

Limited Dependent Variables

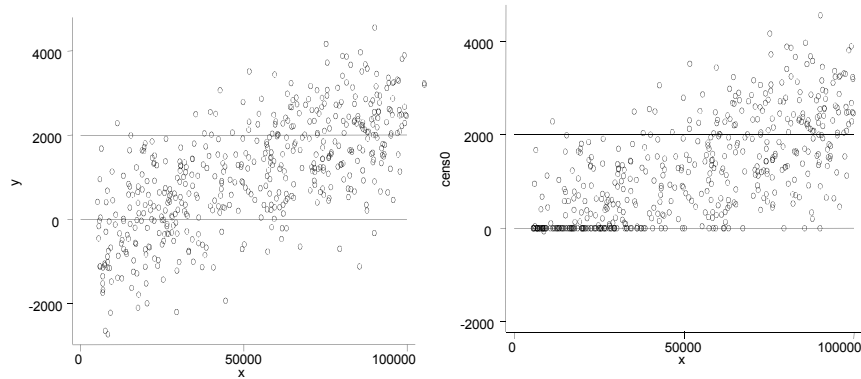
Categorical and Limited
Dependent Variables

Paul A. Jargowsky

Limited Dependent Variables

- Some continuous variables are affected by limits
 - Values pile up at the limit (censoring)
 - Highest SAT score is 800
 - IRA contributions limits
 - Missing values based on limits (truncation)
 - Minimum wage (people who would earn below the minimum are unemployed)
 - People who want to contribute less than \$0 to IRA (in other words take some out)
- These conditions can cause bias in the coefficients, often severe. The bias is towards 0 (no effect).

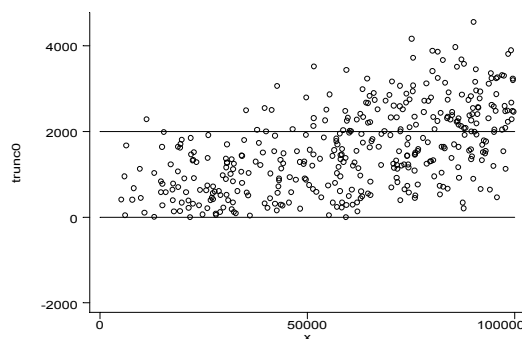
Censored Data: Misleading Values



The true values. For example, this could be desired contributions to an IRA account. Negative values mean that you would withdraw money.

Here is the same data, but withdrawals are not allowed, so the recorded value for these people is a contribution of 0. Note how the values "pile up" at zero.

Truncation: Values are Missing (Selected on Value of the DV)



Here we have truncation at zero. People who can't afford to make an IRA contribution of at least \$1 are simply not in the dataset. Similarly, people who can't get a job of at least minimum wage are unemployed, and have no observed wage.

The Data (tobit.dta)

Censored means that the data contains "lies," that is a value is recorded but it is not the "true" or desired value. Usually this is because the data are top-coded or bottom-coded or some limit prevents the value from going higher or lower.

Truncated means that the data above or below a limit are missing altogether. In this case we have some information (the x values), but in many cases the truncated individuals are not even in the data set at all.

$$\text{PRF: } Y_i = -500 + 0.03X_i + u_i \quad u_i \sim N(0, \sigma^2) \quad \sigma = 1000$$

```
. list x y cens0 cens2000 censboth trunc0 trunc2k ira, noobs nod
```

x	y	cens0	cens2000	censboth	trunc0	trunc2k	ira
43186.99	-298.3602	0	-298.3602	0	.	-298.3602	.
56510.91	1562.408	1562.408	1562.408	1562.408	1562.408	1562.408	1562.408
86320.5	2235.013	2235.013	2000	2000	2235.013	.	2000
53107.77	1359.011	1359.011	1359.011	1359.011	1359.011	1359.011	1359.011
57319.36	1698.989	1698.989	1698.989	1698.989	1698.989	1698.989	1698.989
16162.89	-1248.756	0	-1248.756	0	.	-1248.756	.
81442.45	2244.707	2244.707	2000	2000	2244.707	.	2000
34482.09	-1011.442	0	-1011.442	0	.	-1011.442	.
29066.62	510.9071	510.9071	510.9071	510.9071	510.9071	510.9071	510.9071
83317.7	3132.799	3132.799	2000	2000	3132.799	.	2000
27959.77	-319.5778	0	-319.5778	0	.	-319.5778	.
49503.07	-715.4042	0	-715.4042	0	.	-715.4042	.
89647.03	1998.137	1998.137	1998.137	1998.137	1998.137	1998.137	1998.137
5897.91	949.1215	949.1215	949.1215	949.1215	949.1215	949.1215	949.1215

OLS Regression on Censored Data

```
. reg y x
```

Source	SS	df	MS	Number of obs = 500		
Model	408705226	1	408705226	F(1, 498)	=	430.70
Residual	472568086	498	948931.901	Prob > F	=	0.0000
Total	881273312	499	1766078.78	R-squared	=	0.4638
				Adj R-squared	=	0.4627
				Root MSE	=	974.13

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	.0319565	.0015398	20.753	0.000	.0289312	.0349819
_cons	-607.3557	92.22465	-6.586	0.000	-788.5531	-426.1584

