UNIT 2. Tobit and Selection Models

Instructor: Henry Redondo

From: *Econometric Analysis of Cross-Section and Panel Data*
by J.M. Wooldridge

Dpto. de Fundamentos del Análisis Económico.
Universidad de Alicante

Econometrics II. Academic Year 2023/2024

Contents

1. Censored regression models: The Tobit model
   - Introduction
   - Expected values and marginal effects
   - Inconsistency of OLS
   - Maximum likelihood estimation

2. Selection models
   - Introduction
   - Selection on the basis of an explanatory variable
   - Selection on the basis of the response variable: Truncated regression
   - The generalized selection model

Data censoring

- We are going to study a statistical model that can be used in different types of applications.
- In the first type of applications there is a variable with quantitative meaning, call it y^*, and we are interested in the population regression $E(y^* \,|\, \mathbf{x})$.
- If y^* and $\mathbf{x}$ were observed for a random sample of the population, we could use standard regression methods.
- A data problem arises because y^* is censored above and/or below some value; that is, y^* is not observable for part of the population.
- An example of **data censoring** is top coding in survey data. For example, assume that y^* is family wealth, and, for a randomly drawn family, the actual value of wealth is recorded up to some threshold.

Example 1 (Top Coding of Wealth)

- In the population of all families in the United States, let $wealth^*$ denote actual family wealth, measured in thousands of dollars. Suppose that wealth follows the linear regression model

$$E(wealth^* \mid \mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$$

where $\mathbf{x}$ is a $k \times 1$ vector of conditioning variables.

- If we observed $(wealth^*, \mathbf{x}')'$ for a random sample of the population, we could estimate $\boldsymbol{\beta}$ by OLS.
- However, we observe $wealth^*$ only when $wealth^* < 200$. When $wealth^*$ is greater than 200 we know that $wealth^* \geq 200$, but we do not know the actual value of wealth.
- Define observed *wealth* as

$$wealth = \min(wealth^*, 200)$$

  - The definition of $wealth = 200$ when $wealth^* > 200$ is arbitrary, but it is useful for defining the statistical model that follows.

Example 1 (Top Coding of Wealth)

- To estimate $\boldsymbol{\beta}$ we might assume that $wealth^*$ given $\mathbf{x}$ has a homoskedastic normal distribution.
- Then, the error-form model for $wealth^*$ is

$$wealth^* = \mathbf{x}'\boldsymbol{\beta} + u, \quad u \,|\, \mathbf{x} \sim N(0, \sigma^2)$$

- This is a strong assumption about the conditional distribution of $wealth^*$, something we could avoid entirely if $wealth^*$ were not censored above 200.
- Under these assumptions we can write recorded wealth as

$$wealth = \min(\mathbf{x}'\boldsymbol{\beta} + u, 200) \tag{1}$$

Corner solution models

- In the second kind of application y is an observable choice by some economic agent, such as an individual or a firm, with the following characteristics: y takes value zero with positive probability but is a continuous random variable for strictly positive values.
- There are many examples of variables that, at least approximately, have these features.
  - Amount of life insurance coverage chosen by an individual
  - Family contributions to an individual retirement account
  - Firm expenditures on research and development
- In each of these examples we can imagine economic agents solving an optimization problem, and for some agents the optimal choice will be the corner solution, $y = 0$.
- We will call this kind of models **corner solution models**.
- For corner solution applications, we must understand that the issue is not data observability and we are interested in features of the distribution of y given $\mathbf{x}$, such as $E(y \,|\, \mathbf{x})$, $E(y \,|\, \mathbf{x}, y > 0)$ and $P(y = 0 \,|\, \mathbf{x})$.

Example 2 (Charitable Contributions)

- Suppose that family i chooses annual consumption c_i (in dollars) and charitable contributions q_i (in dollars) to solve the problem

$$\begin{aligned} &\max_{c,q} \{c + a_i \log(1+q)\} \\ &st \quad c + p_i q = m_i \\ &\qquad\ \ c, q \geq 0 \end{aligned}$$

  where a_i determines the marginal utility of charitable contributions, m_i is income, and p_i is the price of one dollar of charitable contributions, where $p_i < 1$ because of the tax deductibility of charitable contributions, and this price differs across families because of different marginal tax rates and different state tax codes.

Example 2 (Charitable Contributions)

- The solution of this maximization problem is

$$\begin{aligned} q_i &= 0 \quad \text{if } a_i/p_i \leq 1 \\ q_i &= a_i/p_i - 1 \quad \text{if } a_i/p_i > 1 \end{aligned}$$

- We can write this relation as

$$1 + q_i = \max(a_i/p_i, 1)$$

- If $a_i = \exp(\mathbf{z}_i'\gamma + u_i)$, where u_i is unobservable, then charitable contributions are determined by

$$\log(1 + q_i) = \max(\mathbf{z}_i'\gamma - \log(p_i) + u_i, 0) \tag{2}$$

Censored Tobit model

- Comparing equations (2) and (1) shows that they have similar statistical structures.
  - In equation (2) we are taking a maximum, and the lower threshold is zero, whereas in equation (1) we are taking a minimum with an upper threshold of 200.
- Each of this two problems can be transformed into the same statistical model:

$$\begin{aligned} y^* &= \mathbf{x}'\boldsymbol{\beta} + u, \quad u \,|\, \mathbf{x} \sim N(0, \sigma^2) \\ y &= \max(y^*, 0) \end{aligned} \tag{3}$$

- These equations constitute what is known as the **standard censored Tobit model** or **type I Tobit model** (after Tobin, 1958).

