Listening to Data

Valentino Dardanoni

University of Palermo

March 2021

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 (and Models) 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?
- What are the models (from economic and statistical theory) that may be useful to understand the data?
- 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 understand some elementary notions of statistics and econometrics

- Typical Data Structures:

  a) Passive, nonexperimental
  b) Active, experimental

- 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**:

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

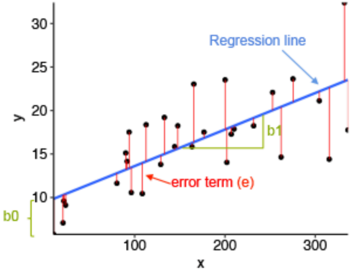
- The data unit i can be, for example, an individual, a firm, an household, a region...

Parameter Estimation

- The model we are considering is composed of a set of parameters (the alphas 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)

Parameter Estimation

- Example: you have some data that follows the linear model $y_i = \beta_0 + \beta_1 x_i + e_i$, where e_i is a random error that has mean zero and is independent of x_i



- In a figure, you can draw the "true" regression line $\beta_0 + \beta_1 x_i$, and the data points (x_i, y_i)
- In practice, we only have the data, so we have to decide both "the model" (e.g. straight line? quadratic? etc.), and the best estimation for the parameters of the model

What can go wrong?

1. The ‘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

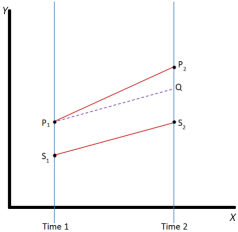
- To use the IV approach with x_K endogenous, we need an observable variable, say z, **not** in regression equation that satisfies two conditions: **Exogeneity** and **Relevance**.
- Informally, in attempting to estimate the causal effect of some variable X on another Y, an instrument is a third variable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

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 than cross section and time series data alone.
- 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 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.

Minimum Wage and Unemployement

- The most well-known application of DiD is Card and Krueger (1994) who investigated the impact of New Jersey's 1992 increase of the minimum hourly wage from \$4.25 to \$5.05.
- Panel of 331 fast food restaurants in New Jersey during the period February-March (before minimum wage increase) and then again during the period May-December 1992 (after).
- Fast food restaurants are a major employer of minimum wage employees. Before the change about 30% of the workers were paid \$4.25.

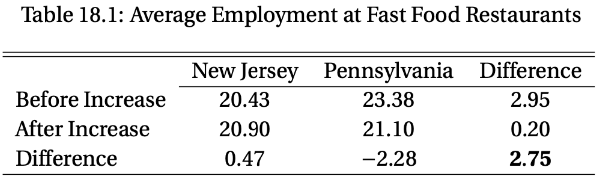


Table 18.1: Average Employment at Fast Food Restaurants

	New Jersey	Pennsylvania	Difference
Before Increase	20.43	23.38	2.95
After Increase	20.90	21.10	0.20
Difference	0.47	−2.28	**2.75**

- Before the increase the average number of employees was 20.4. After was 20.9.
- Employment slightly increased (by 0.5 employees per restaurant) rather than decreased!
- All employment change is attributed to the policy. No direct evidence of the counterfactual (what would have happened if the minimum wage had not been increased).
- A DiD estimator compares the change in the treatment sample with a comparable change in a control sample.
- Card and Krueger selected eastern Pennsylvania (quite similar to New Jersey) for their control sample.
- In the absence of a minimum wage increase we expect the same changes in employment.

- Before the policy change the average number of employees in eastern Pennsylvania was 23.4. After the policy change the average number was 21.1.
- Treating Pennsylvania as a control means comparing the change in New Jersey (0.5) with that in Pennsylvania (-2.3). The difference (2.75 employees per restaurant) is the DiD estimate.
- In complete contradiction to conventional economic theory the estimate indicates an increase in employment rather than a decrease. This surprising estimate has been widely discussed among economists and the popular press.

Do Police Reduce Crime?

- DiTella and Schargrodsky (2004) consider whether the street presence of police officers reduces car theft.
- Rational crime models predict that the the presence of an observable police force will reduce crime rates (at least locally) due to deterrence.
- Causal effect is difficult to measure, as police forces are not allocated exogenously, but rather are allocated in anticipation of need. A DiD estimator requires an exogenous event which changes police allocations. DiTella and Schargrodsky use the police response to a terrorist attack as exogenous variation.
- In July 1994 there was a horrific terrorist attack on the main Jewish center in Buenos Aires, Argentina. Within two weeks the federal government provided police protection to all Jewish and Muslim buildings in the country.

- Data on car thefts in selected neighborhoods of Buenos Aires for April-December 1994, resulting in a panel for 876 city blocks.
- They hypothesized that the terrorist attack and the government's response were exogenous to auto thievery.
- DiD estimator based on the average number of car thefts per block, before and after the terrorist attack, and between city blocks with and without a Jewish institution.

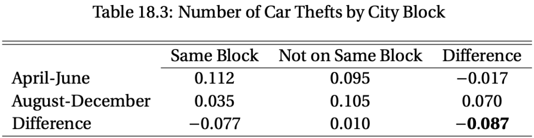


Table 18.3: Number of Car Thefts by City Block

	Same Block	Not on Same Block	Difference
April-June	0.112	0.095	−0.017
August-December	0.035	0.105	0.070
Difference	−0.077	0.010	**−0.087**

- The average number of car thefts dramatically decreased in the protected city blocks, from 0.112 per month to 0.035, while the average number in non protected blocks was near constant, rising from 0.095 to 0.105.
- DiD thus estimates that the effect of police presence decreased car thefts by 0.087, which is about 78%.

Conclusions

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

Thanks!