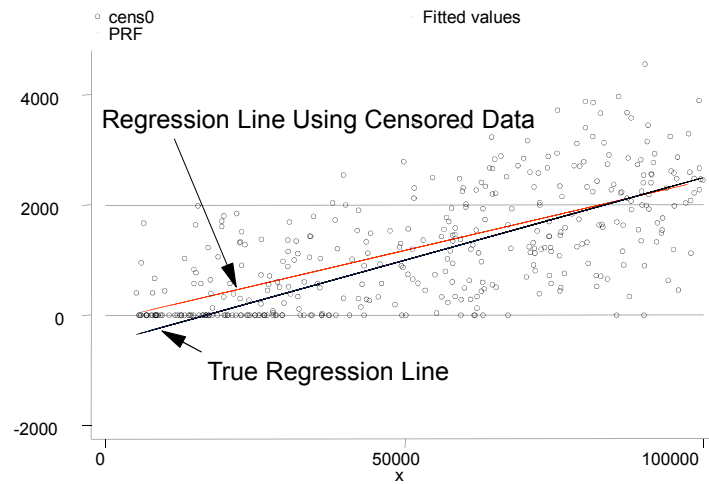
```
. reg cens0 x
```

OLS on the censored data is biased!!

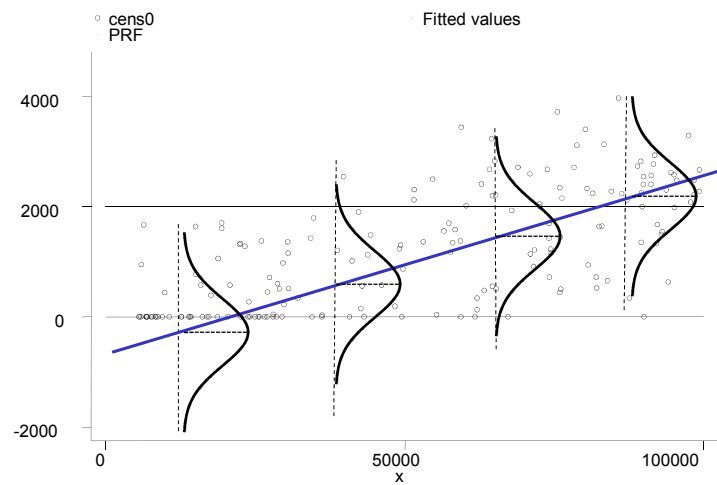
Source	SS	df	MS	Number of obs = 500		
Model	257524250	1	257524250	F(1, 498)	=	399.45
Residual	321062454	498	644703.723	Prob > F	=	0.0000
Total	578586704	499	1159492.39	R-squared	=	0.4451
				Adj R-squared	=	0.4440
				Root MSE	=	802.93

cens0	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	.0253667	.0012692	19.986	0.000	.022873	.0278603
_cons	-90.35696	76.0168	-1.189	0.235	-239.7101	58.99619

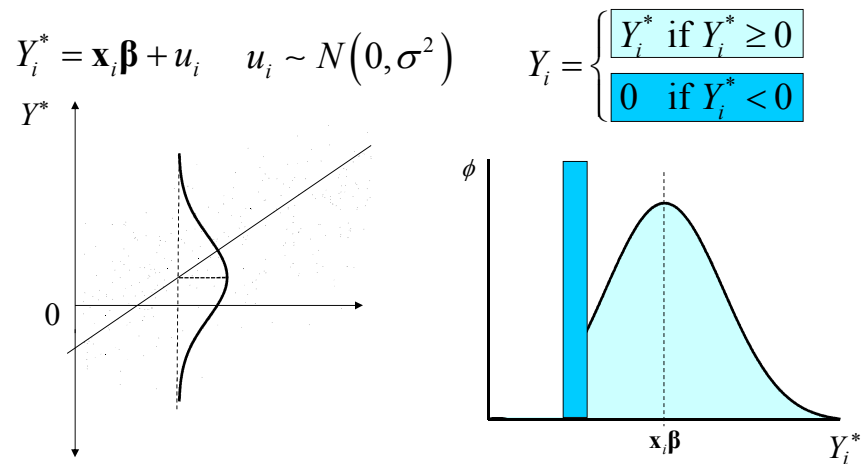
Biased Slope From Censoring at 0



The probability of being at zero depends on the distribution of u_i and X . Is that a problem? If so, why?

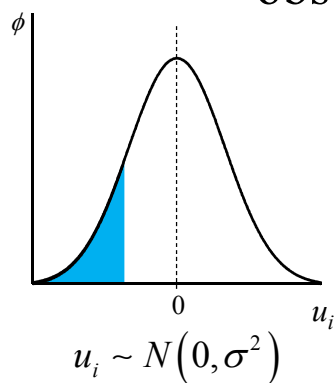


Need a Probability Statement to Use Maximum Likelihood



We are going to need a two-piece likelihood function!

