

Lecture 11 (Revised, 2007)

Corner Solution Outcomes and
Censored Regression Models

Read { Woodridge ch 16
Greene ch 22.1-22.3
Verbeek ch 7.4-7.5

Censored Data

"Values in a certain range are transformed to a single value."

... Data problem . coding problem ; corner solution

eg) # of tickets sold for football games is censored at the max capacity of a stadium. An OLS estimation of the demand function is biased.

eg) reservation wage is (real) observed wage
... we only observe a minimum wage for low skilled workers

eg) WTP for a road construction

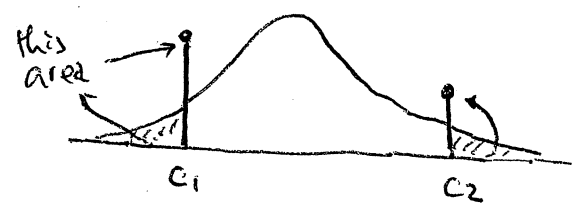
... WTP can be negative, but we observe 0 values for those with negative WTP

eg) Duration models ; time length to finish unemployment spell, but some unemployed still remain unemployed. we just observed time length of unemployment spell.

the sum:

$$P(X=C_1) + P(X=C_2)$$

$$+ \int_{C_1}^{C_2} f(x) dx = 1.$$



$$X = \max(C_1, X^*)$$
$$X = \min(C_2, X^*)$$

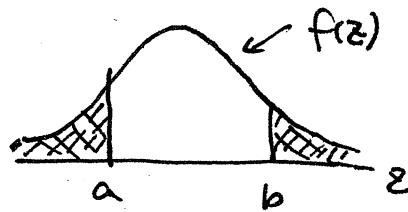
The probability at a point is zero in a continuous dist.

$P(X=C) = 0$. But, this is not the case with censoring
 $P(X \leq C_1) = P(C_1)$, $P(X \geq C_2) = P(C_2)$

Distinguish: censoring vs truncation

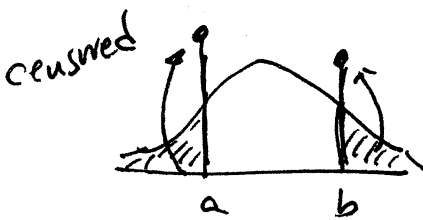
(2)

Suppose $z \sim N(\mu, \sigma^2)$



∴ Suppose that z is observed as censored ($N_1 + N_2 + N_3$ obs)

We observe a if $z \leq a$... N_1 obs
 b if $z \geq b$... N_3 obs
 z if $a < z < b$... N_2 obs



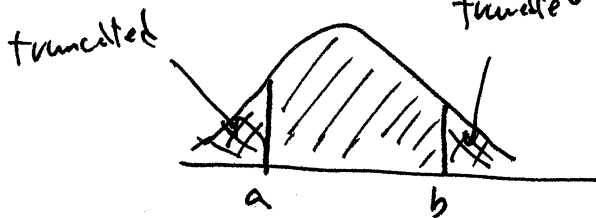
$$\int_{-\infty}^{\infty} f(z) dz = P(z=a) + \int_a^b f(z) dz + P(z=b) = 1$$

each area = height at a & b

thus $\int_a^b f(z) dz = 1 - P(z=a) - P(z=b) < 1$

∴ Suppose that z is observed only if $a < z < b$ (N_2 obs)

Only N_2 obs are observed, and others are truncated. If $z \leq a$ or $z \geq b$, they do Not belong in my sample.)



truncated if so, $f(z)$ is not valid,

since $\int_a^b f(z) dz \neq 1$.

A valid density function is

$$f^*(z) = \frac{f(z)}{\Phi(b) - \Phi(a)}$$

why? $\int_a^b f^*(z) dz = \int_a^b \frac{f(z)}{\Phi(b) - \Phi(a)} dz = \frac{1}{\Phi(b) - \Phi(a)} \int_a^b f(z) dz$
 $= \frac{\Phi(b) - \Phi(a)}{\Phi(b) - \Phi(a)} = 1$.

∴ $f^*(z) = f(z | a < z < b)$ = conditional density function

∴ Truncated regression uses $f^*(z)$ with N_2 obs.

point { Censored Models use $N_1 + N_2 + N_3$ obs.
 Truncated Models use N_2 obs, only

popular cases

$$y_i = \begin{cases} 0 & \text{if } y_i \leq 0 \\ y_i^* & \text{if } y_i > 0 \end{cases} \Leftrightarrow y_i = \max(0, y_i^*)$$

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

$y_i = 0$	N_1 obs
$y_i > 0$	N_2 obs

 $\Rightarrow D_i = \begin{cases} 0 & \text{if } y_i = 0 \\ 1 & \text{if } y_i > 0 \end{cases}$

Probit is possible.

