Microeconometrics Notes

Balázs Faragó

Contents

Chapter 1

Angrist Pischke Ch. 1&2

1.1 Chapter 1 - Questions about questions

Advice: Think about the ideal experiment first given no H.R. Restrictions in a perfect "lab environment". How would you do it and can it be done?
Questions which cannot be answered by an experiment are F.U.Q. - Fundamentally Unidentified Questions.

Identification strategy - the way in which you use observational data

Advice: Ask: What is your mode of statistical inference?
i.e. 1. What is the population? 2. What is the sample? 3. What are the assumptions made when constructing standard errors?

1.2 Chapter 2 - The experimental ideal

RCTs are the ideal.

$$\text{potential outcome } = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases}$$

Ideally, you´d want to be measuring this where D_i is the treatment such as hospitalization for example and Y_i is the effect of (not)hospitalization.

But, this observed outcome can be written as:

$$Y_i = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases}$$

$$= Y_{0i} + (Y_{1i} - Y_{0i})D_i$$

$$\underbrace{E\left[\mathrm{Y}_i|\mathrm{D}_i=1\right]-E[\mathrm{Y}_i|\mathrm{D}_i=0]}_{\text{Observed difference in average health}} = \underbrace{E\left[\mathrm{Y}_{1i}|\mathrm{D}_i=1\right]-E[\mathrm{Y}_{0i}|\mathrm{D}_i=1]}_{\text{average treatment effect on the treated}} + \underbrace{E\left[\mathrm{Y}_{0i}|\mathrm{D}_i=1\right]-E\left[\mathrm{Y}_{0i}|\mathrm{D}_i=0\right]}_{\text{selection bias}}$$

1

2 $E[\mathrm{Y}_{1i}|\mathrm{D}_i=1]-E[\mathrm{Y}_{0i}|\mathrm{D}_i=1]=E[\mathrm{Y}_{1i}-\mathrm{Y}_{0i}|\mathrm{D}_i=1]$

Figure 1.1: 2 is the average causal effect of hospitalization on those who were hospitalized. This term captures the averages diference between the health of the hospilized.

Definition 1.2.1: Selection Bias

The difference in the average Y_{0i} between treated and untreated. (Y_{0i} is the "effect" of no treatment.)
i.e. $E[Y_{0i}|D_i=1]-E[Y_{0i}|D_i=0]$

Random assignment (treatment/no-treatment) solves the selection problem!

Show (This is true because)
$E[Y_i|D_i=1]-E[Y_i|D_i=0]=E[Y_{1i}|D_i=1]-E[Y_{0i}|D_i=0]$

$$=E[Y_{1i}|D_i=1]-E[Y_{0i}|D_i=1]$$

We can do this (swap $D_i=1$ and $D_i=0$) because D_i and Y_{0i} are independent
... And this becomes simply:

$$=E[Y_{1i}-Y_{0i}|D_i=1]=E[Y_{1i}-Y_{0i}]$$

The effect of randomly-assigned hospitalization on the hospitalized is the same as the effect of hospitalization on a randomly chosen patient.

1.2.1 Experiments as Regression

We can reformulate an experiment to a regression:

$$Y_i=E[Y_{0i}]+(Y_{1i}-Y_{0i})D_i+Y_{0i}-E[Y_{0i}]$$

becomes...

$$Y_i=\alpha+\rho D_i+\eta_i$$

With the tratement witched on and off:

$$\begin{aligned} E\left[\mathrm{Y}_i|\mathrm{D}_i=1\right] &= \alpha+\rho+E\left[\eta_i|\mathrm{D}_i=1\right] \\ E\left[\mathrm{Y}_i|\mathrm{D}_i=0\right] &= \alpha+E\left[\eta_i|\mathrm{D}_i=0\right], \end{aligned}$$

so that,

$$E[\mathrm{Y}_i|\mathrm{D}_i=1]-E[\mathrm{Y}_i|\mathrm{D}_i=0]= \underbrace{\rho}_{\text{treatment effect}} + \underbrace{E[\eta_i|\mathrm{D}_i=1]-E[\eta_i|\mathrm{D}_i=0]}_{\text{selection bias}}.$$

Thus, selection bias amounts to correlation between the regression error term, η_i, and the regressor, D_i. Since

$$E\left[\eta_i|\mathrm{D}_i=1\right]-E\left[\eta_i|\mathrm{D}_i=0\right]=E\left[\mathrm{Y}_{0i}|\mathrm{D}_i=1\right]-E\left[\mathrm{Y}_{0i}|\mathrm{D}_i=0\right],$$

This correlation reflects the difference in (no-treatment) potential outcomes between those who get treated and those who don´t.

Note: **1.** If controls X_i are uncorrelated with the treatment D_i, they will NOT affect the estimate of ρ. **2.** They will reduce your standard error. *You can introduce them:* $Y_i=\alpha+\rho X_i'\gamma+\eta_i$.

Chapter 2

Duflo et.al. 2008 - Using randomization in development economics - a toolkit

RCT via random assignment is: **1.** Unbiased **2.** Internally valid

Remember that:

$$E[Y_{1i} - Y_{0i}]$$

(so the average difference in with and with out treatment) estimates only the overall effect of the treatment which may be comprised of many factors.

2.0.1 Other methods to control for selection bias

Non-random-assignment methods basically rely on having assumptions and making comparison groups based on these assumptions.

Controlling for observables

Theoretically, it could be that we have a set of observale variables X and we can condition our data on these: $E[Y_{i0}|X, D_i = 1] - E[Y_{i0}|X, D_i = 0] = 0$. But this basically never happens so we have to control for these variables usually.

Definition 2.0.1: Fully non-parametric matching

If the dimension of X (variables we want to control for) is not large, we compute the difference in between treatment and control within each cell formed by the various values of X and the treatment effect is a weighed average of these.

Not good for many control variables. *Propensity score matching* is better.
Propensity score - The probability of being assigned to treatment conditional on the values of X.

A third approach is to control for X, parametrically or non-parametrically, in a regression framework For both matching and regression, we have different assumptions and estimate different parameters but 1 assumtionis true for both: we have controlled for everything. Controlling for the propensity score leads to unbiased estimate of the treatment effect under assumption/equation from before.
this paper goes further into other topics discussed later in the course.