Given X , what is the probability we observe a zero?



$$\begin{aligned} \Pr(Y_i = 0 | X_i) &= \Pr(Y_i^* < 0 | X_i) \\ &= \Pr(\mathbf{x}_i \boldsymbol{\beta} + u_i < 0) \\ &= \Pr(u_i < -\mathbf{x}_i \boldsymbol{\beta}) \end{aligned}$$

So this is the probability of censoring given X , as a function of the betas and sigma. Can we convert this problem to a standard normal?

Getting to the Standard Normal

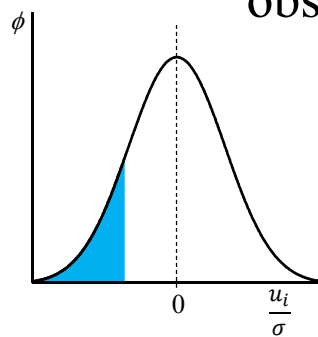
$$\text{var}(cX) = c^2 \text{var}(X)$$

$$\text{var}\left(\frac{u_i}{\sigma}\right) = \text{var}\left(\frac{1}{\sigma}u_i\right) = \frac{1}{\sigma^2} \text{var}(u_i) = \frac{\sigma^2}{\sigma^2} = 1$$

So if $u_i \sim N(0, \sigma^2)$, then $\frac{u_i}{\sigma} \sim N(0, 1)$, which is standard normal.

$$\begin{aligned} \Pr(Y_i = 0 \mid X_i) &= \Pr(u_i < -\mathbf{x}_i\boldsymbol{\beta}) \\ &= \Pr\left(\frac{u_i}{\sigma} < \frac{-\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right) = \Phi\left(\frac{-\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right) \end{aligned}$$

Given X , what is the probability we observe a zero?

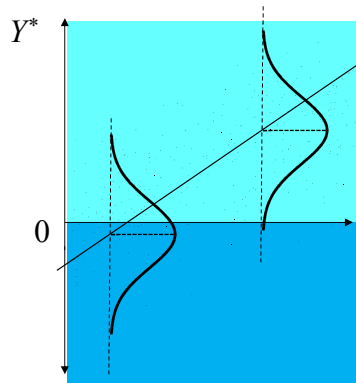


$$\frac{u_i}{\sigma} \sim N(0, 1)$$

$$\begin{aligned} \Pr(Y_i = 0 \mid X_i) &= \Pr(Y_i^* < 0 \mid X_i) \\ &= \Pr(\mathbf{x}_i\boldsymbol{\beta} + u_i < 0) \\ &= \Pr(u_i < -\mathbf{x}_i\boldsymbol{\beta}) \\ &= \Pr\left(\frac{u_i}{\sigma} < \frac{-\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right) \\ &= \Phi\left(\frac{-\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right) \end{aligned}$$

So this is the probability of censoring given X , as a function of the betas and sigma, making use of the machinery of the standard normal.

What about the continuous portion of the distribution of Y?



For $Y_i > 0$, $P_i = ???$

The probability of any *exact* value in a continuous distribution is zero, but some points are more likely than others. We need to use the density function, which gives the relative likelihood of any value.

Normal Density Functions

$$w_i \sim N(\mu, \sigma^2) \rightarrow f(w) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{w_i - \mu}{\sigma}\right)^2}$$

$$z_i \sim N(0,1) \rightarrow f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \equiv \phi(z)$$

The disturbance term is a hybrid.

$$u_i \sim N(0, \sigma^2) \rightarrow f(u_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{u_i}{\sigma}\right)^2}$$

The Continuous Part

$$\begin{aligned}
 f(u_i) &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{u_i}{\sigma}\right)^2} = \frac{1}{\sigma} \left[\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{u_i}{\sigma}\right)^2} \right] \\
 &= \frac{1}{\sigma} \phi\left(\frac{u_i}{\sigma}\right) \\
 &= \frac{1}{\sigma} \phi\left(\frac{Y_i - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)
 \end{aligned}$$

$$\text{So, for } Y_i > 0, \quad P_i = \frac{1}{\sigma} \phi\left(\frac{Y_i - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)$$

Putting the Pieces Together

$$\mathcal{L} = \prod_{i=1}^n [P_i]$$

$$= \prod_{Y_i=0} \left[\Phi\left(\frac{-\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right) \right] \prod_{Y_i>0} \left[\frac{1}{\sigma} \phi\left(\frac{Y_i - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right) \right]$$

Censored Uncensored

$$\ln \mathcal{L} = \sum_{Y_i=0} \ln \left[\Phi\left(\frac{-\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right) \right] + \sum_{Y_i>0} \ln \left[\frac{1}{\sigma} \phi\left(\frac{Y_i - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right) \right]$$

Now apply maximum likelihood to estimate $\boldsymbol{\beta}$ and σ .

