TA Session 3: Censoring, Truncation and Selection

Microeconometrics with Joan Llull

IDEA, Fall 2024

TA: Conghan Zheng

October 11, 2024

Overview

Introduction

Censoring

- When a dependent variable has a mixed discrete/continuous distribution ...
- Problem from the constrained dependent variable: a pile-up of observations on a boundary, therefore, conventional (e.g. least squares) estimators are biased for the population parameters of the uncensored distribution.
- In censoring, we observe the characteristics (regressors) of the sample whose y^* is not observed.

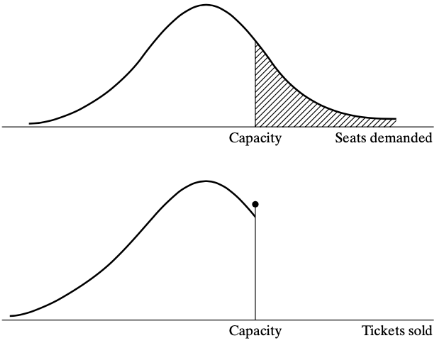

Figure 1: Partially Censored Distribution

Truncation

- Problem: incompletely observed sample, the sample is observed only if y^* is above/below a threshold. Clearly, conventional estimators are inconsistent because a truncated sample is not representative of the population.
- In truncation, we know noting about the missing sample (consider them as *who decided not to buy from me*), even the characteristics (regressors).

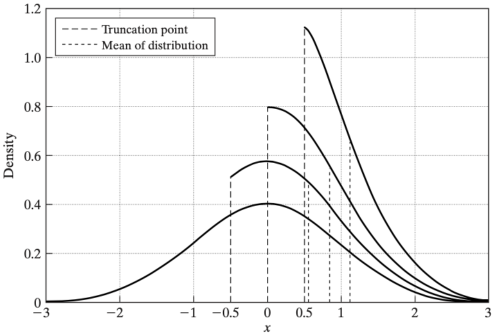

Figure 2: Truncated Normal Distribution

Tobit Regression

Type-I Tobit

- Without loss of generality, we consider the case of censoring from below at zero:

$$y = \begin{cases} y^*, & y^* > 0 \\ 0, & y^* \leq 0 \end{cases}$$

- Tobin(1958) proposed the **censored regression** (also known as **Tobit regression** or **Type-I Tobit**):

$$y^* = X'\beta + \varepsilon$$
$$\varepsilon|X \sim \mathcal{N}(0, \sigma^2)$$
$$y = \max(y^*, 0)$$

  Positive values are uncensored and negative values are transformed to 0.
- Problem: Tobit MLE relies crucially on normality.

$$f(y|X) = \begin{cases} f^*(y|X), & y^* > 0 \\ F^*(0|X), & y^* \leq 0 \end{cases} = \begin{cases} \phi\left(\dfrac{y - X\beta}{\sigma}\right), & y^* > 0 \\ 1 - \Phi\left(\dfrac{X\beta}{\sigma}\right), & y^* \leq 0 \end{cases}$$

Censored data

- **Data** (`TA3.dta`):

  The data on the dependent variable for ambulatory expenditure (`ambexp`) and the regressors (`age`, `female`, `educ`, `blhisp`, `totchr`, `ins`) are taken from the 2001 Medical Expenditure Panel Survey (US).

Variable	Obs	Mean	Std. Dev.	Min	Max
ambexp	3,328	1386.519	2530.406	0	49960
age	3,328	4.056881	1.121212	2.1	6.4
female	3,328	.5084135	.5000043	0	1
educ	3,328	13.40565	2.574199	0	17
blhisp	3,328	.3085938	.4619824	0	1
totchr	3,328	.4831731	.7720426	0	5
ins	3,328	.3650841	.4815261	0	1

  In this sample of 3,328 observations, there are 526 (15.8%) zero values of `ambexp`. Censoring might be an issue.

Tobit Regression with Censored Data

- Linear Tobit model:

```
.   tobit $xlist, ll(0) vce(robust)
```

```
Tobit regression                                    Number of obs     =  3,328
                                                       Uncensored     =  2,802
Limits: Lower =     0                                Left-censored    =    526
        Upper = +inf                                 Right-censored   =      0

                                                    F(6, 3322)        =  59.52
                                                    Prob > F          = 0.0000
Log pseudolikelihood = -26359.424                   Pseudo R2         = 0.0130
```

ambexp	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
age	314.1479	41.19122	7.63	0.000	233.3852	394.9107
female	684.9918	100.1353	6.84	0.000	488.6585	881.325
educ	70.8656	17.25925	4.11	0.000	37.02577	104.7054
blhisp	-530.311	102.8097	-5.16	0.000	-731.8877	-328.7342
totchr	1244.578	98.91188	12.58	0.000	1050.644	1438.513
ins	-167.4714	84.42021	-1.98	0.047	-332.9923	-1.95054
_cons	-1882.591	317.2026	-5.93	0.000	-2504.524	-1260.659
var(e.ambexp)	6635296	1088362			4810499	9152305

- The interpretation of the coefficients is as a partial derivative of the latent variable y^* with respect to X.