Chapter 3

Bertrand and Mullainathan 2004 - A Field Experiment on Labor Market Discrimination

Basic summary about the paper:

- A study was conducted to analyze racial differences in callback rates for job applicants
- Applicants with White names received more callbacks than those with African-American names
- The study found that resume characteristics were less predictive of callback rates for African-Americans. The racial gap in callback rates was consistent across different job categories
- The study suggests that getting a job is more difficult for African-Americans
- The findings indicate statistical discrimination in the labor market based on race
- Locations: Boston & Chicago. Jobs: sales, administrative support, clerical services, and customer services

More specific details about setup, inference etc..

- Used paper resumase to insulate the study from demand effects
- Large sample size was important
- RCTs, randomization, and measurement
- Used Spearman Rank order correlation to analyze the relationship between callback rates and mother's education within each race-gender group
- Regression estimates and the calculation of standard deviations for predicted callback rates
- Control variables: sex, city, occupation dummies, and a vector of job requirements

Spearman´s Rank Correlation Coefficient

$r_s \in \{-1, 1\}$

$$1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} = \frac{cov(R(X), R(Y))}{\sigma_{R(X)} \sigma_{R(Y)}}$$

where d_i = difference between 2 ranks of each observation, n = number of observations, $R_(X_i)$ is the rank of observation X_i

Close to 0 $\implies$ Weak monotonic relationship.

Close to 1 or -1 $\implies$ Strong monotonic relationship

- Robust to outliers
- Like correlation but for ranked variables
- No requirement for linearity
- Useful when the data violates assumptions of parametric correlation methods.

Chapter 4

Chetty et.al. 2016 - he Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment

Main result from study/topic: Moving households in poverty to better neighbourhood is good for them. More specifically, the amount of time spent by children in "bad neighbourhoods" is negatively correlated with outcomes like high school completion or earnings. This experiment was done via some voucher program that helped households in impoverished neighbourhoods to move. The result also suggests that the earlier children move from bad neighbourhoods, the better.

- They estimate the treatment effects of growing up in these very different envi- ronments by replicating the **intent-to-treat (ITT)** specifications used in prior work (e.g., Kling, Liebman, and Katz 2007).**regressing outcomes in adulthood on indicators for assignment to each of the treatment arms.**

Definition 4.0.1: Intend-To-Treat (ITT) / (encouragement design)

In field studies or RCTs, even if participants assigned to a treatment group do not fully follow through with the prescribed treatment in reality, the ITT analysis includes them in the evaluation. This ensures that the analysis reflects the initial treatment assignment, providing a more realistic and unbiased assessment of the treatment's effectiveness under real-world conditions, where not everyone may adhere perfectly to the assigned treatment.

- **treatment-on-the-treated (TOT)** estimate for those who took up the exper- imental voucher of $3,477, a 31 percent increase relative to the control group mean of $ 11,270.

Definition 4.0.2: Treatment-On-The-Treated (TOT)

This refers to an analysis that specifically evaluates the impact of the treatment among those who actually receive it. It contrasts with Intent-to-Treat (ITT), which includes all assigned to the treatment group regardless of adherence. TOT, therefore, focuses on the subset that complies with or "takes" the treatment, providing insights into the treatment's effectiveness when implemented as intended.

- "We examined the robustness of these findings by **modeling age in linear interaction with treatment indicators.** Our results consistently show that the benefits of relocating to lower-poverty areas decrease as the child's age at the move increases. This implies that each additional year of exposure to a low-poverty environment during childhood is advantageous. While we don't identify a distinct "critical age" for moving to a better neighborhood, precise estimates are limited due to small sample sizes at each child age in the MTO data."

- Consistency with exposure effect model, but MTO design doesn't conclusively prove causal link between childhood exposure and long-term outcomes. This is because: Ages at which children move are perfectly correlated with length of exposure, making it difficult to distinguish age-related disruption effects from age-invariant disruption cost with an exposure effect.
- + Treatment effects may differ for families with young versus old children due to variations in the families taking up vouchers and the chosen areas for relocation.
- Despite underlying uncertainties, experimental results support the conclusion that subsidized housing vouchers for moving to lower-poverty areas yield greater benefits for younger children.

- "We find that the MTO treatments had little or no impact on adults' economic outcomes, consistent with prior work"
- The positive effect was only for kids ¡13 age.

- We find no systematic differences in the treatment effects of MTO on children's long-term outcomes by gender, race, or site.

- Concerns exist that our exploration of treatment effects based on a new dimension (child's age at move) in MTO data may be influenced by **multiple hypothesis testing**.
- To address this, we test the null hypothesis that treatment effects for main subgroups (gender, race, site, and age) are all zero using F-tests and a nonparametric permutation test.
- We reject the null hypothesis for most outcomes with $p < 0.05$ using F-tests and $p < 0.01$ with the permutation test, suggesting that significant treatment effects for younger

children are not an artifact of analyzing multiple subgroups.

Definition 4.0.3: Multiple hypothesis testing (Problem)

Multiple hypothesis testing involves conducting many statistical tests simultaneously. The problem is an increased risk of obtaining false-positive results by chance alone when performing numerous tests

4.0.1 Balancing and attrition

The studies compared 195(more or less) variables across randomized groups to see if the mean differences between groups are statistically significant. - they shouldn´t be.

There is a lot of partial compliance in the treatment group- i.e. only about half of the people offered vouchers moved. ITT Estimates:

$$y_i = \alpha + \beta_E^{ITT} Exp_i + \beta_S^{ITT} S8_i + \gamma X_i + s_i\delta + \epsilon_i$$

Note: where Exp and S8 are dummy variables for being randomly assigned to experimental and S8 groups, respectively. X is a vector of baseline covariates and si are dummies for site (city?)
The offer of voucher was used as an instrument IV for the treatment. So ITT as instrument.

$$TOT = ITT/\text{treatment-uptake rate}$$

Some more lingo:
ATE - Average treatment effect
ATET - Average treatment effect on the treated
(kinda like a more specific version of TOT)

Chapter 5

Lecture 1- Causality I.