Tobit Analysis

```
tobit depvar [indepvars] [weight] [if exp] [in range], ll[(#)] ul[(#)]
[ level(#) offset(varname) maximize_options ]
```

```
. tobit cens0 x, ll(0)
```

```
Tobit estimates
```

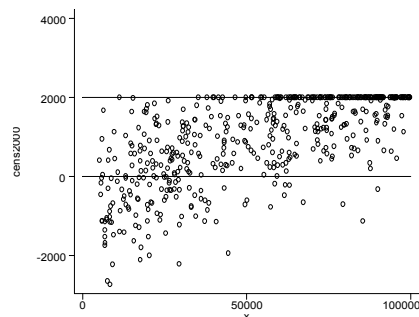
Number of obs	=	500
LR chi2(1)	=	305.98
Prob > chi2	=	0.0000
Pseudo R2	=	0.0437

Log likelihood = -3348.696

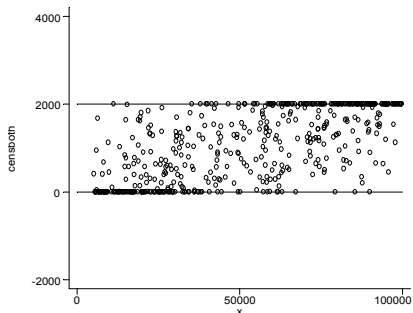
cens0	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	.0313236	.0016149	19.397	0.000	.0281508	.0344964
_cons	-553.965	100.5348	-5.510	0.000	-751.4887	-356.4412
_se	946.287	34.75281	(Ancillary parameter)			

Obs. summary: 105 left-censored observations at cens0<=0
395 uncensored observations

Censoring, continued



The maximum allowable contribution is 2000. So for people who desire to contribute more, we observe 2000.



Censoring from above and below. Note how data piles up at the limits.

Does OLS give biased results for top censoring?

Bias from Top Censoring

```
. reg cens2000 x
```

Source	SS	df	MS	Number of obs = 500		
Model	258948923	1	258948923	F(1, 498)	=	391.21
Residual	329634450	498	661916.566	Prob > F	=	0.0000
				R-squared	=	0.4400
				Adj R-squared	=	0.4388
Total	588583373	499	1179525.8	Root MSE	=	813.58

cens2000	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	.0254367	.001286	19.78	0.000	.02291	.0279635
_cons	-449.9395	77.02489	-5.84	0.000	-601.2733	-298.6057

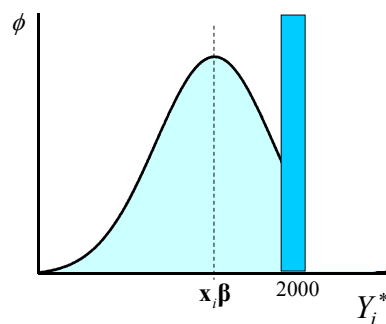
Likelihood for Top Censoring

$$\Pr(Y_i = 2000 | X_i) = \Pr(\mathbf{x}_i\boldsymbol{\beta} + u_i > 2000) = \Pr(u_i > 2000 - \mathbf{x}_i\boldsymbol{\beta})$$

$$= \Pr\left(\frac{u_i}{\sigma} > \frac{2000 - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)$$

$$= 1 - \Pr\left(\frac{u_i}{\sigma} < \frac{2000 - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)$$

$$= 1 - \Phi\left(\frac{2000 - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)$$



$$\mathcal{L} = \prod_{Y_i < 2000} \left[\frac{1}{\sigma} \phi\left(\frac{Y_i - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right) \right] \prod_{Y_i = 2000} \left[1 - \Phi\left[\frac{2000 - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right] \right]$$

The Fix using Tobit

```
. tobit cens2000 x, ul(2000)
```

Tobit regression

Number of obs = 500
 LR chi2(1) = 301.28
 Prob > chi2 = 0.0000
 Pseudo R2 = 0.0457

Log likelihood = -3145.3849

cens2000	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	.0332636	.0017547	18.96	0.000	.0298162	.0367111
_cons	-645.1107	97.89438	-6.59	0.000	-837.4467	-452.7748
/sigma	1006.317	38.64256			930.3951	1082.239

Obs. summary: 0 left-censored observations
 366 uncensored observations
 134 right-censored observations at cens2000>=2000

OLS Regression on Doubly Censored Data

```
. reg censboth x
```

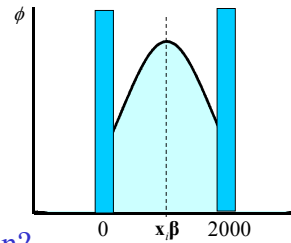
Source	SS	df	MS	Number of obs = 500		
Model	142157758	1	142157758	F(1, 498) =	403.79	
Residual	175324537	498	352057.303	Prob > F =	0.0000	
Total	317482295	499	636237.064	R-squared =	0.4478	
				Adj R-squared =	0.4467	
				Root MSE =	593.34	
censboth	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	.0188469	.0009379	20.09	0.000	.0170041	.0206896
_cons	67.05927	56.1741	1.19	0.233	-43.30818	177.4267

Regression slope is biased toward zero. Intercept is also biased.
 Standard error of the estimate (sigma) biased toward zero.

