

- We focused on Tobit
  - Linear model with normal errors and left-censored t zero.
    - e.g. annual hours worked or annual expenditure on automobiles.
- Extensions include
  - Censoring from above
    - e.g. top-coded income
  - Interval censoring
    - e.g. income reported in ranges
  - Semiparametric methods
    - relax distributional assumptions
  - Nonnormal and nonlinear models
    - e.g. number of doctor visits top-coded
  - Two part model and richer models with selection
    - different (possibly correlated) processes for censoring and outcome.

Top-coded and interval Censored Data

- Top-coded example: if $y > 100000$ only observe this fact
  - straightforward adaptation of previous Tobit MLE
  - Stata command `tobit y x, ul(100000)`
- Interval-censored data example: observe annual income in intervals of \$10,000's
  - $y \leq 0$, $0 < y \leq 10000$, ..., $90000 < y \leq 100000$, $y > 100000$.
  - contribution to the likelihood is probability of being in each interval
  - e.g. $\Pr[90000 < y^* \leq 100000]$ where $y^* \sim \mathcal{N}[\mathbf{x}'\boldsymbol{\beta}, \sigma^2]$.
  - Stata command `intreg`.

Tobit in logs

- Tobit is often applied to right-skewed data, e.g. income or expenditure
  - these are closer to lognormal than normal
  - so should do Tobit model in logs
  - most people do not do this.
- For lognormal y^* we specify

$$\begin{aligned} y^* &= \exp(\mathbf{x}'\boldsymbol{\beta} + \varepsilon) \\ \varepsilon &\sim \mathcal{N}[0, \sigma^2]. \end{aligned}$$

- We observe

$$y_i = \begin{cases} y_i^* & \text{if } \ln y_i^* > \gamma \\ 0 & \text{if } \ln y_i^* \leq \gamma. \end{cases}$$

  - The censoring point for $\ln y$ is no longer 0 but is $\gamma \neq 0$.
  - When data are censored $y = 0$ (and not $\ln \gamma$).
  - Follow Carson and Sun (2007) and let $\widehat{\gamma} = \min(\text{uncensored } \ln y^*)$.

- To implement use Stata command `tobit lny x` with option `ll(#)` where
  - the threshold # is $\widehat{\gamma}$ = the minimum uncensored value of $\ln y$ (or better $\widehat{\gamma} - \Delta$ where Δ is very small)
  - the censored observations set $\ln y$ equal to #
- The censored conditional mean in levels (not logs) is

$$\mathrm{E}[y|\mathbf{x}] = \exp(\mathbf{x}'\boldsymbol{\beta} + \frac{\sigma^2}{2})(1 - \Phi(\frac{\gamma - \mathbf{x}'\boldsymbol{\beta} - \sigma^2}{\sigma})).$$

  - This can be used to get marginal effects in levels.

Nonnormal and nonlinear models

- For count data may only observe positive counts
  - then truncated data with density $\Pr[y = k|y > 0]$
  - for Poisson $\Pr[y > 0] = 1 - \Pr[y = 0] = 1 - \exp(-\lambda)$
    so $\Pr[y = k|y > 0] = [\exp(-\lambda)\lambda^y/y!]/[1 - \exp(-\lambda)]$
  - Stata commands `ztp` and `ztnb`
- For duration data usually have right-censored data
  - e.g. length of an incomplete unemployment spell
  - can use parametric methods analogous to above
  - e.g. Stata command `streg y x, dist(Weibull)`
- For duration data more popular is semiparametric method
  - Cox proportional hazards
  - Model the conditional hazard of spell ending rather than mean duration
  - e.g. Stata command `stcox y x`.

8. Two-Part Model (here in logs)

- Consider separate models for zeroes (nonparticipant) and nonzeroes (participant with outcome observed):
  - 1. Participation: Model for $d = 1$ (participation) or $d = 0$ (nonparticipation).
  - 2. Outcome: Model for outcome y conditional on participation
    - the outcome is 0 for nonparticipants.
- Medical expenditure example
  - 1. Probit (or logit) for whether any expenditure
  - 2. Lognormal for positive expenditures.