Applications

- The charitable contributions example immediately fits into the standard censored Tobit framework by defining

$$\mathbf{x} = (\mathbf{z}', \log(p)) \text{ and } y = \log(1+q).$$

  - This particular transformation of q and the restriction that the coefficient on $\log(p)$ is -1 depends critically on the utility function used in the example.

- The wealth example can be cast into the standard censored Tobit framework after a simple transformation:

$$-(\mathit{wealth} - 200) = \max(200 - \mathbf{x}'\beta - u, 0)$$

  and so the intercept changes, and all slope coefficients have the opposite sign from equation (1).

Applications

- As we saw from the two previous examples, different features of model (3) are of interest depending on the type of application.
  - In examples with true data censoring, such as Example 1, the vector $\boldsymbol{\beta}$ tells us everything we want to know because $E(y^* \mid \mathbf{x})$ is of interest.
  - For corner solution outcomes, such as Example 2, $\boldsymbol{\beta}$ does not give the entire story. Usually, we are interested in $E(y \mid \mathbf{x})$ or $E(y \mid \mathbf{x}, y > 0)$. These certainly depend on $\boldsymbol{\beta}$, but in a nonlinear fashion.
  - For corner solution outcomes, we must avoid placing too much emphasis on the latent variable y^*. Most of the time y^* is an artificial construct, and we are not interested in $E(y^* \mid \mathbf{x})$.
- For the statistical model (3) to make sense, the variable y^* should have characteristics of a normal random variable.
  - In data censoring cases this requirement means that the variable of interest y^* should have a homoskedastic normal distribution. In some cases the logarithmic transformation can be used to make this assumption more plausible.
  - In corner solution examples, the variable y should be (roughly) continuous when $y > 0$.

Expected values and marginal effects

- As we mentioned above, in corner solution applications, interest centers on probabilities or expectations involving y. Most of the time we focus on the expected values $E(y\,|\,\mathbf{x})$ and $E(y\,|\,\mathbf{x}, y > 0)$.
- When u is independent of $\mathbf{x}$ and has a normal distribution, we can find an explicit expression for $E(y\,|\,\mathbf{x})$. We first derive $P(y > 0\,|\,\mathbf{x})$ and $E(y\,|\,\mathbf{x}, y > 0)$, which are of interest in their own right. Then, $E(y\,|\,\mathbf{x})$ can be computed as:

$$\begin{aligned} E(y\,|\,\mathbf{x}) &= P(y = 0\,|\,\mathbf{x})0 + P(y > 0\,|\,\mathbf{x})E(y\,|\,\mathbf{x}, y > 0) \qquad (4) \\ &= P(y > 0\,|\,\mathbf{x})E(y\,|\,\mathbf{x}, y > 0) \end{aligned}$$

- Define the binary variable $w = 1$ if $y > 0$, $w = 0$ if $y = 0$. Then w follows a probit model:

$$\begin{aligned} P(w &= 1\,|\,\mathbf{x}) = P(y > 0\,|\,\mathbf{x}) = P(u > -\mathbf{x}'\boldsymbol{\beta}\,|\,\mathbf{x}) \qquad (5) \\ &= P\left(\frac{u}{\sigma} < \frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\middle|\, x\right) = \Phi\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right) \end{aligned}$$

- One implication of equation (5) is that $\boldsymbol{\beta}/\sigma$, but not $\boldsymbol{\beta}$ and σ separately, can be identified from a probit of w on $\mathbf{x}$.

Expected values and marginal effects

- To derive $E(y \,|\, \mathbf{x}, y > 0)$, we need the following result about the normal distribution: if $z \sim N(0,1)$, then, for any constant c,

$$E(z \,|\, z > c) = \frac{\phi(c)}{1 - \Phi(c)}$$

  where $\phi(\bullet)$ and $\Phi(\bullet)$ are the pdf and cdf of the standard normal.

Expected values and marginal effects

- To derive $E(y \mid \mathbf{x}, y > 0)$, we need the following result about the normal distribution: if $z \sim N(0,1)$, then, for any constant c,

$$E(z \mid z > c) = \frac{\phi(c)}{1 - \Phi(c)}$$

where $\phi(\bullet)$ and $\Phi(\bullet)$ are the pdf and cdf of the standard normal.

- <u>Proof</u>
  - For any $z_0 > c$, the cdf of z conditional on $z > c$ at z_0 is

$$P(z < z_0 \mid z > c) = \frac{P(c < z < z_0)}{P(z > c)} = \frac{\Phi(z_0) - \Phi(c)}{1 - \Phi(c)}$$

  - Taking derivatives with respect to z_0, the pdf of z at z_0 is

$$\frac{\phi(z_0)}{1 - \Phi(c)}$$

  - Then

$$E(z \mid z > c) = \int_c^{\infty} z \frac{\phi(z)}{P(z > c)} dz = \frac{1}{1 - \Phi(c)} \int_c^{\infty} z \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z^2\right) dz$$

Expected values and marginal effects

- Proof
  - If we define $v = \frac{1}{2}z^2$, we have that $dv = zdz$, then

$$\int_c^{\infty} z\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) dz = \frac{1}{\sqrt{2\pi}} \int_{\frac{1}{2}c^2}^{\infty} \exp(-v)\, dv =$$

$$-\frac{1}{\sqrt{2\pi}} \left[\exp(-v)\right]_{\frac{1}{2}c^2}^{\infty} = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}c^2\right) = \phi(c)$$

