Research Methods

Carlos Noton

Term 2 - 2012

Outline

1. Econometrics in a nutshell: Variation and Identification

2. Linear Regression Model in STATA

Main Assumptions

Dependent variable or outcome Y is the result of two forces: 1) Some **observable** characteristics or features denoted by X; and 2) some random **unobservable** shocks denoted by ε.

$$Y = f(X, \varepsilon)$$

Values of X are totally independent of shocks ε.
This is equivalent that some exogenous force (call it Nature or God) set X. Then, a different God played with a roulette and drew the values of ε. Given these two values, Y is uniquely determined through function f. ε is the non-systematic component (everything but X).

Assumption on Linear Specification

Suppose K explanatory variables $X_1, X_2, .., X_K$. Suppose each shock ε_i is **i.i.d.** with mean zero and constant variance $V(\varepsilon_i) = \sigma_\varepsilon^2$.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_K X_{Ki} + \varepsilon_i$$

for $i \in \{1, .., N\}$. Therefore the relationship $f(\cdot)$ is linear and fully determined by a K-vector of parameters β plus the variance σ_ε^2. We can re-write it in matrix notation:

$$Y = X\beta + \varepsilon$$

where Y and ε are column vectors of dimension $N \times 1$,X is a matrix of dimension $N \times K$and β is a vector of dimension $K \times 1$.

In empirical work, we only observe Y and X, and we have to come up with a clever guess (we call estimates) of β and ε.

Exogeneity Assumption

During this class, let us assume perfect exogeneity.

Example in Biology: Apply a vaccine to a random subset of the population to see the effects on Health status.

Some Examples in Economics:

- Impose large inflations in a random subset of countries to see effects on Economic Growth.
- Impose mergers of firms in a random subset of counties to see effects on prices and quantities.
- Impose certain level of education in a random subset of people to see effects on wages.

Critique to the Exogeneity Assumption

Does this exogeneity assumption apply in Economics?
Not really. Actually, we have theoretical models that argues that the exogeneity assumption is wrong. (We think that countries choose inflation, firms choose to merge or not, people choose their level of education.)
This is the so-called Endogeneity problem. Next class we will study how to overcome this sometimes ridiculous assumption.

Through the Exogeneity assumption, we assumed a causality effect. But simple things first. For today, let us assume Exogeneity. Given that, What do you need to identify marginal causal effects?

First: we need Variation!

Usually in Economics we have data where similar agents are facing similar treatments.

- If all the productive sectors within a country are facing the same tax rates, then it is hard to know the marginal effect of a change in tax rates.
- If all the workers within a firm are facing the same bonus policy, then it is hard to know the marginal effect of a change in that policy.
- If all the firms within a country are facing the same antitrust policy, then it is hard to know the marginal effect of a change in that policy.

Marginal Effects: β

Notice that $\frac{\partial Y}{\partial X_j} = \beta_j$. This is the marginal causal effects of X_j on outcome Y.

To infer from the data what is the marginal effect of characteristic X over some outcome Y, you need data on a population that has been exposed to different levels of X.
You could think of X as different levels of treatment (labor examples) or different environments (macro examples): Can I say something about inflation if I only observed low-inflation periods? Can I say something about unemployment if I observe employed people? Can I say something about competition if I only observed monopolies?

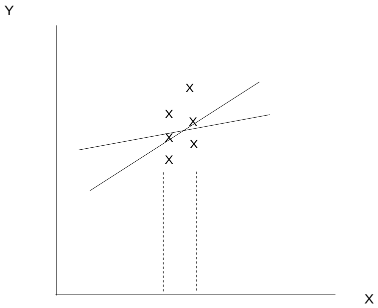
Simplest Linear Model

Suppose only one explanatory variable X

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

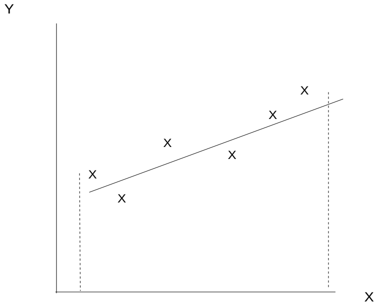
for $i \in \{1, .., N\}$. This estimation tries to fit the best linear function to two variables.

Little Variation in X



Little Variation in X:
Identification Problem

Decent Variation in X



More Variation in X is more informative

Linear Model with K explanatory variables

Suppose K explanatory variables $X_1, X_2, .., X_K$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_K X_{Ki} + \varepsilon_i$$

for $i \in \{1, .., N\}$ or in matrix notation:

$$Y = X\beta + \varepsilon$$

where Y and ε are column vectors of dimension $N \times 1$, X is a matrix of dimension $N \times K$ and β is a vector of dimension $K \times 1$.

Variation and Identification

Now, we have K explanatory variables $X_1, X_2, .., X_K$.
The same notion of variation now is needed in all the possible combinations of $X_1, X_2, .., X_K$.
For example: If whenever X_1 is high, X_2 is also high, then it is hard to distinguish what is really driven the Y outcome.

It is hard to identify the effect.

Economic Identification

Some Economic Examples:

- Suppose the outcome Y is wages, and two characteristics of the workers: gender and race. If I observed in the data that non-white woman have lower wages. Is it because of the race or the gender?
- Suppose the outcome Y is price, and two characteristics of the firm: age and size. If older firms are also larger, then what is really driven the outcome in prices?
- Suppose the outcome Y is GDP, and two characteristics of the country: inflation and tariff (trade openness). If countries with lower inflation also have lower tariffs, then I cannot identify the marginal effect of these characteristics in the Y outcome.

Formal Terms

$$\widehat{\beta}_{OLS} = (X'X)^{-1}(X'Y)$$

and

$$V(\widehat{\beta}_{OLS}) = \sigma_{\varepsilon}^2(X'X)^{-1}$$

Larger Variation is associated with large values in the diagonal of $(X'X)$ (consequently, small diagonal values of $(X'X)^{-1}$)

Showing this phenomena in STATA

See Program “IdentificationExercise.do”