... Type I Tobit
 censored normal (y_i^* has a normal dist)

(more on this, later)

Distinguish

a) using $N_1 + N_2$ obs with $y_i = 0$ (censored)
 \Rightarrow Tobit censored model

b) using $N_1 + N_2$ obs.
 [1st step probit and obtain $\hat{\lambda}_i$ (Inverse Mills ratio)
 2nd step regression with $\hat{\lambda}_i$ added.

i) in the 2nd step, using N_2 obs only

$$y_i = x_i \beta + c \hat{\lambda}_i + \epsilon_i$$

... Heckman's selection model

ii) in the 2nd step, using $N_1 + N_2$ obs

$$y_i = x_i \beta + d D_i + c \hat{\lambda}_i + \epsilon_i$$

... Heckman's treatment effect model.

where $D_i = \begin{cases} 0 & \text{not treated} \\ 1 & \text{treated} \end{cases}$

$\hat{\lambda}_i$ is defined differently
 for the obs with $D_i = 0$ and $D_i = 1$

more on this, later (lecture 12)

Censoring

we observe c_1 if $y \leq c_1$
 c_2 if $y \geq c_2$
 y if $c_1 < y < c_2$

c_1, c_2 known. [This assumption can be relaxed.]

problem: OLS is biased.

Tobit Model (Censored regression at zero)

$$y_i^* = \alpha + \beta'x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

$$y_i = \begin{cases} 0 & \text{if } y_i^* \leq 0 \\ y_i^* & \text{if } y_i^* > 0 \end{cases} \quad \text{simply put, } y_i = \text{Max}(0, y_i^*)$$

$$\begin{aligned} P(y_i = 0) &= P(y_i^* \leq 0) = P(\alpha + \beta'x_i + \epsilon_i \leq 0) \\ &= P(\epsilon_i \leq -\alpha - \beta'x_i) \\ &= 1 - \Phi(\alpha + \beta'x_i) \end{aligned} \quad \text{where } \Phi \text{ is the cdf of std. normal dist.}$$

Note y_i^* is artificial, and we must avoid placing too much emphasis on this latent variable.

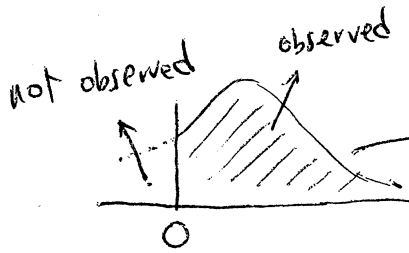
$$P(y_i^* > 0) = \Phi(\alpha + \beta'x_i) = P(y_i > 0)$$

Note Actually β and σ are not separately identified. $(\alpha + \beta'x_i) / \sigma$ is a correct expression.

We're interested in $E(y_i | x_i)$ and $E(y_i | x_i, y_i > 0)$.
(or)

Note we will omit $(\cdot | x_i)$ for simplicity, and let $\alpha + \beta'x_i \Rightarrow \beta'x_i$

① $E(y_i | y_i > 0)$... discarding the observations with $y_i = 0$
(truncation)



$$f(y_i | y_i > 0) = \frac{f(y_i)}{P(y_i > 0)} = \frac{f(y_i)}{1 - \Phi(\beta'x_i)}$$

$$E(y_i | y_i > 0) = \beta'x_i + \sigma \cdot \frac{\phi(\beta'x_i/\sigma)}{\Phi(\beta'x_i/\sigma)}$$

(truncated mean)

←

$$\lambda_i = \frac{\phi(\cdot)}{\Phi(\cdot)} = \frac{\text{pdf}}{\text{cdf}} = \text{inverse Mills ratio}$$

point: when the subsets of the data are truncated, the OLS is biased (inconsistent).

Reason: 2nd term is omitted (ignored)

$$\begin{aligned} \textcircled{2} E(y_i) &= \underbrace{P(y_i = 0)}_0 \cdot 0 + \underbrace{P(y_i > 0)}_{\Phi(\frac{\beta'x_i}{\sigma})} \cdot \underbrace{E(y_i | y_i > 0)}_{\beta'x_i + \sigma \lambda_i \text{ (above)}} \\ \text{(unconditional mean)} & \\ &= \Phi\left(\frac{\beta'x_i}{\sigma}\right) [\beta'x_i + \sigma \lambda_i] \end{aligned}$$

point: Even when using the whole sample including the cases with $y_i = 0$ (censored data will be used), the OLS is biased, due to 2 terms: Φ and λ_i

Note Even though the OLS is biased, the slope coefficients (β) are inconsistent by the same multiplicative factors: $\Phi(\frac{\beta'X_i}{\sigma})$.

- i) they are downward biased. why?
- ii) Relative effects of any two variables are consistently measured from the OLS.
: $\hat{\beta}_i / \hat{\beta}_j$ is not biased.

Note Marginal effect of Tobit models

(next page)
text.
log 16.11)

$$\frac{dE(y_i)}{dx_{ij}} = \Phi\left(\frac{\beta'X_i}{\sigma}\right) \cdot \beta_j \quad \text{or} \quad \left[\frac{dE(y_i | y_i > 0)}{dx_{ij}} \right] = \beta_j + \beta \frac{dy_i}{dx_{ij}}$$

thus the marginal effect of one variable (x_{ij}) depends on its coefficient β_j as well as all other coefficients (β) plus all regressors (X): evaluate at mean

(more on this later)

if x_{ij} is binary, the diff = $E(y_i | x_{ij}=1) - E(y_i | x_{ij}=0)$ is used instead.