  - and

$$E(z \,|\, z > c) = \frac{\phi(c)}{1 - \Phi(c)}$$

Expected values and marginal effects

- We can use this equation to find $E(y\,|\,\mathbf{x}, y > 0)$ when y follows a Tobit model

$$\begin{aligned}E(y\,|\,\mathbf{x}, y > 0) &= E(\mathbf{x}'\boldsymbol{\beta} + u\,|\,\mathbf{x}, u > -\mathbf{x}'\boldsymbol{\beta})\\ &= \mathbf{x}'\boldsymbol{\beta} + \sigma E\left(\frac{u}{\sigma}\ \mathbf{x}\frac{u}{\sigma}\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right)\\ &= \mathbf{x}'\boldsymbol{\beta} + \sigma\frac{\phi\left(-\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right)}{1-\Phi\left(-\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right)}\end{aligned}$$

- Then, since $\phi\left(-\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right) = \phi\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right)$ and $1-\Phi\left(-\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right) = \Phi\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right)$

$$E(y\,|\,\mathbf{x}, y > 0) = \mathbf{x}'\boldsymbol{\beta} + \sigma\frac{\phi\left(\frac{x'\boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right)} = \mathbf{x}'\boldsymbol{\beta} + \sigma\lambda\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right) \qquad (6)$$

where for any c the quantity $\lambda(c) = \phi\left(c\right)/\Phi\left(c\right)$ is called the **inverse Mills ratio**.

Expected values and marginal effects

- If x_j is a continuous explanatory variable, using that $\partial\phi\left(c\right)/\partial c = -c\phi\left(c\right)$ and $\partial\Phi\left(c\right)/\partial c = \phi\left(c\right)$, we have

$$\frac{\partial\lambda\left(c\right)}{\partial c} = \frac{\partial\left(\frac{\phi(c)}{\Phi(c)}\right)}{\partial c} = -\frac{c\phi\left(c\right)}{\Phi\left(c\right)} - \left(\frac{\phi\left(c\right)}{\Phi\left(c\right)}\right)^2 = -c\lambda\left(c\right) - \left(\lambda\left(c\right)\right)^2$$

- Then, the marginal effect of x_j on $E(y \mid \mathbf{x}, y > 0)$ is

$$\frac{\partial E(y \mid \mathbf{x}, y > 0)}{\partial x_j} = \beta_j\left[1 - \frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\lambda\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right) - \left(\lambda\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right)\right)^2\right] \qquad (7)$$

- This equation shows that the partial effect of x_j on $E(y \mid \mathbf{x}, y > 0)$ is not entirely determined by β_j; there is an adjustment factor multiplying β_j. It can be shown that this adjustment factor is between 0 and 1 and therefore the sign of β_j is the same as the sign of the partial effect of x_j.

Expected values and marginal effects

- We can also compute $E(y\,|\,\mathbf{x})$. Using equations (4) and (5):

$$\begin{aligned} E(y\,|\,\mathbf{x}) &= P(y>0\,|\,\mathbf{x})E(y\,|\,\mathbf{x},y>0) \\ &= \Phi\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right)\mathbf{x}'\boldsymbol{\beta}+\sigma\phi\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right) \qquad (8) \end{aligned}$$

- If x_j is a continuous explanatory variable, the marginal effect of x_j on $E(y\,|\,\mathbf{x})$ is

$$\begin{aligned} \frac{\partial E(y\,|\,\mathbf{x})}{\partial x_j} &= \phi\left(\frac{x'\boldsymbol{\beta}}{\sigma}\right)\frac{\beta_j}{\sigma}\mathbf{x}'\boldsymbol{\beta}+\Phi\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right)\beta_j-\sigma\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\phi\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right)\frac{\beta_j}{\sigma} \\ &= \Phi\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right)\beta_j \qquad (9) \end{aligned}$$

- This equation shows that the partial effect of x_j on $E(y\,|\,\mathbf{x})$ is not entirely determined by β_j; there is an adjustment factor multiplying β_j. Since this adjustment factor is between 0 and 1, the sign of β_j is the same as the sign of the partial effect of x_j.

Inconsistency of OLS

- We can use the previous expectation calculations to show that OLS using the entire sample or OLS using the subsample for which $y_i > 0$ are both (generally) inconsistent estimators of $\boldsymbol{\beta}$.
- First consider OLS using the subsample with strictly positive y_i.
  - From equation (6) we can write

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \sigma\lambda_i + \varepsilon_i$$

$$E(\varepsilon_i \,|\, \mathbf{x}_i, y_i > 0) = 0$$

    where $\lambda_i = \lambda\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)$.
  - It follows that if we run OLS of y_i on $\mathbf{x}_i$ using the sample for which $y_i > 0$ we effectively omit the variable λ_i. Correlation between λ_i and $\mathbf{x}_i$ in the selected subpopulation results in inconsistent estimation of $\boldsymbol{\beta}$.
- From equation (8) it is also pretty clear that regressing y_i on $\mathbf{x}_i$ using all of the data will not consistently estimate $\boldsymbol{\beta}$: $E(y \,|\, \mathbf{x})$ is nonlinear in $\mathbf{x}$, $\boldsymbol{\beta}$, and σ, so it would be very unlikely that this linear regression consistently estimates $\boldsymbol{\beta}$.