Double Censoring

Censoring above at 2000 & below at 0.

$$Y_i = \begin{cases} 0 & \text{if } Y_i^* \leq 0 \\ Y_i^* & \text{if } 0 < Y_i^* < 2000 \\ 2000 & \text{if } Y_i^* \geq 2000 \end{cases}$$



How many pieces in the likelihood function?

$$\mathcal{L} = \prod_{Y_i=0} \left[\Phi \left(\frac{-\mathbf{x}_i \boldsymbol{\beta}}{\sigma} \right) \right] \prod_{0 < Y_i < 2000} \left[\frac{1}{\sigma} \phi \left(\frac{Y_i - \mathbf{x}_i \boldsymbol{\beta}}{\sigma} \right) \right] \prod_{Y_i=2000} \left[1 - \Phi \left(\frac{2000 - \mathbf{x}_i \boldsymbol{\beta}}{\sigma} \right) \right]$$

$$\mathbf{L} = \prod_{Y_i=0} \left[\Phi \left(\frac{-\mathbf{x}_i \hat{\boldsymbol{\beta}}}{\hat{\sigma}} \right) \right] \prod_{0 < Y_i < 2000} \left[\frac{1}{\hat{\sigma}} \phi \left(\frac{Y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}}{\hat{\sigma}} \right) \right] \prod_{Y_i=2000} \left[1 - \Phi \left(\frac{2000 - \mathbf{x}_i \hat{\boldsymbol{\beta}}}{\hat{\sigma}} \right) \right]$$

Now maximize $\ln \mathbf{L}$ with respect to $\hat{\boldsymbol{\beta}}, \hat{\sigma}$

“Fixing” Upper and Lower Censoring with Tobit

```
. tobit censboth x, ll(0) ul(2000)
```

Tobit estimates

Number of obs = 500
LR chi2(1) = 294.89
Prob > chi2 = 0.0000
Pseudo R2 = 0.0591

Log likelihood = -2345.4808

censboth	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	.0328139	.0019722	16.638	0.000	.0289391	.0366888
_cons	-614.9721	112.6902	-5.457	0.000	-836.3778	-393.5664
_se	985.9899	48.16598	(Ancillary parameter)			

obs. summary: 105 left-censored observations at censboth<=0
261 uncensored observations
134 right-censored observations at censboth>=2000

Interpreting the results

What is the desired contribution for a person with \$50,000 in income?

$$\hat{Y}_i^* = -615 + 0.0328X_i \quad \hat{\sigma} = 986$$

For $X_i = 50,000$: $\hat{Y}_i^* = ?$

$$\begin{aligned}\hat{Y}_i^* &= \mathbf{x}_i \hat{\boldsymbol{\beta}} \\ &= -615 + 0.0328(50,000) \\ &= \$1,025\end{aligned}$$

But this is not the only possible outcome! Given the disturbance term, any given individual may desire to contribute more or less than the predicted amount.

Probability of a Censored Outcome

$$\widehat{\Pr}(Y_i = 0 \mid X_i = 50000) = ?$$

$$\begin{aligned}\Phi\left(\frac{-\mathbf{x}_i \hat{\boldsymbol{\beta}}}{\hat{\sigma}}\right) &= \Phi\left(\frac{-1025}{986}\right) \\ &= \Phi(-1.04) \\ &= 0.15\end{aligned}$$

$$\widehat{\Pr}(Y_i = 2000 \mid X_i = 50000) = ?$$

$$\begin{aligned}1 - \Phi\left(\frac{2000 - \mathbf{x}_i \hat{\boldsymbol{\beta}}}{\hat{\sigma}}\right) &= 1 - \Phi\left(\frac{2000 - 1025}{986}\right) \\ &= 1 - \Phi(0.99) \\ &= 0.16\end{aligned}$$

Marginal Effect of X

Effect on desired contribution:

$$Y_i^* = \mathbf{x}_i \boldsymbol{\beta} + u_i \quad \frac{\partial Y_i^*}{\partial X_k} = ? \quad \beta_k \quad \dots \text{just like OLS!}$$

Effect on probability of being censored:

$$P_0 = \Pr(Y_i = 0 | X_k) = \Phi\left(\frac{-\mathbf{x}_i \boldsymbol{\beta}}{\sigma}\right) \quad \frac{\partial P_0}{\partial X_k} = \phi\left(\frac{-\mathbf{x}_i \boldsymbol{\beta}}{\sigma}\right) \left(\frac{-\beta_k}{\sigma}\right)$$

$$P_{2000} = \Pr(Y_i = 2000 | X_k) = 1 - \Phi\left(\frac{2000 - \mathbf{x}_i \boldsymbol{\beta}}{\sigma}\right)$$

$$\frac{\partial P_{2000}}{\partial X_k} = -\phi\left(\frac{2000 - \mathbf{x}_i \boldsymbol{\beta}}{\sigma}\right) \left(\frac{-\beta_k}{\sigma}\right) = \phi\left(\frac{2000 - \mathbf{x}_i \boldsymbol{\beta}}{\sigma}\right) \left(\frac{\beta_k}{\sigma}\right)$$

Beron: Two Limit Tobit

The amount people are willing to pay is cs_i . Assume people pay only between \$1 and the amount due, $csdue_i$. If they desire to pay less than zero, they are observed at \$0. If they would be willing to pay more, they still don't pay more than $csdue_i$.

$$cspaid_i = \begin{cases} csdue_i & \text{if } cs_i^* \geq csdue_i \\ cs_i^* & \text{if } 0 < cs_i^* < csdue_i \\ 0 & \text{if } cs_i^* \leq 0 \end{cases}$$

