Advanced Econometrics

Limited Dependent Variable Models

1. Introduction;
2. Truncated data;
3. Censored data;
4. Sample selection;

Introduction

- In some applications, the domain of the variate of interest is limited.
- Although in many cases those bounds are irrelevant, in microeconometric applications the bounds on the domain of Y often have to be taken into account.
- Traditionally, textbooks consider three forms of limited dependent variable models:
  1. *Truncated data*;
  2. *Censored data*;
  3. *Sample selection* (incidental truncation).

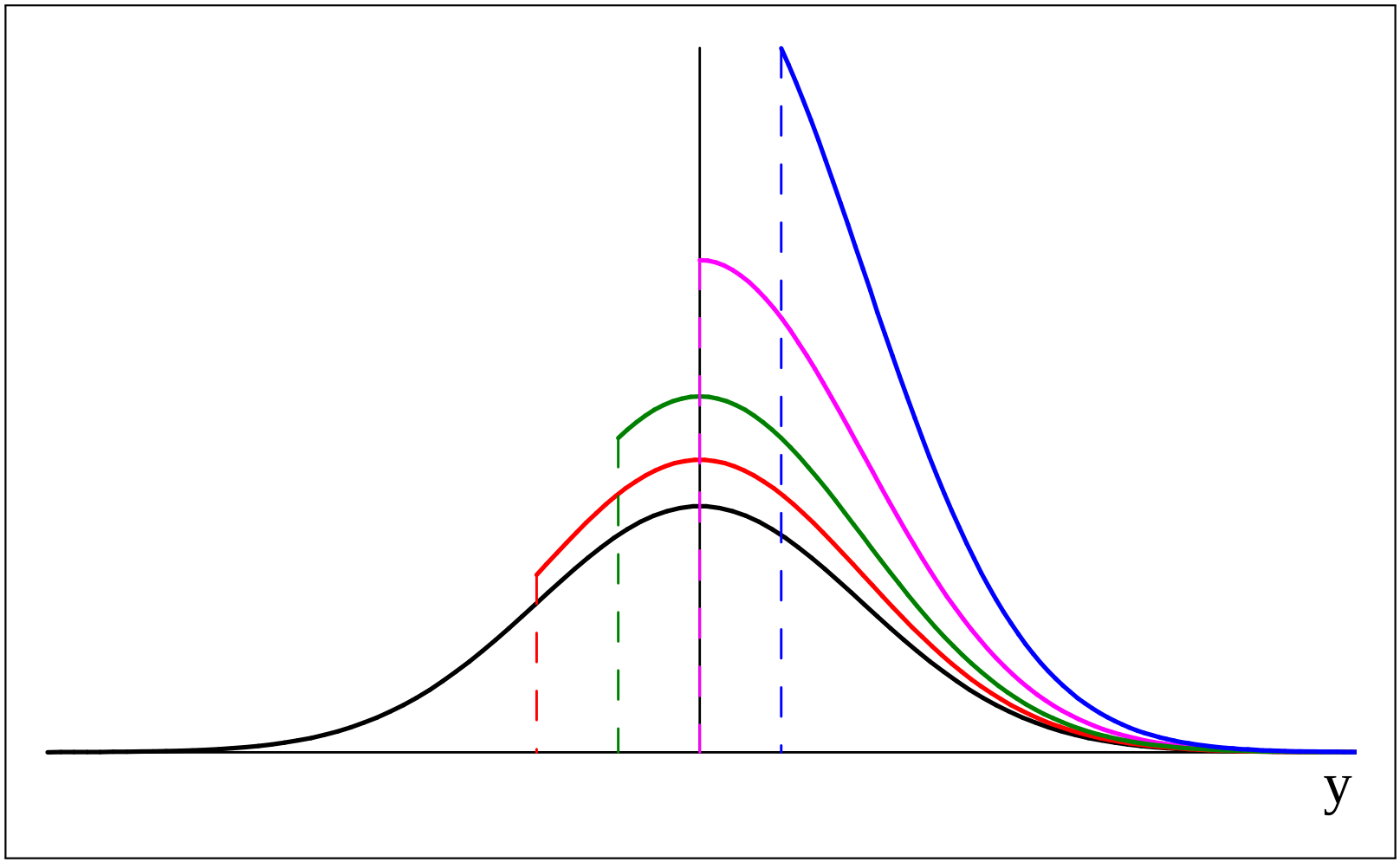
Truncated data

- A sample of Y is said to come from a truncated distribution when it is not possible to obtain observations from part of the domain of Y.
- **Examples:**
  - if we want to study the size of certain fish based on the specimens captured with a net, fish smaller than the net grid will not be present in our sample.
  - Fogel et al. (1978) published a dataset on the height of Royal Marines that extends over two centuries. Trussell and Bloom (1979) point out that the sample is truncated due to minimum height restrictions for the recruits.
- If Y is a continuous random variable with pdf $f(y)$ and a is a constant, we have left-truncation at a if the sample is drawn from the conditional distribution

