Quantitative Economics for the Evaluation of the European Policy

Dipartimento di Economia e Management

Co-funded by the Erasmus+ Programme of the European Union

Project funded by

European Commission Erasmus + Programme –Jean Monnet Action

Project number 553280-EPP-1-2015-1-IT-EPPJMO-MODULE

Irene Brunetti   Davide Fiaschi   Angela Parenti[1]

29/11/2016

[1] ireneb@ec.unipi.it, davide.fiaschi@unipi.it, and aparenti@ec.unipi.it.

Introduction

Recall the definition of the Propensity Score as:

$$e_i(X_i) = Pr(D_i = 1|X_i) = E[D_i|X_i]$$

where D_i is a dummy treatment indicator and X_i a set of observable control variables.

Theorem 1 (**Unconfoundedness**):

$$Y_{0i} \perp D_i|X_i \implies Y_{0i} \perp D_i|e(X_i) \quad (1)$$

In words, *conditional independence of Y_{0i} given X_i, which is the hypothesis of this theorem, implies conditional independence of Y_{0i} given the propensity score $e(X_i)$.*

Theorem 2 (**The Balancing Property**):

$$D_i \perp X_i|e(X_i) \quad (2)$$

In words, *the distributions of the treatment status D_i and the observable control variables X_i are orthogonal to each other, once conditioning on the propensity score $e_i(X_i)$.*

Introduction

- Propensity score matching essentially estimates each individual's propensity to receive a binary treatment (via a probit or logit) as a function of observables and matches individuals with similar propensities.
- Propensity score methods typically assume a common support, i.e. the range of propensities to be treated is the same for treated and control cases, even if the density functions have quite different shapes.
- Often *there are reasons to believe that treated and untreated differ in unobservable characteristics that are associated to potential outcomes* even after controlling for differences in observed characteristics.
- In such cases, treated and untreated may not be directly comparable, even after adjusting for observed characteristics.

Introduction

- Possible solution to handle unobserved heterogeneity so far: **Panel Data**.
- Panel Data offer another powerful way to tackle issues related to omitted variable bias.
- In particular, panel data allow to control for **unobserved** but **fixed** factors that drive participation and that are related to potential outcomes.
- The trick is to exploit to have several observations which all contain the same unobserve information.

Introduction

- As an example, assume potential outcomes for individual i at time t can be written as:

$$Y_{0it} = \beta + \epsilon_{it},$$
$$Y_{1it} = Y_{0it} + \rho.$$

- Observed outcomes are given by:

$$Y_{it} = \beta + \rho D_{it} + \epsilon_{it},$$

but treatment is not as good as randomly assigned (D_{it} is not independent of ϵ_{it}).

- The crucial assumption that allows exploiting the panel structure of the data is that the unobserved component of potential outcomes ϵ_{it} can be decomposed.

Main assumptions

1 ϵ_{it} can be written as: $\epsilon_{it} = \gamma_i + \lambda_t + \eta_{it}$, where:

- γ_i is specific to individual i and fixed over time;
- λ_t is a time trend; and
- η_{it} is a transitory mean zero noise term.

2 Selection into treatment only depends on the individual fixed effect γ_i but is independent of λ_t or η_{it}.

- $E(\lambda_t|D_{it}) = E(\lambda_t)$ and
- $E(\eta_{it}|D_{it}) = E(\eta_{it}) = 0$
- $E(\epsilon_{it}|D_{it}) = E(\gamma_i|D_{it}) + E(\lambda_t)$

Hence, treatment and control group differ only in terms of the individual fixed effect, not in terms of the time trend and transitory shocks to outcomes.

- γ_i is also known as a **fixed effect**. Note that the fixed effect enters **additively** and **linearly**!

Introduction

- Under this assumption we can write observed outcomes as:

$$Y_{it} = \beta + \gamma_i + \rho D_{it} + \lambda_t + \eta_{it}. \tag{3}$$

- Under the assumption above we can take advantage of multiple observations on each unit and eliminate the fixed effect by, for example, differencing the equation above:

$$\Delta Y_{it} = \Delta D_{it}\rho + \Delta\lambda_t + \Delta\eta_{it}. \tag{4}$$

  where the Δ denotes changes of the variable from $t-1$ to t.
- Note that for differencing to work it is necessary that the fixed effect enter additively and linearly!
- $\Delta\eta_{it}$ is uncorrelated to ΔD_{it} and running OLS on the differenced outcome equation yields the causal effect.

N.B. When the level of potential outcomes differs between treatment and control group due to a linear and additive fixed effect, the change of potential outcomes over time does not differ

Examples 1

- Wages, health, happiness...are all outcomes that depend on observable variables but also genetic factors which are unobserved.
- To tackle this kind of unobserved heterogeneity, a lot of studies use data on monozygotic twins.
- The idea is that differences in outcomes between monozygotic twins can solely be attributed to observable factors.