Maximum likelihood estimation

- To use maximum likelihood, we need to derive the distribution of y given $\mathbf{x}$.
- We have already shown that

$$P(y = 0 \,|\, \mathbf{x}) = 1 - \Phi\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right)$$

- For any $y_0 > 0$, since $y^* \,|\, \mathbf{x} \sim N(\mathbf{x}'\boldsymbol{\beta}, \sigma^2)$, the cdf of y at y_0 is

$$\begin{aligned} P(y < y_0 \,|\, \mathbf{x}) &= P(y^* < y_0 \,|\, \mathbf{x}) = P\left(\frac{y^* - \mathbf{x}'\boldsymbol{\beta}}{\sigma} < \frac{y_0 - \mathbf{x}'\boldsymbol{\beta}}{\sigma} \middle| \mathbf{x}\right) \\ &= P\left(\frac{u}{\sigma} < \frac{y_0 - \mathbf{x}'\boldsymbol{\beta}}{\sigma} \middle| \mathbf{x}\right) = \Phi\left(\frac{y_0 - \mathbf{x}'\boldsymbol{\beta}}{\sigma}\right) \end{aligned}$$

- Then, the pdf of y given $\mathbf{x}$ (for y positive) is

$$\frac{1}{\sigma}\phi\left(\frac{\mathrm{y} - \mathbf{x}'\boldsymbol{\beta}}{\sigma}\right)$$

Maximum likelihood estimation

- Let $\{(\mathbf{x}_i', y_i)' : i = 1, 2, ..N\}$ be a random sample following the censored Tobit model. Then the log-likelihood function is

$$\log L(\boldsymbol{\beta}, \sigma; y_1, ..y_N \,|\, \mathbf{x}_1, .., \mathbf{x}_N) = \sum_{i=1}^N \log \left(1 - \Phi\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right) 1(y_i = 0)$$
$$+ \sum_{i=1}^N \left(\log \phi\left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right) - \log \sigma\right) 1(y_i > 0)$$

- The Maximum Likelihood Estimator (MLE) of $\boldsymbol{\beta}$, denoted $\widehat{\boldsymbol{\beta}}$, maximizes this log likelihood. There is no explicit solution for the MLE estimator but the estimator can be found by numerical methods.
- Testing is easily carried out in a standard MLE framework. Single exclusion restrictions are tested using asymptotic t statistics once $\widehat{\beta}_j$ and its asymptotic standard error have been obtained. Multiple exclusion restrictions are easily tested using a LR, a Wald or an LM statistic.

Example 3 (Annual Hours Equation for Married Women)

- We use MROZ.DTA to estimate a reduced form annual hours equation for married women.
- The equation is a reduced form because we do not include hourly wage offer as an explanatory variable.
- The hourly wage offer is unlikely to be exogenous, and, just as importantly, we cannot observe it when $hours = 0$.
- The explanatory variables are the same ones appearing in the labor force participation probit in Unit 1, Example 1.
- We report OLS estimates using all the observations and Tobit estimates.

Example 3 (Annual Hours Equation for Married Women)

OLS and Tobit estimation of Annual Hours Worked

Dependent variable: *hours*		
Explanatory variables	OLS	Tobit (MLE)
nwifeinc	-3.45 (2.54)	-8.81 (4.46)
educ	28.76 (12.95)	80.65 (21.58)
exper	65.67 (9.96)	131.56 (17.28)
$exper^2$	-0.700 (0.325)	-1.86 (0.54)
age	-30.51 (4.36)	-54.41 (7.42)
*kidslt*6	-442.09 (58.85)	-894.02 (111.88)
*kidsge*6	-32.78 (23.18)	-16.22 (38.64)
Log-likelihood		-3819.09
Pseudo R^2	0.266	0.275
$\widehat{\sigma}$	750.18	1122.02

Example 3 (Annual Hours Equation for Married Women)

- Not surprisingly, the Tobit coefficient estimates have the same sign as the corresponding OLS estimates, and the statistical significance of the estimates is similar.
- Second, though it is tempting to compare the magnitudes of the OLS estimates and the Tobit estimates, such comparisons are not very informative.
  - For example, we must not think that, because the Tobit coefficient on kidslt6 is roughly twice that of the OLS coefficient, the Tobit model somehow implies a much greater response of hours worked to young children.

Example 3 (Annual Hours Equation for Married Women)

- To obtain the average partial effects (APE) for roughly continuous variables, we estimate the factor in equation (9) for each woman in the sample using the Tobit estimates. The average estimated factor is 0.589. Then, an additional year of education is estimated to increase expected hours by about $0.589 \times 80.65 = 47.5$ hours on average, which is well above the OLS estimate of 28.76.
- The APE of education conditional on $hours > 0$ is obtained by evaluating the factor in equation (7) for all woman in the sample with positive hours. The average estimated factor is 0.525. Conditional on $hours > 0$, an additional year of education is estimated to increase expected hours by about $0.525 \times 80.65 = 42.4$ hours on average, slightly smaller than the unconditional average partial effect.

Example 3 (Annual Hours Equation for Married Women)

- We can also compute the estimated partial effect at the average by evaluating the factor in equation (9) at the estimates and the mean values of the x_j (but where we square $\overline{exper}$ rather than use the average of the $exper_i^2$). The estimated factor is 0.645. Then, for a women with average characteristics, an additional year of education is estimated to increase expected hours by about $0.645 \times 80.65 = 52$ hours, which is also well above the OLS estimate of 28.76.
- The factor in equation (7) at the estimates and the mean values of the x_j is about 0.452. Then, conditional on hours being positive, for a women with average characteristics, an additional year of education is estimated to increase expected hours by about $0.452 \times 80.65 = 36.5$ hours.

Example 3 (Annual Hours Equation for Married Women)

- For the discrete variable kidslt6, we use equation (8) to compute the difference in estimated expected values at $kidslt6 = 0$ and $kidslt6 = 1$, and average these differences. We have that the first child is estimated to decrease expected hours by about 467.6 hours on average, which is a bit larger in magnitude than OLS estimate.
- We have that the second child is estimated to decrease expected hours by about 246.2 hours on average, not surprisingly, the effect on expected hours of having a second young child is less than having a first young child.
- We could also compute the APE from zero to one small child (or from one small child to two small children) conditional on $hours > 0$ or the estimated partial effects at the average.