Estimation

MLE

$$L = \prod_{i=1}^n f(y_i) = \prod_{y_i=0} [1 - \Phi(\beta'X_i/\sigma)] \cdot \prod_{y_i>0} \left(\frac{1}{\sigma}\right) \phi\left(\frac{y_i - \beta'X_i}{\sigma}\right)$$

(part 1) (part 2)

$y_i=0$ $y_i>0$

$$\log L = \sum [o(y_i=0) \log [1 - \Phi(\beta'X_i/\sigma)] + o(y_i>0) [\log \phi(\frac{y_i - \beta'X_i}{\sigma}) - \log(\sigma^2)^{1/2}]$$

More in technical details

⑦

Exercise (Poirier, 1995, p. 115, A3.3.2) H/W

Consider the truncated normal random variable z with the pdf

$$f(z | a < z < b) = \frac{\phi(z | \mu, \sigma^2)}{\Phi(b | \mu, \sigma^2) - \Phi(a | \mu, \sigma^2)} \quad \text{if } a < z < b$$

Define $a^* = \frac{a - \mu}{\sigma}$, $b^* = \frac{b - \mu}{\sigma}$

a) show that $E(z) = \mu + \sigma \frac{\phi(a^*) - \phi(b^*)}{\Phi(b^*) - \Phi(a^*)}$

Hint: $E(z) = \int_a^b z \cdot f(z | a < z < b) dz$

b) Show that

$$\text{Var}(z) = \sigma^2 \left[1 + \frac{a^* \phi(a^*) - b^* \phi(b^*)}{\Phi(b^*) - \Phi(a^*)} - \left(\frac{\phi(a^*) - \phi(b^*)}{\Phi(b^*) - \Phi(a^*)} \right)^2 \right]$$

Hint: $\text{Var}(z) = E(z^2) - [E(z)]^2$

More on Marginal effects

Note $E(y_i | y_i > 0) = \beta' x_i + \sigma \lambda_i$

$E(y_i) = \underbrace{E(\beta' x_i)}_{\text{het. } 0 \leq 1} [\beta' x_i + \sigma \lambda_i] = P(y > 0) E(y | y > 0)$

Show
H/W

i) $\frac{dE(y_i | y_i > 0)}{dx_{ij}} = \beta_j + \beta_j \left[\frac{d\lambda_i}{dx_{ij}} \right] = \beta_j \left[1 - \lambda_i \left(\frac{\beta_j x_j}{\sigma} + \lambda_i \right) \right]$ Wooldridge p 522

⇒ thus, the sign of β_j is the same as the partial effect.

ii) $\frac{dE(y_i)}{dx_{ij}} = \underbrace{\frac{dP(y_i > 0)}{dx_{ij}}}_{\left(\frac{\beta_j}{\sigma}\right) \phi\left(\frac{\beta_j x_j}{\sigma}\right)} \cdot \underbrace{E(y_i | y_i > 0)}_{\text{above}} + \underbrace{P(y_i > 0)}_{\Phi\left(\frac{\beta_j x_j}{\sigma}\right)} \cdot \underbrace{\frac{dE(y_i | y_i > 0)}{dx_{ij}}}_{\text{above}} : \text{chain rule}$

$= \Phi\left(\frac{\beta_j x_j}{\sigma}\right) \beta_j$ after simplification.

⇒ Again, the sign of $\hat{\beta}_j$ is the same as the partial effect.

Extended Discussion of Tobit Models

(8)

(1) Endogeneity issue

$$y_1 = \max(0, z_1 \delta_1 + \delta_1 y_2 + u_1) \quad \text{Tobit} \quad \left(\begin{array}{l} \text{IV for } y_2 \\ = z_2 \end{array} \right)$$

$$y_2 = z_1 \delta_{21} + z_2 \delta_{22} + v_2 \quad \text{reduced form}$$

- Testing for endogeneity

1st OLS of y_2 on z_1 and z_2 (reduced form).

2nd Tobit estimation with z_1 , y_2 and \hat{v}_2
(residuals from the 1st stage)

Test H_0 : coeff of $\hat{v}_2 = 0$ (t-test)
(no endogeneity)

Note when the null is rejected (evidence of endogeneity), the std. errors and test statistics are invalid. They need to be corrected, by accounting for correlation between u_1 and v_2 .

"Murphy & Toppe
adjustment"

(A recent version of Cindex P.O. does this!)

Note when y_2 is binary, the reduced form estimation is probit. ...

- Estimation

① The above 2-step estimation with \hat{v}_2
(But std. errors need to be corrected)

② FIML (MLE) with $f(y_1, y_2) = \underbrace{f(y_1 | y_2)}_{\text{Tobit}} \underbrace{f(y_2)}_{\text{normal}}$

(2) Heteroskedasticity and non-normality

As in the probit, both problems result in inconsistent tobit estimators. (as in probit)

the source of the problem is that the assumption $y^* | x \sim N(\beta'x, \sigma^2)$ is not satisfied.

i) Heteroskedasticity

usual test (LM). If the heteroskedasticity exists, then allow for it and modify the MLE ... not a serious problem.

ii) Non-normality

not clear solution.

(3) Panel Tobit Models

i) Pooled tobit

easy extension
$$Q = \sum_{i=1}^N \sum_{t=1}^T Q_{it}$$

ii) FE tobit ;

- Conditional likelihood method is not feasible.
- Demeaning procedures as in linear FE models are not feasible - using dummy variables can lead biased estimator.

iii) RE tobit (if x_i and ϵ_i are correlated,

Add \bar{x}_i and do the RE treatment.)