Marginal Effects

- Marginal effect varies according to whether interest lies in the latent variable mean or the the truncated or censored means:
  1. on latent variable mean

$$E(y^*|x) = x\beta$$
$$\Rightarrow \frac{\partial E(y^*|x)}{\partial x} = \beta$$

  2. on left-truncated (at 0) mean (check the Appendix for derivations)

$$E(y|x, y > 0) = x\beta + E[\varepsilon|\varepsilon > -x\beta]$$
$$\Rightarrow \frac{\partial E(y|x, y > 0)}{\partial x} = \left[1 - \frac{x\beta}{\sigma}\frac{\phi(\frac{x\beta}{\sigma})}{\Phi(\frac{x\beta}{\sigma})} - \left(\frac{\phi(\frac{x\beta}{\sigma})}{\Phi(\frac{x\beta}{\sigma})}\right)^2\right] \cdot \beta$$

  3. on left-censored (at 0) mean

$$E(y|x) = P(\varepsilon > -x\beta)[x\beta + E(\varepsilon|\varepsilon > -x\beta)]$$
$$\Rightarrow \frac{\partial E(y|x)}{\partial x} = \Phi(\frac{x\beta}{\sigma}) \cdot \beta$$

Three Means

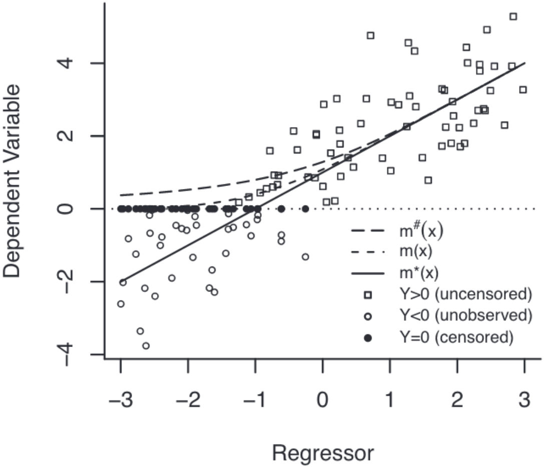

Figure 3: The conditional mean (m) of censored distributions

- Uncensored (y^*); Censored (y); and Truncated ($y^\#$)

Marginal Effects

When censoring is the case ...

- Example for using the ME on latent variable mean: income (usually top-coded)
- Example for using the ME on censored mean: hours of work for workers (participation, cersored from below)
- Example for using the ME on truncated mean: if a subsample of individuals (who has hours of work exceeds 20 hours per week) is of interest.

Marginal Effects

- ME for **left-truncated (at 0) mean** $E(y|x, y>0)$

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

```
Marginal effects after tobit
      y  = E(ambexp|ambexp>0) (predict, e(0, .))
         =  2494.4777
```

variable	dy/dx	Std. err.	z	P>|z|	[95%	C.I.]	X
age	145.524	18.794	7.74	0.000	108.689	182.359	4.05688
female*	317.1037	44.117	7.19	0.000	230.636	403.572	.508413
educ	32.82734	7.79096	4.21	0.000	17.5573	48.0973	13.4056
blhisp*	-240.2953	46.59	-5.16	0.000	-331.61	-148.98	.308594
totchr	576.5307	44.95	12.83	0.000	488.43	664.632	.483173
ins*	-77.19554	38.288	-2.02	0.044	-152.238	-2.15296	.365084

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

- The MEs here are smaller than the linear Tobit coefficient estimates $\hat{\beta}$ (= ME on latent variable mean) given previously, as expected given the relatively small variation in the range of y being considered.

