Models for Censored and Truncated Data

Ass.Prof. Andriy Stavytskyy

Agenda

- Difference between censored and truncated data
- Truncated regression
- Tobit-model
- Heckman Selection Model

Difference between censored and truncated data

Censored Data: Definitions

- Y is censored when we observe X for all observations, but we only know the true value of Y for a restricted range of observations. Values of Y in a certain range are reported as a single value or there is significant clustering around a value, say 0.
  - If Y=k or Y>k for all Y =>Y is censored from below or left-censored.
  - If Y=k or Y<k for all Y =>Y is censored from above or right-censored.

- We usually think of an uncensored Y, Y*, the true value of Y when the censoring mechanism is not applied. We typically have all the observations for {Y,X}, but not {Y*,X}.

Truncated Data

- Y is truncated when we only observe X for observations where Y would not be censored. We do not have a full sample for {Y,X}, we exclude observations based on characteristics of Y.

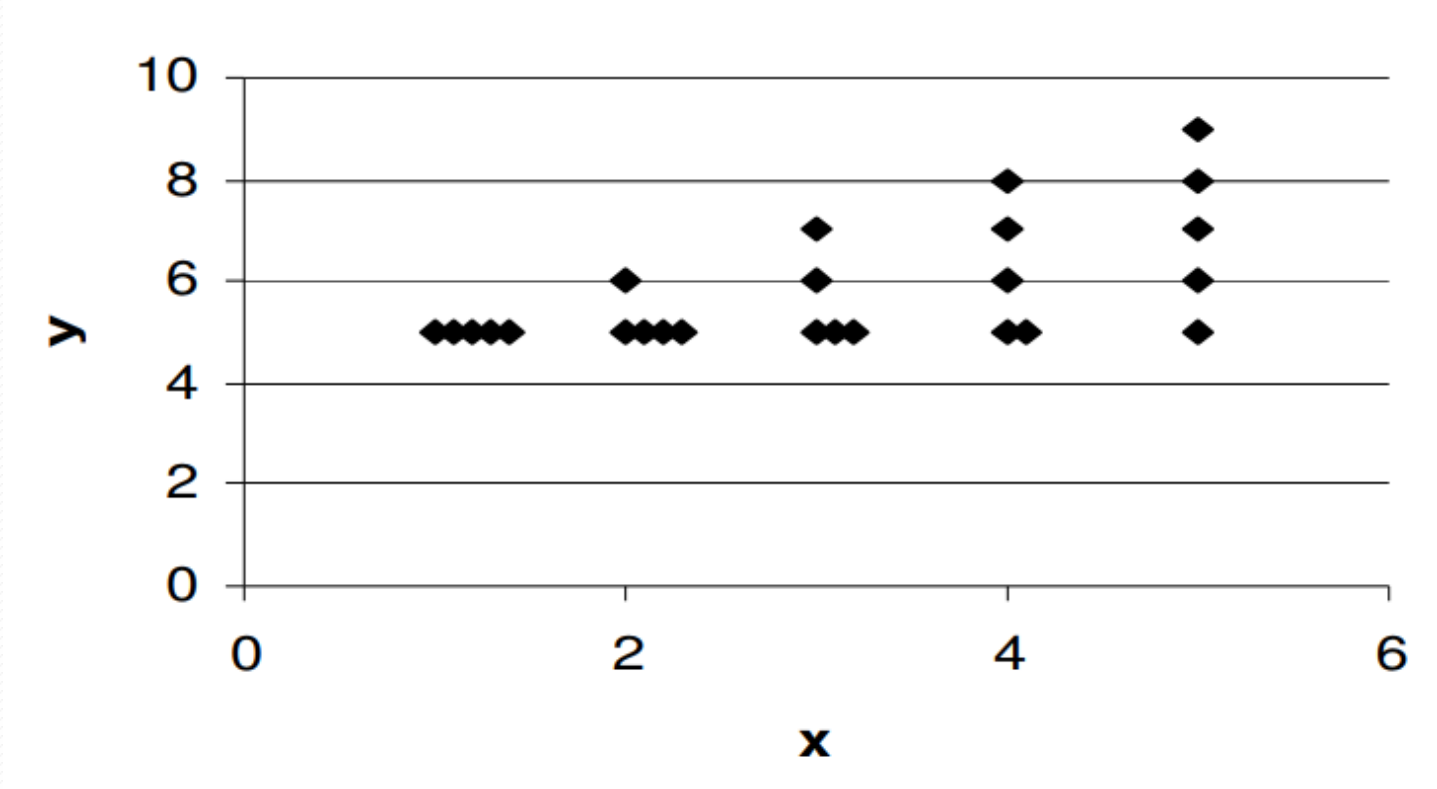
Censored from below: Example - 1

- A Central Bank intervenes if the exchange rate hits the band's lower limit.

$$\text{If } S_t \leq \bar{E} \Rightarrow S_t = \bar{E}.$$

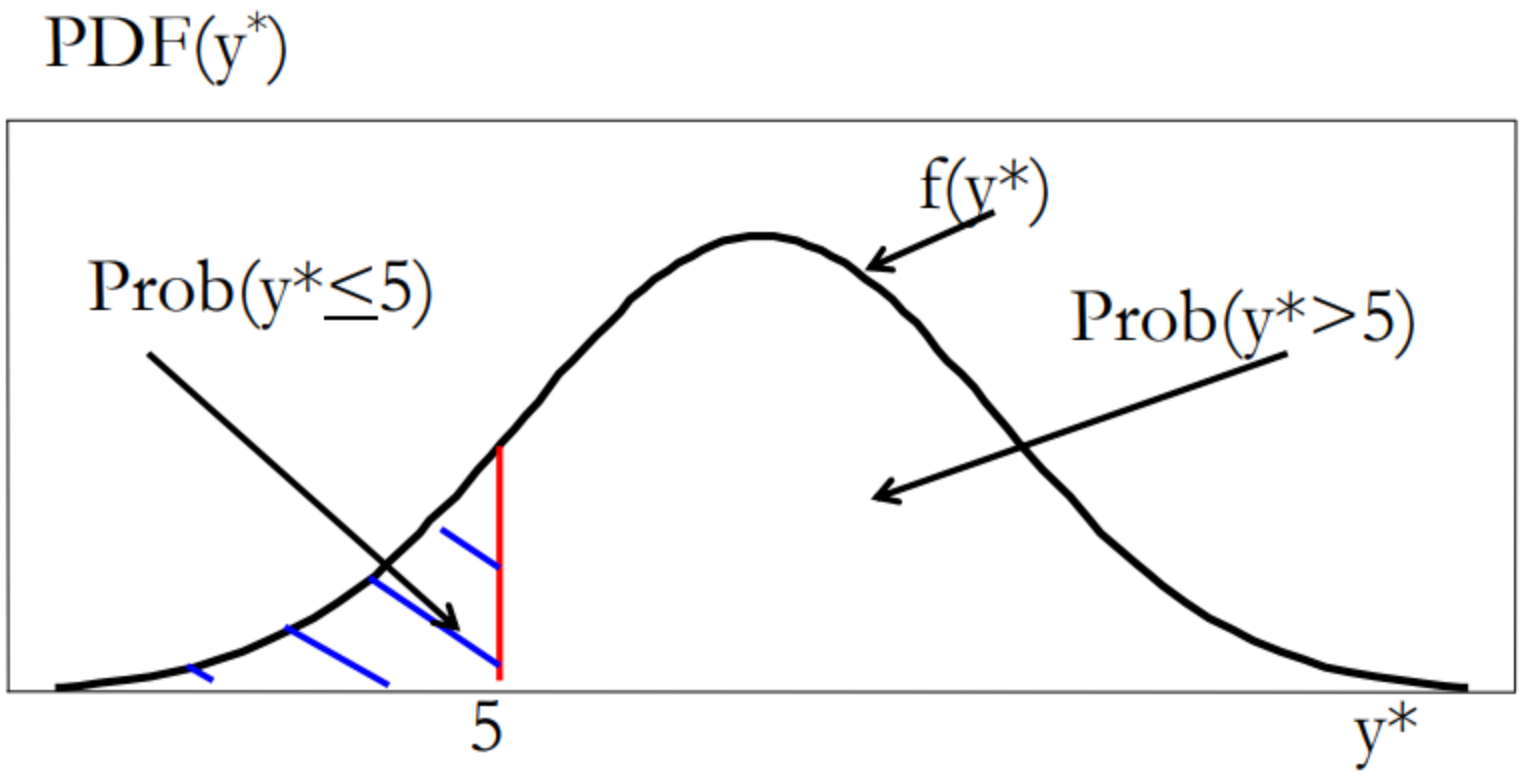
- If $Y \leq 5$, we do not know its exact value.