MLE can be done. (Adding \bar{x}_i allows that ϵ_i and x_i are correlated)

... see Wooldridge, p. 540

Usual RE (integrating out) is also possible. →

v) Dynamic Panel Tobit

$$y_{it} = \max(0, \alpha_i + \beta y_{i,t-1} + \epsilon_{it})$$

Note $\beta y_{i,t-1}$ can be replaced with $\lambda_1 r_{it} + \beta_i (1 - r_{it}) y_{i,t-1}$
 where r_{it} is a dummy variable equal to 1 if $y_{it} = 0$.

Note Initial condition matters then modify the model as:

$$y_{it} = \max(0, \psi + \alpha_i + \beta y_{i,t-1} + \beta_0 y_{i0} + \epsilon_{it})$$

and do RE treatment

Note Honoré : GMM.

(4) Robust Variance

Huber/white sandwich estimator can be used.

OMLE ..

Note stata options

, robust

, robust cluster() .. allowing for correlation within cluster

(5) Prediction

stata : $y^*(0, \cdot) \Rightarrow E(y^* | y > 0) = P(y=0) \cdot 0 + P(y > 0) E(y | y > 0)$
 $= P(y > 0) E(y | y > 0)$

$e(0, \cdot) \Rightarrow E(y_i | y_i > 0)$

```

*** Tobit, Cnreg, Intreg

set memory 30m
set more off

log using tobit_cnreg_intreg.log, replace

** Tobit Models

use http://www.stata-press.com/data/r8/auto, clear

generate wgt = weight/1000
regress mpg wgt

replace mpg = 17 if mpg <= 17
* this replacement is not needed, though.

tobit mpg wgt, ll nolog
* ll (lower limit), ul (upper limit)
tobit mpg wgt, ul(24) nolog
* ul (upper limit 24 is imposed)

```

```

tobit mpg wgt, ll (17) ul(24) nolog
. ** Tobit Models
.
. use http://www.stata-press.com/data/r8/auto, clear
(1978 Automobile Data)
.
. generate wgt = weight/1000
. regress mpg wgt

```

Source	SS	df	MS	Number of obs =	74
Model	1591.99024	1	1591.99024	F(1, 72) =	134.62
Residual	851.469221	72	11.8259614	Prob > F =	0.0000
				R-squared =	0.6515
				Adj R-squared =	0.6467
Total	2443.45946	73	33.4720474	Root MSE =	3.4389

mpg	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
wgt	-6.008687	.5178782	-11.60	0.000	-7.041058	-4.976316
_cons	39.44028	1.614003	24.44	0.000	36.22283	42.65774

```

. replace mpg = 17 if mpg <= 17
(14 real changes made)

. * this replacement is not needed, though.
.

```


se | 2.886337 .3952143 (Ancillary parameter)

Obs. summary: 18 left-censored observations at mpg<=17
33 uncensored observations
23 right-censored observations at mpg>=24

```
** Tobit Models and marginal effects
use http://fmwww.bc.edu/ec-p/data/wooldridge/MROZ, clear
regress hours nwifeinc educ exper expersq age kidslt6 kidsge6
tobit hours nwifeinc educ exper expersq age kidslt6 kidsge6, ll(0) nolog
* -- fixup for expersq : take square of mean rather than mean of square per JMW
summ exper,meanonly
local exp2=r(mean)^2
mfx compute, at(mean expersq=`exp2') predict(ystar(0,.))
* -- marginal effects conditional on positive hours
mfx compute, at(mean expersq=`exp2') predict(e(0,.))
* e(a,b) gives the conditionl expectation given a < Xb < b.
```

```
. ** Tobit Models and marginal effects
. use http://fmwww.bc.edu/ec-p/data/wooldridge/MROZ, clear
. regress hours nwifeinc educ exper expersq age kidslt6 kidsge6
```

OLS

Table with 5 columns: Source, SS, df, MS, and a summary of statistics (Number of obs, F, Prob > F, R-squared, Adj R-squared, Root MSE).

Table with 7 columns: Variable, Coef., Std. Err., t, P>|t|, and [95% Conf. Interval].

```
. tobit hours nwifeinc educ exper expersq age kidslt6 kidsge6, ll(0) nolog
```

```
Tobit estimates                               Number of obs   =       753
                                                LR chi2(7)      =       271.59
                                                Prob > chi2     =       0.0000
Log likelihood = -3819.0946                    Pseudo R2      =       0.0343
```

hours	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
nwifeinc	-8.814243	4.459096	-1.98	0.048	-17.56811	-.0603725
educ	80.64561	21.58322	3.74	0.000	38.27453	123.0167
exper	131.5643	17.27938	7.61	0.000	97.64231	165.4863
expersq	-1.864158	.5376615	-3.47	0.001	-2.919667	-.8086479
age	-54.40501	7.418496	-7.33	0.000	-68.96862	-39.8414
kidslt6	-894.0217	111.8779	-7.99	0.000	-1113.655	-674.3887
kidsge6	-16.218	38.64136	-0.42	0.675	-92.07675	59.64075
_cons	965.3053	446.4358	2.16	0.031	88.88531	1841.725

_se	1122.022	41.57903	(Ancillary parameter)			

```
Obs. summary:      325 left-censored observations at hours<=0
                   428 uncensored observations
```

```
. ** fixup for expersq : take square of mean rather than mean of square per JMW
. summ exper,meanonly
```

```
. local exp2=r(mean)^2
```

```
. mfx compute, at(mean expersq=`exp2') predict(ystar(0,..))
```

$(\bar{X})^2$ vs $\overline{X^2}$
 ↑
 this is used.