Note the subscript “i” – *the upper limit varies* depending on the child support order in a particular case. Therefore:

$$\mathcal{L} = \prod_{Y_i=0} \left[\Phi\left(\frac{-\mathbf{x}_i \boldsymbol{\beta}}{\sigma}\right) \right] \prod_{\substack{0 < cspaid_i < csdue_i}} \left[\frac{1}{\sigma} \phi\left(\frac{cspaid_i - \mathbf{x}_i \boldsymbol{\beta}}{\sigma}\right) \right] \prod_{cspaid_i = csdue_i} \left[1 - \Phi\left(\frac{csdue_i - \mathbf{x}_i \boldsymbol{\beta}}{\sigma}\right) \right]$$

Beron's Notation (page 656)

Beron defines $\frac{\mathbf{x}_i\boldsymbol{\beta}}{\sigma} = z$, so....

$$\mathcal{L} = \prod_{Y_i=0} [\Phi(-z)] \prod_{\substack{0 < \\ cspaid_i \\ < csdue_i}} \left[\frac{1}{\sigma} \phi\left(\frac{cspaid_i}{\sigma} - z\right) \right] \prod_{\substack{cspaid_i = \\ csdue_i}} \left[1 - \Phi\left(\frac{csdue_i}{\sigma} - z\right) \right]$$

Can be estimate using “interval regression” -- `intreg`

`intreg depvar1 depvar2 [indepvars] [if] [in] [weight] [, options]`

depvar1 and depvar2 should have the following form:

Type of data		depvar1	depvar2
point data	a = [a,a]	a	a
interval data	[a,b]	a	b
left-censored data	(-inf,b]	.	b
right-censored data	[a,inf)	a	.

Intreg:

You need to create two dependent variables to specify they range which includes the true value: {Lower < DV < Upper}

	<u>lower</u>	<u>upper</u>
A single value, e.g. \$1,317, in continuous data:	1317	1317
A value in a range, say \$1-\$500:	1	500
A left censored value, -∞ up to \$0:	.	0
A right censored value, \$2000 to + ∞:	2000	.

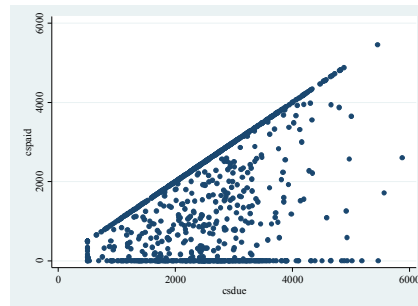
```
gen lower = cspaid if (cspaid>0) & (cspaid<=csdue)
gen upper = cspaid if (cspaid>=0) & (cspaid<csdue)
```

Beron Pseudo data

```
clear
set obs 1000
set seed 1

gen csdue = max(500, 2500 + 1000*invnorm(uniform()))
gen randnum = uniform()
gen cspaid = csdue
replace cspaid = 0 if randnum<0.25
replace cspaid = csdue*uniform() if randnum>0.75

scatter cspaid csdue
gen lower = cspaid if (cspaid>0) & (cspaid<=csdue)
gen upper = cspaid if (cspaid>=0) & (cspaid<csdue)
list csdue cspaid lower upper in 1/20
```



csdue	cspaid	lower	upper
1391.749	1391.749	1391.749	.
1403.395	1403.395	1403.395	.
2377.407	2166.737	2166.737	2166.737
500	251.4147	251.4147	251.4147
1733.313	0	.	0
2294.699	0	.	0
1824.811	0	.	0

intreg lower upper x1 x2 ...

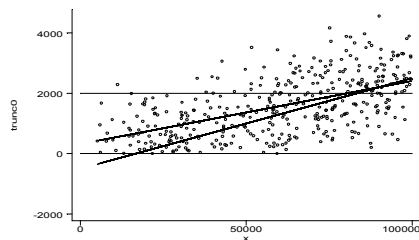
Regression on Data Truncated from Below

```
. reg trunc0 x
```

Source	SS	df	MS
Model	113701840	1	113701840
Residual	257630493	393	655548.329
Total	371332333	394	942467.851

Number of obs = 395
F(1, 393) = 173.45
Prob > F = 0.0000
R-squared = 0.3062
Adj R-squared = 0.3044
Root MSE = 809.66

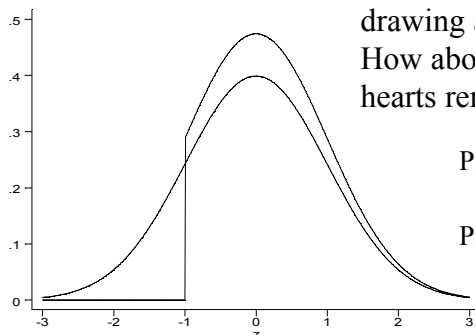
trunc0	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
x	.0209047	.0015873	13.170	0.000	.0177841 .0240254
_cons	319.3285	104.0796	3.068	0.002	114.7061 523.9509



Notice: 1) bias towards zero, 2) only 395 observations (Y is missing for the rest).