Censored from below: Example – 2



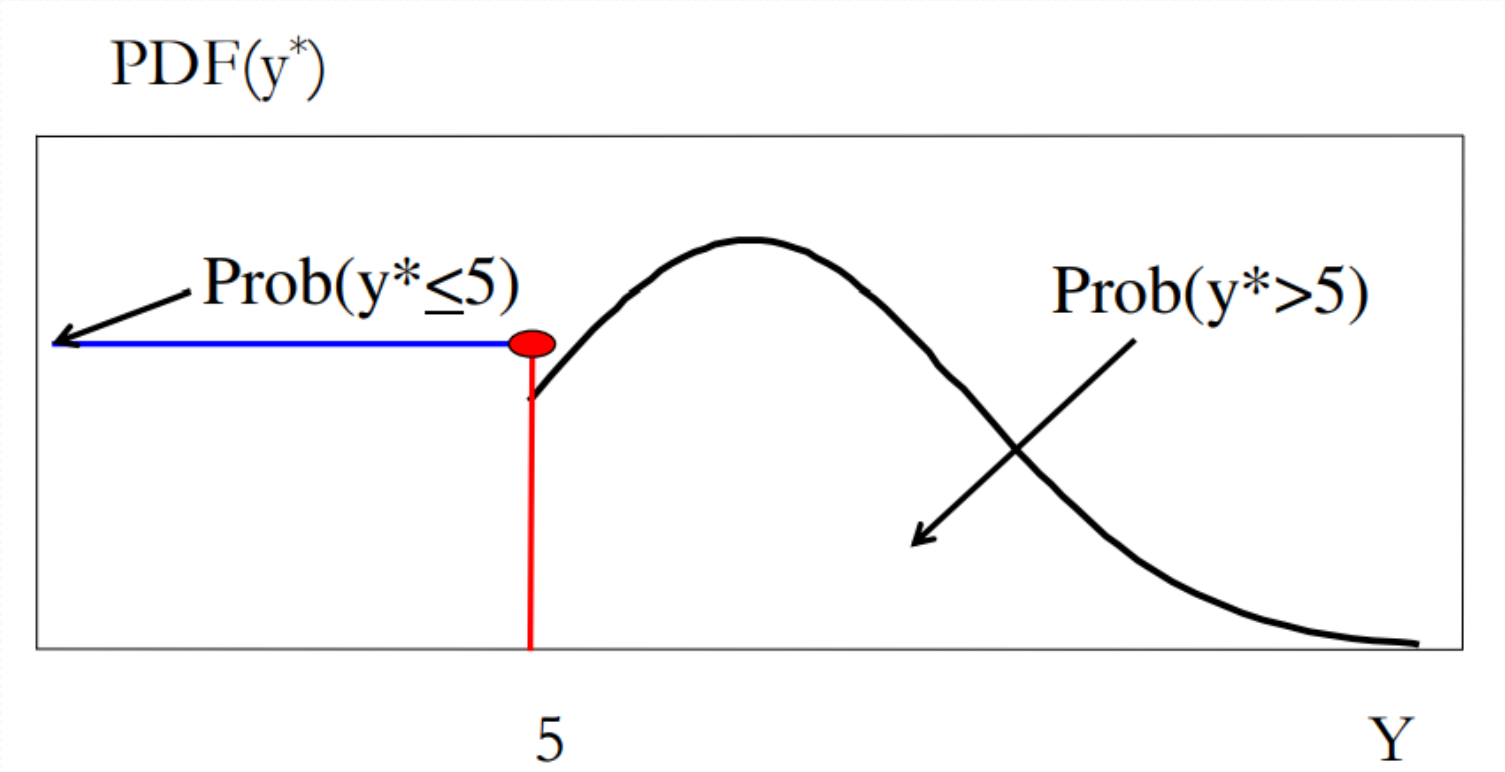
Censored from below: Example – 3

- The pdf of the observable variable, y, is a mixture of discrete (prob. mass at Y=5) and continuous (Prob[Y*>5]) distributions.



Censored from below: Example – 4

- Under censoring we assign the full probability in the censored region to the censoring point, 5.



New example

- Consumer maximizes utility by purchasing durable goods under constraint that total expenditures do not exceed income

  expenditure of durables $\geq$ cost of least expensive durable good

- If available income is less than least expensive durable good then no expenditure is observed. Don't know how much a household would have spent if a durable good could be purchased for less than the least expensive item.

One more

- Model how much an individual spends on alcohol in a given month.
- A significant fraction have zero expenditure.