Marginal effects after tobit

```
y = E(hours*|hours>0) (predict, ystar(0,..))
= 687.31745
```

variable	dy/dx	Std. Err.	z	P> z	[95% C.I.]		X
nwifeinc	-5.687381	2.87788	-1.98	0.048	-11.3279	-.046836	20.129
educ	52.03649	13.82	3.77	0.000	24.9495	79.1234	12.2869
exper	84.89173	12.398	6.85	0.000	60.593	109.19	10.6308
expersq	-1.202846	.36661	-3.28	0.001	-1.92139	-.484297	113.014
age	-35.10478	4.66947	-7.52	0.000	-44.2568	-25.9528	42.5378
kidslt6	-576.8666	70.93	-8.13	0.000	-715.887	-437.847	.237716
kidsge6	-10.46465	24.94	-0.42	0.675	-59.3456	38.4163	1.35325

```
. * marginal effects conditional on positive hours
```

```
. mfx compute, at(mean expersq=`exp2') predict(e(0,..))
```

Marginal effects after tobit

```
y = E(hours|hours>0) (predict, e(0,..))
= 1065.1973
```

variable	dy/dx	Std. Err.	z	P> z	[95% C.I.]	X
nwifeinc	-3.987413	2.01764	-1.98	0.048	-7.94192	-.032909		20.129
educ	36.48269	9.68927	3.77	0.000	17.4921	55.4733		12.2869
exper	59.51744	8.68378	6.85	0.000	42.4975	76.5373		10.6308
expersq	-.843313	.25692	-3.28	0.001	-1.34686	-.339765		113.014
age	-24.6119	3.27362	-7.52	0.000	-31.0281	-18.1957		42.5378
kidslt6	-404.4402	49.722	-8.13	0.000	-501.893	-306.987		.237716
kidsge6	-7.336744	17.485	-0.42	0.675	-41.607	26.9335		1.35325

. ** e(a,b) gives the conditionl expectation given $a < Xb < b$.

```

*** Panel RE Tobit (xttobit)

use http://www.stata-press.com/data/r8/nlswork, clear
xttobit ln_wage union age grade not_smsa south occ_code, i(id) ul(1.9) tobit
nolog
*quadchk

```

```

. *** Panel RE Tobit (xttobit)
.
. use http://www.stata-press.com/data/r8/nlswork, clear
(National Longitudinal Survey. Young Women 14-26 years of age in 1968)

. xttobit ln_wage union age grade not_smsa south occ_code, i(id) ul(1.9) tobit
nolog

```

```

Random-effects tobit regression
Group variable (i): idcode
Random effects u_i ~ Gaussian
Log likelihood = -6672.7585

Number of obs = 19151
Number of groups = 4140
Obs per group: min = 1
                avg = 4.6
                max = 12
Wald chi2(6) = 3303.29
Prob > chi2 = 0.0000

```

ln_wage	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
union	.1545505	.0069135	22.35	0.000	.1410002	.1681008
age	.0099367	.000414	24.00	0.000	.0091252	.0107482
grade	.0788248	.0022308	35.33	0.000	.0744525	.0831972
not_smsa	-.1276947	.0088914	-14.36	0.000	-.1451215	-.1102679
south	-.0868263	.0086892	-9.99	0.000	-.1038569	-.0697957
occ_code	-.0190243	.0010974	-17.34	0.000	-.0211751	-.0168735
_cons	.521857	.0320114	16.30	0.000	.4591158	.5845982
/sigma_u	.2847095	.0044262	64.32	0.000	.2760343	.2933848
/sigma_e	.2497528	.0018149	137.61	0.000	.2461956	.25331
rho	.5651268	.0082516			.5489039	.5812407

Likelihood-ratio test of sigma_u=0: chibar2(01)= 5920.66 Prob>=chibar2 = 0.000

```

Observation summary:    12288    uncensored observations
                          0      left-censored observations
                          6863    right-censored observations

```

```

. *quadchk

```

Alternatives to Tobit models

(17)

(Two-tier models; hurdle models)

$$y=0 \rightarrow P(y=0) = 1 - \Phi(X\gamma)$$

$$y>0 \rightarrow \log(y | y>0) \sim N(X\beta, \sigma^2) \quad \dots \text{log-normal}$$

$$\text{let } d_i = \begin{cases} 0 & \text{if } y_i = 0 \\ 1 & \text{if } y_i > 0 \end{cases} \Rightarrow d_i = I(y_i > 0) = \begin{cases} 1 & \text{if } y_i > 0 \\ 0 & \text{if } y_i = 0 \end{cases}$$

