Advanced Econometrics I

Jürgen Meinecke

Lecture 11 of 12

Research School of Economics, Australian National University

Roadmap

- Limited Dependent Variable (LDV) Models
  - Binary Choice Models
  - Sample Selection Models

Limited dependent variables occur often

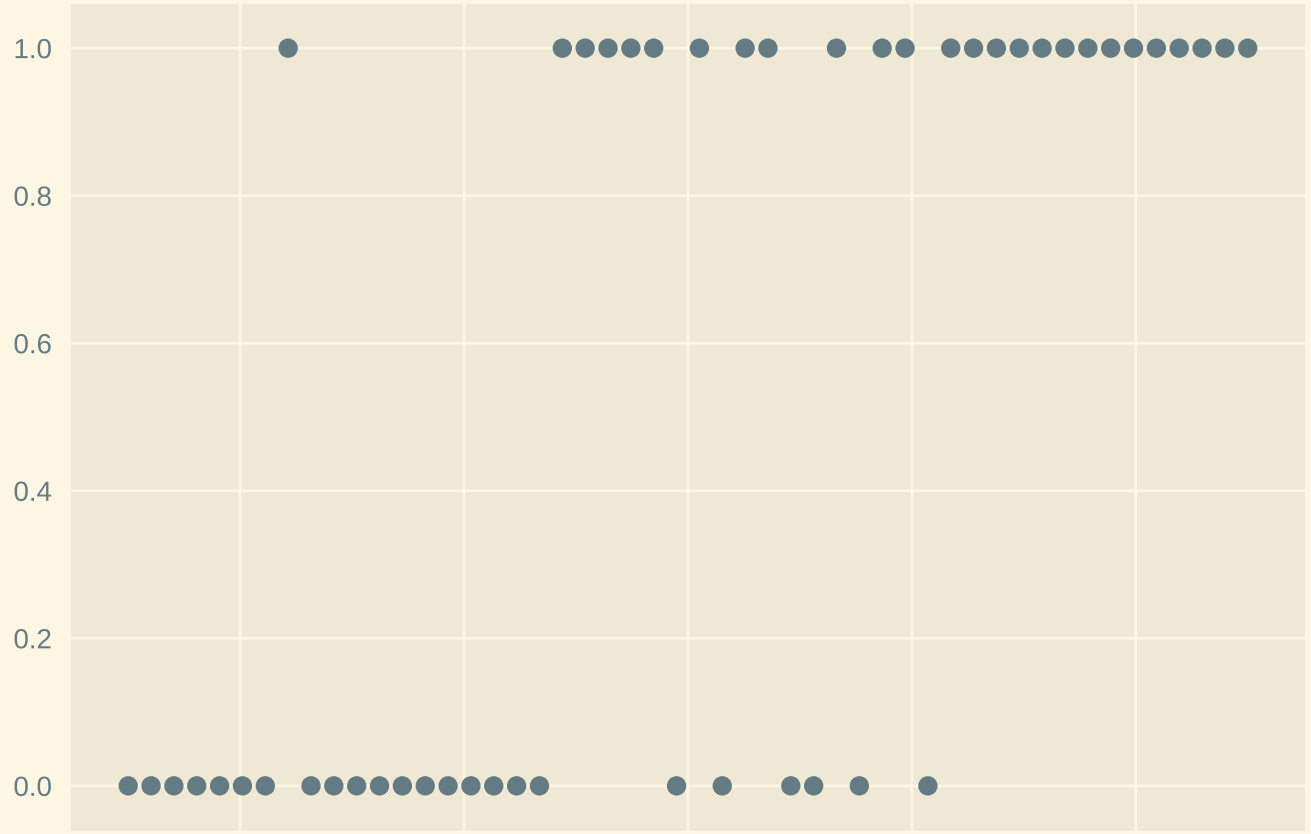
Examples

- binary outcome: $Y \in \{0, 1\}$
- multinomial outcome: $Y \in \{0, 1, \ldots, s\}$
- integer outcome: $Y \in \{0, 1, \ldots\}$
- censored outcome: $Y \in \mathbb{R}^+$