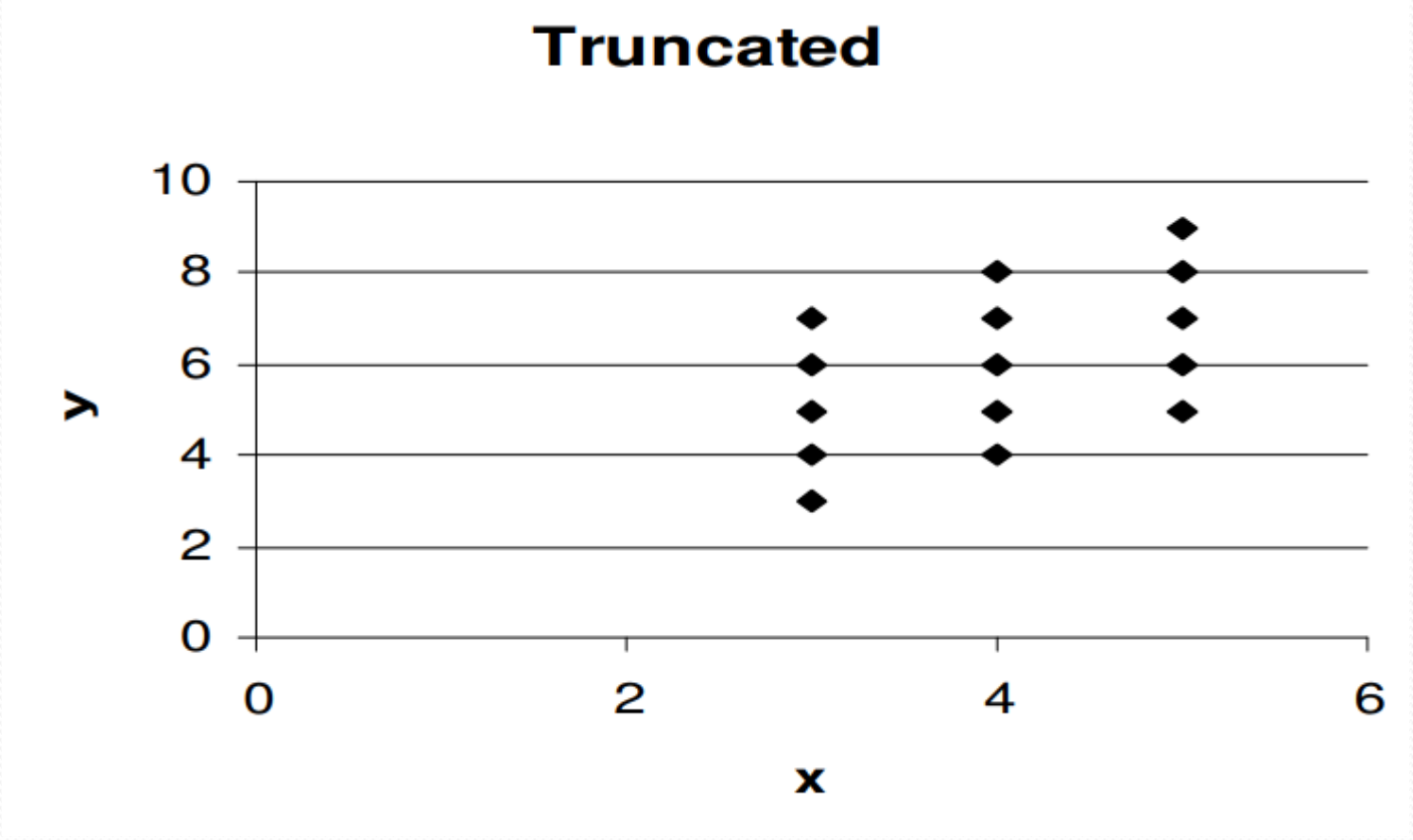
Truncated Data: Example – 1

- If a family's income is below certain level, we have no information about the family's characteristics.

If $Y < 3$, the value of X (or Y) is unknown.

(Truncation from below.)

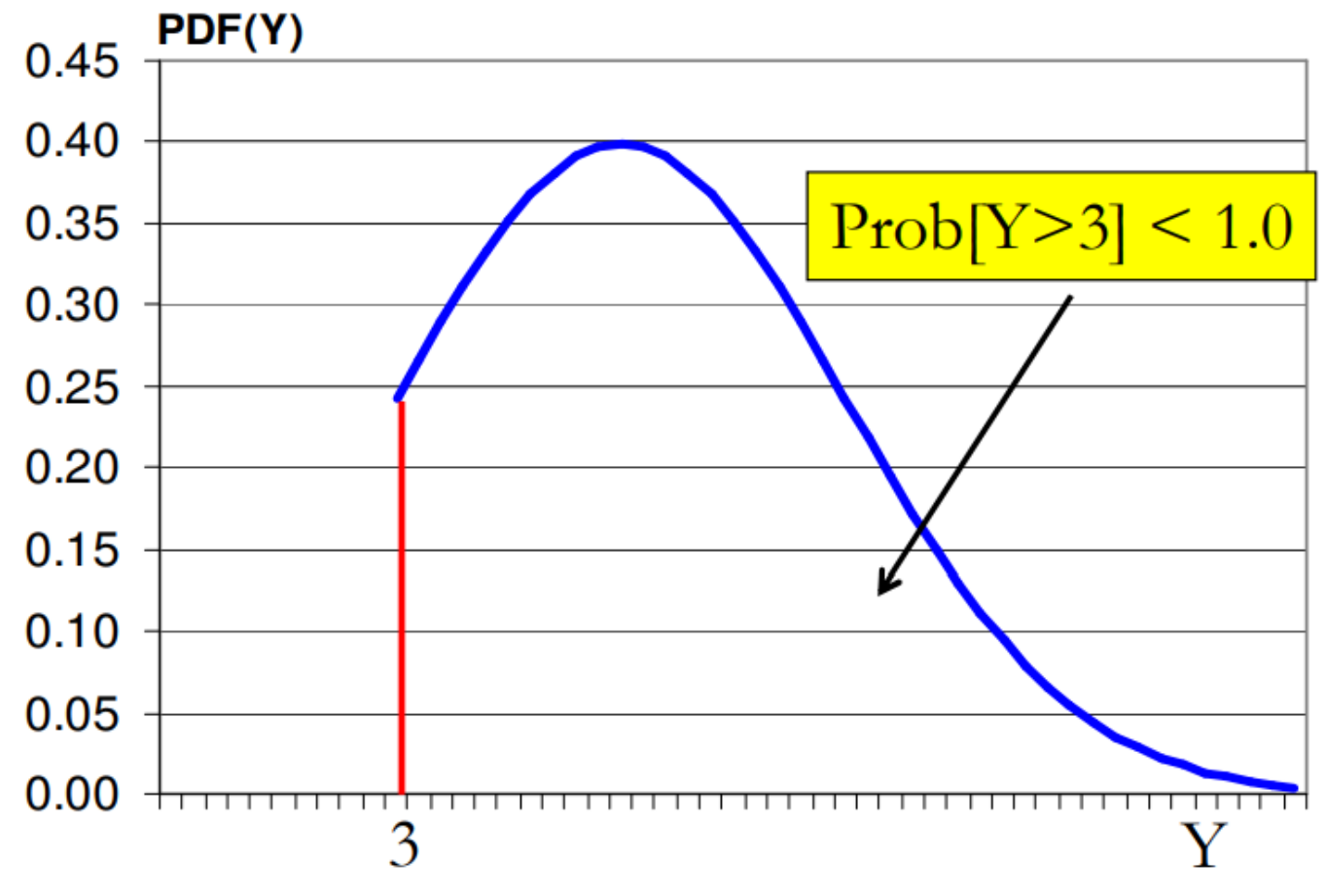
Truncated Data: Example – 2



Truncated Data: Example – 3

- Under data censoring, the censored distribution is a combination of a pmf plus a pdf. They add up to 1. We have a different situation under truncation. To create a pdf for Y we will use a conditional pdf.