Marginal Effects

- ME for **left-censored (at 0) mean** $E(y|x)$

```
.  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	26.802	7.74	0.000	154.994	260.058	4.05688
female*	451.6399	62.751	7.20	0.000	328.651	574.629	.508413
educ	46.81378	11.116	4.21	0.000	25.0261	68.6015	13.4056
blhisp*	-342.4803	66.293	-5.17	0.000	-472.412	-212.549	.308594
totchr	822.1678	64.078	12.83	0.000	696.577	947.758	.483173
ins*	-110.0883	54.609	-2.02	0.044	-217.119	-3.05739	.365084

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

- The MEs for the censored mean are larger in absolute value than those for the truncated mean and smaller than those for the latent mean (the coefficient estimates from the Tobit regression).

Model prediction

- Data:

ambexp

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	3,328
25%	113	0	Sum of Wgt.	3,328
50%	534.5		Mean	1386.519
		Largest	Std. Dev.	2530.406
75%	1618	28269		
90%	3585	30920	Variance	6402953
95%	5451	34964	Skewness	6.059491
99%	11985	49960	Kurtosis	72.06738

- Prediction:

Linear prediction

	Percentiles	Smallest		
1%	-968.7247	-1564.703		
5%	-557.2417	-1464.055		
10%	-281.8153	-1376.214	Obs	3,328
25%	192.9728	-1292.367	Sum of wgt.	3,328
50%	819.2401		Mean	1066.683
		Largest	Std. dev.	1257.455
75%	1742.236	7116.928		
90%	2750.839	7199.602	Variance	1581194
95%	3497.282	7524.147	Skewness	1.13039
99%	5082.921	8027.957	Kurtosis	4.955689

Model prediction

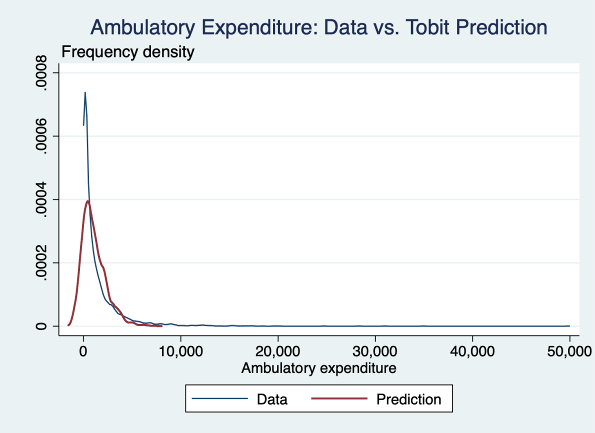

Figure 4: Data vs. fitted values of y^*

- The Tobit model fits especially poorly in the upper tail of the distribution.

Normality

- Detailed summary of `ambexp`:

ambexp

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	3,328
25%	113	0	Sum of Wgt.	3,328
50%	534.5		Mean	1386.519
		Largest	Std. Dev.	2530.406
75%	1618	28269		
90%	3585	30920	Variance	6402953
95%	5451	34964	Skewness	6.059491
99%	11985	49960	Kurtosis	72.06738

- The `ambexp` variable is heavily skewed (normal skewness = 0, positive skewness = concentrated on the left) and has considerable non-normal kurtosis (normal kurtosis = 3).
- Tobit MLE, which relies crucially on normality, might be a flawed estimator for the model.

Normality

- To see if these characteristics persist when the zero observations are ignored (which creates a truncated distribution), we summarize for positives only:

ambexp

	Percentiles	Smallest		
1%	22	1		
5%	67	2		
10%	107	2	Obs	2,802
25%	275	4	Sum of Wgt.	2,802
50%	779		Mean	1646.8
		Largest	Std. Dev.	2678.914
75%	1913	28269		
90%	3967	30920	Variance	7176579
95%	6027	34964	Skewness	5.799312
99%	12467	49960	Kurtosis	65.81969

- The skewness and non-normal kurtosis are reduced only a little if the zeros are ignored.

Normality

- Could the skewness and non-normal kurtosis of `ambexp` be due to regressors that are skewed? Let's try an OLS.
- The OLS residuals have a skewness statistic of 6.6 and a kurtosis statistic of 92.2.

```
Mean          -8.43e-07
Std. dev.       2319.71

Variance        5381056
Skewness       6.602534
Kurtosis       92.24478
```

- The skewness and non-normal kurtosis of `ambexp` are not due to regressors that are skewed. Even after conditioning on regressors, the dependent variable is very non-normal.
- Possible solution: use log-normal transformation to reduce skewness.

Log-normal transformation

- Summary of `ln(ambexp)`:

lambexp

	Percentiles	Smallest		
1%	3.091043	0		
5%	4.204693	.6931472		
10%	4.672829	.6931472	Obs	2,802
25%	5.616771	1.386294	Sum of Wgt.	2,802
50%	6.65801		Mean	6.555066
		Largest	Std. Dev.	1.41073
75%	7.556428	10.24952		
90%	8.285766	10.33916	Variance	1.990161
95%	8.704004	10.46207	Skewness	-.3421614
99%	9.43084	10.81898	Kurtosis	3.127747

- `ln(ambexp)` is almost symmetrically distributed. We expect that Tobit model is better suited to modeling `ln(ambexp)` than `ambexp`.

Tobit for lognormal data

- The Tobit model relies crucially on normality, but expenditure data are often better modeled as log-normal, as we have just seen.
- Introduce log-normality:

$$y^* = e^{x\beta+\varepsilon}, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$$\text{when we observe that } y = \begin{cases} y^*, \text{ if } \ln y^* > \gamma \\ 0, \text{ if } \ln y^* \geq \gamma \end{cases}$$

- In general $\gamma \neq 0$ (see later pages for why).

Tobit for lognormal data

- **Setting the censoring point γ for data in logs**
  - Why it is a problem: After the tranformation of the dependent variable to logs, zero values will become missing.[1]
  - To avoid this loss, we set all censored observations of $\ln y$ to an amount slightly smaller (relative to the scale of the vairable) than the minimum noncensored value of $\ln y$.

[1] Another complication if you are using Stata is that the smallest value of `ambexp` is 1, in which case `ln(ambexp)` equals zero. Stata will mistakenly treats this observation as censored, leading to a shrinkage in the sample size for noncensored observations.

Tobit for lognormal data

- Compare Tobit and OLS of the Log-normal data on the regressors:

	(1) tobit_log		(2) ols_log	
main				
age	0.363***	(0.0457)	0.325***	(0.0388)
female	1.342***	(0.0991)	1.145***	(0.0832)
educ	0.138***	(0.0201)	0.114***	(0.0165)
blhisp	-0.873***	(0.117)	-0.734***	(0.0974)
totchr	1.161***	(0.0538)	1.059***	(0.0463)
ins	0.261**	(0.0989)	0.208*	(0.0841)
_cons	0.924**	(0.355)	1.729***	(0.287)
/				
var(e.lny)	7.735***	(0.284)		
N	3328		3328	

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

- All OLS coefficients but the intercept are smaller in absolute terms, which is the impact of censoring. The larger the proportion of censored observations, the more biased the OLS estimates.

Truncated Tobit

