TA Session 6

Bruno Amat

November 6, 2024

Outline

1. Instrumental Variables

2. Identification

3. Limited Dependent Variable and Selection Models

Instrumental Variables

INTRODUCTION

One of the assumptions to estimate an OLS model is the exogeneity condition. This condition implies that the observable elements are not related to the error component.

Mathematically, this condition is given by:

$$E[X\epsilon] = 0$$

EXAMPLE

However, sometimes this hypothesis is not satisfied. One example is the education variable. Suppose we want to estimate the following model like Angrist and Krueger (1991):

$$\log(W_i) = X_i\beta + \rho E_i + \epsilon$$

where $\log(W_i)$ is the log income of individual i, X_i is a vector of covariates and E_i is the education of the individual i.

We want to estimate the return of education on wages. Hence, we will focus on ρ

EXAMPLE

Note that there might be other variables related to education that are not accounted in the model.

For example richer parents are able to finance best education to their students when compared to poorer ones. Hence, the variable of parents income is present on the error term and it is correlated with the education of the son.

Hence, we have an endogeneity problem, i.e. $E[\epsilon|X_i, E_i] \neq 0$.

How can we solve the Endogeneity problem?

We need to define an **Instrumental Variable (IV)** to capture the variation of the education at the same time that it is not correlated with the error term.

Hence, the instrumental variable should satisfy two conditions:

Exogeneity: $E[\epsilon|Z] = 0$
Relevance: $E[XZ] \neq 0$

How can we solve the Endogeneity problem?

In the Example, various papers have tried to determine good instruments for education such as: education of the parents, distance to the school and the quarter of births.

Let Z_2 be the instrumental variable for E and let X be exogenous. Hence, the vector of instrumental variables is given by $Z = [X, Z_2]'$

CONDITIONS

Additionally to the conditions of Exogeneity and Relevance, we need one more condition to identify the parameter: the dimension of the instrument vector should be at least equal to endogenous variable vector.

If it is equal we have a **just-identified** model. If we have more instruments than endogenous variables we have a **over-identified** model.

Reduced Form:

To simplify the problem let's look at the model with only E_i. We already know that:

$$\log(W_i) = \rho E_i + \epsilon$$

Moroever,

$$E_i = \Gamma' Z_2 + u$$

Hence, the reduced form can be written as:

$$\log(W_i) = \rho(\Gamma' Z_2 + u) + \epsilon$$
$$\log(W_i) = \rho\Gamma' Z_2 + \rho * u\epsilon$$
$$\log(W_i) = \lambda Z_2 + v_i$$

IV Estimator

Suppose we are on the just-identified case.

Estimating $E_i = \Gamma' Z_2 + u$ we obtain that $\hat{\Gamma} = (Z'Z)^{-1}(Z'E)$.

When estimating $\log(W_i) = \lambda Z_2 + v_i$, we obtain $\hat{\lambda} = (Z'Z)^{-1}(Z'Y)$.

Hence,

$$
\begin{gathered}
\hat{\lambda} = \hat{\beta}\hat{\Gamma}' \\
(Z'Z)^{-1}(Z'Y) = \hat{\beta}(Z_2'Z)^{-1}(Z'E) \\
\hat{\beta} = (Z'E)^{-1}(Z'Y)
\end{gathered}
$$

This is the IV estimator.

2SLS

Now, suppose we are on the over-identified case.

We know that $\log(W_i) = \rho\Gamma' Z_2 + v$. Let $w = \Gamma Z_2$. The estimator for ρ is given by:

$$\hat{\rho} = (w'w)^{-1}(w'Z)$$

We can obtain the predicted value of the regression of E_i on Z_2. which is given by: $\hat{w} = Z'(Z'Z)^{-1}Z$.

Combining the two we have:

$$\hat{\lambda} = (E'Z(Z'Z)^{-1}Z'E)(E'Z(Z'Z)^{-1}Z'y)$$

This is the 2sls estimator.

2SLS

How to estimate it?

- Reg E on Z and obtain $\hat{E}$
- Reg Y on $\hat{E}$

Identification

What is Identification?

During class you learned the formal definition of Identification. Here I am going to explain my understand about Identification and its importance.

Identification is a set of necessary restriction/hypothesis made to identify parameters and consequently have economic understand about this variables.

What is Identification?

Look at the reduced form of the IV estimation.

$$\log(W_i) = \rho\Gamma' Z_2 + v_i$$

We are interested on estimating ρ. Can you guess what is the Identification hypothesis behind this estimation?

What is Identification?

The main hypothesis to identify the return of education is the Relevance hypothesis: $E[ZE] \neq 0$.

Note that if $E[ZE] = 0$ then $\Gamma = 0$ and the estimation is given by:

$$\log(W_i) = 0'Z_2 + v_i$$

Hence, we cannot know if there is return for education since our instrument is not relevant we cannot evaluate the parameter ρ.

Parametric Identification

In structural econometrics we need to define the utility function of the agent to obtain the predicted results.

For example the household utility:

$$U(c^W, l^W, c^H, l^H) = \mu\, u(c^W, l^W) + (1-\mu)\, u(c^H, l^H)$$

the household utility is given by the weighted sum of the utility of the husband and the wife

We need to come up with a parametric identification to find μ and the elasticities between leisure and consumption for both wife and husband.