Truncated Data: Example – 4



Truncated regression

Truncated regression

- Truncated regression is different from censored regression in the following way:
  - Censored regressions: The dependent variable may be censored, but you can include the censored observations in the regression
  - Truncated regressions: A subset of observations are dropped, thus, only the truncated data are available for the regression.

Why do we have truncation?

- Truncation by survey design:
  - Studies of poverty. By survey's design, families whose incomes are greater than that threshold are dropped from the sample.
- Incidental Truncation:
  - Wage offer married women. Only those who are working has wage information. It is the people's decision, not the survey's design, that determines the sample selection.

Example

- Hausman and Wise's analyse the New Jersey negative income tax experiment
- Goal: Estimate earnings function for low income individuals
- Truncation: Individuals with earnings greater than 1.5×poverty level were excluded from the sample.
- Two types of inference:
  - Inference about entire population in presence of truncation
  - Inference about sub-population observed

What happens when we apply OLS to a truncated data?

- Suppose that you consider the following regression:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i ,$$

- We have a random sample of size N.
- All assumptions are satisfied. (The most important assumption is $E(\varepsilon_i|x_i)=0$.)
- Instead of using all the N observations, we use a subsample.
- Then, run OLS using this sub-sample (truncated sample) only.

Under what conditions, does sample selection matter to OLS?

- OLS is Unbiased
- Sample selection is randomly done.
- Sample selection is determined solely by the value of **x-variable**. For example, suppose that x is age. Then if you select sample if age is greater than 20 years old, this OLS is unbiased.

Truncation and OLS – 1

- OLS is Biased
- Sample selection is determined by the value of y-variable.
  - Example: Y is family income. We select the sample if y is greater than certain threshold. Then this OLS is biased.
- Sample selection is correlated with ε_i.
  - Example: We run a wage regression $w_i = \beta_0 + \beta_1 \, educ_i + \varepsilon_i$, where ε_i contains unobserved ability. If sample is selected based on the unobserved ability, this OLS is biased.
  - In practice, this situation happens when the selection is based on the survey participant's decision. Since the decision to participate is likely to be based on unobserved factors which are contained in ε, the selection is likely to be correlated with ε_i.

Truncation and OLS – 2

- Consider the previous regression:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

- All CLM assumptions are satisfied.
- Instead of using all the N observations, we use a subsample. Let s_i be a selection indicator: If $s_i=1$, then person i is included in the regression. If $s_i=0$, then person i is dropped from the data.
- If we run OLS using the selected subsample, we use only the observation with $s_i=1$. That is, we run the following regression:

$$s_i y_i = \beta_0 s_i + \beta_1 s_i x_i + s_i \varepsilon_i$$

- Now, $s_i x_i$ is the explanatory variable, and $u_i = s_i \varepsilon_i$ is the error term.
- OLS is unbiased if $E(u_i = s_i \varepsilon_i | s_i x_i) = 0$.
- We need check under what conditions the new condition is satisfied.

Truncation and OLS – 3

- It is sufficient to check: $E(u_i|s_i x_i)=0$. (If this is zero, then new condition is also zero.)
- $E(u_i|x_i,s_i) = s_i E(\varepsilon_i|x_i,s_i)$ - s_i is in the conditional set.
- It is sufficient to check the condition which ensures $E(u_i|x_i, s_i)=0$.

Cases – 1

- Sample selection is done randomly
  - s is independent of ε and x. => $E(\varepsilon|x,s)=E(\varepsilon|x)$. Since the assumptions are satisfied => we have $E(\varepsilon|x)=0$. => OLS is unbiased

Cases – 2

- Sample is selected based solely on the value of x-variable.
  - Example: We study trading in stocks, y_i. One of the dependent variables, x_i, is wealth, and we select person i, if wealth is greater than 50K. Then,

$$s_i=1 \text{ if } x_i \geq 50K,$$

$$s_i=0 \text{ if } x_i.$$

  - Now, s_i is a deterministic function of x_i .
  - Since s is a deterministic function of x, it drops out from the conditioning set. Then,

$$E(\varepsilon|x, s) = E(\varepsilon|x, s(x)) = E(\varepsilon|x) = 0$$

  - CLM assumptions satisfied.
  - OLS is unbiased.

Cases – 3

- Sample selection is based on the value of y-variable
  - Example: We study determinants of wealth, Y. We select individuals whose wealth is smaller than 150K. Then, $s_i=1$ if y_i <150K.
  - Now, s_i depends on y_i (and ε_i). It cannot be dropped out from the conditioning set like we did before. Then, $E(\varepsilon|x, s)\neq E(\varepsilon|x) = 0$.
  - OLS is biased.

Cases – 4

- Sample selection is correlated with u_i.
  - The inclusion of a person in the sample depends on the person's decision, not the surveyor's decision. This type of truncation is called the incidental truncation. The bias that arises from this type of sample selection is called the Sample Selection Bias.
  - Example: wage offer regression of married women:

$$wage_i = \beta_0 + \beta_1 edu_i + \varepsilon_i.$$

  - Since it is the woman's decision to participate, this sample selection is likely to be based on some unobservable factors which are contained in ε_i. s cannot be dropped out from the conditioning set:
  - $E(\varepsilon|x, s) \neq E(\varepsilon|x) = 0$
  - OLS is biased.

Cases – 5

- The selection rule based on the x-variable may be correlated with ε_i.
  - Example: X is IQ. A survey participant responds if IQ > v. Now, the sample selection is based on x-variable and a random error v.
  - Two cases:
    - If v is independent of ε, then it does not cause a bias.
    - If v is correlated with ε, then this OLS will be biased.

Estimation with Truncated Data

- Under cases (1), (2), (5) OLS is appropriate.
- Under case (3), we use Truncated regression.
- Under case (4) –i.e., incidental truncation-, we use the Heckman Sample Selection Correction method. This is also called the Heckit model.