```
. truncreg lny age female educ blhisp totchr ins,
ll(gamma01) vce(robust)
```

	(1) truncat~g		(2) censor_~g	
main				
age	0.217***	(0.0221)	0.363***	(0.0457)
female	0.379***	(0.0489)	1.342***	(0.0991)
educ	0.0222*	(0.00965)	0.138***	(0.0201)
blhisp	-0.239***	(0.0560)	-0.873***	(0.117)
totchr	0.562***	(0.0282)	1.161***	(0.0538)
ins	-0.0208	(0.0487)	0.261**	(0.0989)
_cons	4.908***	(0.172)	0.924**	(0.355)
/				
sigma	1.268***	(0.0192)		
var(e.lny)			7.735***	(0.284)
N	2802		3328	

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Model Prediction

Conditional and Unconditional Means for Models in Logs

- In the Tobit regression for lognormal data, the dependent variable is $\ln y$ instead of y. But the interest is still in predicting spending in levels rather than logs.
- Because of the strict convexity of the exponential function, by Jensen Inquality,

$$exp\left[\mathbb{E}(x)\right] \geq \mathbb{E}\left[exp(x)\right]$$

Therefore in model predictions, we see a correction for the convexity: $-\frac{\sigma^2}{2}$.

Moment	Model	Prediction function
$E(y\|\mathbf{x}, y > 0)$	Tobit	$\exp(\mathbf{x}'\boldsymbol{\beta} + \sigma^2/2)[1 - \Phi\{(\gamma - \mathbf{x}'\boldsymbol{\beta})/\sigma\}]^{-1}$ $[1 - \Phi\{(\gamma - \mathbf{x}'\boldsymbol{\beta} - \sigma^2)/\sigma\}]$
$E(y\|\mathbf{x})$	Tobit	$\exp(\mathbf{x}'\boldsymbol{\beta} + \sigma^2/2)[1 - \Phi\{(\gamma - \mathbf{x}'\boldsymbol{\beta} - \sigma^2)/\sigma\}]$
$E(y_2\|\mathbf{x}, y_2 > 0)$	Two-part	$\exp(\mathbf{x}_2'\boldsymbol{\beta}_2 + \sigma_2^2/2)$
$E(y_2\|\mathbf{x})$	Two-part	$\exp(\mathbf{x}_2'\boldsymbol{\beta}_2 + \sigma_2^2/2)\Phi(\mathbf{x}_1'\boldsymbol{\beta}_1)$
$E(y_2\|\mathbf{x}, y_2 > 0)$	Selection	$\exp(\mathbf{x}_2'\boldsymbol{\beta}_2 + \sigma_2^2/2)\{1 - \Phi(-\mathbf{x}_1'\boldsymbol{\beta}_1)\}^{-1}$ $\{1 - \Phi(-\mathbf{x}_1'\boldsymbol{\beta}_1 - \sigma_{12}^2)\}$
$E(y_2\|\mathbf{x})$	Selection	$\exp(\mathbf{x}_2'\boldsymbol{\beta}_2 + \sigma_2^2/2)\{1 - \Phi(-\mathbf{x}_1'\boldsymbol{\beta}_1 - \sigma_{12}^2)\}$

Source: Cameron and Trivedi (2022) Book

Model Specification

- Concerns:
  1. If we perform tests for normality and homoskedasticity to our censored regression, we see a failure in both assumptions, even though the expenditure `ambexp` was transformed to logarithms (if the error is either heteroskedastic or nonnormal the MLE not even inconsistent).
  2. The censoring mechanism and outcome may be modeled using separate processes (e.g., one process determine hospitalization, another on consequent hospital expenses).
- Next step: a more general model.
- Two approaches to such generalization:
  1. Two-part model: specifies one model for the censoring mechanism and a second distinct model for the outcome conditional on the outcome being observed.
  2. Sample-selction model: specifies a joint distribution for the censoring mechanism and outcome, and then finds the implied distribution conditional on the outcome observed.

Two-part Model

Two-part model

- The tobit regression makes a strong assumption that the same probability mechanism generates both the zeros (censoring point) and the positives.
- *Two-part model*[2]: a more flexible model which allows for the possibility that the zero and positive values are generated by different mechanisms, and thus can provide a better fit. Again, we apply it to a model in logs rather than in levels.
    - **1st part**: a binary outcome model that models $\mathbb{P}(y > 0)$, using any binary outcome model considered in previous chapter (usually probit), all observations are used for estimation;
    - **2nd part**: a linear regression that models $\mathbb{E}(y|y > 0)$, only observations with $y > 0$ are used.
- The two parts are assumed to be independent and are usually estimated separately.

[2] also known as a hurdle model, since crossing a hurdle or a threshold leads to participation

Two-part model

- Let y denote `ambexp`.
  - **1st part**: define a binary indicator d such that

$$d = \begin{cases} 1, & y > 0 \\ 0, & y = 0 \end{cases}$$

  - **2nd part**: For those with $y > 0$, let $f(y|d = 1)$ be the conditional density of y. When $y = 0$, we observe only $\mathbb{P}(d = 0)$.
- The two-part model for y is then given by

$$f(y|\mathbf{x}) = \begin{cases} P(d = 1|\mathbf{x}) \cdot f(y|d = 1, \mathbf{x}), & y > 0 \\ P(d = 0|\mathbf{x}), & y = 0 \end{cases}$$

Often the same regressors appear in both parts of the model, but this can and should be relaxed if there are obvious exclusion restrictions.

Two-part model

- **Part 1**: MLE of a discrete choice model using all observations.