Partial Identification

Until know we always adopted moment equality in our models. However, a common practice on Industrial Organization problems is to work with moment inequalities.

For example in a Dynamic discrete choice it is impossible to account for all possible decisions. Hence we use revealed preferences to identify the density of the distribution.

Therefore, instead of obtaining a point estimation we obtain a set of possible results for the estimation

Limited Dependent Variable and Selection Models

PROBIT

First, let's review the Probit model. We want to estimate the following model

$$y^* = X\beta + \epsilon$$

However, we are not able to observe y^*. We observe y that can that takes value equal to 0 or 1

$$y = \begin{cases} 1 \text{ if } y^* > 0 \\ 0 \text{ otherwise} \end{cases}$$

Probit

Hence to estimate the model we rely on the hypothesis that ϵ follows a $N(0, \sigma^2)$.

$$\begin{aligned} P(y=1) = P(y^* > 0) &= P(X\beta + \epsilon > 0) \\ = P(\epsilon > -X\beta) &= P\left(\frac{\epsilon}{\sigma} > \frac{-X\beta}{\sigma}\right) \\ &= P\left(\frac{\epsilon}{\sigma} < \frac{X\beta}{\sigma}\right) \\ &= \Phi\left(\frac{X\beta}{\sigma}\right) \end{aligned}$$

Probit

Therefore,

$$P(y = 1) = \Phi\left(\frac{X\beta}{\sigma}\right)$$
$$P(y = 0) = 1 - \Phi\left(\frac{X\beta}{\sigma}\right)$$

Hence, we estimate the model using by Maximizing the Log Likelihood

$$L(\beta) = \sum y \log\left(\Phi\left(\frac{X\beta}{\sigma}\right)\right) + (1-y)\log\left(1 - \Phi\left(\frac{X\beta}{\sigma}\right)\right)$$

Censored Data

- A regression model is censored when the recorded data on the dependent variable cuts off outside a certain range with multiple observations at the endpoints of that range
- Consequently, variation in the dependent variable will understate the effect of the regressors and the ordinary least squares using censored data will be biased

The Censored Variable Problem

Let y be the observed dependent variable, while y^* is the true dependent variable. we observe y^* in the range [a,b]. Therefore,

$$y = \begin{cases} a \text{ if } x'\beta + \epsilon < a \\ b \text{ if } x'\beta + \epsilon > b \\ x'\beta + \epsilon \text{ otherwise} \end{cases}$$

Assuming $a = 0$ and $b \to \infty$. we have $y = max(X\beta + \epsilon, 0)$

How to solve it?

We can use the Tobit estimation. Note that the distribution of our data is truncated in 0. Consequently, when assuming $\epsilon \sim N(0, \sigma)$ we have

1) If $y^* < 0$ The contribution for the likelihood is given by

$$P(y^* < 0) = P(-X\beta < \epsilon) = 1 - \Phi\left(\frac{X\beta}{\sigma}\right)$$

2) If $y^* > 0$ the contribution of the likelihood is given by

$$P(y^* > 0)\phi(y^*|y^* > 0) = \Phi\left(\frac{X\beta}{\sigma}\right)\frac{1}{\sigma}\frac{\phi((y - X\beta)/\sigma)}{\Phi(X\beta/\sigma)}$$

How to solve it?

The Log-likelihood function is given by:

$$L(\beta) = \sum (1 - D_i) \log \left(1 - \phi\left(\frac{X\beta}{\sigma}\right)\right) + D_i \left(\log\left(\phi\left(\frac{y - X\beta}{\sigma}\right)\right) - \log\sigma\right)$$

How to solve it?

We can use a Semiparametric approach:

- Censored Least Absolute Deviation (CLAD)
- Symmetrically Censored Least Squares (SCLS)
- Identically Censored Least Absolute Deviation (ICLAD)

We are not going to focus on that

Truncated Data

Suppose we have data on how much each student can lift on the chest press.

The students that go to the gym have positive value for the weight they can lift. However, the students who do not go to the gym have value equal to 0.

Hence, we have a problem of selection bias. Since we are not able to capture how all students can lift on the chest press, but only the students that go to the gym.

Truncated Data

Hence, we want to estimate the following model

$$y_1^* = X_1\beta_1 + \epsilon_1$$

However, we are not able to observe y^*. We observe y defined as

$$y_1 = \begin{cases} y_1^* \text{ if } y_2^* = 1 \\ 0 \text{ otherwise} \end{cases}$$

where $y_2^* = X_2\beta_2 + \epsilon_2$

Heckit

The idea of the Heckit estimation is to introduce the inverse mills ratio into the equation that would account for the probability of a student to go the gym.

$$y_1^* = X_1\beta_1 + \sigma_{12}\lambda(X_2\beta_2) + \epsilon_1$$

where $\lambda(X_2\beta_2)$ is the inverse mills ratio defined as:

$$\frac{\phi(X_2\hat{\beta}_2)}{\Phi(X_2\hat{\beta}_2)}$$

Heckit

Two-step Heckit:

1) Estimate β_2 by Probit

2) Obtain the inverse mills ratio and add it as a parameter on the estimation of y_1^*