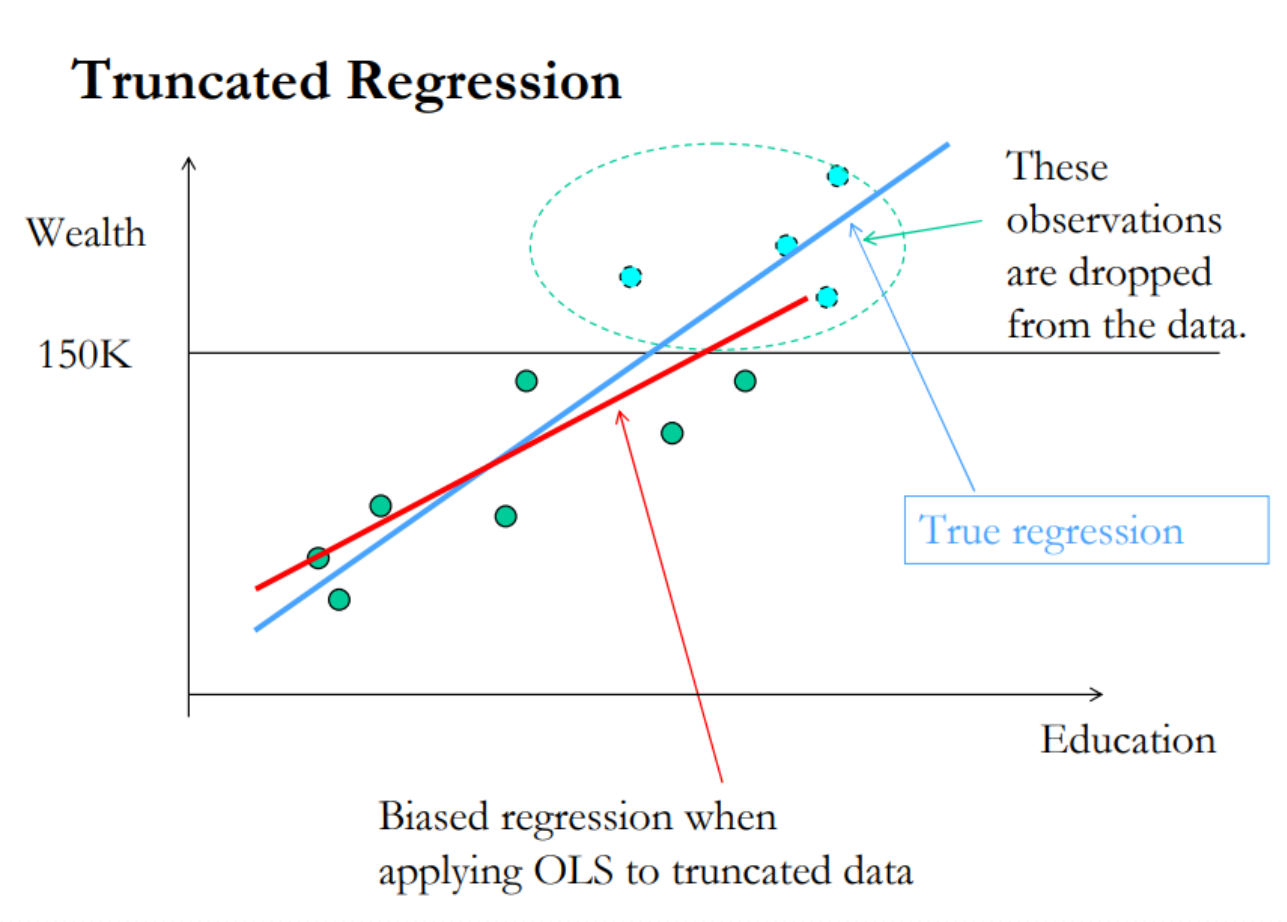
Truncated Regression – 1

- The truncation is based on the y-variable
- We have the following regression satisfies all CLM assumptions:

$$y_i = x_i'\beta + \varepsilon_i \,,\ \varepsilon_i \sim N(0, \sigma^2)$$

- We sample only if $y_i < c_i$
- Observations dropped if $y_i \geq c_i$ by design.
- We know the exact value of c_i for each person.
- We know that OLS on the truncated data will cause biases. The model that produces unbiased estimate is based on the ML Estimation.

Truncated Regression – 2

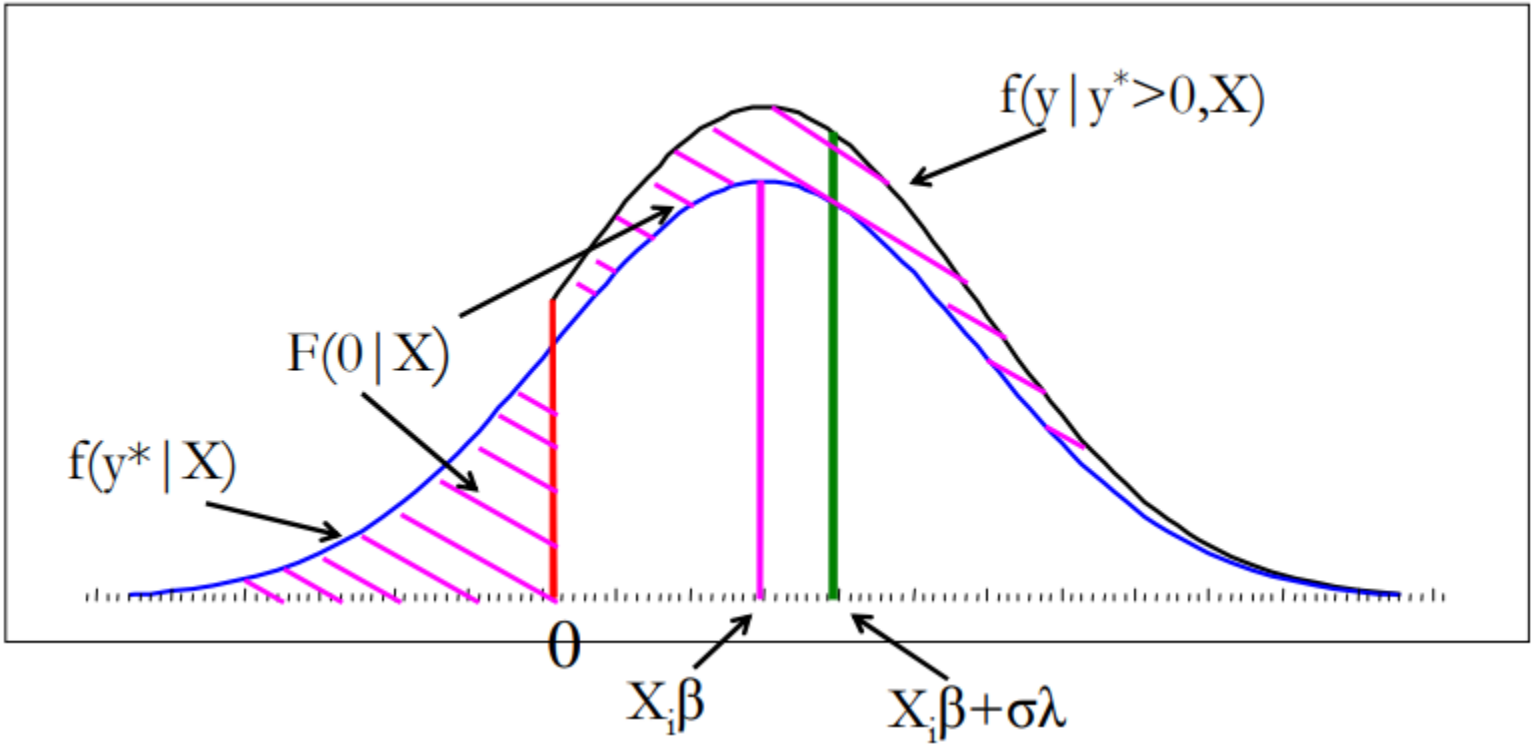


Truncated Normal: Moments

- Let $y^* \sim N(\mu^*, \sigma^2)$ and $\alpha = (c - \mu^*)/\sigma$.
- First moment: $E[y^* | y > c] = \mu^* + \sigma \lambda(\alpha)$ <= This is the truncated regression.
  - If $\mu^* > 0$ and the truncation is from below, i.e., $\lambda(\alpha) > 0$, the mean of the truncated variable is greater than the original mean
  - Note: For the standard normal distribution $\lambda(\alpha)$ is the mean of the truncated distribution.
- Second moment:
  - $Var[y^* | y > c] = \sigma^2[1 - \delta(\alpha)]$, where $\delta(\alpha) = \lambda(\alpha)[\lambda(\alpha) - \alpha]$
  - Truncation reduces variance! This result is general, it applies to upper or lower truncation given that $0 \leq \delta(\alpha) \leq 1$