$$f(y|Y > a) = \frac{f(y)}{\mathcal{P}(Y > a)}.$$

- We have also truncation if **the regressors are also not observed** when $Y < a$.

Truncated data



Truncated data

- For a discrete random variable with support on $0, 1, \ldots, \infty$, we have left-truncation at a if the sample is drawn from

$$\mathcal{P}\left(Y = j | Y > a\right) = \frac{\mathcal{P}\left(Y = j\right)}{1 - \sum_{k=0}^{\lfloor a \rfloor} \mathcal{P}\left(Y = k\right)},$$

  where $\lfloor a \rfloor$ denotes the integer part of a.
- For example, we have already seen the zero-truncated Poisson defined by

$$\mathcal{P}\left(Y = j | Y > 0\right) = \frac{\exp\left(-\lambda_i\right) \lambda_i^j}{\left(1 - \exp\left(-\lambda_i\right)\right) j!}.$$

- **Notice that** truncation is only problematic if it depends on Y or on another variable that is not independent of Y and that is not used as a regressor.
- For example, in general, standard inference can be performed if the sample is obtained from $\mathcal{P}\left(Y = j | \mathbf{X}, \mathbf{X} > a\right)$.

Truncated data

- Under certain conditions, it is possible to perform inference about the entire population using truncated samples.
- The problem is that this inference tends to be sensitive to *distributional assumptions*.
- For example, for the normal distribution with mean μ and variance σ^2,

$$\mathrm{E}\left(Y_i | Y_i > a\right) = \mu + \sigma \frac{\phi\left(\frac{a-\mu}{\sigma}\right)}{1 - \Phi\left(\frac{a-\mu}{\sigma}\right)}.$$

- For the case of count data with $\mathrm{E}\left(Y_i | \mathbf{X}_i\right) = \exp\left(\mathbf{X}_i'\boldsymbol{\beta}_0\right)$,

$$\mathrm{E}\left(Y_i | \mathbf{X}_i, Y_i > a\right) = \frac{\exp\left(\mathbf{X}_i'\boldsymbol{\beta}_0\right)}{1 - \sum_{k=0}^{\lfloor a \rfloor} \mathcal{P}\left(Y_i = k | \mathbf{X}_i\right)}.$$

- In both cases, the expectation for the truncated data depends on the shape of the conditional distribution.
- Because of this, estimation is often performed by maximum likelihood.

Truncated data

- An example is the *truncated normal regression model* defined by

$$\begin{aligned} Y_i^* &= \mathbf{X}_i'\boldsymbol{\beta}_0 + u_i^*, \qquad u_i^* \sim \mathcal{N}\left(0, \sigma^2\right), \\ \left(Y_i, \mathbf{X}_i\right) &= \begin{cases} \left(Y_i^*, \mathbf{X}_i\right) & \text{if } Y_i^* > 0 \\ \text{non-observable} & \text{if } Y_i^* \leq 0 \end{cases}. \end{aligned}$$

- This implies (using the results of the truncated normal random variable) that

$$E\left(Y_i | \mathbf{X}_i, Y_i > 0\right) = \mathbf{X}_i'\boldsymbol{\beta}_0 + \sigma \frac{\phi\left(-\mathbf{X}_i'\boldsymbol{\beta}_0/\sigma\right)}{1 - \Phi\left(-\mathbf{X}_i'\boldsymbol{\beta}_0/\sigma\right)}.$$

- The parameters of interest can be estimated from the likelihood function

$$L\left(\boldsymbol{\beta}, \sigma\right) = \prod_{i=1}^{n} \frac{1}{\sigma} \frac{\phi\left(\frac{Y_i - \mathbf{X}_i'\boldsymbol{\beta}}{\sigma}\right)}{1 - \Phi\left(-\mathbf{X}_i'\boldsymbol{\beta}/\sigma\right)}.$$