$D_i = 1$ or 0 - Treatment or not
Y_{1i} or Y_{0i} - Potential outcomes (treatment or not)
Note: since it is indexed by i, we mean for 1 unit. And this is potential outcomes, so this is not dependent wether someone is ***actually*** *treated.*

$\Delta_i = Y_{1i} - Y_{0i}$ - The effect of participating in treatment. This is the difference in potential outcomes. This is never observable.

5.0.1 Statistical solution to the counterfactual problem

$ATE = E[\Delta_i]$= - Average treatment effect (How much on average, a populaiton is affected by a treatment)
$ATET = E[\Delta_i|D_i = 1] = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1]$ - Average treatment effect on the treated.

Selection bias

Main issue: Potential outcomes are not independent of actual treatment status.

s.p.s. We naively yry to find ATET by: $E[Y_{0i}|D_i = 0] - E[Y_{1i}|D_i = 1]$.
Now we use a trick : $E[Y_{0i}|D_i = 0] - E[Y_{0i}|D_i = 1] + E[Y_{0i}|D_i = 1] - E[Y_{1i}|D_i = 1]$.
Simplifying: $E[Y_{1i} - Y_{0i}|D_i = 1] + E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]$
Which can be written as $E[Y_{1i} - Y_{0i}|D_i = 1] +$ a bias term . *This is ATET + bias.*

If decision making is not random: "I am in it if it is worth it":
$D = 1$ if $Y_{1i} - Y_{0i} = c > 0$
Leads to some "natural" selection bias. It can tell us about decision making as well.
We need some exogenous variation to make treatment status independent of outcomes.

5.0.2 Randomization

We randomize: $(Y_{1i}, Y_{0i}) \perp D$
With randomization, treatment status is independent of potential outcomes. See why.

Chapter 6

Lecture 2- Causality & II.

Note: RA (Randomized (assignment)) studies tend to come up with very different results than for example an observational regression with control variables.

6.1 Analyzing data from experiments

If well done, you just compare 2 sample means:

$$\hat{ATE} = \hat{ATET} = \frac{\sum_{i=1}^{n} D_i Y_i}{\sum_{i=1}^{n} D_i} - \frac{\sum_{i=1}^{n} (1 - D_i) Y_i}{\sum_{i=1}^{n} (1 - D_i)}$$

s.t. $Y_i = (1 - D_i)Y_{0i} + D_i Y_{1i}$
Alternatively, just a regression *Beta will be the difference between the sample means*

$$Y_i = \beta_0 + \beta_1 D_i + U_i$$

6.2 Examples

Not including lab experiments
- Field experiments
- Natural experiments

6.2.1 Bertrand paper on discrimination

Measuring Discrimination on the labour market. The effect of being "black" Y_{1i} or "white" Y_{0i}. Treatment effect:

$$E[Y_{1i} - Y_{0i}]$$

People just used to include a lot of dummies, but nowadays, it is not common since we

don´t trust we can use all the right dummies.
They do a regression.$Y_i = \alpha_i + \gamma N_i + \epsilon_i$
From randomization $Cov(N_i, \epsilon : i) = 0$**.** Treatment effect: γ.

N_i is the binary variable describing wether a name is black or white stereotypical.

Can´t distinguish between statistical and preference based discrimination.

They also distinguished between high vs low quality CVs sent and having high quality CV didn´t make any difference if black but yes if white.

6.2.2 Natural experiment paper: "Chattopadhyay and Duflo (2004) Women as Policy Makers"

Question: Do

6.3 Problems with experiments

Definition 6.3.1: Internal validity

Does the experiment provide an estimate of the causal effect in the population under the study?

Definition 6.3.2: External validity

The extent to which the result can be generalized outside of the experimental framework. (Placebo effects go here)

Internal validity problems

Definition 6.3.3: Partial compliance

Threat to internal validity. When not all the treatment group ends up treated or if someone outside (in the control) do take the treatment by themselves.

If therre is *partial compliance*, you should consider the **original** treatment groups **ITT** to avoid selection bias. The experiment this way, also shows the effect on randomization, so the result includes the contamination across treatment and control groups - which can be interesting in itself especially given the imperfections of actual policy.
- Z is the offer of the treatment and D is the actual treatment. The measured effect is that of Z now (the ITT).

Definition 6.3.4: Attrition

When individuals drop out of the experiment.

Random dropout - Just a problem for statistical power.
non-Random dropout - can lead to over or underestimation given some optimizing behaviour of participants.

External validity problems

- Maybe your volunteers are non representative or otherwise sample is not representative.
- People changing their behaviour because they are part of an experiment

Definition 6.3.5: Hawthorne effects

Some subjects get excited by being in an experiment. They might perform "better".

Definition 6.3.6: John Henry effects

People are "offended" by being in in some comparison group. Negative performance, basically the opposite of the Hawthorne effects.

Definition 6.3.7: General equilibrium effects

Small scale doesn´t translate to large scale because of perhaps, market mechanisms.

Chapter 7

OLS Regression & Causality

7.1 Omitted variables bias OVB

Zero conditional mean assumption: $E[u|X] = 0$

1. Let´s say there is this model: $y_i = \alpha + \rho S_i + \gamma A_i + \epsilon_i$
2. But the researcher mistakenly specifies the model as: $y_i = \rho S_i + \epsilon_i$
3. We can use the bivariate regression formula to d**erive the bias of** ρ in the incorrectly specified model: $\rho_{OLS} = \dfrac{Cov(S_i, y_i)}{Var(S_i)}$

4.1 Substituting the formula for Y_i from the correctly specified model:

$$\rho_{OLS} = \frac{Cov(S_i, \alpha + \rho S_i + \gamma A_i + \epsilon_i)}{Var(S_i)}$$

$$= \frac{Cov(S_i, \alpha) + \rho Cov(S_i, S_i) + \gamma Cov(S_i, A_i) + Cov(S_i, \epsilon_i)}{Var(S_i)}$$