Introduction

- Up to this point we have assumed the availability of a random sample from the underlying population.
- This assumption is not always realistic: because of the way some economic data sets are collected, and often because of the behavior of the units being sampled, random samples are not always available.
- A selected sample is a general term that describes a nonrandom sample. There are a variety of **selection mechanisms** that result in nonrandom samples.
- Before we describe the sample selection model in detail, there is an important general point to remember: sample selection can only be an issue once the population of interest has been carefully specified.
  - If we are interested in a subset of a larger population, then the proper approach is to specify a model for that part of the population, obtain a random sample from that part of the population, and proceed with standard econometric methods.

Example 4 (Saving Function)

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

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

  where *age* is the age of the household head.

- However, we only have access to a survey that included families whose household head was 45 years of age or older.
- This limitation raises a sample selection issue because we are interested in the saving function for all families, but we can obtain a random sample only for a subset of the population.

Example 5 (Truncation Based on Wealth)

- We are interested in estimating the effect of worker eligibility in a particular pension plan on family wealth. Let the population model be

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

  where plan is a binary indicator for eligibility in the pension plan.
- However, we can only sample people with a net wealth less than \$200,000, so the sample is selected on the basis of wealth.
- As we will see, sampling based on a response variable is much more serious than sampling based on an exogenous explanatory variable.

Example 6 (Wage Offer Function)

- Consider estimating a wage offer equation for people of working age.
- This equation is supposed to represent all people of working age, whether or not a person is actually working at the time of the survey.
- Because we can only observe the wage offer for working people, we effectively select our sample on this basis.
- This is sometimes called incidental truncation because wage is missing as a result of the outcome of another variable, labor force participation.

Selection on the basis of an explanatory variable

- We are going to see that selection on the basis of an explanatory variable (like in example 4) does not affect the consistency of the OLS estimator.
- We can consider a more general framework: The population model is the standard single-equation linear model with possibly endogenous explanatory variables:

$$y = \mathbf{x}'\boldsymbol{\beta} + u, \quad E(u \,|\, \mathbf{z}) = 0 \tag{10}$$

  where $\mathbf{x}$ is a $k \times 1$ vector of explanatory variables and $\mathbf{z}$ is an $L \times 1$ vector of instruments $(L \geq k)$.
- If we could obtain a random sample from the population, equation (10) could be estimated by 2SLS under the condition $\text{rank}(E(\mathbf{x}\mathbf{z}')) = k$.
- Our general treatment includes the leading special case when $\mathbf{z} = \mathbf{x}$, so that the explanatory variables are exogenous and equation (10) is a model of the conditional expectation

$$E(y \,|\, \mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$$

Selection on the basis of an explanatory variable

- Rather than obtaining a random sample we only use data points that satisfy certain conditions. Let s be a binary selection indicator representing a random draw from the population. By definition, $s = 1$ if we use the draw in the estimation, and $s = 0$ if we do not. Usually, we do not use observations when $s = 0$ because data on at least some elements of $(\mathbf{x}', y, \mathbf{z}')'$ are unobserved—because of survey design, nonresponse or incidental truncation.
- It can be shown that the key assumption underlying the validity of 2SLS on the selected sample is

$$E(u \,|\, \mathbf{z}, s) = 0 \tag{11}$$

- Then, if s is a deterministic function of $\mathbf{z}$, then $E(u \,|\, \mathbf{z}, s) = E(u \,|\, \mathbf{z}) = 0$.
- In the example 4 $\mathbf{x} = \mathbf{z}$ and s is a deterministic function of $\mathbf{x}$, so OLS based on the selected sample is a consistent estimator.

Selection on the basis of the response variable: Truncated regression

- In this section we explicitly treat the case where the sample is selected on the basis of the dependent variable.
- In applying the following methods it is important to remember that there is an underlying population of interest, often described by a linear conditional expectation: $E(y \mid \mathbf{x}) = \mathbf{x}'\beta$. If we could observe a random sample from the population, then we would just use standard regression analysis.
- The problem arises because the sample we can observe is chosen at least partly based on the value of y.
- Unlike in the case where selection is based only on $\mathbf{x}$, selection based on y causes problems for standard OLS analysis on the selected sample.

Selection on the basis of the response variable: Truncated regression

- An example of selection based on y is example 5 where we do not observe data on families with wealth above \$200,000.
  - This case is different from the top coding on wealth (example 1).
  - Here, we observe nothing about families with high wealth: they are entirely excluded from the sample.
  - In the top coding case, we have a random sample of families, and we always observe $\mathbf{x}$; the information on $\mathbf{x}$ is useful even if wealth is top coded.
- We assume that y is a continuous random variable and that the selection rule takes the form
$$s = 1(a_1 < y < a_2) \tag{12}$$
where a_1 and a_2 are known constants such that $a_1 < a_2$.
- A good way to think of the sample selection is that we draw $(\mathbf{x}_i', y_i)'$ randomly from the population. If y_i falls in the interval (a_1, a_2), then we observe both y_i and $\mathbf{x}_i$. If y_i is outside this interval, then we do not observe y_i or $\mathbf{x}_i$.