Typical method use to estimate LDV models: parametric MLE

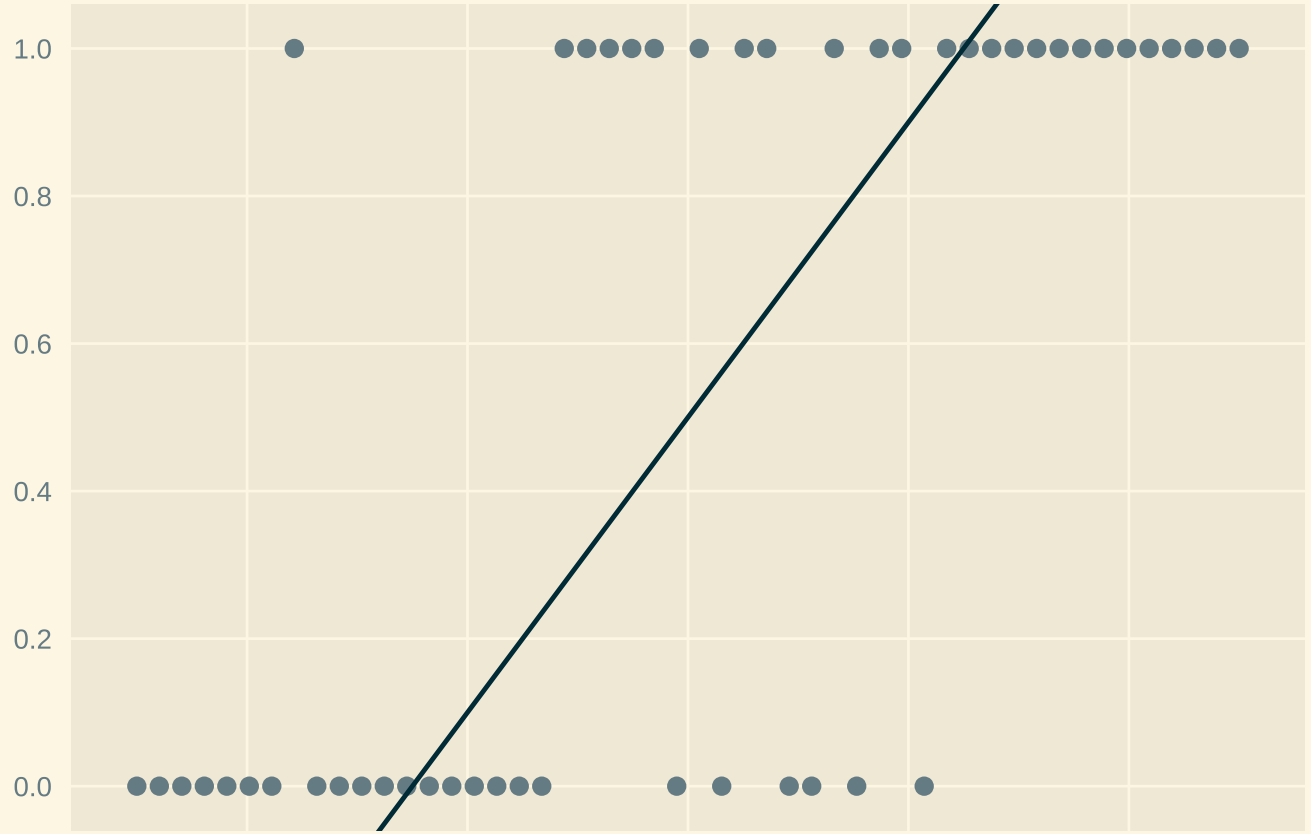
Note: we will disregard endogeneity as problem here

Binary outcomes: illustration of scatterplot



How about fitting a line?

Binary outcomes: fitting a line



Good idea?

The dependent variable is $Y \in \{0, 1\}$

It's easy to see that $\mathrm{E}(Y_i|X_i) = \Pr(Y_i = 1|X_i)$

At the same time $\mathrm{E}(Y_i|X_i) = X_i'\beta$ in the linear regression model

Combining gives $\Pr(Y_i = 1|X_i) = X_i'\beta$

This explains the term *linear probability model (lpm)*

The lpm is often a solid starting point for binary choice analysis

Before you do probit/logit estimation, you should always estimate a lpm first

Interpretation of β in the lpm:
the effect of X_i on the *probability of success* $\Pr(Y_i = 1|X_i)$

Limitations of the lpm

- Possibly $\widehat{\Pr}(Y_i = 1|X_i) < 0$ and $\widehat{\Pr}(Y_i = 1|X_i) > 1$
- linearity of the probability restrictive