Note: Since $Cov(S_i, \alpha) = Cov(S_i, \epsilon_i) = 0$, *we know:*

$$= \frac{\rho Cov(S_i, S_i) + \gamma Cov(S_i, A_i)}{Var(S_i)}$$

$$= \frac{\rho Var(S_i) + \gamma Cov(S_i, A_i)}{Var(S_i)}$$

$$= \rho + \frac{Cov(S_i, A_i)}{Var(S_i)} = \rho + \gamma\delta_{AS}$$

Definition 7.1.1: Omitted Variables OVB formula

$\rho + \dfrac{Cov(S_i, A_i)}{Var(S_i)} = \rho + \gamma\delta_{AS}$ where δ_{AS} is the regression coefficient of S_i on A_i

Fact 7.1.2

ρ_{OLS} will not be biased when:
$\gamma = 0$ - when the model was not misspecified in the 1st place
$\delta_{AS} = 0$ - if S_i and A_i are unrelated.

di

7.2 Why still use controls?

1. Gain efficiency (smaller variance)
2. Check for random assignment.If randomization works, result should be independent of control variables. Instead of checking all the mean differences like in the Chetty paper, you can try to predict the treatment indicator with control variables- use an F-test to see if they jointly are 0.

7.3 Good & Bad controls

7.3.1 Lundborg et.al. 2019 - The effect of military service in Denmark

Military service was supposed to be randomly assigned so they checked this. Checked the treatment indicator as a function of a bunch of predetermined variables. They did this and these variables were not predictive so all good.

Controls also allow slicing of the data.

Military service was found to have an effect on earnings for those with the highest IQ-s.

Definition 7.3.1: Bad controls

Variables that themselves might be effected by the treatment.

Definition 7.3.2: Good controls

Variables that you can think of as being fixed at the time the treatment variable was determined.

7.3.2 Bad control example

Effect of collage degree on earnings.
Without occupation, it seems like this has omitted variables but it seems, that this actually would create selection bias if we included it.
This troubling phenomenon is the composition effect. If you limit the study to white collar jobs only, by control, in the control group you get people who manage to get a white collar job without collage, in the treatment group you have those + those who only can get a white collar job because they have a degree. Tho composition is therefore very different for the 2 groups.

Consider 3 groups, AB,AW,and BW i.e. always blue color job no matter the degree, Always white, and Blue white (need the degree). In the end, this lowers the seeming effect of a collage degree in this example because of the compositional difference in the treatment and control groups.
- If the treatment changes the control, it is not a control.

Chapter 8

Instrumental variables I.V.

IV is a kind of quasi-experiment.
Zero conditional mean assumption is unlikely to be fulfilled in many cases.
Sps. We have the following (real) relationship:

$$y_{si} = f_i(s) = \pi_0 + \pi_1 \underbrace{s}_{\text{Observed independent variable (ex: schooling)}} + \underbrace{\eta_i}_{\text{error/unmeasured}}$$

but $E[\eta_i|s_i] \neq 0$

$$\eta_i = \underbrace{A_i'}_{\text{Unobserved effect (ex: ability)}} \gamma + \upsilon_i$$

Note: A_i and υ_i are uncorrelated

Try estimating:

$$y_i = \alpha + \rho s_i + \eta_i$$

$\implies$ OVB Since A_i in η_i is correlated with s

Use IV when:
$\exists$ a variable z_i that is ...
1. Correlated with s_i [1st stage]
2. Uncorrelated with all other determinants of y_i [Exclusion restriction]
$[Cov(\eta_i, z_i) = 0]$

//Essentially, IV breaks the variation in s into two parts://
1. Correlated w. η (problematic)
2. Unorrelated w. η (not-problematic) - *z_i helps detect this to estimate π_1*

Exclusion restriction:
z_i only effects y_i through s_i [1st stage channel]
z_i is as good as RA (Random assignment) $\implies$ independent of potential outcome and conditional on covariates

$s_i = X_i'\pi_{10} + \pi_{11}z_i + \zeta_1 i =$ **1st Stage, s reg** z_i
$y_i = X_i'\pi_{20} + \pi_{21}z_i + \zeta_1 i =$ **Reduced form,** y_i **reg** z_i
By the exclusion restriction...

$$\rho = \frac{Cov(y_i, z_i)}{Cov(s_i, z_i)} = \frac{Cov(y_i, z_i/V(z_i)}{Cov(s_i, z_i)/V(z_i)} = \frac{\text{reduced form reg. coefficient}}{\text{1st stage reg. coefficient}}$$

WALD ESTIMATION is this essentially. If z_i is {0,1} then $Cov(y_i, z_i) = p(1-p)$

2SLS -Basically you are just allowed to have many instruments (overidentified model).These other exogenous variables are used only in the second stage.
The parameter of interest is estimated the same way as before basically, it is that ratio or via the cov-s.

8.0.1 Angrist and Krueger, quarter of birth instrument

They were worried about OVB. So they started to think about factors in the variation that are nor related to ability so institutional factors. Some kids start somewhat before or after they turn 6 when they start school. This is just based on the specific date of birth (quarter of the year you´re born in). Schooling is compulsory until age 16. $\Longrightarrow$ some people will be held in school for longer just because they were born later in the year so maybe those born earlier in the year may be prone to drop out earlier from school. This seems true.
So quarter of birth is uncorrelated with ability but is with education so it could be an instrument. The quarter of birth also predicts earnings the same way. People in the 1st quarter earn less than the 4th etc...

Chapter 9

IV Part 2

Though it may be hard to test the Exclusion restriction, you could try to if you have a decent estimate of the unobserved: $\eta_i = y_i - \rho s_i$. But this is impossible because ρ must be consistent unbiased, which is also what we want to find etc... so yeah testing this is not possible.
- overidentifications are not very useful because it still relies on the same assumption
- For the 1st stage, as a rule of thumb you need F-stat of 10.

9.0.1 IV with treatment heterogeneity

Can´t capture ATE or ATET but LATE. Think about treatment as a chain. where there is heterogeneity in people being treated and the effect of the treatment on the outcome may be heterogeneous as well.
- Now the instrument must also be independent of potential treatment status. (Basically random potential treatment status)
So the 1st stage also shows the effect of treatment assignment on treatment status.
Monotonicity assumption: if you are assigned treatment it should on average increase the probability of being treated.