Selection on the basis of the response variable: Truncated regression

- Thus all we know is that there is some subset of the population that does not enter our data set because of the selection rule. We know how to characterize the part of the population not being sampled because we know the constants a_1 and a_2.
- In most applications we are still interested in estimating $E(y \,|\, \mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$. However, because of sample selection based on y, we must specify a full conditional distribution of y given $\mathbf{x}$ and estimate $\boldsymbol{\beta}$ by maximum likelihood.
- In most applications we assume that $y \,|\, \mathbf{x} \sim N(\mathbf{x}'\boldsymbol{\beta}, \sigma^2)$. Then, for any value y_0, $a_1 < \ y_0 \ < a_2$. The cdf of y at y_0 given $\mathbf{x}$ and $s = 1$ is

$$P(y < y_0 \,|\, \mathbf{x}, s = 1) = \frac{P(y < y_0, s = 1 \,|\, \mathbf{x})}{P(s = 1 \,|\, \mathbf{x})} = \frac{P(a_1 < \ y \ < y_0 \,|\, \mathbf{x})}{P(a_1 < \ y \ < a_2 \,|\, \mathbf{x})}$$

$$= \frac{P(\frac{a_1 - \mathbf{x}'\boldsymbol{\beta}}{\sigma} < \ \frac{y - \mathbf{x}'\boldsymbol{\beta}}{\sigma} \ < \frac{y_0 - \mathbf{x}'\boldsymbol{\beta}}{\sigma} \,|\, \mathbf{x})}{P(\frac{a_1 - \mathbf{x}'\boldsymbol{\beta}}{\sigma} < \ y \ < \frac{a_2 - \mathbf{x}'\boldsymbol{\beta}}{\sigma} \,|\, \mathbf{x})} = \frac{\Phi\left(\frac{y_0 - \mathbf{x}'\boldsymbol{\beta}}{\sigma}\right) - \Phi\left(\frac{a_1 - \mathbf{x}'\boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\frac{a_2 - \mathbf{x}'\boldsymbol{\beta}}{\sigma}\right) - \Phi\left(\frac{a_1 - \mathbf{x}'\boldsymbol{\beta}}{\sigma}\right)}$$

Selection on the basis of the response variable: Truncated regression

- The density of y at y_0 given $\mathbf{x}$ and $s = 1$ is obtained by differentiating with respect to y_0

$$\frac{\frac{1}{\sigma}\phi\left(\frac{y_0 - \mathbf{x}'\boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\frac{a_2 - \mathbf{x}'\boldsymbol{\beta}}{\sigma}\right) - \Phi\left(\frac{a_1 - \mathbf{x}'\boldsymbol{\beta}}{\sigma}\right)}$$

for $a_1 < \; y_0 \; < a_2$.

- Then, the log-likelihood function is

$$\log L(\boldsymbol{\beta}, \sigma; y_1, ..y_N \,|\, \mathbf{x}_1, .., \mathbf{x}_N) =$$

$$\sum_{i=1}^{N}\left[\log\phi\left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right) - \log\sigma - \log\left(\Phi\left(\frac{a_2 - \mathbf{x}'\boldsymbol{\beta}}{\sigma}\right) - \Phi\left(\frac{a_1 - \mathbf{x}'\boldsymbol{\beta}}{\sigma}\right)\right)\right.$$

- This is called the **truncated Tobit model** or **truncated normal regression model**.

Selection on the basis of the response variable: Truncated regression

- The truncated Tobit model is related to the censored Tobit model but there is a key difference: in censored regression, we observe the covariates **x** for all people, even those for whom the response is not known. If we drop observations entirely when the response is not observed, we obtain the truncated regression model.
- If in Example 1 we use the information in the top coded observations, we are in the censored regression case. If we drop all top coded observations, we are in the truncated regression case.
- Given a choice, we should use a censored regression analysis, as it uses all of the information in the sample.

Labor Force Participation and the Wage Offer

- The generalized selection model is motivated by Gronau's (1974) model of the wage offer and labor force participation.
- Interest lies in estimating $E(w_i^o \mid \mathbf{x}_i)$, where w_i^o is the hourly wage offer for a randomly drawn individual i.
- If w_i^o were observed for everyone in the (working age) population, we would proceed in a standard regression framework.
- However, a potential sample selection problem arises because w_i^o is observed only for people who work.
- We can cast this problem as a weekly labor supply model:

$$\max_h u_i(w_i^o h + a_i, h) \qquad \text{subject to } 0 \leq h \leq 168$$

where h is hours worked per week, a_i is nonwage income and $q_i = w_i^o h + a_i$ is total income.

Labor Force Participation and the Wage Offer

- Let $s_i(h) = u_i(w_i^o h + a_i, h)$, and assume that we can rule out the solution $h = 168$. Then the solution can be $h = 0$ or $0 < h < 168$.
- If $\partial s_i(h)/\partial h \leq 0$ at $h = 0$, then the optimum is $h = 0$.
- We have that

$$\frac{\partial s_i(h)}{\partial h} = mu_i^q(w_i^o h + a_i, h)w_i^o + mu_i^h(w_i^o h + a_i, h)$$

where $mu_i^q(\cdot,\cdot)$ is the marginal utility of income and $mu_i^h(\cdot,\cdot)$ is the marginal disutility of working.

- Then

$$\left.\frac{\partial s_i(h)}{\partial h}\right|_{h=0} = mu_i^q(a_i, 0)w_i^o + mu_i^h(a_i, 0)$$

and $h = 0$ if and only if

$$w_i^o \leq -\frac{mu_i^h(a_i, 0)}{mu_i^q(a_i, 0)} \tag{13}$$

- Gronau (1974) called the right-hand side of equation (13) the reservation wage, w_i^r, which is assumed to be strictly positive.