Censored data

- Censoring is also a problem of partial observability. For example, if Y^* is a random variable and a is a constant, we have left-censoring at a if we can only observe

$$Y = \max\{Y^*, a\} = \begin{cases} a & \text{if } Y^* \leq a \\ Y^* & \text{if } Y^* > a \end{cases} .$$

- **Examples:**
  - Childhood learning (Time-to-event): the age at which a child learns to accomplish certain tasks in children learning centers. Left censoring occurs if children can already perform the tasks when they start their study at the centers.
  - Suppose you're conducting a study on pregnancy duration. Now suppose you survey some women in your study at the 250-day mark, but they already had their babies. You know they had their babies before 250 days, but don't know exactly when. These are therefore left-censored observations.

Censored data

- We can have also right truncation, in this case $Y = \min\{Y^*, a\}$.
  - **Example:** Survey data are often topcoded before release to the public to preserve the anonymity of respondents. For example, if a survey answer reported a respondent with self-identified wealth of $79 billion, it would not be anonymous because people would know there is a good chance the respondent was Bill Gates.
- For simplicity we focus on left-censored data.
- If Y^* is a discrete random variable with support on $0, 1, \ldots, \infty$, the probability function of Y is given by

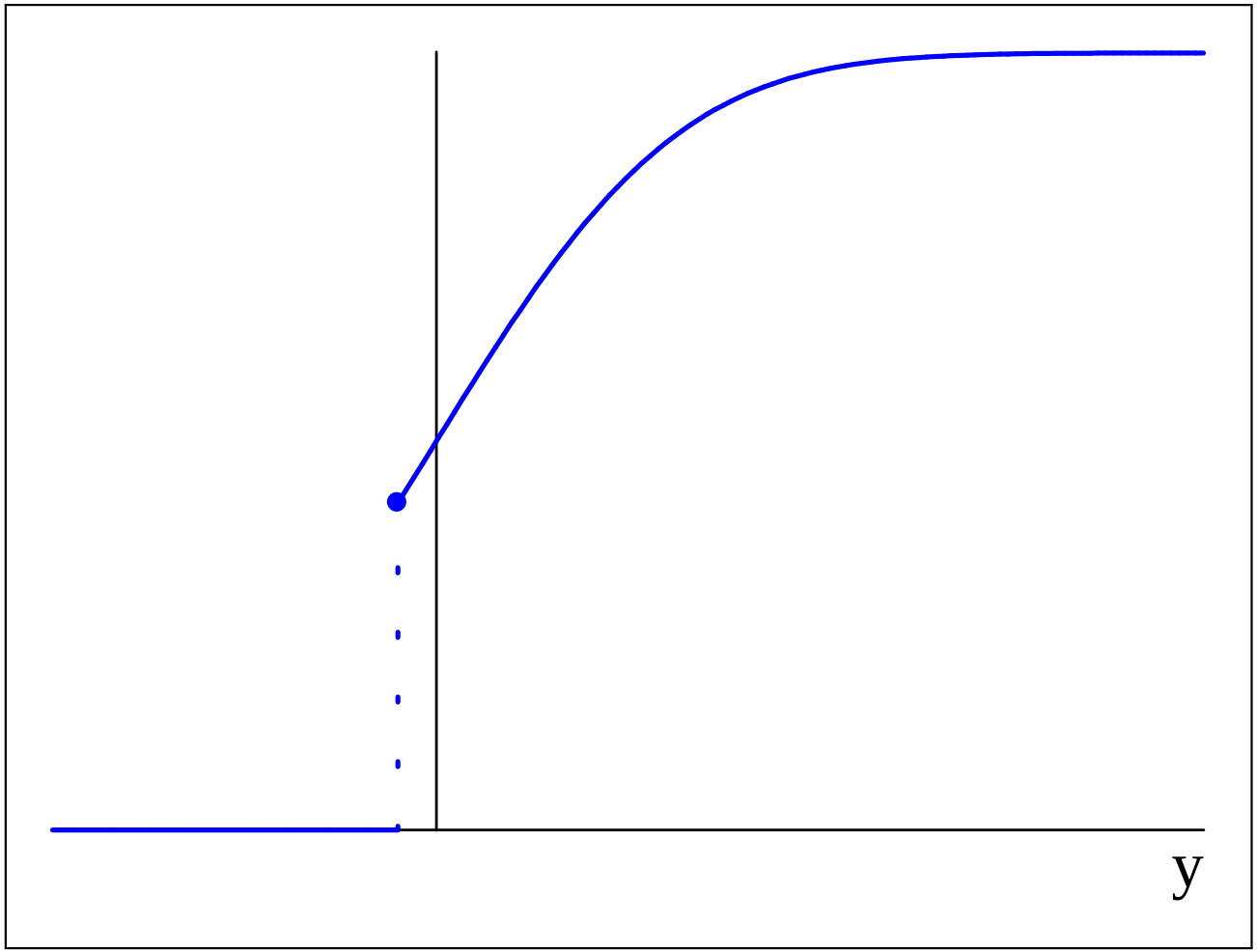
$$\mathcal{P}(Y = j) = \begin{cases} 0 & \text{if } j < a \\ \mathcal{P}(Y^* \leq a) & \text{if } j = a \\ \mathcal{P}(Y^* = j) & \text{if } j > a \end{cases}.$$