```
. probit dy $xlist, vce(robust)
```

- **Part 2**: estimation of the parameters of the conditional density using only the observations with $y > 0$.

```
. reg lny $xlist if dy==1, vce(robust)
```

	(1) part1_p~t		(2) part2_ols	
main				
age	0.0973***	(0.0273)	0.217***	(0.0221)
female	0.644***	(0.0610)	0.379***	(0.0490)
educ	0.0702***	(0.0109)	0.0222*	(0.00966)
blhisp	-0.374***	(0.0610)	-0.239***	(0.0560)
totchr	0.794***	(0.0740)	0.562***	(0.0282)
ins	0.181**	(0.0612)	-0.0208	(0.0488)
_cons	-0.718***	(0.186)	4.908***	(0.172)
N	3328		2802	

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

- The coefficients in the two parts have the same sign, aside from the `ins` variable, which is highly statistically insignificant in the second part.

Two-part model

- Given the assumption that the two parts are independent, the joint likelihood for the two parts is the sum of the two log likelihoods.

```
.  scalar lltwopart = llprobit + lllognormal
.  display "lltwopart = " lltwopart
lltwopart = -5838.8218
```

- For comparison, the log likelihood for the previous log-normal Tobit is -7494.29.
- The two-part model fits the data considerably better, even if AIC or BIC is used to penalize the two-part model for its additional parameters.
- Concern: no link allowed between the two parts.
- Solution: to allow for the possible dependence in the two parts, we shall adopt a bivariate sample-selection model.

Selection

Selection

- If the reason the observations are missing is appropriately exogenous, using the subsample has no serious consequences.
- Selection occurs when sampling is endogenous. Pure random samples are rare.
- Problem: the sample drawn from a subset of the population is used to estimate unknown population parameters.
- Some mechanisms are due to sample design, while others are due to the behavior of the units being sampled. Examples:
  1. migration: self-selection to migrate
  2. wage regression: the decision to work
  3. survey data: nonresponse/noncompletion

Sample Selection

Example: sampling based on an explanatory variable

Suppose we wish to estimate a saving function for all families in a given country, and the population saving function is

$$saving = \beta_0 + \beta_1 income + \beta_2 age + \beta_3 married + \beta_4 kids + \varepsilon$$

where age the age of the household head and the other vairables are self-explanatory. Now we only have a restricted sample of which the household head was 45 years old or older. A sample selection issues is then raised here since we can obtain a random sample only for a subset of the population.

Sample Selection

Example: sampling based on a response variable

We are interested in esimating the effect of worder eligibility in a particular pension plan on family wealth. The population model is

$$wealth = \beta_0 + \beta_1 plan + \beta_2 educ + \beta_3 age + \beta_4 income + \varepsilon$$

where $plan$ is a binary variable indicator for the eligibility in the pension plan. However, we can sample people with a net wealth less than 200k USD, so the sample is selected on the basis of $wealth$. Sampling based on a response variable is much more serious than sampling based on an exogenous variable.

Sample Selection Model

Example: labor force participation and the wage offer (Gronau, 1974)

- Interest lies in estimating $\mathbb{E}(w_i^o|x_i)$, where w_i^o is the wage offer for a randomly drawn individual i.
- A potential sample selection problem arises because w_i^o is observed only for people who work.
- Assume an individual i has a reservation wage level w_i^r, she decides to work only if $w_i^o > w_i^r$.
- Now we make parametric assumptions:

$$w_i^o = e^{x_{i1}\beta_1 + u_{i1}}, \quad w_i^r = e^{x_{i2}\beta_2 + u_{i2}}$$

- Then the wage offer is observed only if the individual works, that is only if

$$\ln w_i^o - \ln w_i^r = x_{i1}\beta_1 - x_{i2}\beta_2 + u_{i1} - u_{i2} > 0$$

Sample Selection Model

Example: labor force participation and the wage offer (Gronau, 1974)

- Then there is a potential sample selection problem if we want to estimate the wage equation