However: if in an application you have reason to believe that your probabilities are not at the boundary and that things are locally linear, then lpm could be really good

If that is not the case, something else is needed

The following trick solves the above two limitations: let

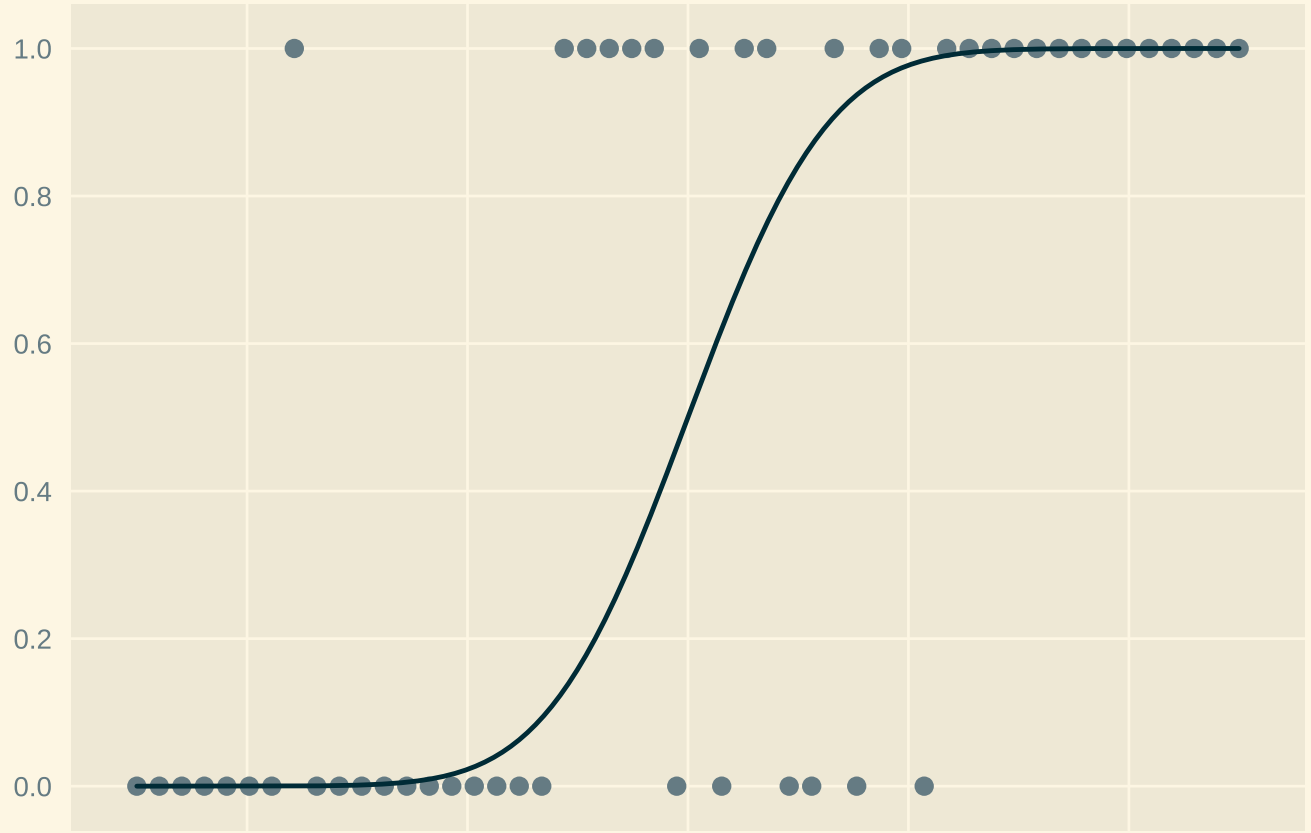
$$\Pr(Y_i = 1|X_i) = G(X_i'\beta)$$

where G is a known cumulative distribution function

Because G is a cdf, probabilities are bounded between 0 and 1

It also makes the probability of success evolve nonlinearly

Binary outcomes: fitting a curve instead



Good candidates: cumulative distribution functions!

Which specific functional form should we choose for G?

People do this

$$G(s) := \begin{cases} \Phi(s) & \text{probit model} \\ \Lambda(s) & \text{logit model,} \end{cases}$$

where

- $\Phi(s)$ is the cdf of the standard normal distribution
- $\Lambda(s) := \frac{\exp(s)}{1+\exp(s)}$ is the cdf of the logistic distribution with zero mean and unit variance

Both cdfs look similar, the logistic is heavier in the tails

How would you estimate β?