Truncated Normal – 1

- Model: $y_i^* = X_i\beta + \varepsilon_i$
- Data: $y = y^* \mid y^* > 0$



Truncated Normal – 2

- Truncated (from below –i.e., y*>0) regression model:

$$E(y_i | y_i^* > 0, X_i) = X_i\beta + \sigma\lambda_i > E(y_i|X_i)$$

Truncated Regression: ML Estimation

- The likelihood contribution for ith observation is given by

$$L_i = \frac{\frac{1}{\sigma}\phi\left(\frac{y_i - x_i'\beta}{\sigma}\right)}{\Phi\left(\frac{x_i'\beta}{\sigma}\right)}$$

- The likelihood function is given by

$$\ln L(\beta, \sigma) = \sum_{i=1}^{N} \ln L_i \rightarrow \max$$

- The values of (β, σ) that maximizes LnL are the ML estimators of the Truncated Regression.

The partial effects

- The estimated β_k shows the effect of x_{ki} on y_i. Thus,

$$\frac{\partial E\left(y_i \mid X_i, y_i^* > 0\right)}{\partial X_{k,i}} = \beta_k + \frac{\partial E\left(\varepsilon_i \mid y_i^* > 0\right)}{\partial X_{k,i}} =$$

$$= \beta_k + \sigma \frac{\partial \lambda}{\partial X_{k,i}} = \beta_k + \sigma\left(\lambda_i^2 - \alpha_i \lambda_i\right)\left(-\frac{\beta_k}{\sigma}\right) =$$

$$= \beta_k \left(1 - \lambda_i^2 - \alpha_i \lambda_i\right) = \beta_k \left(1 - \delta_i\right),$$

$$where\ \delta_i = \lambda\left(\alpha_i\right)\left(\lambda\left(\alpha_i\right) + \alpha_i\right), 0 < \delta_i < 1.$$

Tobit-model

Example – 1

- A popularly used model in these situations is the **Tobit model**, which was originally developed by James Tobin, a Nobel laureate economist.
- For this purpose we use the data collected by Mroz. His sample gives data on 753 married women, 428 of whom worked outside the home and 325 of whom did not work outside the home, and hence had zero hours of work.

OLS estimation of the hours worked function

Dependent Variable: HOURS
Method: Least Squares
Sample: 1 753
Included observations: 753

	Coefficient	Std. Error	t-Statistic	Prob.
C	1298.293	231.9451	5.597413	0.0000
AGE	−29.55452	3.864413	−7.647869	0.0000
EDUC	5.064135	12.55700	0.403292	0.6868
EXPER	68.52186	9.398942	7.290380	0.0000
EXPERSQ	−0.779211	0.308540	−2.525480	0.0118
FAMINC	0.028993	0.003201	9.056627	0.0000
KIDSLT6	−395.5547	55.63591	−7.109701	0.0000
HUSWAGE	−70.51493	9.024624	−7.813615	0.0000

R-squared	0.338537	Mean dependent var	740.5764
Adjusted R-squared	0.332322	S.D. dependent var	871.3142
S.E. of regression	711.9647	Akaike info criterion	15.98450
Sum squared resid	3.78E+08	Schwarz criterion	16.03363
Log likelihood	−6010.165	Hannan−Quinn criter.	16.00343
F-statistic	54.47011	Durbin−Watson stat	1.482101
Prob(F-statistic)	0.000000		

Analysis of OLS

- The results in this table are to be interpreted in the framework of the standard linear regression model.
- For example, if husband's wages go up by a dollar, the average hours worked by married women declines by about 71 hours, ceteris paribus.
- Except for education, all the other coefficients seem to be highly statistically significant.
- But beware of these results, for in our sample 325 married women had zero hours of work.

OLS estimation of hours function for working women only

Dependent Variable: HOURS
Method: Least Squares
Sample: 1 428
Included observations: 428

	Coefficient	Std. Error	t-Statistic	Prob.
C	1817.334	296.4489	6.130345	0.0000
AGE	−16.45594	5.365311	−3.067100	0.0023
EDUC	−38.36287	16.06725	−2.387644	0.0174
EXPER	49.48693	13.73426	3.603174	0.0004
EXPERSQ	−0.551013	0.416918	−1.321634	0.1870
FAMINC	0.027386	0.003995	6.855281	0.0000
KIDSLT6	−243.8313	92.15717	−2.645821	0.0085
HUSWAGE	−66.50515	12.84196	−5.178739	0.0000

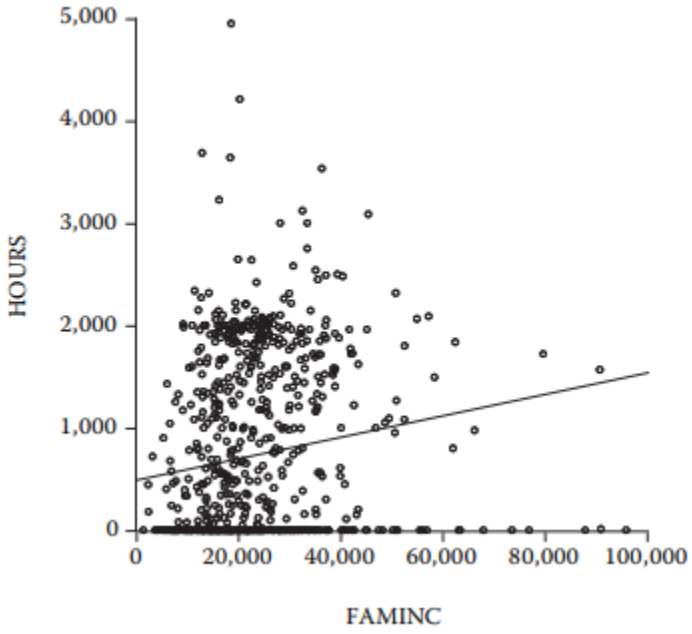
R-squared	0.218815	Mean dependent var	1302.930
Adjusted R-squared	0.205795	S.D. dependent var	776.2744
S.E. of regression	691.8015	Akaike info criterion	15.93499
Sum squared resid	2.01E+08	Schwarz criterion	16.01086
Log likelihood	−3402.088	Hannan−Quinn criter.	15.96495
F-statistic	16.80640	Durbin−Watson stat	2.107803
Prob(F-statistic)	0.000000		