$$\ln w_i^o = x_{i1}\beta_1 + u_{i1}$$

 but use only the data on working people.
- This example differs in an important respect from **top-coding, where the censoring rule is known for each unit in the population.**
- In the current example, **we do not know the individual reservation wage**, so we cannot use the wage offer in a censored regression analysis.
- More importantly, the reservation wage is allowed to depend on unobservables, so we need a new framework.

Bivariate Sample Selection Model

Type II Tobit Model (Amemiya, 1985)

- Sample selection bias can be corrected if we have a sample which includes the non-selected observations (Heckman, 1979).
- A **bivariate sample selection model** comprises
  - a **participation equation** that

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

  - a resultant **outcome equation** that

$$y_2 = \begin{cases} y_2^*, & y_1^* > 0 \\ \text{missing}, & y_1^* \leq 0 \end{cases}$$

Bivariate Sample Selection Model

Type II Tobit Model (Amemiya, 1985)

- The standard model specifies a linear model with additive errors for the latent variables,

$$y_1^* = \mathbf{x}_1\beta_1 + \varepsilon_1$$
$$y_2^* = \mathbf{x}_2\beta_2 + \varepsilon_2$$

  with problem arising in estimating β_2 if ε_1 and ε_2 are correlated. The Tobit model is a special case where $y_1^* = y_2^*$.
- It's assumed that the correlated errors $\{\varepsilon_1, \varepsilon_2\}$ are jointly normally distributed and homoskedastic.
- The likelihood function for this model is

$$\mathcal{L} = \Pi_i \Big\{ \underbrace{[P(y_{1i}^* \leq 0)]^{1-y_{1i}}}_{\text{contribution when } y_{1i}^* \leq 0} \underbrace{[f(y_{2i}|y_{1i}^* > 0) \times P(y_{1i}^* > 0)]^{y_{1i}}}_{\text{contribution when } y_{1i}^* > 0} \Big\}$$

Heckman Two-Step Estimator

Step 1

- **Step 1**: estimate y_1^* on x_1 using Probit

$$\begin{aligned}
\mathbb{P}(y_1^* > 0) &= \mathbb{P}(x_1\beta_1 + \varepsilon_1 > 0) \\
&= \mathbb{P}\left(\varepsilon_1 > -x_1\beta_1\right) \\
&= \mathbb{P}\left(\frac{\varepsilon_1}{\sigma_1} > -\frac{x_1\beta_1}{\sigma_1}\right) \\
&= 1 - \underbrace{\Phi\left(-\frac{x_1\beta_1}{\sigma_1}\right)}_{\varepsilon_1 \sim \mathcal{N}(0,\sigma_1^2)}
\end{aligned}$$

- In this step, $\frac{\beta_1}{\sigma_1}$ is identified.

Heckman Two-Step Estimator

Step 2

- **Step 2**: y_2^* on x_2

$$\begin{aligned} \mathbb{E}(y_2|\mathbf{x}, y_1^* > 0) &= x_2\beta_2 + \mathbb{E}(\varepsilon_2|y_1^* > 0) \\ &= x_2\beta_2 + \underbrace{\mathbb{E}(\varepsilon_2|\varepsilon_1 > -x_1\beta_1)}_{\equiv\Delta} \end{aligned} \tag{1}$$

- Without information on the selection process (correlation between ε_1 and ε_2) there is little that can be done to "correct" the selection bias (Δ) other than to be aware of its presence.
- Heckman(1979) on the correlated errors (the projection of ε_1 on ε_2):

$$\varepsilon_2 = \delta\varepsilon_1 + \eta \tag{2}$$

Heckman Two-step Estimator

Step 2

- Endogenous selection changes the conditional mean: Derivations

$$(1)\&(2) \Rightarrow \ \mathbb{E}(y_2|\mathbf{x}, y_1^* > 0) = x_2\beta_2 + \mathbb{E}(\delta\varepsilon_1 + \eta|\varepsilon_1 > -x_1\beta_1) \\ = x_2\beta_2 + \delta\lambda\left(\frac{x_1\beta_1}{\sigma_1}\right)$$

  where $\varepsilon_1 \sim \mathcal{N}(0, \sigma_1^2)$ is assumed.

- In Heckman's two-step procedure, step 2 uses positive values of y_2 to estimate by OLS the model

$$y_2 = x_2\beta_2 + \delta\lambda\left[x_1\widehat{\left(\frac{\beta_1}{\sigma_1}\right)}\right] + \nu \tag{3}$$

  where $\widehat{\left(\frac{\beta_1}{\sigma_1}\right)}$ comes from step 1.

- In step 2, β_2 and δ are identified.

FIML without Exclusion Restrictions

- Heckman FIML without exclusion restrictions ($x_1 = x_2$ for the two steps):

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

- The log likelihood for this model (-5838.397) is only slightly higher than that for the two-part model (-5838.822).

rho	-.1242024	.0934546	-.3012415	.0611142
sigma	1.270739	.019318	1.233435	1.309171
lambda	-.1578287	.1190885	-.3912379	.0755805