Consider the conditional mean functions

$$\mathrm{E}\left(Y_i|X_i\right) = \begin{cases} X_i'\beta & \text{linear probability model} \\ G(X_i'\beta) & \text{probit or logit model} \end{cases}$$

Looks like we could do linear regression in case of the lpm and, perhaps, nonlinear regression in the probit/logit case

Using OLS in the lpm case is good

But in the probit/logit case, people tend to use MLE

Let's study MLE now

Big picture: You've got an iid random sample (X_i, Y_i)

The joint likelihood function of the observed data is

$$L(\beta) = \prod_{i=1}^{N} \Pr\left(X_i = x_i, Y_i = y_i\right)$$

$$= \prod_{i=1}^{N} \Pr\left(Y_i = y_i | X_i = x_i\right) \Pr\left(X_i = x_i\right)$$

Here's how you think about the involved probabilities:
$\Pr\left(Y_i = y | X_i = x\right) = f_{Y|X}\left(y|x, \beta\right)$ and $\Pr\left(X_i = x\right) = f_X\left(x\right)$

Key here: β doesn't play a role in the pmf or pdf of X_i

Then

$$L(\beta) = \prod_{i=1}^{N} f_{Y|X}\left(y_i|x_i, \beta\right) f_X(x_i)$$

And the log likelihood function becomes

$$\ln L(\beta) = \sum_{i=1}^{N} \ln f_{Y|X}(y_i|x_i, \beta) + \sum_{i=1}^{N} \ln f_X(x_i)$$

You can see that

$$\operatorname*{argmax}_{\beta \in B} \ln L(\beta) = \operatorname*{argmax}_{\beta \in B} \sum_{i=1}^{N} \ln f_{Y|X}(y_i|x_i, \beta)$$

The function $\sum_{i=1}^{N} \ln f_{Y|X}(y_i|x_i, \beta)$ is called the **conditional log likelihood function**

I'm overloading the "$\ln L$" symbol:

$$\ln L(\beta) := \sum_{i=1}^{N} \ln f_{Y|X}(y_i|x_i, \beta)$$

(I've snipped off the $\sum_{i=1}^{N} \ln f_X(x_i)$ bit without causing harm)

In the binary outcome model, what is $f_{Y|X}\left(y|x,\beta\right)$ equal to?

Economists like the so called *latent variable representation*:

$$Y_i^* = X_i'\beta + e_i,$$

where e_i are randomly drawn and have cdf G

Notice that the *latent outcome* Y_i^* is not observed, but instead

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{else} \end{cases}$$

is observed

With $Y_i^* = X_i'\beta + e_i$, it follows

$$\begin{aligned}\Pr(Y_i = 1|X_i = x) &= \Pr(Y_i^* > 0|X_i = x) \\ &= \Pr(e_i > -x'\beta|X_i = x) \\ &= 1 - G(-x'\beta) \\ &= G(x'\beta)\end{aligned}$$

and consequently

$$\Pr(Y_i = 0|X_i = x) = 1 - G(x'\beta)$$

Notice: G needs to be symmetric at zero, $G(a) = 1 - G(-a)$

It follows that for $y \in \{0, 1\}$

$$f_{Y|X}(y|x, \beta) = \Pr(Y_i = y|X_i = x) = G(x'\beta)^y \cdot (1 - G(x'\beta))^{1-y}$$

This is a curious construction: we have shown that the conditional pmf of Y_i can be expressed in terms of the cdf G

The big picture is this: Y_i, given $X_i = x$ has a Bernoulli distribution with probability of success $G(x'\beta)$

The conditional pmf is $f_{Y|X}(y|x,\beta) = G(x'\beta)^y \cdot (1 - G(x'\beta))^{1-y}$ for $y \in \{0,1\}$

The log likelihood for the whole sample becomes

$$\begin{aligned}
\ln L(\beta) &= \sum_{i=1}^{N} \ln f_{Y|X}(y_i|x_i,\beta) \\
&= \sum_{i=1}^{N} \ln \left(G(x_i'\beta)^{y_i} \cdot (1 - G(x_i'\beta))^{1-y_i} \right) \\
&= \sum_{i=1}^{N} y_i \ln G(x_i'\beta) + \sum_{i=1}^{N} (1 - y_i) \ln(1 - G(x_i'\beta))
\end{aligned}$$

Repeating for convenience

$$\ln L(\beta) = \sum_{i=1}^{N} y_i \ln G(x_i'\beta) + \sum_{i=1}^{N} (1 - y_i) \ln(1 - G(x_i'\beta))$$

The log likelihood is well behaved

Proposition

Let G be the standard normal cdf or the logistic cdf. Then $\ln L(\beta)$ is globally concave.