Labor Force Participation and the Wage Offer

- We now make the parametric assumptions

$$\begin{aligned} w_i^0 &= \exp(\mathbf{x}_{1i}'\boldsymbol{\beta}_1 + u_{1i}) \\ w_i^r &= \exp(\mathbf{x}_{2i}'\boldsymbol{\beta}_2 + \gamma_2 a_i + u_{2i}) \end{aligned} \tag{14}$$

  where $(u_{1i}, u_{2i})'$ is independent of $(\mathbf{x}_{1i}', \mathbf{x}_{2i}', a_i)'$.

- Here, $\mathbf{x}_{1i}$ contains productivity characteristics, and possibly demographic characteristics, of individual i, and $\mathbf{x}_{2i}$ contains variables that determine the marginal utility of leisure and income; these may overlap with $\mathbf{x}_{1i}$.
- From (14) we have the log wage equation

$$\log w_i^0 = \mathbf{x}_{1i}'\boldsymbol{\beta}_1 + u_{1i} \tag{15}$$

- But the wage offer w_i^0 is observed only if the person works, that is, only if $w_i^0 \geq w_i^r$, which is equivalent to

$$\log w_i^0 - \log w_i^r = \mathbf{x}_{1i}'\boldsymbol{\beta}_1 - \mathbf{x}_{2i}'\boldsymbol{\beta}_2 - \gamma_2 a_i + u_{1i} - u_{2i} \equiv \mathbf{x}_i'\boldsymbol{\delta}_2 + v_{i2} > 0$$

  - This introduces a potential sample selection problem if we use data only on working people to estimate equation (15).

The generalized selection model

- This example differs in an important respect from top coding examples. With top coding, the censoring rule is known for each unit in the population. In Gronau's example, we do not know w_i^r, so we cannot use w_i^0 in a censored regression analysis.
- If w_i^r were observed and exogenous and $\mathbf{x}_{i1}$ were always observed, then we would be in the censored regression framework (but with thresholds varying across individuals).
- If w_i^r were observed and exogenous but $\mathbf{x}_{i1}$ were observed only when w_i^0 is, we would be in the truncated Tobit framework (also with thresholds varying across individuals).
- But w_i^r is not observable, and so we need a new framework.

The generalized selection model

- If we drop the i subscript, let $y_1 \equiv \log w^0$ and let y_2 be the binary labor force participation indicator, Gronau's model can be written for a random draw from the population as

$$
\begin{aligned}
y_1 &= \mathbf{x}_1'\boldsymbol{\beta}_1 + u_1 \quad (16)\\
y_2 &= 1(\mathbf{x}'\delta_2 + v_2 > 0) \quad (17)
\end{aligned}
$$

- Equation (17) is called the selection equation.
- We discuss estimation of this model under the following set of assumptions:
    - (a) $(\mathbf{x}', y_2)'$ are always observed and y_1 is observed when $y_2 = 1$
    - (b) $(u_1, v_2)'$ is independent of $\mathbf{x}$ and has zero mean
    - (c) $v_2 \sim N(0, 1)$
    - (d) $E(u_1 \mid v_2) = \gamma_1 v_2$

The generalized selection model

- Assumption (a) emphasizes the sample selection nature of the problem.
- Assumption (b) is a strong, but standard, form of exogeneity of $\mathbf{x}$.
- Assumption (c) is needed to derive a conditional expectation given the selected sample. It is probably the most restrictive assumption because it is an explicit distributional assumption. Assuming $Var(v_2) = 1$ is without loss of generality because y_2 is a binary variable.
- Assumption (d) requires linearity in the population regression of u_1 on v_2. It always holds if $(u_1, v_2)'$ is bivariate normal but Assumption (d) holds under weaker assumptions. In particular, we do not need to assume that u_1 itself is normally distributed.

The generalized selection model

- Since y_1 is observed only when $y_2 = 1$, what we can hope to estimate is $E(y_1 \,|\, \mathbf{x}, y_2 = 1)$ [along with $P(y_2 = 1 \,|\, \mathbf{x})$].
- How does $E(y_1 \,|\, \mathbf{x}, y_2 = 1)$ depend on the vector of interest $\boldsymbol{\beta}_1$? First, under Assumption (a) to (d) and equations (16) and (17),

$$\begin{aligned} E(y_1 \,|\, \mathbf{x}, v_2) &= \mathbf{x}_1' \boldsymbol{\beta}_1 + E(u_1 \,|\, \mathbf{x}, v_2) \\ &\underset{\text{by (b)}}{=} \mathbf{x}_1' \boldsymbol{\beta}_1 + E(u_1 \,|\, v_2) \underset{\text{by (d)}}{=} \mathbf{x}_1' \boldsymbol{\beta}_1 + \gamma_1 v_2 \end{aligned} \tag{18}$$

- Equation (18) is very useful. The first thing to note is that, if $\gamma_1 = 0$, which implies that u_1 and v_2 are uncorrelated, then $E(y_1 \,|\, \mathbf{x}, v_2) = \mathbf{x}_1' \boldsymbol{\beta}_1$. Then, since y_2 is a function of $\mathbf{x}$ and v_2, it follows immediately that $E(y_1 \,|\, \mathbf{x}, y_2) = \mathbf{x}_1' \boldsymbol{\beta}_1$.
  - In other words, if $\gamma_1 = 0$, there is no sample selection problem, and $\boldsymbol{\beta}_1$ can be consistently estimated by OLS using the selected sample.