Truncated Normal Density Function

Normal curve after
truncating at $z = -1$
and renormalizing



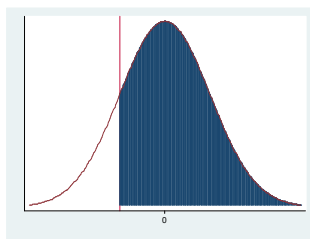
$$f(z|z > -1) = \frac{\phi(z)}{1 - \Phi(-1)} = \frac{\phi(z)}{\Phi(1)} = \frac{\phi(z)}{0.8413}$$

Example: what is the probability of
drawing a spade from a deck of cards?
How about a deck of cards with all the
hearts removed?

$$\Pr(\text{spade}) = \frac{1}{4} = 0.25$$

$$\Pr(\text{spade} | \text{no hearts}) = \frac{0.25}{0.75} = \frac{1}{3}$$

Truncation at 0 Likelihood Function



$$P_i = \frac{\frac{1}{\sigma} \phi\left(\frac{Y_i - \mathbf{x}_i \boldsymbol{\beta}}{\sigma}\right)}{1 - \Phi\left(\frac{-\mathbf{x}_i \boldsymbol{\beta}}{\sigma}\right)} = \frac{\frac{1}{\sigma} \phi\left(\frac{Y_i - \mathbf{x}_i \boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\frac{\mathbf{x}_i \boldsymbol{\beta}}{\sigma}\right)}$$

$$\mathcal{L} = \prod_{i=1}^n [P_i] = \prod_{i=1}^n \left[\frac{\frac{1}{\sigma} \phi\left(\frac{Y_i - \mathbf{x}_i \boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\frac{\mathbf{x}_i \boldsymbol{\beta}}{\sigma}\right)} \right]$$

Only one piece. All the
observed points are in
the continuous part of
the distribution.

Estimating a "Truncit"

help for truncreg (manual: [R] truncreg)

```
truncreg depvar [varlist] [weight] [if exp] [in range]
[, ll(varname | #) ul(varname | #)
noconstant offset(varname) marginal at(matname)
robust cluster(varname) score(newvarlist)
level(#) constraints(numlist) noskip nolog
maximize_options ]
```

```
. truncreg trunc0 x, ll(0)
```

Truncated regression

Limit: lower = 0

Number of obs = 395

upper = +inf

wald chi2(1) = 126.25

Log likelihood = -3169.2288

Prob > chi2 = 0.0000

	trunc0	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
<hr/>							
eq1							
	x	.029533	.0026284	11.24	0.000	.0243814	.0346845
	_cons	-422.5624	199.0235	-2.12	0.034	-812.6412	-32.48358
<hr/>							
sigma							
	_cons	939.7842	47.10231	19.95	0.000	847.4654	1032.103
<hr/>							

OLS on Upper Truncation

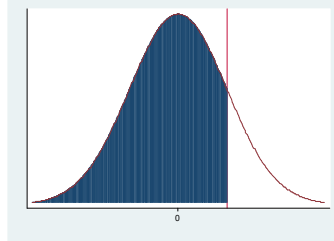
```
. reg trunc2k x
```

Source	SS	df	MS	Number of obs	=	366
Model	105189846	1	105189846	F(1, 364)	=	147.83
Residual	259003007	364	711546.723	Prob > F	=	0.0000
-----				R-squared	=	0.2888
-----				Adj R-squared	=	0.2869
Total	364192853	365	997788.639	Root MSE	=	843.53

trunc2k	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	.0206337	.001697	12.16	0.000	.0172964	.0239709
_cons	-412.3078	86.14517	-4.79	0.000	-581.7125	-242.9031

The bias is even worse than with censoring.

Upper Truncation at 2000 Likelihood Function



$$P_i = \frac{\frac{1}{\sigma} \phi\left(\frac{Y_i - \mathbf{x}_i \boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\frac{2000 - \mathbf{x}_i \boldsymbol{\beta}}{\sigma}\right)}$$

$$\mathcal{L} = \prod_{i=1}^n [P_i] = \prod_{i=1}^n \left[\frac{\frac{1}{\sigma} \phi\left(\frac{Y_i - \mathbf{x}_i \boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\frac{2000 - \mathbf{x}_i \boldsymbol{\beta}}{\sigma}\right)} \right]$$

Only one piece.
All the observed
points are in the
continuous part of
the distribution.

Fix using truncreg

```
. truncreg trunc2k x, ul(2000)
(note: 0 obs. truncated)
```

Truncated regression

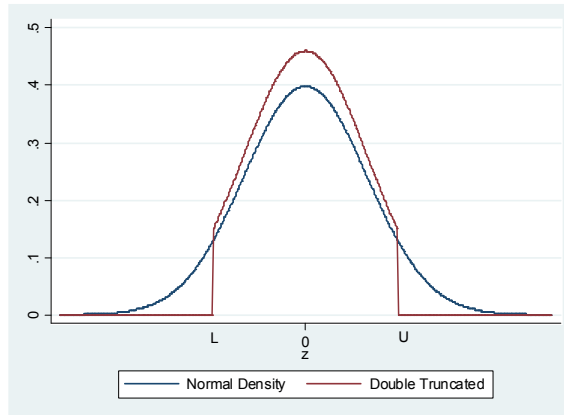
Limit: lower = -inf
upper = 2000
Log likelihood = -2939.6585