- If Y^* is a continuous random variable with pdf $f(y^*)$, Y has the **mixed distribution**

$$\mathcal{P}(Y \leq k) = \begin{cases} 0 & \text{if } k < a \\ \mathcal{P}(Y^* \leq k) & \text{if } k \geq a \end{cases}.$$

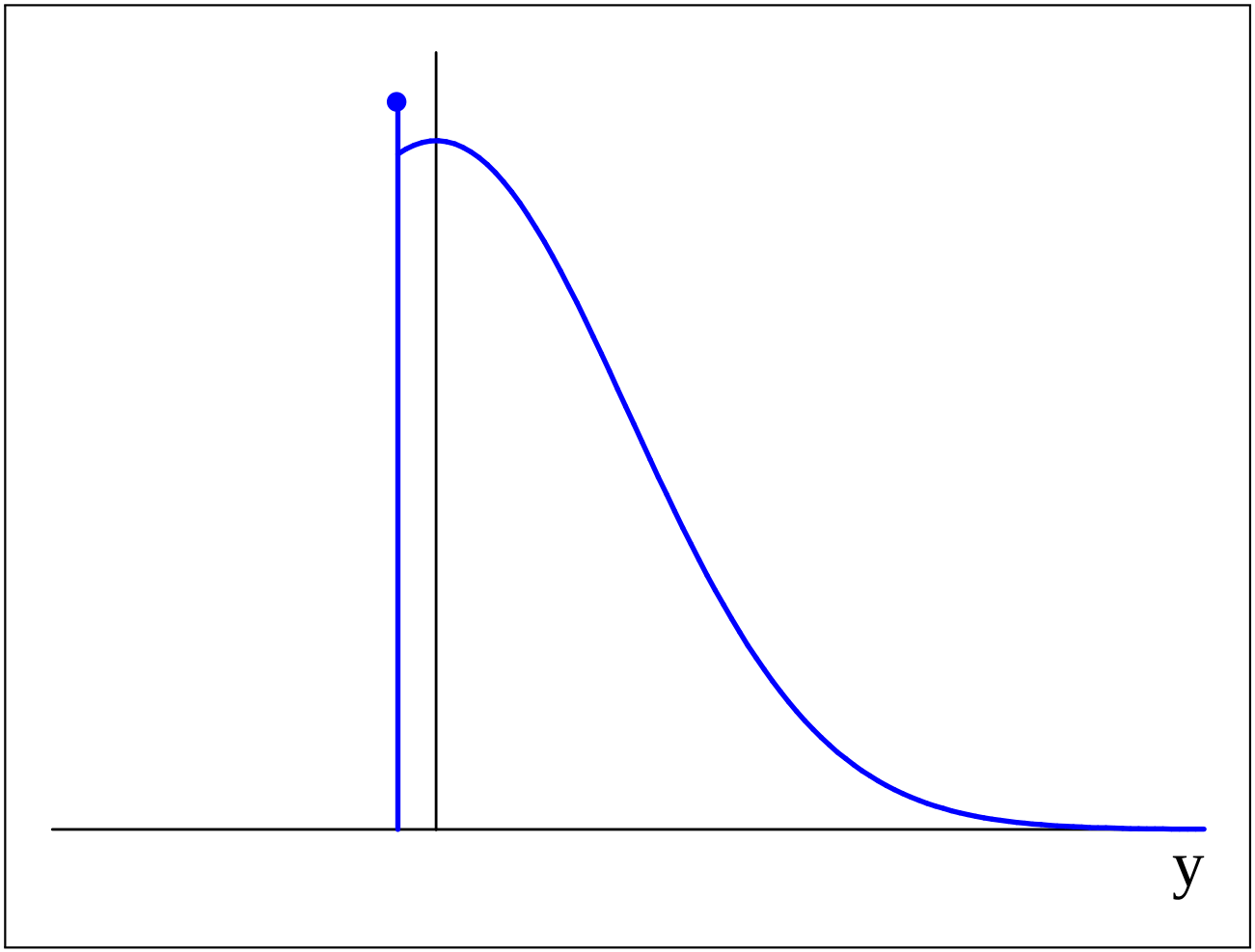
- It is assumed that the *regressors are fully observed*, even for cases with censored Y^*.