The generalized selection model

- What if $\gamma_1 \neq 0$? Using the law of iterated expectations on equation (18),

  $$E(y_1 \,|\, \mathbf{x}, y_2 = 1) = E(E(y_1 \,|\, \mathbf{x}, v_2) \,|\, \mathbf{x}, y_2 = 1) = \mathbf{x}_1'\boldsymbol{\beta}_1 + \gamma_1 E(v_2 \,|\, \mathbf{x}, y_2 = 1)$$

  and since

  $$E(v_2 \,|\, \mathbf{x}, y_2 = 1) = E(v_2 \,|\, \mathbf{x}, v_2 > -\mathbf{x}'\boldsymbol{\delta}_2) \underset{\substack{\text{by(c) and the} \\ \text{prop in slide 13}}}{=} \lambda(\mathbf{x}'\boldsymbol{\delta}_2)$$

  where $\lambda(\cdot)$ is the inverse Mills ratio, we can write

  $$E(y_1 \,|\, \mathbf{x}, y_2 = 1) = \mathbf{x}_1'\boldsymbol{\beta}_1 + \gamma_1 \lambda(\mathbf{x}'\boldsymbol{\delta}_2) \tag{19}$$

- Equation (19) makes it clear that an OLS regression of y_1 on $\mathbf{x}_1$ using the selected sample omits the term $\lambda(\mathbf{x}'\boldsymbol{\delta}_2)$ and generally leads to inconsistent estimation of $\boldsymbol{\beta}_1$.

The generalized selection model

- Equation (19) also suggests a way to consistently estimate $\boldsymbol{\beta}_1$ and γ_1, that is by regressing y_{1i} on $\mathbf{x}_{1i}$ and $\lambda(\mathbf{x}_i'\boldsymbol{\delta}_2)$ using the selected sample.
- The problem is that $\boldsymbol{\delta}_2$ is unknown, so we cannot compute the additional regressor $\lambda(\mathbf{x}'\boldsymbol{\delta}_2)$.
- Nevertheless, a consistent estimator of $\boldsymbol{\delta}_2$, $\widehat{\boldsymbol{\delta}}_2$, is available from the first-stage probit estimation of the selection equation.
- Then, using the results for generated regressors we studied in Econometrics I, we can use the following procedure to consistently estimate $\boldsymbol{\beta}_1$ and γ_1 using the selected sample.
    - Step 1. Obtain the probit estimate $\widehat{\boldsymbol{\delta}}_2$ from the model

      $$P(y_2 = 1 \,|\, \mathbf{x}) = \Phi(\mathbf{x}_i'\boldsymbol{\delta}_2)$$

      using all the observations and obtain the estimated inverse Mills ratios $\lambda(\mathbf{x}_i'\widehat{\boldsymbol{\delta}}_2)$.
    - Step 2. Obtain $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\gamma}_1$ from the OLS regression

      $$y_{1i} \text{ on } \mathbf{x}_{1i} \text{ and } \lambda(\mathbf{x}_i'\widehat{\boldsymbol{\delta}}_2) \text{ using the selected sample} \tag{20}$$

The generalized selection model

- The only problem in the regression in step 2 is that we have to adjust the variance matrix of the OLS estimator since we have a generated regressor.
  - However, a very simple test for selection bias is available from regression (20).
  - Under the null of no selection bias, $H_0 : \gamma_1 = 0$, from the results on generated regressors, the asymptotic variance of $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\gamma}_1$ is not affected by the fact that we have a generated regressor and therefore we can use a standard t-test.

The generalized selection model

- We do not need $\mathbf{x}_1$ to be a strict subset of $\mathbf{x}$ for $\boldsymbol{\beta}_1$ to be identified, and the 2-step procedure does carry through when $\mathbf{x}_1 = \mathbf{x}$.
  - However, if $\mathbf{x}_i'\widehat{\boldsymbol{\delta}}_2$ does not vary much in the sample, then $\lambda(\mathbf{x}_i'\widehat{\boldsymbol{\delta}}_2)$ can be approximated well by a linear function of $\mathbf{x}$. Then, if $\mathbf{x}_1 = \mathbf{x}$, regression (20) can suffer from a severe collinearity problem which can lead to large standard errors.
  - Moreover, when $\mathbf{x}_1 = \mathbf{x}$, $\boldsymbol{\beta}_1$ is identified only due to the nonlinearity of the inverse Mills ratio, and we would have to wonder whether a statistically significant inverse Mills ratio term is due to sample selection or functional form misspecification in the population model (16).
- If we replace assumptions (b) to (c) by the strongest assumption

$$\left(\begin{array}{c} u_1 \\ v_2 \end{array} \right) \Bigg| \; x \sim N\left(\left(\begin{array}{c} 0 \\ 0 \end{array} \right), \left(\begin{array}{cc} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & 1 \end{array} \right) \right)$$

we can estimate the parameters by maximum likelihood.

Example 7 (Wage Offer Equation for Married Women)

- We use the data in MROZ.DTA to estimate a wage offer function for married women, accounting for potential selectivity bias into the workforce.
- The labor force participation equation contains the variables in Example 1 of Unit 1, including other income, age, number of young children, and number of older children—in addition to $educ$, $exper$, and $exper2$.
- The 2-step method use four exclusion restrictions in the structural equation, because $nwifeinc$, age, $kidslt6$, and $kidsge6$ are all excluded from the wage offer equation.
- The results of OLS on the selected sample and the 2-step procedure described above are presented in the table below.
- The differences between the OLS and 2-step estimates are practically small, and the inverse Mills ratio term is statistically insignificant.

Example 7 (Wage Offer Equation for Married Women)

Wage Offer Equation for Married Women

Dependent variable: log(*wage*)

Explanatory variables	OLS	2-step
educ	0.108 (0.014)	0.109 (0.016)
exper	0.042 (0.012)	0.044 (0.016)
$exper^2$	-0.00081 (0.00039)	-0.00088 (0.00044)
$\widehat{\lambda}$	–	0.032 (0.134)
Sample size	428	428
R^2	0.157	0.157