Your computer can find a unique solution for you!

To find the maximizer, we need to obtain the score function:

$$S(y|x,\beta) = \frac{\partial \ln f_{Y|X}}{\partial \beta}(y|x,\beta)$$

Recall

$$f_{Y|X}(y|x,\beta) = G(x'\beta)^y \cdot (1 - G(x'\beta))^{1-y}$$
$$\ln f_{Y|X}(y|x,\beta) = y \ln G(x'\beta) + (1-y)\ln(1 - G(x'\beta))$$

Therefore

$$\begin{aligned} S(y|x,\beta) &= \begin{cases} \frac{1}{G(x'\beta)} \cdot g(x'\beta) \cdot x & \text{if } y = 1 \\ -\frac{1}{1-G(x'\beta)} \cdot g(x'\beta) \cdot x & \text{if } y = 0 \end{cases} \\ &= \frac{y - G(x'\beta)}{G(x'\beta)\,(1 - G(x'\beta))} \cdot g(x'\beta) \cdot x, \end{aligned}$$

where $g(z) = \frac{\partial G}{\partial z}(z)$

How would you use the score to find the minimizer?

Tell your computer to find β such that $\sum_{i=1}^{N} S(y_i|x_i, \beta) = 0$

(there's no closed form solution, you need a computer)

That will be the MLE

Of course you need to let your computer know what the functional forms of g and G are, but that's easy

Let's say you have obtained $\hat{\beta}^{\text{ML}}$

What is its distribution?

Luckily we've done all the hard work last week

For ML estimators, we obtained the generic result

$$\sqrt{N}\left(\hat{\theta}^{\mathrm{ML}} - \theta\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, I(\theta)^{-1}\right)$$

Translating this to the current setting

$$\sqrt{N}\left(\hat{\beta}^{\mathrm{ML}} - \beta\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, I(\beta)^{-1}\right)$$

where

$$I(\beta) = \mathrm{E}\left(S(Y_i|X_i, \beta) \cdot S(Y_i|X_i, \beta)'\right)$$

Working this out...

$$
\begin{aligned}
I(\beta) &= \mathsf{E}\left(S(Y_i|X_i,\beta)S(Y_i|X_i,\beta)'\right) \\
&= \mathsf{E}\left(\mathsf{E}\left(S(Y_i|X_i,\beta)S(Y_i|X_i,\beta)'|X_i\right)\right) \\
&= \mathsf{E}\left(\mathsf{E}\left(\frac{(Y_i - G(X_i'\beta))^2}{G(X_i'\beta)^2\left(1-G(X_i'\beta)\right)^2}\cdot g(X_i'\beta)^2 \cdot X_iX_i' \middle| X_i\right)\right) \\
&= \mathsf{E}\left(\left(\frac{g(X_i'\beta)^2}{G(X_i'\beta)^2\left(1-G(X_i'\beta)\right)^2}\cdot X_iX_i'\right)\mathsf{E}\left((Y_i - G(X_i'\beta))^2|X_i\right)\right) \\
&= \mathsf{E}\left(\frac{g(X_i'\beta)^2}{G(X_i'\beta)\left(1-G(X_i'\beta)\right)}\cdot X_iX_i'\right)
\end{aligned}
$$

because $Y_i|X_i$ is Bernoulli with $\Pr(Y_i = 1|X_i) = G(X_i'\beta)$, so that $\mathsf{E}\left((Y_i - G(X_i'\beta))^2|X_i\right) = \mathsf{Var}\left(Y_i|X_i\right) = G(X_i'\beta)\cdot(1-G(X_i'\beta))$

How would you estimate the asymptotic variance?

