Listening to Data

Valentino Dardanoni

University of Palermo

March 2020

What we are going to do

- Data and Models: hand in hand?
- We will discuss some basic preliminary notions that will help understanding what Data tell us.

Empirical analysis

- You want to study how a given variable(s) y (dependent variable, response) depends on another variable(s) x (explanatory variables, regressors, covariates).
- Interest is, for example, in understanding: Does Immigration Increase Crime?
- There is an excellent book with this title:
  Does Immigration Increase Crime? Migration Policy and the Creation of the Criminal Immigrant by Francesco Fasani, Giovanni Mastrobuoni, Emily G. Owens, Paolo Pinotti, Cambridge University Press, 2019.
- To understand it you may want to have unerstand some elementary notions of statistics and econometrics.

- Typical Data Structures:
  a) Passive, nonexperimental;
  b) Active, experimental;
  c) The 'natural experiment'.

- Data types:
  a) Cross section;
  b) Pure time series;
  c) Panel or Longitudinal data.

Conditional Expectation

- Much applied econometric studies estimate or test hypotheses about the expectation of y conditional on x.
- The key model to analyze conditional expectations is the **linear model**:

$$y_i = \alpha + \beta_1 x_i^1 + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + \varepsilon_i.$$

- The data $\{(y_i, \mathbf{x}_i)\}, i = 1, 2, ..., n$ can be, for example, an individual, a firm, an household, a region...

Parameter Estimation

- The model we are considering is composed of a set of parameters (the alpha and the betas).
- Statistical inference means find an estimate of the parameter, taking into account that uncertainty that characterize this process.
- The oldest and more common way to estimate the parameters is by using ordinary least squares (OLS):

  We have the data, but we don't know where the true line lies. What is the straight line that best fits the data points?

What can go wrong?

1. The conditional expectation model is wrong!
2. The nature of the error term...
3. Endogeneity: an explanatory variable x_j is said to be *endogenous* simply if it is correlated with ε. Usually arises in one of three ways: Omitted Variables, Measurement Error, Simultaneity.
4. Sample selection....

Can we fix it?

- Most of applied economics and econometrics deals with "fixing" these issues...
- We will look at a few techniques and methods of interest.

Instrumental Variable

- To use the IV approach with x_K endogenous, we need an observable variable, say z, **not** in regression equation that satisfies two conditions: **Exogeneity** and **Relevance**.
- When z satisfies these two conditions it is said to be an instrumental variable (IV) candidate for x_K.
- Informally, in attempting to estimate the causal effect of some variable X on another Y, an instrument is a third variable Z which affects Y only through its effect on X.
- Under some conditions, we using IV we get correct estimations.

Example

- Suppose a researcher wishes to estimate the causal effect of smoking on general health. Correlation between health and smoking does not imply that smoking causes poor health because other variables, such as depression, may affect both health and smoking, or because health may affect smoking.
- The researcher may attempt to estimate the causal effect of smoking on health from observational data by using the tax rate for tobacco products as an instrument for smoking.
- The tax rate for tobacco products is a reasonable choice for an instrument because the researcher assumes that it can only be correlated with health through its effect on smoking. If the researcher then finds tobacco taxes and state of health to be correlated, this may be viewed as evidence that smoking causes changes in health.

Fixed Effects

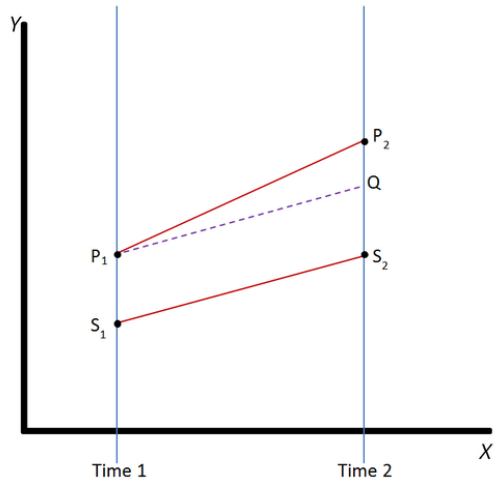
- Panel data combines cross sectional and time series data: we have a time series for each of the agents observed in a cross section.
- Panel data are much more informative of cross section and time series data.
- They allow use of a powerful tool to address the endogeneity problem: **fixed effects**.
- The basic idea is to allow variables to have two indices, $i = 1, 2, ..., n$ and $t = 1, 2, ..., T$. The simple linear model becomes

$$y_{it} = \alpha_i + \beta^1 x_{it}^1 + \cdots + \beta^k x_{it}^k + \varepsilon_{it}$$

- If endogeneity is caused be the correlation of the unobservable individual constants α_i with the $\boldsymbol{x}_{it}$ ("fixed effects"), we can estimate correctly by taking various techniques.

Difference in Differences

DiD requires data measured from a treatment group and a control group at two or more different time periods (at least one time period before "treatment" and one after).



Outcome in the treatment group is the line P and outcome in the control group is S. DID calculates the "normal" difference in the outcome variable between the two groups represented by Q. The treatment effect is the difference between P2 and Q.

Example

- One of the most famous DID is the Card and Krueger article on minimum wage in New Jersey, published in 1994.
- Card and Krueger compared employment in the fast food sector in New Jersey and in Pennsylvania, in February 1992 and in November 1992, after New Jersey's minimum wage rose from $4.25 to 5.05$ in April 1992.
- Observing a change in employment in New Jersey only, before and after the treatment, would fail to control for omitted variables such as weather and macroeconomic conditions of the region. By including Pennsylvania as a control in a difference-in-differences model, any bias caused by variables common to New Jersey and Pennsylvania is implicitly controlled for, even when these variables are unobserved.
- The evidence suggested that the increased minimum wage did not induce a decrease in employment in New Jersey, contrary to what simplistic economic theory would suggest.

Conclusions

- What have we learned? How do we read and write an applied paper (a paper that uses data)?
- The concepts and tool we have used may be very useful!

Thanks!