Definition 9.0.1: LATE

Given:
1. independence assumption
2. exclusion restriction
3. monotonicity
4. the existence of the 1st stage

IV estimate can be interpreted as the effect of the treatment status on those whose status was changed by the instrument.
This parameter is called the local average treatment effect.

The LATE Theorem

$$\frac{E[Y_i|Z_i=1]-E[Y_i|Z_i=0]}{E[D_i|Z_i=1]-E[Y_i|Z_i=0]} = E[Y_{1i}-Y_{0i}|D_{1i}>D_{0i}]$$

This is the average treatment effect for the group $D_{1i} > D_{0i}$
Proof:

To give a proof, start with the first bit of the Wald estimator:

$$E[Y_i|Z_i=1]$$

This can be written as weighted average of the outcome among compliers, never-takers,always-takers, and defiers:

$$\begin{aligned} E[Y_i|Z_i &= 1] = \\ E[Y_i|Z_i &= 1, \text{complier}] \cdot \Pr(\text{complier}|Z_i=1) \\ +E[Y_i|Z_i &= 1, \text{never-taker}] \cdot \Pr(\text{never-taker}|Z_i=1) \\ +E[Y_i|Z_i &= 1, \text{always-taker}] \cdot \Pr(\text{always-taker}|Z_i=1) \\ +E[Y_i|Z_i &= 1, \text{defier}] \cdot \Pr(\text{defier}|Z_i=1) \end{aligned}$$

We rule out defiers and re-write:

$$E[Y_i|Z_i=1] = E[Y_{1i}|C]\cdot\pi_c + E[Y_{0i}|N]\cdot\pi_n + E[Y_{1i}|A]\cdot\pi_a,$$

where π_c, π_n, and π_a are the *fraction* of compliers, never-takers, and always-takers, respectively and where C, N, and A denotes compliers, never-takers, and always-takers respectively.

Consider now the second term:

$$E[Y_i|Z_i=0]$$

In a similar vain, this can be written as:

$$E[Y_i|Z_i=0] = E[Y_{0i}|C]\cdot\pi_c + E[Y_{0i}|N]\cdot\pi_n + E[Y_{1i}|A]\cdot\pi_a,$$

The difference:

$$E[Y_i|Z_i=1]-E[Y_i|Z_i=0]$$

can be written as:

$$\begin{aligned} & E[Y_{1i}|C]\cdot\pi_c + E[Y_{0i}|N]\cdot\pi_n + E[Y_{1i}|A]\cdot\pi_a \\ & -E[Y_{0i}|C]\cdot\pi_c + E[Y_{0i}|N]\cdot\pi_n + E[Y_{1i}|A]\cdot\pi_a \\ = \;& E[Y_{1i}|C]\cdot\pi_c - E[Y_{0i}|C]\cdot\pi_c \\ = \;& E[Y_{1i}-Y_{0i}|C]\cdot\pi_c \end{aligned}$$

[H]

Next, turn to the denominator, where we can use similar arguments:

$$E[D_i|Z_i = 1] = E[D_{1i}|C] \cdot \pi_c + E[D_{0i}|N] \cdot \pi_n + E[D_{1i}|A] \cdot \pi_a,$$

and

$$E[D_i|Z_i = 0] = E[D_{0i}|C] \cdot \pi_c + E[D_{0i}|N] \cdot \pi_n + E[D_{1i}|A] \cdot \pi_a,$$

The difference is

$$\begin{aligned} E[D_i|Z_i &= 1] - E[D_i|Z_i = 0] \\ &= E[D_{1i}|C] \cdot \pi_c - E[D_{0i}|C] \cdot \pi_c = E[D_{1i} - D_{0i}|C] \cdot \pi_c \\ &= \pi_c \end{aligned}$$

We can now write:

$$\begin{aligned} & \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} \\ = & \frac{E[Y_{1i} - Y_{0i}|C] \cdot \pi_c}{\pi_c} = E[Y_{1i} - Y_{0i}|C], \end{aligned}$$

or:

$$= E[Y_{1i} - Y_{0i}|D_{1i} > D_{0i}],$$

This is the treatment effect among compliers

So ATET

Chapter 10

RD

10.1 Sharpe RD

$D_i = 1[x_i \geq x_0]$- is the treatment indicator. If something passes threshold, it will be treated.
threshold = x_0
x_i is the "forcing variable"
The regression becomes: $y_i = \alpha + \beta x_i + \rho D_i + \eta_i$
You can also use some p-order polynomial. You can also have different ones below and above the threshold.- You do this by including interactions between Forcing variable with the threshold indicator.

10.1.1 Non-Parametric RD

You only consider data very close to the threshold. δ = size of the neighbourhood.
Compare means between treatment and control.
However you need a lot of data.
There are bandwidth tests for δ but you can and maybe should just change delta and plot the results as you do that.

Those to the left and right of the threshold should be same, also, the control group should not be reacting to the treatment (Lucas-critique?).

10.2 Fuzzy RD

I a reduced form version, the actual treatment is unobserved, but we know there is a jump in the probability of treatment.Reaching the discontinuity can be like an instrument for the actual treatment. In the end you get LATE (but it is even more "local" because you only get it for those close to the threshold.).
In an RD, you should always do this graph: Also, plot the distribution of observations across bins:

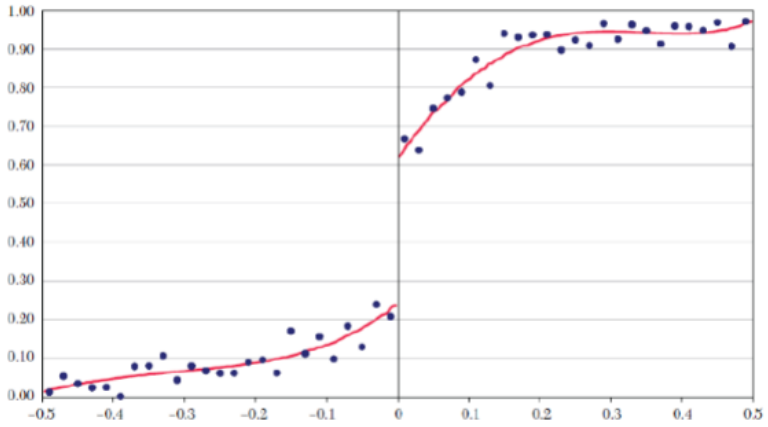
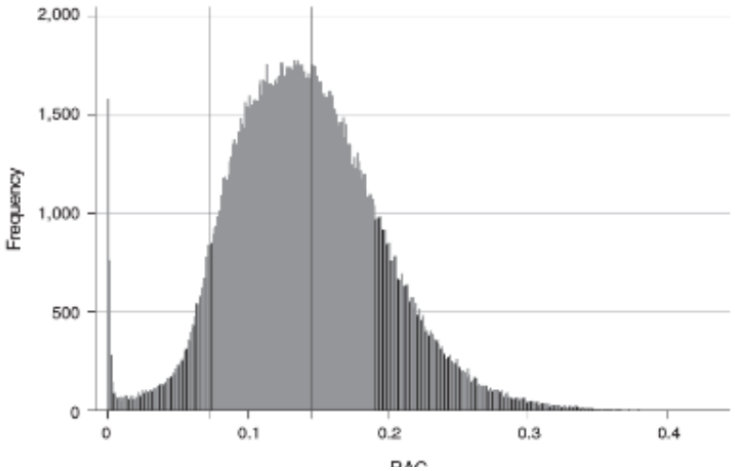


Figure 10.1: RD scatter plot x axis - forcing variable, y axis - the outcome variable. The points are the actual data binned.



There shouldn´t be any weird manipulation around the threshold for treatment.

10.3 Fixed effects and panel data

For causality, RD & IV are preferred but if we can´t use those, we can still do panel data. A way of using panel is through fixed effects. Typically, our No. of individuals observed is larger than the time dimension but this does not necessarily have to be the case.

Fixed effects refer to the variation in individuals that don´t change over time η_i. When using fixed effects, we allow $E[\eta_i|X_{i1}...X_{iT}] \neq 0$ There may be correlation between X_{iT} and η_i and we deal with it by introducing "dummy" variables for the fixed effects $\forall$ individuals. But this could be a lot of individuals

10.3.1 Within estimation

We use the trick that implementing all those dummies is like using the difference from the means.

To implement the trick, we first calculate the *individual-specific* averages over time, so that:

$$\bar{Y}_i = \bar{X}_i\beta + \overline{\eta_i} + \bar{\varepsilon}_i$$

where

$$\bar{Y}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it} \qquad \bar{X}_i = \frac{1}{T}\sum_{t=1}^{T} X_{it} \qquad \bar{\varepsilon}_i = \frac{1}{T}\sum_{t=1}^{T} \varepsilon_{it} \qquad \overline{\eta_i} = \frac{1}{T}\sum_{t=1}^{T} \eta_i$$

Next subtract $\bar{Y}_i$ from Y_{it},

$$Y_{it} - \bar{Y}_i = X_{it}\beta + \eta_i + \varepsilon_{it} - \bar{X}_i\beta - \overline{\eta_i} - \bar{\varepsilon}_i = (X_{it} - \bar{X}_i)\beta + (\varepsilon_{it} - \bar{\varepsilon}_i).$$

Note know that $\overline{\eta_i} = \eta_i$, since η_i is always the same across time periods. This implies that we get the specification

$$\tilde{Y}_{it} = \tilde{X}_{it}\beta + \tilde{\varepsilon}_{it} \qquad i = 1, \ldots, N \quad t = 1, \ldots, T$$

with

$$\tilde{Y}_{it} = Y_{it} - \bar{Y}_i \qquad \tilde{X}_{it} = X_{it} - \bar{X}_i \qquad \tilde{\varepsilon}_{it} = \varepsilon_{it} - \bar{\varepsilon}_i$$

The within estimator $\hat{\beta}_{\text{within}}$ is then obtained by applying OLS.

In this specification, we got rid of η_i, i.e. the fixed effects!

On the properties of within estimation

1. The parameters β are identfied due to (within) variation in X_{it} over time.
2. Estimators for η_i and β are consistent if the asymptotics imply that T becomes large.
3. If instead T is fixed and N goes to infinity, only $\hat{\beta}$ within is consistent, but $\hat{\eta}_i$ is not (so called incidental parameters).
4. If N is not too large, one could simply include dummy variables for each individual and estimate the original model by OLS. This provides the within estimators and $\hat{\eta}_i$ in a single step

10.3.2 1st-Differences

We can also take first differences instead. If T=2 then this is equivalent to doing the Fixed effects.

$$\begin{aligned} Y_{it} - Y_{it-1} &= X_{it}\beta + \eta_i + \varepsilon_{it} - X_{it-1}\beta - \eta_i - \varepsilon_{it-1} \\ &= (X_{it} - X_{it-1})\beta + (\varepsilon_{it} - \varepsilon_{it-1}) \qquad t = 2 \ldots, T \end{aligned}$$

or $\Delta Y_{it} = \Delta X_{it}\beta + \Delta\varepsilon_{it}$, where taking first-differences eliminates η_i from the model.

More assumptions:

For both the first-differences and within estimator to provide consistent estimates, we now need the regressors to be *strictly exogenous*:

$$\mathrm{E}[\varepsilon_{it}|X_{i1}, \ldots, X_{iT}, \eta_i] = 0 \qquad i = 1, \ldots, N \quad t = 1, \ldots, T$$

The part of the error term that vary over time ϵ_{it} , must be unrelated to the value of the treatment indicator or other control variables in any time period. It would typically fail, if there is some time-specifc unobserved shock that affects both the outcome and our X variable of interest.

10.3.3 Pitfalls of Fixed Effects

Measurement error problem

Increase in **measurement error** compared to OLS because of only having variation in X_{it} over time. Specifically, **downward bias**.The downward bias gets stronger, the stronger the correlation between the x-variables in different periods.

Impossible to estimate time-invariant regressors

The deviation from the individual-specific mean will always be zero for such a variable. We can therefore not estimate the effect of time-invariant factors such as gender, ethnicity, education (at least as an adult), etc. Random effects could do this but has unrealistic assumptions.

The effect is only identified for those who actually change treatment status

We are relying on a sample that changes treatment status. Since we are relying on within-individual variation. This makes results difficult to compare with OLS.