Repeating from the previous slide

$$I(\beta) = \mathsf{E}\left(\frac{g(X_i'\beta)^2}{G(X_i'\beta)\left(1 - G(X_i'\beta)\right)} \cdot X_i X_i' \right)$$

Proposition

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{g(X_i'\hat{\beta}^{ML})^2}{G(X_i'\hat{\beta}^{ML})\left(1 - G(X_i'\hat{\beta}^{ML})\right)} \cdot X_i X_i' \right) = I(\beta) + o_p(1)$$

Now that we know how to estimate β and how to obtain its asymptotic distribution and have a way to estimate it, what do we do with all that information?

We are not actually interested in β per se

To see this, recall from lecture 4 that we mostly care about estimating causal effects $C_k(X_1, \ldots, X_M)$

We saw that in the linear regression model, by imposing conditional mean independence, the causal effect was equal to $\partial\mu(X)/\partial X_k$, where $\mu(X) = \mathrm{E}(Y_i|X_i)$

The β's are the causal effects and $\hat{\beta}^{\mathrm{OLS}}$ are their estimators

Yet in the binary choice model

$$\mathrm{E}(Y_i|X_i = x) = \Pr(Y_i = 1|X_i = x) = G(x'\beta) \neq x'\beta$$

The function G is nonlinear and the β's are not the causal effects

We still have conditional mean independence, so presumably we're still interested in $\partial\mu(X)/\partial X_k$

Our research objective therefore is

$$\frac{\partial \mathrm{E}\left(Y_i|X_i\right)}{\partial X_i} = \frac{\partial \Pr\left(Y_i = 1|X_i\right)}{\partial X_i} = g(X_i'\beta)\beta \neq \beta$$

Due to nonlinearity of G, the causal effects are functions of X_i

Now, of course, we cannot evaluate an *individual* causal effect but look at this instead

$$\psi = \mathrm{E}\left(\frac{\partial \Pr\left(Y_i = 1|X_i\right)}{\partial X_i}\right) = \mathrm{E}\left(g(X_i'\beta)\beta\right)$$

It should be easy to estimate ψ via analog estimation:

$$\hat{\psi} = \frac{1}{N}\sum_{i=1}^{N}\left(g(X_i'\hat{\beta}^{\mathrm{ML}})\hat{\beta}^{\mathrm{ML}}\right)$$

What is the asymptotic distribution of $\hat{\psi}$?

Lemma (Delta Method)

Let $\sqrt{N}\left(\hat{\theta}-\theta\right) \xrightarrow{d} \mathcal{N}(0,\Omega)$ *with* $\dim\theta = K$. *Take a continuously differentiable function* $C:\Theta\to\mathbb{R}^Q$ *where* $Q \leq K$. *Then*

$$\sqrt{N}\left(C(\hat{\theta})-C(\theta)\right) \xrightarrow{d} \mathcal{N}\left(0, c(\theta)\cdot\Omega\cdot c(\theta)'\right),$$

where $c(\theta) := \frac{\partial C}{\partial \theta'}(\theta)$ *and* $\dim c(\theta) = Q\times K$.

The delta method is useful for establishing limiting distributions for functions of estimators

In our case: $C(b) = g(X_i'b)b$

Roadmap

Limited Dependent Variable (LDV) Models

- Binary Choice Models
- Sample Selection Models

The traditional, Heckman style, sample selection model is a combination of a linear regression and a binary choice model

$$Y_i^* = X_i'\beta + e_i$$
$$D_i = 1 \cdot \left(Z_i'\gamma + v_i > 0\right),$$

where X_i and Z_i could be identical (but don't have to)

You are interested in β but you do not observe (X_i, Y_i^*)

Instead you observe (D_i, X_i, Y_i, Z_i) where

$$Y_i := \begin{cases} Y_i^* & \text{if } D_i = 1 \\ \text{unobserved} & \text{if } D_i = 0 \end{cases}$$

Typical example:

- Y_i^* are earnings
- D_i is a work dummy

Earnings are observed only for people who work

A lot of people don't work

Ideally you would like to know, *what would these people earn if they worked*?

If you had that information, you could regress Y_i^* on X_i

But you do not have that information

Instead, you observe earnings only for the sub-sample of workers

That sub-sample may be a selective subset of the entire population

It seems possible that people who work differ systematically from people who do not work

To be more concrete, let X_i be education
(and ignore other regressors)

So you're interested in the effect of education on earnings

Hopefully you believe that $\beta > 0$

Suppose that β is increasing in education
(returns to education are increasing)

At the same time, let's suppose people with more education face higher opportunity cost of not working

This is to say that the sample of workers is not a representative sample from the entire population

Instead, workers are, on average, more educated

Running a regression of earnings on education would result in biased estimates

How do you estimate a sample selection model?