Censored data



Censored distribution

Censored data



Censored density

Censored data

- The standard regression model for continuous censored data is the *Tobit* (normal censored regression model) which was proposed by Tobin (1958).
- The term Tobit was derived from Tobin's name by truncating and adding -it by analogy with the probit model.
- The Tobit for censoring at zero is defined by

$$\begin{aligned} Y_i^* &= \mathbf{X}_i'\boldsymbol{\beta}_0 + u_i^*, \qquad u_i^* \sim \mathcal{N}\left(0, \sigma^2\right). \\ Y_i &= \begin{cases} 0 & \text{if } Y_i^* \leq 0 \\ Y_i^* & \text{if } Y_i^* > 0 \end{cases}. \end{aligned}$$

- The model is appropriate when we are interested in making inference about Y_i^*, not Y_i.
- Note that in this case

$$\mathrm{E}\left(Y_i|\mathbf{X}_i\right) = \mathbf{X}_i'\boldsymbol{\beta}_0\Phi\left(\frac{\mathbf{X}_i'\boldsymbol{\beta}_0}{\sigma}\right) + \sigma\phi\left(\frac{\mathbf{X}_i'\boldsymbol{\beta}_0}{\sigma}\right).$$

Censored data

- ML inference is based on the likelihood function

$$L(\beta,\sigma) = \prod_{i=1}^{n} \left[1 - \Phi\left(\frac{\mathbf{X}_i'\beta}{\sigma}\right)\right]^{\mathbf{1}(Y_i=0)} \left[\frac{1}{\sigma}\phi\left(\frac{y_i - \mathbf{X}_i'\beta}{\sigma}\right)\right]^{\mathbf{1}(Y_i>0)}.$$

- Reparameterizing $L(\theta)$ with $\gamma = \beta/\sigma$ and $\omega = 1/\sigma$, the Hessian is *negative definite* and estimation is easy, but in some situations the estimates may not exist.
- Consistency depends on the **distributional assumptions** (but independence is not required) and appropriate specification tests are available.

Correcting OLS

- There is a simpler estimator based on OLS.

Note the following results:

- Use sample with $Y_i > 0$: We now from the results on truncated data that

$$\begin{aligned} E\left(Y_i|\mathbf{X}_i, Y_i > 0\right) &= \mathbf{X}_i'\boldsymbol{\beta}_0 + \sigma\frac{\phi\left(-\mathbf{X}_i'\boldsymbol{\beta}_0/\sigma\right)}{1-\Phi\left(-\mathbf{X}_i'\boldsymbol{\beta}_0/\sigma\right)} \\ &= \mathbf{X}_i'\boldsymbol{\beta}_0 + \sigma\frac{\phi\left(\mathbf{X}_i'\boldsymbol{\beta}_0/\sigma\right)}{\Phi\left(\mathbf{X}_i'\boldsymbol{\beta}_0/\sigma\right)}. \end{aligned}$$

- The function $\lambda\left(\mathbf{v}\right) = \phi\left(\mathbf{v}\right)/\Phi\left(\mathbf{v}\right)$ is known as the inverse Mills ratio.
- Define a dummy variable D_i as:

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

Correcting OLS

- Then

$$\begin{aligned} P[D_i = 1 \,|\mathbf{X}_i] &= P[Y_i > 0 \,|\mathbf{X}_i] &&= P[Y_i^* > 0 \,|\mathbf{X}_i] \\ &= P[u_i^* > -\mathbf{X}_i'\boldsymbol{\beta}_0 \,|\mathbf{X}_i] &&= \Phi\left(\frac{\mathbf{X}_i'\boldsymbol{\beta}_0}{\sigma}\right) \end{aligned}$$