Analysis

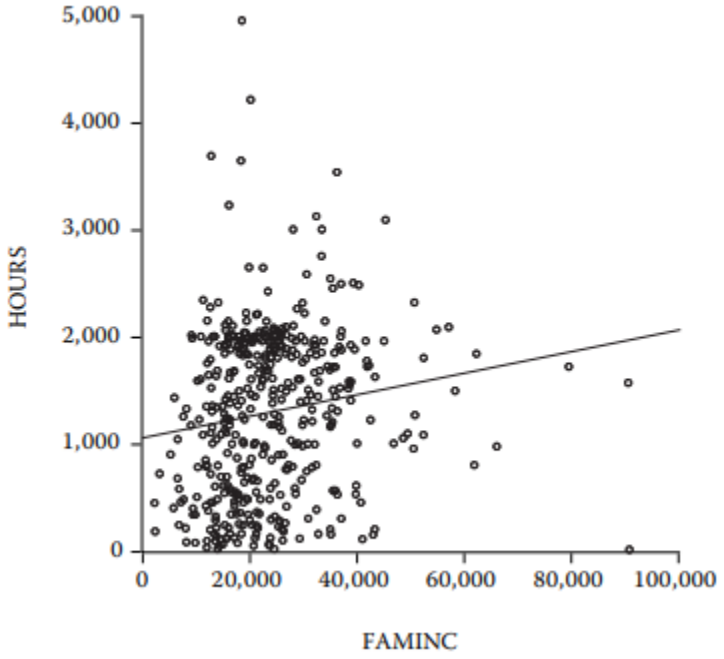
- The education variable now seems to be highly significant, although it has a negative sign.
- This is because OLS estimates of censored regression models, whether we include the whole sample or a subset of the sample, are biased as well as inconsistent – that is, no matter how large the sample size is, the estimated parameters will not converge to their true values.

Graphical analysis

- Hours worked and family income, full sample



- Hours vs. family income for working women



In the left figure there are several observations (actually 325) that lie on the horizontal axis because for these observations the hours worked are zero. In the right figure none of the observations lie on the horizontal axis, for these observations are for 428 working women. The slope coefficients of the regression lines in the two figures will obviously be different.

The Tobit model

- The Y_i^* are desired hours of work. Now

$$Y_i = \begin{cases} 0, if\ Y_i^* \le 0, \\ Y_i^*, if\ Y_i^* \ge 0. \end{cases}$$

- The variable Y_i^* is called a latent variable, the variable of primary interest.

ML estimation of the censored regression model

Dependent Variable: HOURS
Method: ML - Censored Normal (TOBIT) (Quadratic hill climbing)
Sample: 1 753
Included observations: 753
Left censoring (value) at zero
Convergence achieved after 6 iterations
Covariance matrix computed using second derivatives

	Coefficient	Std. Error	z-Statistic	Prob.
C	1126.335	379.5852	2.967279	0.0030
AGE	–54.10976	6.621301	–8.172074	0.0000
EDUC	38.64634	20.68458	1.868365	0.0617
EXPER	129.8273	16.22972	7.999356	0.0000
EXPERSQ	–1.844762	0.509684	–3.619422	0.0003
FAMINC	0.040769	0.005258	7.754009	0.0000
KIDSLT6	–782.3734	103.7509	–7.540886	0.0000
HUSWAGE	–105.5097	15.62926	–6.750783	0.0000

Error Distribution				
SCALE:C(9)	1057.598	39.06065	27.07579	0.0000
Mean dependent var	740.5764	S.D. dependent var	871.3142	
S.E. of regression	707.2850	Akaike info criterion	10.08993	
Sum squared resid	3.72E+08	Schwarz criterion	10.14520	
Log likelihood	–3789.858			
Avg. log likelihood	–5.033012			
Left censored obs	325	Right censored obs	0	
Uncensored obs	428	Total obs	753	

Interpretation of the Tobit estimates

- For example, if the husband's wages go up, on average, a woman will work less in the labor market, ceteris paribus.
- The education variable is not significant in OLS, but it is in censored OLS, although it has a negative sign. Now it is significant and has a positive sign, which makes sense.

But…

- We cannot interpret the Tobit coefficient of a regressor as giving the marginal impact of that regressor on the mean value of the observed regressand. This is because in the Tobit type censored regression models a unit change in the value of a regressor has two effects:
  - the effect on the mean value of the observed regressand,
  - the effect on the probability that Y_i * is actually observed.

Age impact

- Take for instance the impact of age. The coefficient for age of about –54, it means that, holding other variables constant, if age increases by a year, its direct impact on the hours worked per year will be a decrease by about 54 hours per year and the probability of a married woman entering the labor force will also decrease.
- So we have to multiply –54 by the probability that this will happen. Unless we know the latter, we will not able to compute the aggregate impact of an increase in age on the hours worked. And this probability calculation depends on all the regressors in the model and their coefficients.
- Interestingly, the slope coefficient gives directly the marginal impact of a regressor on the latent variable Y_i^*. Thus, the coefficient of the age variable of –54 means if age increases by a year, the desired hours of work will decrease by 54 hours, ceteris paribus. Of course, we do not actually observe the desired hours of work, for it is an abstract construct.
- In our example we have 753 observations. It is a laborious task to compute the marginal impact of each regressor for all the 753 observations. In practice, one can compute the marginal impact at the average value of each regressor.

Non-normality of error term

- In the censored regression models under non-normality of the error term the estimators are not consistent.
- Again, some remedial methods are suggested in the literature. One is to change the error distribution assumption. For example, Eviews can estimate such regression models under different probability distribution assumptions for the error term (such as logistic and extreme value).

Heteroscedasticity

- In the usual linear regression model, if the error term is heteroscedastic, the OLS estimators are consistent, though not efficient. In Tobit-type models, however, the estimators are neither consistent nor efficient.

Robust estimation of the Tobit model

Dependent Variable: HOURS
Method: ML – Censored Normal (TOBIT) (Quadratic hill climbing)
Sample: 1 753
Included observations: 753
Left censoring (value) at zero
Convergence achieved after 6 iterations
QML (Huber/White) standard errors & covariance

	Coefficient	Std. Error	z-Statistic	Prob.
C	1126.335	386.3109	2.915618	0.0035
AGE	−54.10976	6.535741	−8.279056	0.0000
EDUC	38.64634	20.30712	1.903094	0.0570
EXPER	129.8273	17.27868	7.513728	0.0000
EXPERSQ	−1.844762	0.536345	−3.439505	0.0006
FAMINC	0.040769	0.005608	7.269982	0.0000
KIDSLT6	−782.3734	104.6233	−7.478004	0.0000
HUSWAGE	−105.5097	16.33276	−6.460007	0.0000

Error Distribution

SCALE:C(9)	1057.598	42.80938	24.70482	0.0000
Mean dependent var	740.5764	S.D. dependent var		871.3142
S.E. of regression	707.2850	Akaike info criterion		10.08993
Sum squared resid	3.72E+08	Schwarz criterion		10.14520
Log likelihood	−3789.858	Avg. log likelihood		−5.033012
Left censored obs	325	Right censored obs		0
Uncensored obs	428	Total obs		753

Truncated sample regression models

- In our illustrative example, we do not have data on hours worked for 325 women. Therefore we may not consider information about socio-economic variables for these observations, even though we have that information on them in the current example.
- However, the OLS estimators are inconsistent in this situation.