- Separately estimate probit and lognormal models

$$\begin{array}{ll} \Pr[d=1] & = \Phi(\mathbf{x}_1'\boldsymbol{\beta}_1) \\ \ln y|d=1 & \sim \mathcal{N}[\mathbf{x}_2'\boldsymbol{\beta}_2, \sigma_2^2] \end{array}$$

- These can then be combined to predict y using

$$\begin{aligned} \mathrm{E}[y|\mathbf{x}] &= \Pr[d=0|\mathbf{x}] \times 0 + \Pr[d=1|\mathbf{x}] \times \mathrm{E}[y|\mathbf{x}, d=1] \\ &= \Phi(\mathbf{x}_1'\boldsymbol{\beta}_1) \times \exp(\mathbf{x}_2'\boldsymbol{\beta}_2 + \sigma_2^2/2). \end{aligned}$$

Two-Part Model: Data example

- First part is probit.

```
. * Two-part model
. * First part is probit
. probit dy age female educ blhisp totchr ins, nolog

Probit regression                                 Number of obs   =       3328
                                                  LR chi2(6)      =     509.53
                                                  Prob > chi2     =     0.0000
Log likelihood = -1197.6644                       Pseudo R2       =     0.1754
```

dy	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.097315	.0270155	3.60	0.000	.0443656	.1502645
female	.6442089	.0601499	10.71	0.000	.5263172	.7621006
educ	.0701674	.0113435	6.19	0.000	.0479345	.0924003
blhisp	-.3744867	.0617541	-6.06	0.000	-.4955224	-.2534509
totchr	.7935208	.0711156	11.16	0.000	.6541367	.9329048
ins	.1812415	.0625916	2.90	0.004	.0585642	.3039187
_cons	-.7177087	.1924667	-3.73	0.000	-1.094937	-.3404809

- Second part is lognormal for positives

```
. * Second part is lognormal regression for positives
. regress lny age female educ blhisp totchr ins if dy==1

      Source |       SS       df       MS              Number of obs =    2802
-------------+------------------------------           F(  6,  2795) =  110.58
       Model |  1069.37332     6  178.228887           Prob > F      =  0.0000
    Residual |  4505.06629  2795  1.61183051           R-squared     =  0.1918
-------------+------------------------------           Adj R-squared =  0.1901
       Total |  5574.43961  2801  1.99016052           Root MSE      =  1.2696

------------------------------------------------------------------------------
         lny |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .2172327   .0222225     9.78   0.000     .1736585    .2608069
      female |   .3793756   .0485772     7.81   0.000     .2841247    .4746265
        educ |   .0222388   .0097615     2.28   0.023     .0030983    .0413793
      blhisp |  -.2385321   .0551952    -4.32   0.000    -.3467597   -.1303046
      totchr |   .5618171   .0305078    18.42   0.000      .501997    .6216372
         ins |   -.020827   .0500062    -0.42   0.677    -.1188797    .0772258
       _cons |   4.907825   .1681512    29.19   0.000     4.578112    5.237538
------------------------------------------------------------------------------
```

- Now predict using $\mathrm{E}[y|\mathbf{x}] = \Phi(\mathbf{x}_1'\boldsymbol{\beta}_1) \times \exp(\mathbf{x}_2'\boldsymbol{\beta}_2 + \sigma^2 2)$.

```
. * Two-part model prediction
. quietly probit dy age female educ blhisp totchr ins

. predict dyhat, pr

. quietly regress lny age female educ blhisp totchr ins if dy==1

. predict xbpos, xb

. generate yhatpos = exp(xbpos+0.5*e(rmse)^2)

. generate yhat2step = dyhat*yhatpos

. summarize yhat2step y
```

Variable	Obs	Mean	Std. Dev.	Min	Max
yhat2step	3328	1680.978	2012.084	87.29432	40289.03
y	3328	1386.519	2530.406	0	49960

- Predicts conditional mean much better than Tobit or log Tobit.

9. Heckman Sample Selection Model: Definition

- Similar to two-part model except errors correlated across the two parts.
- Define two latent variables as follows:

$$\begin{array}{ll} \text{Participation:} & y_1^* = \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1 \\ \text{Outcome:} & y_2^* = \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon_2 \end{array}$$

- Neither y_1^* nor y_2^* are completely observed.
  - Participation: We observe whether y_1^* is positive or negative

$$y_1 = \begin{cases} 1 & \text{if } y_1^* > 0 \\ 0 & \text{if } y_1^* \leq 0. \end{cases}$$

  - Outcome: Only positive values of y_2^* are observed

$$y_2 = \begin{cases} y_2^* & \text{if } y_1^* > 0 \\ - & \text{if } y_1^* \leq 0. \end{cases}$$

- Specified the error to be joint normal (with normalization $\sigma_1^2 = 1$)

$$\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \sim \mathcal{N}\left[\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 = 1 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}\right]$$

- Then can estimate by MLE
  - Stata command `heckman y x`
- The problem is that the MLE is very sensitive to misspecification
  - inconsistent if ε is nonnormal or is heteroskedastic.
  - so use estimator that relies on weaker assumptions.

Sample Selection Model: Heckman 2-step estimator

- Assume that the errors $(\varepsilon_1, \varepsilon_2)$ satisfy

$$\varepsilon_2 = \delta \times \varepsilon_1 + v,$$

where $\varepsilon_1 \sim \mathcal{N}[0, 1]$ and v is independent of ε_1.