This motivates the following 2-step estimator

Step 1: Estimate $\boldsymbol{\beta}_0/\sigma$ by using a Probit model for $P[D_i = 1 \,|\mathbf{X}_i]$ on the full sample. Use to construct consistent estimates of

$$\lambda_i = \frac{\phi\left(\mathbf{X}_i'\boldsymbol{\beta}_0/\sigma\right)}{\Phi\left(\mathbf{X}_i'\boldsymbol{\beta}_0/\sigma\right)}$$

(denoted by $\hat{\lambda}_i$) for each observation.

Step 2: OLS regression of Y_i on $\mathbf{X}_i$ and $\hat{\lambda}_i$ using observations for which $Y_i > 0$. Gives consistent estimates of $\boldsymbol{\beta}_0$ and σ.

Sample selection

- In many cases, the sample available depends on individual decisions (self-selection).
- **Example** (wage equation): Suppose that a researcher wants to estimate the determinants of wages, but has access to wage observations for only those who work. Since people who work are selected non-randomly from the population, estimating the determinants of wages from the subpopulation who work may introduce bias.
- Heckman (1979) considered the following model: Suppose that we are interested in the regression

$$Y_i = \mathbf{X}_i'\boldsymbol{\beta}_0 + \varepsilon_i, \qquad \mathrm{E}\left(\varepsilon_i|\mathbf{X}_i\right) = 0,$$

but Y_i is observable only when $Z_i = 1$, where

$$Z_i = \mathbf{1}\left(\mathbf{X}_i'\gamma_0 + u_i > 0\right).$$

- In the example of the wage equation Z_i is 1 for those that work and 0 otherwise. $\mathcal{P}(Z_i = 1|\mathbf{X}_i)$ gives the probability of working.

Sample selection

- If ε_i and u_i are not independent, OLS estimation (based on the observation with $Z_i = 1$) of $\boldsymbol{\beta}_0$ is inconsistent because $\mathrm{E}\left(\varepsilon_i | \mathbf{X}_i, u_i > -\mathbf{X}_i'\boldsymbol{\gamma}_0\right) \neq 0$.
- Heckman (1976) popularized an estimator of $\boldsymbol{\beta}_0$ which is consistent under the assumption

$$\begin{bmatrix} u_i \\ \varepsilon_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma^2 \end{bmatrix} \right).$$

Sample selection

- Under this assumption, β_0, γ_0, ρ and σ can be estimated by maximum likelihood.
- Alternatively, Heckman (1976) used a two-step estimator (*Heckit*) based on the result

$$\begin{aligned} \mathrm{E}\left(\varepsilon_i | \mathbf{X}_i, u_i > -\mathbf{X}_i'\gamma_0\right) &= \rho\sigma \frac{\phi\left(\mathbf{X}_i'\gamma_0\right)}{\Phi\left(\mathbf{X}_i'\gamma_0\right)}, \\ \mathrm{E}\left(Y_i | \mathbf{X}_i, Z_i = 1\right) &= \mathbf{X}_i'\beta_0 + \rho\sigma \frac{\phi\left(\mathbf{X}_i'\gamma_0\right)}{\Phi\left(\mathbf{X}_i'\gamma_0\right)} \end{aligned}$$

- Therefore, in the **first step**, $\mathbf{X}_i'\gamma_0$ can be estimated by a probit. Use to construct consistent estimates of

$$\lambda_i = \frac{\phi\left(\mathbf{X}_i'\gamma_0\right)}{\Phi\left(\mathbf{X}_i'\gamma_0\right)}$$

  (denoted by $\hat{\lambda}_i$) for each observation.
- In the **second step**, β and $\rho\sigma$ are estimated in the OLS regression of Y_i on $\mathbf{X}_i$ and $\hat{\lambda}_i$.

Sample selection

- A standard t-test for the significance of $\rho\sigma$ is a **valid test** for correlation between u_i and ε_i.
- In case u_i and ε_i are correlated, the covariance matrix of the second step has to be corrected to account for the **estimated regressor**.
- Identification is easier if there are **exclusion restrictions**.

Remark: Heckman shared the 2000 Nobel Memorial Prize in Economic Sciences with James McFadden. Heckman's share of the prize was for "*his development of theory and methods for analyzing selective samples*".