... index function

then

$$f(y_i) = \underbrace{P(y_i=0)}_{1 - \Phi(X_i'\gamma)} \cdot d_i + (1-d_i) \underbrace{P(y_i>0)}_{\Phi(X_i'\gamma)} \cdot \underbrace{\phi\left[\frac{\log y_i - X_i'\beta}{\sigma}\right] \cdot \frac{1}{y_i \sigma}}_{\text{log-normal density}}$$

$$\begin{aligned} \mathcal{L}_i &= d_i \log(1 - \Phi(X_i'\gamma)) + (1-d_i) \left[\log \Phi(X_i'\gamma) - \log y_i \right. \\ &\quad \left. - \frac{1}{2} \log \sigma^2 - \frac{1}{2} \log 2\pi - \frac{1}{2} \left[\log y_i - X_i'\beta \right]^2 / \sigma^2 \right] \end{aligned}$$

then

$$E(y_i | y_i > 0) = \exp(X_i'\beta + \frac{1}{2}\sigma^2)$$

$$E(y_i) = \Phi(X_i'\gamma) \exp(X_i'\beta + \frac{\sigma^2}{2})$$

Note this method is similar to a regime switching model.
[Also used in count data model.]

Censored Normal Regression

Censoring values may vary from observation to observation
 (stata: "cnreg") ... needs an indicator to show each obs is censored or not.

Note this type of censoring is useful for duration models.

$$y_i \text{ is observed. } y_i = \begin{cases} L_i & \text{if } y_i^* \leq L_i \\ R_i & \text{if } y_i^* \geq R_i \\ y_i & \text{if } L_i < y_i^* < R_i \end{cases}$$

$$\begin{aligned} \mathcal{L} = & \sum_{j \in L} \log \Phi \left(\frac{y_{Lj} - X_j \beta}{\sigma} \right) \quad \dots L_i \\ & + \sum_{j \in R} \log \left[1 - \Phi \left(\frac{y_{Rj} - X_j \beta}{\sigma} \right) \right] \quad \dots R_i \\ & + \sum_{j \in C} \left[-\frac{1}{2} \left(\frac{y_{ij} - X_j \beta}{\sigma} \right)^2 - \frac{1}{2} \log(2\pi\sigma^2) \right] \quad \dots C_i \end{aligned}$$

Interval Regression

if y_j is observed as an interval data $\{y_{1j}, y_{2j}\}$ rather than a point. (stata: "intreg")

eg)

wage1	wage2	
.	4	.. left censored (L_i)
5	10] interval data * (I_i)
5	10	
10	15	
10	10	... point data (usual) (C_i)
15	.	.. right censored (R_i)

$$\mathcal{L} = (\text{the above 3 terms}) + \sum_{j \in I} \log \left[\Phi \left(\frac{y_{2j} - X_j \beta}{\sigma} \right) - \Phi \left(\frac{y_{1j} - X_j \beta}{\sigma} \right) \right]$$

* Panel : RE version

```

** Censored regression CNREG (each can be censored at a different point)

use http://www.stata-press.com/data/r8/news, clear

generate cnsrd = 0
replace cnsrd = -1 if before82
replace date = mdy(1,1,1982) if before82
replace cnsrd = 1 if date>= .
replace date = mdy(1,1,1991) if date >= .
list date cnsrd in 1/12

cnreg date lncltn famown, censored(cnsrd) nolog

```

```

. ** Censored regression CNREG (each can be censored at a different point)
.
. use http://www.stata-press.com/data/r8/news, clear
.
. generate cnsrd = 0
.
. replace cnsrd = -1 if before82
(24 real changes made)
.
. replace date = mdy(1,1,1982) if before82
(24 real changes made)
.
. replace cnsrd = 1 if date>= .
(11 real changes made)
.
. replace date = mdy(1,1,1991) if date >= .
(11 real changes made)
.
. list date cnsrd in 1/12

```

	date	cnsrd
1.	8036	-1
2.	8036	-1
3.	8036	-1
4.	8927.786	0
5.	10466.89	0
6.	8823.7	0
7.	9356.509	0
8.	10784.71	0
9.	10990.53	0
10.	9033.669	0
11.	8036	-1
12.	11290.81	0

```
. cnreg date lncltn famown, censored(cnsrd) nolog
```

Censored normal regression	Number of obs	=	100
	LR chi2(2)	=	201.09
	Prob > chi2	=	0.0000
Log likelihood = -519.74678	Pseudo R2	=	0.1621

date	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lncltn	-1377.138	76.92723	-17.90	0.000	-1529.797	-1224.478
famown	576.8444	185.5287	3.11	0.002	208.6687	945.0201
_cons	24439.46	815.7107	29.96	0.000	22820.71	26058.21

_se	607.9846	53.19381	(Ancillary parameter)			

```
Obs. summary:      24 left-censored observations
                   65 uncensored observations
                   11 right-censored observations
```

```
** Intreg (interval regression; even point data; needs depvar1 depvar2)
* point          a  a
* interval       a  b
* left censored  .  b
* right censored a  .

use http://www.stata-press.com/data/r8/womenwage, clear

tab wagecat
by wagecat: keep if _n == 1
generate wage1 = wagecat[_n-1]
keep wagecat wage1
save lagwage, replace

use http://www.stata-press.com/data/r8/womenwage, clear
merge wagecat using lagwage

generate wage2 = wagecat
replace wage2 = . if wagecat == 51
sort age, stable
list wage1 wage2 in 1/10

intreg wage1 wage2 age age2 nev_mar rural school tenure, nolog

oprobit wage1 wage2 age age2 nev_mar rural school tenure, nolog
```

```
. ** Intreg (interval regression; even point data; needs depvar1 depvar2)
. * point          a  a
. * interval       a  b
. * left censored  .  b
```

. * right censored a .