You make a restrictive (yet helpful) assumption about the joint distribution of the error terms, namely

$$\begin{pmatrix} e_i \\ v_i \end{pmatrix} \Big| (X_i, Z_i) \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_e^2 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

So we're assuming that the joint distribution of the errors

- is *exactly* normal;
- with correlation ρ;
- and variances σ_e^2 and $\sigma_v^2 = 1$ (normalization)

How does this help in estimating β?

Aside: we will show that we can estimate β, that is, the coefficient that is representative for the *entire* population, not merely the sub-sample for which $D_i = 1$

Joint normality implies

$$e_i = \rho v_i + w_i,$$

where w_i is independent of $v_i \sim \mathcal{N}(0,1)$

Random fact about standard normal distribution:

$$\mathsf{E}\left(v_i | v_i > -c\right) = \frac{\phi(c)}{\Phi(c)} =: \lambda(c)$$

The function λ is called the *inverse Mills ratio*

Notice that

$$\begin{aligned}
&\mathsf{E}\left(e_i | D_i = 1, X_i = x, Z_i = z\right) \\
&\quad = \mathsf{E}\left(e_i | v_i > -z'\gamma\right) \\
&\quad = \rho \mathsf{E}\left(v_i | v_i > -z'\gamma\right) + \mathsf{E}\left(w_i | v_i > -z'\gamma\right) \\
&\quad = \rho\lambda\left(z'\gamma\right)
\end{aligned}$$

Therefore

$$\begin{aligned}\mathsf{E}\left(Y_i|X_i,Z_i\right) &= \mathsf{E}\left(Y_i^*|D_i=1,X_i,Z_i\right)\\ &= X_i'\beta+\mathsf{E}\left(e_i|D_i=1,X_i,Z_i\right)\\ &= X_i'\beta+\rho\lambda\left(Z_i'\gamma\right)\end{aligned}$$

Given an iid random sample (D_i,Y_i,X_i,Z_i) we face the following regression model

$$Y_i = X_i'\beta+\rho\lambda\left(Z_i'\gamma\right)+w_i,$$

Notice that we do know the functional form of λ

So the above model is a regression model with two sets of regressors: X_i and $\lambda(Z_i'\gamma)$

The corresponding coefficients are β and ρ

So the equation

$$Y_i = X_i'\beta + \rho\lambda\left(Z_i'\gamma\right) + w_i,$$

suggests running OLS of Y_i on X_i and $\lambda(Z_i'\gamma)$

Only problem: we don't know γ

But we can estimate it via simple probit ML estimation!

Connecting the dots, here is Heckman's two step estimator

(i) run a probit estimation of D_i on Z_i and obtain $\hat{\gamma}^{\text{ML}}$

(ii) run OLS of Y_i on X_i and $\lambda(Z_i'\hat{\gamma}^{\text{ML}})$ and obtain $\hat{\beta}^{\text{OLS}}$ and $\hat{\rho}^{\text{OLS}}$

While you can follow these two stages literally in, for example, Stata, the resulting standard errors will be incorrect

Source of the problem: you are using $\hat{\gamma}^{\text{ML}}$ instead of γ

This is similar to two stage least squares estimation where you are using $\hat{\pi}^{\text{OLS}}$ instead of π

The correct asymptotic approximation for $\hat{\beta}^{\text{OLS}}$ needs to take into account the added sampling uncertainty that comes from using $\hat{\gamma}^{\text{ML}}$ instead of γ

Another comparison to IV estimation is interesting:

You should NOT be using $X_i = Z_i$

While it is mechanically possible to do so, because you are sending Z_i through a nonlinear function λ, in practice $\lambda(Z_i'\hat{\gamma}^{\text{ML}})$ can be highly correlated with X_i

And even if it isn't highly correlated, it is poor practice because your identification of β is solely based on making a very restrictive assumption on the joint distribution of the error terms

Econometricians say that β is only identified through imposing a specific *functional form* on the joint error distribution

Instead, it is preferrable to have at least one variable in Z_i that is not included in X_i

In the IV estimation parlance, this is referred to as an *exclusion restriction*

If you do have an exclusion restriction, then your identification rests on two things: functional form and exclusion restriction