```
Wald test of indep. eqns. (rho = 0): chi2(1) = 1.73       Prob > chi2 = 0.1884
```

- `rho`: the estimated correlation ($\rho_{12} = \frac{cov(\varepsilon_1, \varepsilon_2)}{\sigma_{\varepsilon_1}\sigma_{\varepsilon_2}}$) between the errors.
- The Wald test on H_0 : $\texttt{rho} = 0$ implies we can't reject the null that the two parts of the model are independent.

FIML without Exclusion Restrictions

- The bivariate sample selection model with normal errors is theoretically identified without any restriction on the regressors. But there are some practical concerns...
- Without exclusion restrictions, we rely on the **nonlinearity** (by Probit, which automatically generates exclusion restrictions) of the selection regression to generate the needed source of variation in the probability of a positive outcome.
- If the nonlinearity implied by the Probit model is small (small variation in $x_1\hat{\beta}_1$ across observations), then identification will be fragile.

Heckman Two-Step Estimator

LIML without Exclusion Restrictions

- The one-step FIML estimation is based on a bivariate normality assumption $((\varepsilon_1, \varepsilon_2) \sim \mathcal{N}_2)$ that is itself suspect.
- The Heckit (LIML) with a univariate normality assumption $(\varepsilon_1 \sim \mathcal{N}, \varepsilon_2 = \delta\varepsilon_1 + \eta)$ is expected to be more robust.
- Heckit without exclusion restrictions:

```
.  heckman lny $xlist, select(dy = $xlist) twostep
```

/mills						
lambda	-.4801696	.2906565	-1.65	0.099	-1.049846	.0895067
rho	-0.37130					
sigma	1.2932083					

- The coefficient for `lambda` is the estimated δ (-0.48, $p = 0.099$). When it is significant, we should obtain the corrected standard errors.

Exclusion Restrictions

	(1)		(2)		(3)	
	HKM_FIML		HKM_LIML		HKM_LIML_ex	
lny						
age	0.212***	(0.0230)	0.202***	(0.0243)	0.202***	(0.0242)
female	0.350***	(0.0597)	0.289***	(0.0737)	0.292***	(0.0726)
educ	0.0189	(0.0105)	0.0120	(0.0117)	0.0124	(0.0116)
blhisp	-0.220***	(0.0595)	-0.181**	(0.0659)	-0.183**	(0.0653)
totchr	0.541***	(0.0391)	0.498***	(0.0495)	0.501***	(0.0486)
ins	-0.0295	(0.0510)	-0.0474	(0.0532)	-0.0465	(0.0530)
_cons	5.037***	(0.226)	5.303***	(0.294)	5.289***	(0.289)
dy						
age	0.0984***	(0.0270)	0.0973***	(0.0270)	0.0868**	(0.0275)
female	0.644***	(0.0601)	0.644***	(0.0601)	0.664***	(0.0610)
educ	0.0702***	(0.0113)	0.0702***	(0.0113)	0.0619***	(0.0120)
blhisp	-0.373***	(0.0617)	-0.374***	(0.0618)	-0.366***	(0.0619)
totchr	0.795***	(0.0710)	0.794***	(0.0711)	0.796***	(0.0712)
ins	0.182**	(0.0625)	0.181**	(0.0626)	0.169**	(0.0629)
income					0.00268*	(0.00131)
_cons	-0.724***	(0.192)	-0.718***	(0.192)	-0.669***	(0.194)

- The standard errors from LIML are in general larger than those from the FIML, both without exclusion restriction. Usually this imprecision is due to the collinearity that comes from the outcome equation.

Exclusion Restrictions

- The model is theoretically identified without any restriction on the regressors x_1 and x_2.
- But notice that when $x_1 = x_2$, β_2 is indentified only due to the nonlinearity of the inverse Mills ratio $\lambda(x_1\beta_1)$.
- The collinearity happens when there is not much variation in $x_1\beta_1$, and the inverse mills ratio $\lambda(x_1\beta_1)$ can be approximated well by a linear function of x_1.
- If this is the case, then $\lambda(x_1\hat{\beta}_1)$ is collinear with the other regressors (x_2) in the outcome equation.
- Having exclusion restrictions, so that $x_1 \neq x_2$, will reduce the collinearity problem and provide more robust identification, especially in small samples.
- How? Usually we include extra regressors in x_1.
- Why? x_2 would only need to be observed whenever y_2 is (positive values of y_1), whereas x_1 must always be observed (all values of y_1), which implies that x_1 may contain elements that cannot also appear in x_2.

Heckman Two-Step Estimator

LIML with Exclusion Restrictions

- This requires that the participation (selection) equation (step 1) have an exogenous variable that is excluded from the outcome equation (step 2).
- Heckit with exlusion restriction:

```
.  heckman lny $xlist, select(dy = $xlist income) twostep
```

income	.0026773	.0013105	2.04	0.041	.0001088	.0052458
_cons	-.6686471	.1941247	-3.44	0.001	-1.049125	-.2881698
/mills						
lambda	-.4637133	.2825997	-1.64	0.101	-1.017598	.090172

- $\hat{\beta}_{\texttt{income}} = 0.003, \;\; p = 0.041.$
- But the use of this exclusion restriction is debatable as there are reasons to expect that income should also appear in the outcome equation. It's often very difficult to make defensible exclusion restrictions.

Appendix

Truncated First Moment of Normal

The truncated first moment used in Heckman's approach step 2:

$$
\begin{aligned}
\mathbb{E}(y_2|x_1,x_2,y_1^*>0) =& x_2\beta_2+\mathbb{E}(\varepsilon_2|x_1\beta_1+\varepsilon_1>0)\\
\overset{(2)}{=}& x_2\beta_2+\mathbb{E}\left(\delta\varepsilon_1|\varepsilon_1>-x_1'\beta_1\right)\\
=& x_2\beta_2+\delta\mathbb{E}\left(\frac{\varepsilon_1}{\sigma_1}\middle|\frac{\varepsilon_1}{\sigma_1}>\frac{-x_1'\beta_1}{\sigma_1}\right),\quad \frac{\varepsilon_1}{\sigma}\sim\mathcal{N}(0,1)\\
=& x_2\beta_2+\delta\int_{\frac{-x_1\beta_1}{\sigma_1}}^{\infty}u\cdot f\left(u\middle|u>\frac{-x_1'\beta}{\sigma_1}\right)du,\quad u\sim\mathcal{N}(0,1)\\
=& x_2\beta_2+\delta\frac{1}{1-\Phi\left(\frac{-x_1\beta_1}{\sigma_1}\right)}\int_{\frac{-x_1\beta_1}{\sigma_1}}^{\infty}u\cdot\phi(u)du\\
=& x_2\beta_2+\delta\frac{1}{1-\Phi\left(\frac{-x_1\beta_1}{\sigma_1}\right)}\left[u\cdot\Phi(u)\Big|_{\frac{-x_1\beta_1}{\sigma_1}}^{\infty}-\int_{\frac{-x_1\beta_1}{\sigma_1}}^{\infty}\Phi(u)du\right]
\end{aligned}
$$

Truncated First Moment of Normal

$$\cdots = x_2\beta_2 + \delta \frac{1}{1-\Phi\left(\frac{-x_1\beta_1}{\sigma_1}\right)} \left[u \cdot \Phi(u) \Big|_{\frac{-x_1\beta_1}{\sigma_1}}^{\infty} - [u \cdot \Phi(u) + \phi(u)] \Big|_{\frac{-x_1\beta_1}{\sigma_1}}^{\infty} \right]$$

$$= x_2\beta_2 + \delta \frac{1}{1-\Phi\left(\frac{-x_1\beta_1}{\sigma_1}\right)} \left[-\phi(u) \Big|_{\frac{-x_1\beta_1}{\sigma_1}}^{\infty} \right]$$

$$= x_2\beta_2 + \delta \frac{1}{1-\Phi\left(\frac{-x_1\beta_1}{\sigma_1}\right)} \left[-\underbrace{\phi\left(\infty\right)}_{=0} + \phi\left(\frac{-x_1\beta_1}{\sigma_1}\right) \right]$$

$$= x_2\beta_2 + \delta \frac{\phi\left(\frac{-x_1\beta_1}{\sigma_1}\right)}{1-\Phi\left(\frac{-x_1\beta_1}{\sigma_1}\right)} = x_2\beta_2 + \delta \underbrace{\frac{\phi\left(\frac{x_1\beta_1}{\sigma_1}\right)}{\Phi\left(\frac{x_1\beta_1}{\sigma_1}\right)}}_{\equiv \lambda\left(\frac{x_1\beta_1}{\sigma_1}\right)}$$

Back to Heckit

References

- Cameron, A. C., & Trivedi, P. K. (2005). Microeconometrics: methods and applications. Cambridge university press. Chapter 16.
- Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data. MIT press. Chapter 19.
- Hansen, B. E. (2022). Econometrics. Chapter 27.
- Cameron, A. C., & Trivedi, P. K. (2022). Microeconometrics using stata (Second Edition). Stata press. Chapter 19.