Example 2

- Assume you would like to estimate how business taxes impact on FDI.
- Probably, the attractiveness of a region is partly determined by unobserved factors. Which?
- At least within a short period of time, these unobserved factors are likely to be fix.
- Hence, the unobserved factors influence only the level of FDI but not its change over time.
- Estimating an equation that relates the change in tax rates to the change in FDI is therefore less likely to suffer from an omitted variable bias.

Difference in Differences

- The Difference in Differences (DD) estimator is the simplest estimator that makes use of data with a time dimension.
- The DD estimator can be interpreted as a fixed effect estimator that uses only aggregate (group level) data.
- That means that the DD estimator uses differencing at the group level, not at the individual level as in the introductory example.
- This can be done if treatment status varies only at the group level (e.g. state, cohort).
- Non random assignment of treatment must therefore come from unobserved variables at the group level.

Difference in Differences

- The DD approach captures these unobserved variables by a group level fixed effect.
- Since the DD estimator does not use data at the individual level it can also be used in a repeated cross section.
- We will make the following two additional assumptions:
  - there are only 2 periods: "before treatment" $(t = 0)$ and "after treatment" $(t = 1)$; and
  - the treatment variable is binary.

Example: Card and Krueger (1994)

- Card and Krueger (1994) want to estimate the impact of a minimum wage on employment.
- In 1992, the state of New Jersey increased its minimum wage by roughly 20%. In Pennsylvania (neighboring state) the minimum wage did not change.
- Card and Krueger have data on employment at fast food restaurants close to the state border for both states.
- There are two approaches to estimate the employment effect of the minimum wage:
    1. A comparison of employment levels in New Jersey and Pennsylvania after the introduction of the new minimum wage.
    2. A comparison of employment levels in New Jersey before and after the introduction of the new minimum wage.
- Both approaches are unlikely to uncover the causal effect of the minimum wage. Why?

Formal exposition

- Let us assume a constant treatment effect and abstract from any covariates so we write potential outcomes as:

$$Y_{0st} = \epsilon_{st} \quad \textit{and} \quad Y_{1st} = Y_{0st} + \rho, \tag{5}$$

  where the index s indicates the state.
- Moreover, we assume that ϵ_{st} can be decomposed into:
  - a group level (state) effect γ_s (that is the fixed effect);
  - a time trend λ_t common to all states; and
  - a transitory mean zero noise term η_{st}
- While γ_s can be different for the two states, the time trend and the idiosyncratic noise term do not vary systematically between states $\Longrightarrow$ Hence, treatment is as good as randomly assigned conditional on the state effect γ_s.
- The observed outcome can be written as:

$$Y_{st} = \gamma_s + \lambda_t + \rho D_{st} + \eta_{st}$$

  where $E(\eta_{st}|s,t,D_{st}) = E(\eta_{st}) = 0$.

Diff in Diffs

- Now consider what we would get by comparing average employment in both states before ($t = 0$) and after treatment ($t = 1$).

$$E(Y_{st}|s = Penn, t = 1) - E(Y_{st}|s = Penn, t = 0) = \lambda_1 - \lambda_0;$$
$$E(Y_{st}|s = NJ, t = 1) - E(Y_{st}|s = NJ, t = 0) = \rho + \lambda_1 - \lambda_0;$$

- Hence, the treatment effect ρ is given by the difference in differences:

$$E(Y_{st}|s = NJ, t = 1) - E(Y_{st}|s = NJ, t = 0) - E(Y_{st}|s = Penn, t = 1) - E(Y_{st}|s = Penn, t = 0) = \rho.$$

- This can easily be estimated using sample means.

Diff in Diffs

Let us summarize the key idea:

- The comparison over time within a state eliminates the state fixed effect.
  $\Longrightarrow$ We can remove differences in employment levels in the two states that have nothing to do with the minimum wage, by considering the difference in employment levels before and after the introduction of the new minimum wage.
  $\Longrightarrow$ In case of NJ, this difference captures the general time trend in employment and the treatment effect.
- To eliminate the general time trend we compare the development in employment levels in NJ with the development of employment in the control state which consists only of the time trend.

The following assumption is therefore crucial.

Diff in Diffs

Assumption 1

The time trend $\lambda_1 - \lambda_0$ is the same in both states.

Card's results

- Some of Card's results relating to the average employment levels in fast-food restaurants are shown below (with standard errors in parentheses).