Number of obs = 366
wald chi2(1) = 98.32
Prob > chi2 = 0.0000

trunc2k	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
x	.0330026	.0033283	9.92	0.000	.0264793	.0395258
_cons	-607.5885	121.5944	-5.00	0.000	-845.9091	-369.2678
/sigma	1027.542	58.65219	17.52	0.000	912.5859	1142.498

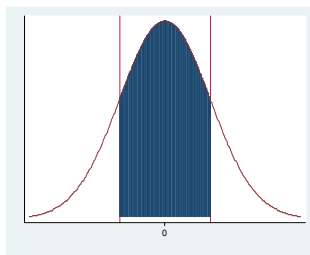
Double Truncation

- Truncation from above (2000) and below (0).
- Observations missing beyond limits.
- What is the likelihood function?



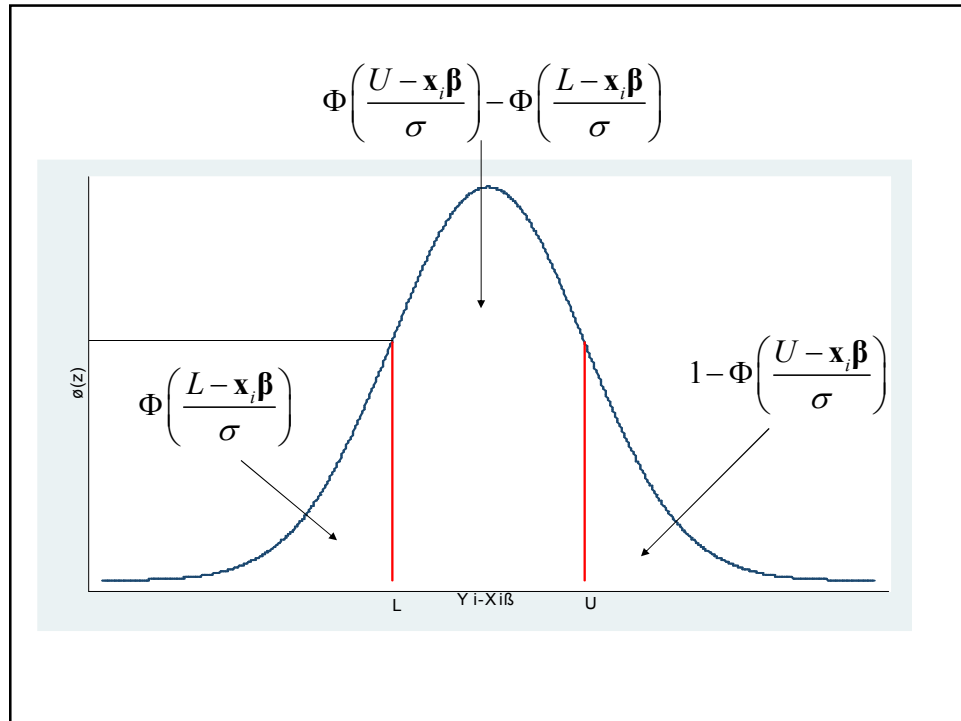
Variable	Obs	Mean	Std. Dev.	Min	Max
truncboth	261	1007.618	556.0427	1.898094	1998.137

Likelihood for Double Truncation



$$\mathcal{L} = \prod_{i=1}^n \left[\frac{\frac{1}{\sigma} \phi\left(\frac{Y_i - \mathbf{x}_i \boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\frac{U - \mathbf{x}_i \boldsymbol{\beta}}{\sigma}\right) - \Phi\left(\frac{L - \mathbf{x}_i \boldsymbol{\beta}}{\sigma}\right)} \right]$$

$$= \prod_{i=1}^n \left[\frac{\frac{1}{\sigma} \phi\left(\frac{Y_i - \mathbf{x}_i \boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\frac{2000 - \mathbf{x}_i \boldsymbol{\beta}}{\sigma}\right) - \Phi\left(\frac{-\mathbf{x}_i \boldsymbol{\beta}}{\sigma}\right)} \right]$$



Results

. reg truncboth x				Number of obs = 261	
Source	SS	df	MS	F(1, 259)	= 32.93
Model	9067352	1	9067352	Prob > F	= 0.0000
Residual	71320352.6	259	275368.157	R-squared	= 0.1128
Total	80387704.6	260	309183.479	Adj R-squared	= 0.1094
				Root MSE	= 524.76
truncboth	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
x	.0075317	.0013125	5.74	0.000	.0049471 .0101162
_cons	620.947	74.80434	8.30	0.000	473.6449 768.2492
. truncreg truncboth x, ll(0) ul(2000)				Number of obs = 261	
Limit: lower = 0				wald chi2(1)	= 5.31
upper = 2000				Prob > chi2	= 0.0212
Log likelihood = -1967.4219					
truncboth	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
x	.0270894	.0117534	2.30	0.021	.0040531 .0501257
_cons	-361.9088	604.5535	-0.60	0.549	-1546.812 822.9943
/sigma	1003.06	228.4255	4.39	0.000	555.3546 1450.766

IRA Data

What is the likelihood for the IRA data analysis
(truncated at 0, censored at 2000)? Write it down...