  - This is implied by $(\varepsilon_1, \varepsilon_2)$ joint normal.
  - But it is a weaker assumption.
  - Intuitively it is just a regression of ε_2 on ε_1.

- Then $y_2 = \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon_2$ if $y_1^* > 0$ implies

$$\begin{aligned} \mathrm{E}[y_2|y_1^* > 0] &= \mathbf{x}_2'\boldsymbol{\beta}_2 + \mathrm{E}[\varepsilon_2|\mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1 > 0] \\ &= \mathbf{x}_2'\boldsymbol{\beta}_2 + \mathrm{E}\left[(\delta \times \varepsilon_1 + v)|\varepsilon_1 > -\mathbf{x}_1'\boldsymbol{\beta}_1\right] \\ &= \mathbf{x}_2'\boldsymbol{\beta}_2 + \delta \times \mathrm{E}[\varepsilon_1|\varepsilon_1 > -\mathbf{x}_1'\boldsymbol{\beta}_1] \\ &= \mathbf{x}_2'\boldsymbol{\beta}_2 + \delta \times \lambda(\mathbf{x}_1'\boldsymbol{\beta}_1) \end{aligned}$$

  where third equality uses v independent of ε_1 and $\lambda(c) = \phi(c)/\Phi(c)$.

- For the observed outcomes:

$$\mathrm{E}[y_2|y_1^* > 0] = \mathbf{x}_2'\boldsymbol{\beta}_2 + \delta\lambda(\mathbf{x}_1'\boldsymbol{\beta}_1).$$

  - OLS of y_2 on $\mathbf{x}_2$ only is inconsistent as regressor $\lambda(\mathbf{x}_1'\boldsymbol{\beta}_1)$ is omitted.
  - Heckman included an estimate of $\lambda(\mathbf{x}_1'\boldsymbol{\beta}_1)$ as an additional regressor.

- Heckman's two-step procedure:
  - **1.** Estimate $\boldsymbol{\beta}_1$ by probit for $y_1^* > 0$ or $y_1^* < 0$ with regressors $\mathbf{x}_{1i}$.
  - Calculate $\widehat{\lambda}_i = \lambda(\mathbf{x}_{1i}'\widehat{\boldsymbol{\beta}}_1) = \phi(\mathbf{x}_{1i}'\widehat{\boldsymbol{\beta}}_1)/\Phi(\mathbf{x}_{1i}'\widehat{\boldsymbol{\beta}}_1)$.
  - **2.** For observed y_2 estimate $\boldsymbol{\beta}_2$ and δ in the OLS regression

$$y_{2i} = \mathbf{x}_{2i}'\boldsymbol{\beta}_2 + \delta\widehat{\lambda}_i + w_i.$$

  - Need standard errors that correct for w_i heteroskedastic and $\widehat{\lambda}_i$ estimated.
  - Use Stata command `heckman y x, twostep`.

- Exclusion restriction:
  - desirable to include some regressors in participation equation $(\mathbf{x}_1)$ that can be excluded from the outcome equation $(\mathbf{x}_2)$
  - otherwise identification comes solely from nonlinearity
    - furthermore $\lambda(\mathbf{x}_{1i}'\boldsymbol{\beta}_1)$ is not vary nonlinear.
- Selection on observables only
  - If $\text{Cov}[\varepsilon_1, \varepsilon_2] = 0$ model then there is no longer selection on unobservables
  - Then can use a two-part model.
- Logs for the outcome
  - Often the outcome is expenditure
  - Then better to use a log model for the outcome
  - But will then need to transform to levels for prediction.

10. Sample Selection Model: Data Example

- Selection MLE: LR test does not reject $H_0 : \rho = 0$ at level .05.

```
. * Heckman MLE without exclusion restrictions
. global xlist age female educ blhisp totchr ins

. heckman lny $xlist, select(dy = $xlist) nolog

Heckman selection model                         Number of obs      =      3328
(regression model with sample selection)        Censored obs       =       526
                                                Uncensored obs     =      2802

                                                Wald chi2(6)       =    294.42
Log likelihood = -5838.397                      Prob > chi2        =    0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lny          |
         age |   .2122921    .022958     9.25   0.000     .1672952     .257289
      female |    .349728   .0596734     5.86   0.000     .2327704    .4666856
        educ |   .0188724   .0105254     1.79   0.073    -.0017569    .0395017
      blhisp |  -.2196042   .0594788    -3.69   0.000    -.3361804   -.103028
      totchr |   .5409537   .0390624    13.85   0.000     .4643929    .6175145
         ins |  -.0295368    .051042    -0.58   0.563    -.1295772    .0705037
       _cons |   5.037418   .2261901    22.27   0.000     4.594094    5.480743
-------------+----------------------------------------------------------------
dy           |
         age |   .0984482   .0269881     3.65   0.000     .0455526    .1513439
      female |   .6436686   .0601399    10.70   0.000     .5257966    .7615407
        educ |   .0702483   .0113404     6.19   0.000     .0480216     .092475
      blhisp |  -.3726284   .0617336    -6.04   0.000    -.4936241   -.2516328
      totchr |   .7946708   .0710278    11.19   0.000     .6554588    .9338827
         ins |   .1821233   .0625485     2.91   0.004     .0595305    .3047161
       _cons |  -.7244413    .192427    -3.76   0.000    -1.101591   -.3472913
-------------+----------------------------------------------------------------
     /athrho |   -.124847   .1466391    -0.85   0.395    -.4122544    .1625604
    /lnsigma |   .2395983   .0143319    16.72   0.000     .2115084    .2676882
-------------+----------------------------------------------------------------
         rho |  -.1242024   .1443771                     -.3903852    .1611435
```