	Before Increase	After Increase	Difference
New Jersey	20.44	21.03	0.59
(Treatment)	(0.51)	(0.52)	(0.54)
Pennsylvania	23.33	21.17	-2.16
(Control)	(1.35)	(0.94)	(1.25)
Difference	-2.89	-0.14	2.76
	(1.44)	(1.07)	(1.36)

- The difference in difference estimator shows a small increase in employment in New Jersey where the minimum wage increased.
- The study has been very controversial but helped to change the common presupposition that a small change in the minimum wage from a low level was bound to cause a significant decrease in employment.

Regression implementation of DD

- The DD estimator can easily be implemented using regression. This is a convenient way to obtain estimates and the corresponding standard errors in one step.
- Let:
  - NJ denote a dummy equal to one for restaurants in New Jersey and let
  - d_1 be a dummy variable equal to one for observations after the introduction of the new minimum wage.
- Then the DD estimate equals the coefficient ρ from the following regression:

$$Y_{st} = \alpha + \gamma NJ + \lambda d_1 + \rho(NJ \cdot d_1) + \eta_{st}. \tag{6}$$

- Note that:
  - the variable $NJ \cdot d_1$ equals to D_{st} and
  - ρ is identical to the DD estimator.

Multiple groups

- Of course, the control group can consist of more than a single control state.
- Including several states as controls is beneficial since it provides a hedge against idiosyncratic shocks in a control state which might make less effective the common trend assumption.
- Assume we also had data on Connecticut in the example above. We could still work with the same regression function.

$$Y_{st} = \alpha + \pi Conn + \gamma NJ + \lambda d_1 + \rho(NJ \cdot d_1) + \eta_{st}. \tag{7}$$

- Now λ would capture an average time trend for Pennsylvania and Connecticut. In particular, λ captures average employment differences between:
  - establishments which are either in Penns. or Conn. in $t = 1$
  - establishments which are either in Penn. or Conn. in $t = 0$.
- The treatment effect ρ would now be obtained by using the average of Pennsylvania and Connecticut as a "control" state.

The Difference in Differences in Differences estimator

A still more convincing analysis than just using multiple control groups would be possible if we could define a "treatment" and a "control" group within each state.

- In the minimum wage example, assume we also had data on employment in sectors not affected by minimum wage legislation.
- Then we could think about two possible DD strategies:
  - we could use employment in the non affected sector in the treatment state as the control group; or
  - We would use employment in the fast food sector in a control state as the control group (approach so far).

The Difference in Differences in Differences estimator

- There is a pro and a con for each approach:
  - The first strategy would be immune to different time trends across states but would depende on the assumption that the time trend in employment is the same for different sectors.
  - The second strategy would control for employment trends in the fast food sector but would be vulnerable to different time trends across the treatment and the control state.
- The DDD approach combines both strategies and computes 2 DD estimators:

The Difference in Differences in Differences estimator

- the DD estimator using the non affected sector in the same state as control group:

$$DD_{NJ} = (E(Y_{st}|s = NJ, t = 1, affected) - E(Y_{st}|s = NJ, t = 0, affected)) \\ -(E(Y_{st}|s = NJ, t = 1, unaffected) - E(Y_{st}|s = NJ, t = 0, unaffected))$$

- and, in order to control for different time trends in the affected versus the non affected sector:

$$DD_{Penn} = (E(Y_{st}|s = Penn, t = 1, affected) - E(Y_{st}|s = Penn, t = 0, affected)) \\ -(E(Y_{st}|s = Penn, t = 1, unaffected) - E(Y_{st}|s = Penn, t = 0, unaffected))$$

- The DDD estimator is given by the difference between the two DD estimators:

$$DDD = DD_{NJ} - DD_{Penn}$$

The Difference in Differences in Differences estimator

- By simply rearranging the expression above, we see that the DDD estimator could also be calculated as the difference between:

$$DDD_{Aff} = (E(Y_{st}|s = NJ, t = 1, aff) - E(Y_{st}|s = NJ, t = 0, aff)) - (E(Y_{st}|s = Penn, t = 1, aff) - E(Y_{st}|s = Penn, t = 0, aff)) \quad (8)$$

- and, in order to control for different time trends between the treatment and the control state:

$$DDD_{Nnaff} = (E(Y_{st}|s = NJ, t = 1, Nnaff) - E(Y_{st}|s = NJ, t = 0, Nnaff)) - (E(Y_{st}|s = Penn, t = 1, Nnaff) - E(Y_{st}|s = Penn, t = 0, Nnaff)) \quad (9)$$

- The DDD estimator is given by the difference between the two DD estimators:

$$DDD = DD_{Aff} - DD_{Nnaff}$$

The Difference in Differences in Differences estimator

- Note that DDD is different from just adding a control group since now we define an affected and non affected group within each state:

  Additional Control group: T,C $\Longrightarrow$ T,(C1, C2)

  $$\text{DDD: T,C} \Longrightarrow (T_{aff}, T_{Nnaff}), (C_{aff}, C_{Nnaff}).$$

- The DDD estimator thus controls at the same time for a state specific and a sector specific trend.
- It can also be implemented via a regression function. Let AF be a dummy equal to one if the sector is affected. Note that the following regression function contains eight parameters, one for each group (NJ affected, NJ non affected, Penn affected, Penn non affected) - time combination.

$$\begin{aligned} Y_{st} = \alpha + \gamma_0 NJ + \gamma_1 AF + \gamma_2 (NJ \cdot AF) + \lambda_0 d_1 \\ + \lambda_1 (d_1 \cdot NJ) + \lambda_2 (d_1 \cdot AF) + \rho (d_1 \cdot NJ \cdot AF) \end{aligned} \quad (10)$$

- The coefficient ρ equals the DDD estimator.

Additional controls

- The regression formulation of DD also allows to include additional control variables. For example, you could estimate:

$$Y_{st} = \gamma_s + \lambda d_1 + X_{st}^{'}\beta + \rho(NJ \cdot d_1) + \eta_{st}. \tag{11}$$

  where:
  - γ_s is a separate dummy for each state; and
  - X_{st} are observable characteristics for each state (e.g. industry structure).
- In this specification
  - λ would capture an average time trend (across all states); and
  - the inclusion of X_{st} would allow for differences in the time trend across states based on observables X_{st}.
- Hence, the estimate of ρ would isolate the treatment effect from a general time trend and state specific trends due to observable differences.

Variable treatment intensity

- The DD estimator can also be used when several groups were treated with differing intensity.
- In the minimum wage example, there might be two reasons for that:
  - The minimum wage changes could be different in each state.
  - Even if the minimum wage changes are the same we might expect a different impact across states if, for example, states differ in the fraction of individuals earning minimum wages before the increase.
- In the former case we could use a continous minimum wage regressor w_{st} instead of the binary treatment D_{st}.

Variable treatment intensity

- In the latter case, a natural specification would be

$$Y_{st} = \gamma_s + \lambda d_1 + \rho(d_1 \cdot FA_s) + \eta_{st},$$

  where:

  - FA_s is a variable measuring the fraction of individuals likely to be affected by the change in minimum wage laws; and
  - the interaction $d_1 \cdot FA_s$ is the treatment variable that accounts for differing treatment intensities.

More than 2 time periods

- One advantage of more than two time periods is that it is possible to shed light on the validity of the common trend assumption.
- If the common trend assumption does not hold exactly, a longer time horizon allows to control for different time trends across groups.
- One possibility would be to include linear, state specific time trends into the model and estimate.
- In addition, many periods offer the opportunity to examine lagged or anticipatory effects of treatment.

Validity

- The most important condition for the validity of DD is the common trend assumption. We have just seen, how data over a longer time horizon can be used to assess (or weaken in case of state specific trends) this assumption.
- We have said in the beginning that DD can be applied in repeated cross sections as well since all we need are group averages.

Caveat:

- The composition of treatment and control groups must not change. If it does, the group "fixed" effect changes over time and can no longer be differenced out.

Validity

Caveat:

- Example: A higher minimum wage induces more able and motivated individuals to work in the fast food industry which makes it more attractive to hire more workers.
- As long as the composition changes along observable dimensions, one can control for it.
- However, if observable group characteristics change by a large amount, we might suspect the same for unobservable characteristics as well.
- If group composition changes over time it is thus a good idea to examine observable group characteristics pre- and post-treatment in practice (see e.g. Gruber (1994), table 2).
- It might also help to examine observable characteristics across groups. If those are similiar one can be more confident that the time trend of both groups is similiar as well.

Validity

- DD also fails to uncover the causal effect if treatment and control group differ in their idiosyncratic (transitory) shocks prior to treatment. Formally, if the transitory component η_{ist} of the error:

$$\epsilon_{ist} = \gamma_s + \lambda_t + \eta_{ist} \tag{12}$$

  differs between the treatment and the control group, the DD estimator has no causal interpretation.

- An Example is Ashenfelter's famous study: Evaluation of a job training program where participants entered the program (or were selected) when earnings were particularly low.

- That is, there is a dip in earnings prior to treatment but we would expect earnings to recover anyway (since the dip is transitory) even without the program.

Validity

- Ashenfelter's dip would correspond to a different expected value of η_{ist} for the treatment and the control group in the period before treatment.
- What is the problem caused by the dip?
- Ashenfelter's Dip can often be detected graphically. If you see a dip, dynamic models are more appropriate.