Truncated normal distribution

Dependent Variable: HOURS
Method: ML – Censored Normal (TOBIT) (Quadratic hill climbing)
Sample (adjusted): 1 428
Included observations: 428 after adjustments
Truncated sample
Left censoring (value) at zero
Convergence achieved after 6 iterations
QML (Huber/White) standard errors & covariance

	Coefficient	Std. Error	z-Statistic	Prob.
C	1864.232	397.2480	4.692867	0.0000
AGE	–22.88776	7.616243	–3.005125	0.0027
EDUC	–50.79302	20.77250	–2.445205	0.0145
EXPER	73.69759	22.42240	3.286784	0.0010
EXPERSQ	–0.954847	0.575639	–1.658761	0.0972
FAMINC	0.036200	0.006947	5.210857	0.0000
KIDSLT6	–391.7641	193.4270	–2.025385	0.0428
HUSWAGE	–93.52777	19.11320	–4.893360	0.0000

Error Distribution			
SCALE:C(9)	794.6310	56.36703	14.09744
0.0000			
Mean dependent var	1302.930	S.D. dependent var	776.2744
S.E. of regression	696.4534	Akaike info criterion	15.78988
Sum squared resid	2.03E+08	Schwarz criterion	15.87524
Log likelihood	–3370.035	Avg. log likelihood	–7.873913
Left censored obs	0	Right censored obs	0
Uncensored obs	428	Total obs	428

- If we compare the results of the censored regression with the truncated regression here, we will see differences in the magnitude and statistical significance of the coefficients.
- Notice particularly that the education coefficient is positive in the censored regression model, but is negative in the truncated regression model.

Interpretation of the truncated regression coefficients

- As in the Tobit model, an individual regression coefficient measures the marginal effect of that variable on the mean value of the regressand for all observations – that is, including the non-included observations.
- But if we consider only the observations in the (truncated) sample, then the relevant (partial) regression coefficient has to be multiplied by a factor which is smaller than 1. Hence, the within-sample marginal effect of a regressor is smaller (in absolute value) than the value of the coefficient of that variable, as in the case of the Tobit model.

Tobit vs truncated regression model

- Which is preferable?
- Since the Tobit model uses more information (753 observations) than the truncated regression model (428 observations), estimates obtained from Tobit are expected to be more efficient.

Heckman Selection Model

Heckman Selection Model

- The Heckman (1976) selection model, sometimes called the Heckit model, is a method for estimating regression models which suffer from sample selection bias.
- Under the Heckman selection framework, the dependent variable is only observable for a portion of the data.
- A classic example, in economics, of the sample selection problem is the wage equation for women, whereby a woman's wage is only observed if she makes the decision to enter the work place, and is unobservable if she does not.
- Heckman's (1976) paper that introduced the Heckman Selection model worked on this very problem.

The model

- The wage equation

$$W_i = \boldsymbol{\beta}\mathbf{X}_i + \varepsilon_i$$

where W_i is the wage, $\mathbf{X}_i$ observed variables relating to the i-th person's productivity and ε_i is an error term. W is observed only for workers, i.e. only people in work receive a wage.

- Sample selection (i.e. being in the labour force so W is observed). There is a second equation relating to employment:

$$E^*_i = \mathbf{Z_i}\boldsymbol{\gamma} + u_i$$

$E^*_i = W_i - E'_i$ is the difference between the wage and the reservation wage E'_i. The reservation wage is the minimum wage at which the i-th individual is prepared to work. If the wage is below that they choose not to work. We observe only an indicator variable for employment defined as E=1 if $E^*_i>0$ and E=0 otherwise.

Assumptions

- $(\varepsilon,u) \sim N(0,0,\sigma^2_{\varepsilon}, \sigma^2_{u},\rho_{\varepsilon u})$

  That is both error terms are normally distributed with mean 0, variances as indicated and the error terms are correlated where $\rho_{\varepsilon u}$ indicates the correlation coefficient.

- (ε,u) is independent of **X** and **Z.**

  The error terms are independent of both sets of explanatory variables.

- $\text{Var}(u) = \sigma^2_{u} = 1$

  This is not so much an assumption as a simplification it normalises the variance of the error term in what will be a probit regression.

The sample selection problem – 1

- **The key problem** is that in regressing wages on characteristics for those in employment we are not observing the equation for the population as a whole.
- Those in employment will tend to have higher wages than those not in the labour force would have (that is why they are not in the labour force).
- Hence the results will tend to be biased (**sample selection bias**) and e.g. we are likely to get biased results when estimating say the returns to education.

The sample selection problem – 2

- For example two groups of people (i) industrious; (ii) lazy. Industrious people get higher wages and have jobs, lazy people do not. In effect we are doing the regression in this simplified example on the industrious part of the labour force. The returns to education will be estimated on them alone not the whole of the population (which includes the lazy people).
- Those individuals who do not satisfy this are excluded from the regression. But this becomes a problem because of the last assumption that the error terms are correlated where $\rho_{\varepsilon u}$ indicates the correlation coefficient. Hence a lower bound on u suggests it too is restricted.

Heckman's methodology

- Heckman's first insight in his 1979 *Econometrica* paper was that this is can be approached as an omitted variables problem. An estimate of the omitted variable would solve this problem and hence solve the problem of sample selection bias.

Estimation methods

- Heckman’s original two-step method
- Maximum Likelihood method.

Comparison of Estimates (different data)

Covariate	OLS w/ All data	OLS w/ Selected sample	MLE of Heckman SS model Function form Ident.
Educ	0.0803	0.0703 [-12.5%]	0.065 [-19.2%]
Age	0.0122	0.0119 [-2.5%]	0.0115 [-5.7%]

[% difference from OLS w/ all data]

Review

Summary – 1

- The nature of censored regression models.
- OLS estimators are biased as well as inconsistent.
- The slope coefficients estimated by ML need to be interpreted carefully.
- The truncated regression model differs from the censored regression model. In the censored regression model, we have data on the regressors for all the values of the regressand including those values of the regressand that are not observed or set to zero or some such limit.

Summary – 2

- In practice, censored regression models may be preferable to the truncated regression models because in the former we include all the observations in the sample, whereas in the latter we only include observations in the truncated sample
- The Heckman (1976) selection model, sometimes called the Heckit model, is a method for estimating regression models which suffer from sample selection bias. Under the Heckman selection framework, the dependent variable is only observable for a portion of the data.

Questions?

Thank you for your attention!