- Selection 2-step: Wald test does not reject $H_0 : \rho = 0$ at level .05.

```
.
. * Heckman 2-step without exclusion restrictions
. heckman lny $xlist, select(dy = $xlist) twostep

Heckman selection model -- two-step estimates   Number of obs      =      3328
(regression model with sample selection)        Censored obs       =       526
                                                Uncensored obs     =      2802

                                                Wald chi2(6)       =    189.46
                                                Prob > chi2        =    0.0000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lny						
age	.202124	.0242974	8.32	0.000	.1545019	.2497462
female	.2891575	.073694	3.92	0.000	.1447199	.4335951
educ	.0119928	.0116839	1.03	0.305	-.0109072	.0348928
blhisp	-.1810582	.0658522	-2.75	0.006	-.3101261	-.0519904
totchr	.4983315	.0494699	10.07	0.000	.4013724	.5952907
ins	-.0474019	.0531541	-0.89	0.373	-.151582	.0567782
_cons	5.302572	.2941363	18.03	0.000	4.726076	5.879069
dy						
age	.097315	.0270155	3.60	0.000	.0443656	.1502645
female	.6442089	.0601499	10.71	0.000	.5263172	.7621006
educ	.0701674	.0113435	6.19	0.000	.0479345	.0924003
blhisp	-.3744867	.0617541	-6.06	0.000	-.4955224	-.2534509
totchr	.7935208	.0711156	11.16	0.000	.6541367	.9329048
ins	.1812415	.0625916	2.90	0.004	.0585642	.3039187
_cons	-.7177087	.1924667	-3.73	0.000	-1.094937	-.3404809
mills						
lambda	-.4801696	.2906565	-1.65	0.099	-1.049846	.0895067
rho	-0.37130					
sigma	1.2932083					
lambda	-.4801696	.2906565				

11. Sample Selection Model: Generalizations

- Heckman two-step method relies on weaker assumptions than MLE.
  - Specifically, outcome equation error is a multiple of the participation equation error plus some noise.
  - This noise is independent of the participation decision.
- Given $\varepsilon_2 = \delta\varepsilon_1 + v$ with $v \perp \varepsilon_1$ we obtain

$$\mathrm{E}[y_2|y_1^* > 0] = \mathbf{x}_2'\boldsymbol{\beta}_2 + \delta\mathrm{E}[\varepsilon_1|\varepsilon_1 > -\mathbf{x}_1'\boldsymbol{\beta}_1].$$

- So Heckman's two-step method can be adapted to
  - distributions for ε_1 other than the normal
  - semiparametric methods that do not impose a functional form for $\mathrm{E}[\varepsilon_1|\varepsilon_1 > -\mathbf{x}_1'\boldsymbol{\beta}_1]$.
    - ★ e.g. add a polynomial in $\mathbf{x}_1'\widehat{\boldsymbol{\beta}}_1$.
- But more common is other treatment evaluation methods.

Truncated, censored and selected data: Stata commands

- Stata commands

Command	Model
`tobit`	Tobit MLE (censored)
`clad`	Censored least absolute deviations (Stata add-on)
`truncreg`	Tobit MLE (truncated)
`cnreg`	Tobit (varying known threshold)
`intreg`	Interval normal data (e.g. $1-$100, $101-$200,..)
`heckman, mle`	Sample selection MLE
`heckman, 2step`	Sample selection two step
`ztp`	Truncated MLE for Poisson counts
`ztnb`	Truncated MLE for Negative binomial counts
`streg`	Censored MLE for duration data
`stcox`	Cox proportional hazards for censored duration data

12. Some References

- The material is covered in graduate level texts including
  - CT(2005) MMA chapter 16 and CT(2009) MUS chapter 16
  - Wooldridge, J.M. (2002), *Econometric Analysis of Cross Section and Panel Data*, MIT Press.
  - Greene, W.H. (2007), *Econometric Analysis*, Prentice-Hall, Sixth edition.
- A classic book is
  - Maddala, G.S. (1986), *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge University Press.