Violation of strict exogeneity assumption

see above.
Note: The fixed effects approach does not require a time dimension! As long as important unobserved variables are shared by some group of individuals, they can be cancelled out using an FE approach.

10.4 Difference in Differences DiD

In fixed effects, we looked at individual level data.However, it is possible that some treatments are at an aggregate level on some group of individuals,

DiD assumes that in the absence of a minimum wage change, employment is determined by the sum of a time-invariant state effect γ_s and a year effect λ_t that is common across states.

We can then write observed employment as:

$$y_{ist} = \gamma_s + \lambda_t + \delta D_{st} + \varepsilon_{ist}, \tag{1}$$

where D_{st} is an indicatator of being "treated" with an increase in minumum wage.

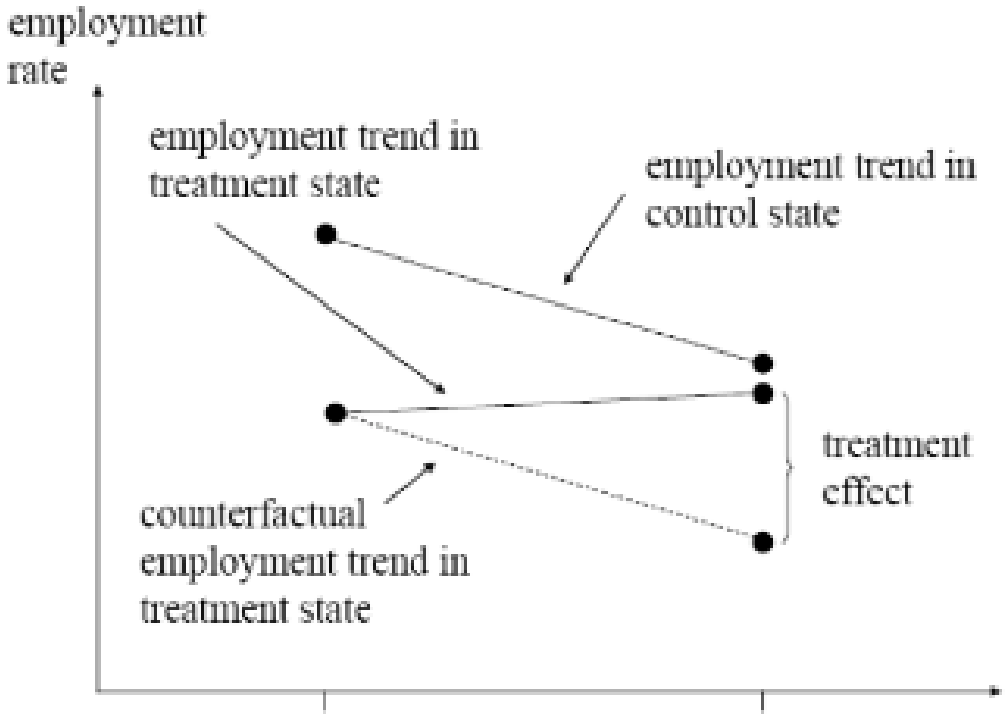
Treatment is at the state level,

DiD estimate:

$$\hat{\Delta} = \left(\bar{Y}_{\text{treatment,after}} - \bar{Y}_{\text{treatment,before}}\right) - \left(\bar{Y}_{\text{control,after}} - \bar{Y}_{\text{control,before}}\right)$$

where $\overline{Y}$ represents sample means

Parallel trend assumption.
Note: It can never be proven right or wrong, it requires a counterfactual statement.



10.4.1 2x2 DiD

You can also estimate this in a regression.

Taking the minimum wage paper as an example, a common way of writing such a 2 x 2 DiD-model is:

$$y_{ist} = \alpha + \gamma NJ_s + \lambda D_t + \delta(NJ_s * D_t) + \varepsilon_{ist}, \qquad (14)$$

where NJ_s is a dummy for being a New Jersey restaurant and D_t is a time dummy that switches on for observations obtained in November, i.e. after the minimum wage change

10.4.2 Two-way fixed effects

When treatment doesn´t turn on at the same time for all individuals: So there are group

For instance, US states may increase their minimum wages at different points in time

We can handle this with the more general DiD specification:

$$y_{ist} = \gamma_s + \lambda_t + \delta D_{st} + \varepsilon_{ist}, \tag{16}$$

where γ and t control for group and time effects and D represent treatment (like in Eq. 1 above)

and time dummies.

10.4.3 Dynamic DiD

This is possible when the data includes several pre and post treatment period. We can then extend the standard DiD with:

$$y_{ist} = \gamma_s + \lambda_t + \sum_{\tau=-q}^{-1} \gamma_\tau D(lead)_{s\tau} + \sum_{\tau=0}^{m} \delta_\tau D(lag)_{s\tau} + \varepsilon_{ist},$$

where we allow for q leads, $(\gamma_{-1}, \gamma_{-2}, ..., \gamma_{-q})$ or pre-treatment effects, and m lags $(\delta_{|1}, \delta_{|2}, ..., \delta_{|m})$., or post-treatment effects

For example, with periods running from -3 to +3, the dummy $D_{s\tau-3}$ would indicate 3 periods before the treatment takes place.

Recall the D takes the value 1 only for those that become treated and is always 0 for the "control" group. The leads and lags are thus always zero for the control group

The associated coefficient, $\gamma_{\tau-3}$, would thus measure how the outcome of the "treatment group" evolves between -3 and some reference period, *compared to the control group*.

Typically -1 is used as the reference period

Using this, we ca create an even study graph to:
1. Study whether the effects change over time (effect dynamics). We look at the post-treatment (lags) effects for this.
2. Study if the parallel assumption makes sense. We look at the pre-treatment (leads) effects for this.