. use http://www.stata-press.com/data/r8/womenwage, clear
(Wages of women)

. tab wagecat

Wage category (\$1000s)	Freq.	Percent	Cum.
5	14	2.87	2.87
10	83	17.01	19.88
15	158	32.38	52.25
20	107	21.93	74.18
25	57	11.68	85.86
30	30	6.15	92.01
40	19	3.89	95.90
50	14	2.87	98.77
51	6	1.23	100.00
Total	488	100.00	

. by wagecat: keep if _n == 1
(479 observations deleted)

. generate wage1 = wagecat[_n-1]
(1 missing value generated)

. keep wagecat wage1

. save lagwage, replace
file lagwage.dta saved

. use http://www.stata-press.com/data/r8/womenwage, clear
(Wages of women)

. merge wagecat using lagwage

. generate wage2 = wagecat

. replace wage2 = . if wagecat == 51
(6 real changes made, 6 to missing)

. sort age, stable

. list wage1 wage2 in 1/10

```

+-----+
| wage1  wage2 |
+-----+

```

1.	.	5
2.	5	10
3.	5	10
4.	10	15
5.	.	5

6.	.	5
7.	.	5
8.	5	10
9.	5	10
10.	5	10

+-----+		

```
. intreg wage1 wage2 age age2 nev_mar rural school tenure, nolog
```

```
Interval regression                               Number of obs =      488
                                                LR chi2(6)       =    221.61
Log likelihood = -856.33293                    Prob > chi2      =    0.0000
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age	.7914438	.4433604	1.79	0.074	-.0775265	1.660414
age2	-.0132624	.0073028	-1.82	0.069	-.0275757	.0010509
nev_mar	-.2075022	.8119581	-0.26	0.798	-1.798911	1.383906
rural	-3.043044	.7757324	-3.92	0.000	-4.563452	-1.522637
school	1.334721	.1357873	9.83	0.000	1.068583	1.600859
tenure	.8000664	.1045077	7.66	0.000	.5952351	1.004898
_cons	-12.70238	6.367117	-1.99	0.046	-25.1817	-.2230583

/lnsigma	1.987823	.0346543	57.36	0.000	1.919902	2.055744

sigma	7.299626	.2529634			6.82029	7.81265

```
Observation summary:      0 uncensored observations
                        14 left-censored observations
                        6 right-censored observations
                        468 interval observations
```

```
. oprobit wage1 wage2 age age2 nev_mar rural school tenure, nolog
```

```
Ordered probit estimates                               Number of obs =      468
                                                LR chi2(7)       =   1570.91
Log likelihood = 0                                Prob > chi2      =    0.0000
                                                Pseudo R2       =    1.0000
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
wage1						
wage2	3.387503	515650.1	0.00	1.000	-1010652	1010659
age	.0079048	1656081	0.00	1.000	-3245858	3245858

age2	-.0001439	27116.31	-0.00	1.000	-53147	53147
nev_mar	-.0011112	2911632	-0.00	1.000	-5706694	5706694
rural	-.0114906	2859866	-0.00	1.000	-5605235	5605235
school	.0022328	558215.5	0.00	1.000	-1094082	1094082
tenure	.000831	400255.5	0.00	1.000	-784486.4	784486.4

_cut1	42.45669	2.46e+07	(Ancillary parameters)			
_cut2	59.43884	2.54e+07				
_cut3	76.37362	2.63e+07				
_cut4	93.30152	2.77e+07				
_cut5	110.3099	2.97e+07				
_cut6	144.063	3.13e+07				