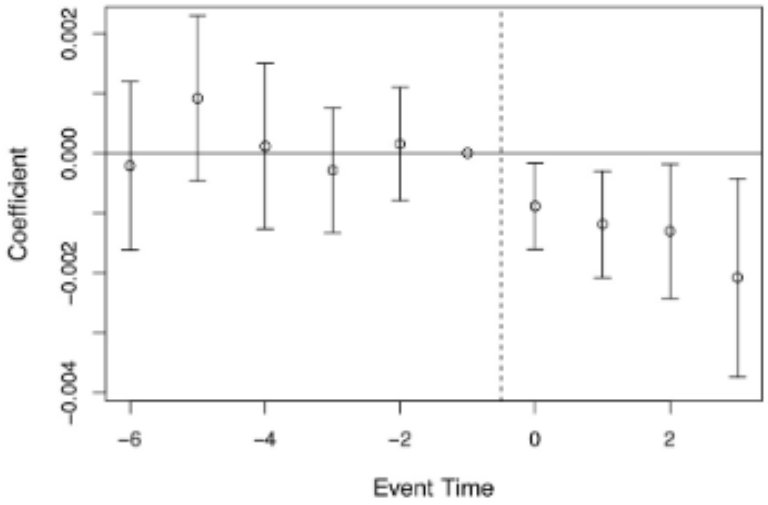


Figure 10.2: Enter Caption

1. Pre-treatment coefficients are around zero

This shows that pre-treatment trends are parallell

Note: *we do not know if the trends would have been parallell also in the absence of treatment!*

use "cluster(state)" ins stata for standard erros & no bootstrapping

10.5 Matching estimators

Matching is useful when you can´t use RD IV or DiD burt ranks lower in causality.
Assumptions:

- The selection into treatment is completely determined by variables that can be observed by the researcher;
- Conditional on these observable variables, the assignment to treatment is (assumed) random.
- Thus, two persons with the same x, one treated, one not, are assumed to have the same counterfactuals

Definition 10.5.1:

All variables that are relevant for jointly determining treatment and outcomes are observed and included in Xi . We write this as $(Y_{0i}^*, Y_{1i}^*) \perp D_i | X_i$.Not Testable!

Definition 10.5.2: Overlap/Common support assumption

All treatment have a control counterpart in the population.Testable!

General matching estimator to get ATET:

Suppose we have N_T individuals with treatment and N_C control individuals without treatment:

$$ATET = \frac{1}{N_T} \sum_{i \in T} \left[Y_i^T - \sum_{j \in C} \omega_{i,j} Y_j^C \right]$$

$$= \frac{1}{N_T} \sum_{i \in T} Y_i^T - \frac{1}{N_T} \sum_{j \in C} \omega_j Y_j^C$$

$\omega_{i,j}$ is the weight that individual i in the treatment group attaches to individual j in the control group

The formula says: for each treated individual, we subtract from his outcome the outcome of one or several untreated individuals. If each treated individual is compared to several untreated, we can use weights that reflects the importance of these untreated individuals. There is a curse of dimensionality and to solve this problem we summarize characteristics of individuals in:

Definition 10.5.3: Propensity score

The propensity score is the estimated probability of participating in a treatment, given observed characteristics X_i

Now CIA: $(Y_{0i}^*, Y_{1i}^*) \perp D_i | p(X_i)$ where $p(X_i)$ is the propensity score. So we match to treated, the control individuals with the same/ as close as possible, propensity score.

What variables should be included when estimating the propensity scores? Only variables that influence simultaneously the treatment decision and the outcome variable should be included It should also be clear that only variables that are unaffected by treatment (or the anticipation of it) should be included in the model. To ensure this, variables should either be **fixed over time** or **measured before participation**.

10.5.1 Nearest Neighbour matching

The most straightforward matching estimator is nearest neighbour (NN) matching:

$$ATET = \frac{1}{N_T} \sum_{i \in T} \left[Y_i^T - \sum_{j \in C} \omega_{i,j} Y_j^C \right]$$

Here, $\omega_{i,j}$ is equal to 1 for the closest control unit and zero for all other control units

Oversampling can be used, where several "closest" neighbours are used. The weights are then 1/Nc, for the included "closest neighbours

How many neighbours to match? - Bias Variance trade-off.
NN depends on who we start the matching with, so we can do this with or without replacement. i.e. If someone has already been matched, can that person also be matched to someone else.- Bias Variance

NN matching faces the risk of bad matches, if the closest neighbour is far away. This can be avoided by imposing a tolerance level on the maximum propensity score distance (caliper). If there are no controls found within the caliper, the treatment individual is discarded.

10.5.2 Kernel matching

Kernel matching uses all control observations but puts some weight on them depending on how close they are to the treatment to be matched to. Close = higher weight obviously. Benefit: low variance (buuut bad matches may also be used)

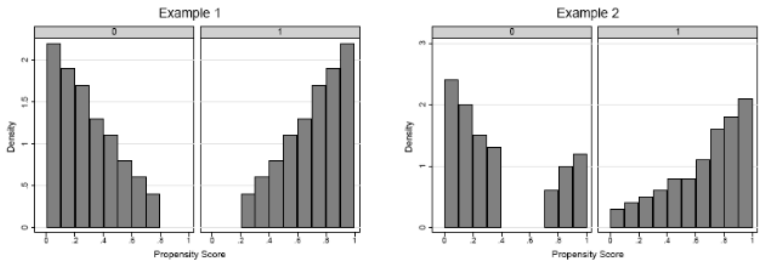


Figure 10.3: Common support problem displayed in example 2

A more formal test:
Compare the minima and maxima of the propensity score in the treatment and control groups Example: assume that the propensity score lies within the interval [0.07;0.94] in the treatment group and within [0.04; 0.89] in the control group. With the minima and maxima criterion, the over lap or common support is given by [0:07; 0:89]. We then delete all observations whose propensity score is smaller than the minimum and larger than the maximum in the opposite group.

More generally, just indicate propensities where there is positive density for both treatment and control.

Balancing: If the distribution of X -variables in both groups is similar, we say that the X -variables are balanced, which means that the propensity score matching did a good job. Some say this is like checking $X_i \perp D_i | p(X_i)$ - after conditioning on the propensity score, X_i should not provide new information about the treatment decision.

A simple approach is to use a two-sample t-test to check if there are significant differences in co-variate means for both groups. f the balancing is not satisfactory, one reason might be misspecification of the propensity score model Try then to include interaction or higher-order terms in the propensity score estimation and test the balancing once again. If still not satisfactory, it may indicate a failure of the CIA

No easy way to get SE-s for matching estimators so we use bootstrapping.