note: 468 observations completely determined. Standard errors questionable.

```
*** Panel RE INTREG (xtintreg)

use http://www.stata-press.com/data/r8/nlswork3.dta, replace

xtintreg ln_wage1 ln_wage2 union age grade not_smsa south southXt occ_code,
i(id) noskip intreg nolog
*quadchk
```

```
. *** Panel RE INTREG (xtintreg)
. use http://www.stata-press.com/data/r8/nlswork3.dta, replace
(National Longitudinal Survey. Young Women 14-26 years of age in 1968)

. xtintreg ln_wage1 ln_wage2 union age grade not_smsa south southXt occ_code,
i(id) noskip intreg nolog

Random-effects interval regression          Number of obs      =      19095
Group variable (i): idcode                 Number of groups   =       4139

Random effects u_i ~ Gaussian              Obs per group: min =         1
                                           avg =         4.6
                                           max =         12

Log likelihood = -14856.934                LR chi2(7)         =      3549.46
                                           Prob > chi2        =       0.0000
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
union	.1409746	.0068364	20.62	0.000	.1275755	.1543737
age	.012631	.0005148	24.53	0.000	.0116219	.01364
grade	.0783789	.0020912	37.48	0.000	.0742802	.0824777
not_smsa	-.1333091	.0089209	-14.94	0.000	-.1507938	-.1158243
south	-.1218994	.0121087	-10.07	0.000	-.145632	-.0981669
southXt	.0021033	.0008314	2.53	0.011	.0004738	.0037328
occ_code	-.0185603	.001033	-17.97	0.000	-.020585	-.0165355

_cons		.4567546	.032493	14.06	0.000	.3930695	.5204398

/sigma_u		.282881	.0038227	74.00	0.000	.2753886	.2903734
/sigma_e		.2696119	.0015957	168.96	0.000	.2664843	.2727394

rho		.524003	.0075625			.5091676	.5388052

Likelihood-ratio test of sigma_u=0: chibar2(01) = 6629.90 Prob>=chibar2 = 0.000

Observation summary: 14372 uncensored observations
 157 left-censored observations
 718 right-censored observations
 3848 interval observations

. *quadchk

Note use of Tobit models for fractional data.
(interval)

If censoring matters (i.e. point prob $P(X=a)$ or $P(X=b)$ is not zero so that the data has a large proportion of observations with values a or b), then using Tobit makes sense. If not, using Tobit models just because the data is bounded between a and b, may not be appealing.

Suggested methods

i) Fractional logit : Wooldridge

ii) Weighted logit regression : Greene

$$\log\left(\frac{p_i}{1-p_i}\right) = X_i\beta + \epsilon_i \quad \epsilon_i \sim \text{heteroskedasticity}$$

Truncated Regression

$$\text{Based on } f(z^*) = f(z) / P(a < z < b) = \frac{f(z)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}$$
$$= \frac{\frac{1}{\sigma} \phi\left(\frac{z-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}$$

$$\log \mathcal{L} = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum (y_i - X_i\beta)^2 - \sum \log\left[\Phi\left(\frac{b-X_i\beta}{\sigma}\right) - \Phi\left(\frac{a-X_i\beta}{\sigma}\right)\right]$$

$$\frac{dE(y_i | y_i > a)}{dX_i} = \beta (1 - \lambda_i^2 + d_i X_i) \quad \text{where } d_i = (a - X_i\beta) / \sigma$$
$$\lambda_i = \phi(d_i) / [1 - \Phi(d_i)]$$

$$\frac{dE(y_i | y_i < b)}{dX_i} = \beta [1 - \lambda_i^2 + d_i X_i] \quad \text{where } d_i = (b - X_i\beta) / \sigma$$
$$\lambda_i = -\phi(d_i) / \Phi(d_i)$$

$$\frac{d E(y_i | a < y_i < b)}{d x_i} = \beta \left[1 - \lambda_i^2 + d_{i2} \lambda_i - \frac{(b-a) \phi(d_{i1})}{\sigma [\Phi(d_{i2}) - \Phi(d_{i1})]} \right] \quad (2b)$$

$$\text{where } d_{i1} = (a - x_i \beta) / \sigma$$

$$d_{i2} = (b - x_i \beta) / \sigma$$

$$\lambda_i = \frac{\phi(d_{i2}) - \phi(d_{i1})}{\Phi(d_{i2}) - \Phi(d_{i1})}$$

Exercises

① Wooldridge Ex 16.1, p. 544

$$t_i^* = \exp(x_i \beta + u_i) \quad u_i \sim N(0, \sigma^2)$$

$$t_i = \min(t_i^*, c)$$

a) Find $P(t_i = c | x_i)$. What happens as $c \rightarrow \infty$?

$$\text{Hint } P(t_i = c) = P(\log t_i = \log c)$$

$$= P(\log t_i > \log c) : \text{right censored}$$

$$= P(x_i \beta + u_i > \log c)$$

$$= 1 - \Phi[(\log c - x_i \beta) / \sigma] \quad \dots \text{answer}$$

b) ~ e)

② Wooldridge Ex 16.3, p. 544

```

*** Truncated Regression

use http://www.stata-press.com/data/r8/laborsub, replace

truncreg whrs kl6 k618 wa we, ll(0) nolog
tobit whrs kl6 k618 wa we, ll(0) nolog

```

```

. *** Truncated Regression
.
. use http://www.stata-press.com/data/r8/laborsub, replace

. 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]	

eq1							
	kl6	-803.0042	321.3614	-2.50	0.012	-1432.861	-173.1474
	k618	-172.875	88.72898	-1.95	0.051	-346.7806	1.030579
	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	-201.9234	3374.442

sigma							
	_cons	983.7262	94.44303	10.42	0.000	798.6213	1168.831

```

. tobit whrs kl6 k618 wa we, ll(0) nolog

```

```

Tobit estimates          Number of obs =    250
Log likelihood = -1367.0903  Pseudo R2      =    0.0084

```

	whrs	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	

	kl6	-827.7657	214.7407	-3.85	0.000	-1250.731	-404.8009
	k618	-140.0192	74.22303	-1.89	0.060	-286.2128	6.174543
	wa	-24.97919	13.25639	-1.88	0.061	-51.08969	1.131316
	we	103.6896	41.82393	2.48	0.014	21.31093	186.0683
	_cons	589.0001	841.5467	0.70	0.485	-1068.556	2246.556

	_se	1309.909	82.73335			(Ancillary parameter)	

```

Obs. summary:    100 left-censored observations at whrs<=0
                 150 uncensored observations

```