Applied Microeconometrics I

Lecture 4: Identification based on observables

Cristina Bratu

Aalto University

September 14, 2023
Lecture Slides

Background

- RCTs solve the selection problem
- It is not possible to run a controlled experiment with every research question
- We need to rely on observational data

Causality without experiments

The identification strategy refers to the manner in which a researcher uses observational data (i.e. data not generated by a randomized trial) to approximate a real experiment.

- *Selection based on observables*
- *Instrumental variables*
- *Difference-in-differences*
- *Regression discontinuity design*

- The goal is to arrive at a situation where:

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

Selection based on observables

- We may not have a controlled experiment, but maybe the treated group and the non-treated group differ only by a set of **observable** characteristics.
- This assumption, which would justify the causal interpretation of our estimates, is known as the **Conditional Independence Assumption** (CIA), also called selection-on-observables

The CIA: an example

- To understand the CIA let's begin with an example: master thesis grade (Y_i) and taking this course (C_i), in particular if you take the course ($C_i = 1$) and if you do not take it ($C_i = 0$)
- Two possible outcomes Y_{0i}, Y_{1i}
- But we observe only

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

- A naive comparison of observed averages yields:

$$\begin{aligned} E[Y_{1i}|C_i = 1] - E[Y_{0i}|C_i = 0] \quad = \quad & E[Y_{1i} - Y_{0i}|C_i = 1] + \\ & E[Y_{0i}|C_i = 1] - E[Y_{0i}|C_i = 0] \end{aligned}$$

- Why do you think the bias is not zero?

Causality and the CIA

- We would like to keep constant relevant observable characteristics (e.g. GPA and affiliation)
- Let us compare the treatment and control group, taking into account observable characteristics:

$$E[Y_{1i}|X_i, C_i = 1] - E[Y_{0i}|X_i, C_i = 0] = E[Y_{1i} - Y_{0i}|X_i, C_i = 1] + E[Y_{0i}|X_i, C_i = 1] - E[Y_{0i}|X_i, C_i = 0]$$

- The CIA is valid when, conditioning on a set of observed characteristics X_i (in the example GPA and affiliation), the bias disappears

$$E[Y_{0i}|X_i, C_i = 1] = E[Y_{0i}|X_i, C_i = 0]$$

- Hence,

$$E[Y_{1i}|X_i, C_i = 1] - E[Y_{0i}|X_i, C_i = 0] = E[Y_{1i} - Y_{0i}|X_i, C_i = 1]$$

- Is the CIA plausible in this case?

Example

EXAMPLE: Case where the CIA holds

		Osku	Mia	Heikki	Maija
Potential grade without the course	Y_{oi}	**3**	**5**	**3**	**5**
Potential grade with the course	Y_{1i}	**4**	**5**	**4**	**5**
Male	X_i	**1**	**0**	**1**	**0**
Treatment (took the course)	D_i	**1**	**0**	**0**	**0**
Realized thesis grade	Y_i	**4**	**5**	**3**	**5**
Treatment effect	$Y_{1i} - Y_{0i}$	**1**	**0**	**1**	**0**

What is the observed difference between treated and non-treated?

What is the effect of treatment on the treated?

What is the observed difference between treated and non-treated among men?

Causality and the CIA

- In practice, how relevant is the selection problem?
- Three possible types of factors that affect the outcome variable:
  1. observable factors ✓
  2. unobservable factors not correlated with the treatment ✓
  3. unobservable factors correlated with the treatment NO
- What drives selection in the previous example? Do these factors affect the outcome variable?
  - Information, differences in preferences...
  - Do any of these (unobserved) selection factors affect the outcome variable?
- Note: why is this not a problem in an RCT?

Matching: Brief Introduction

- Idea: Compare individuals that are similar in observable characteristics
- Implementation of matching
  1. Divide workers into different categories on the basis of the observable characteristic
  2. Compare means in outcomes over these different categories
- Propensity score matching
  1. Estimate the propensity of the treatment using rich set of observational characteristics (propensity score): $P(D_i|X)$
  2. Compare means within cells defined on the basis of the propensity score : $E[Y_i|D_i = 1, P_i = p] - E[Y_i|D_i = 0, P_i = p]$

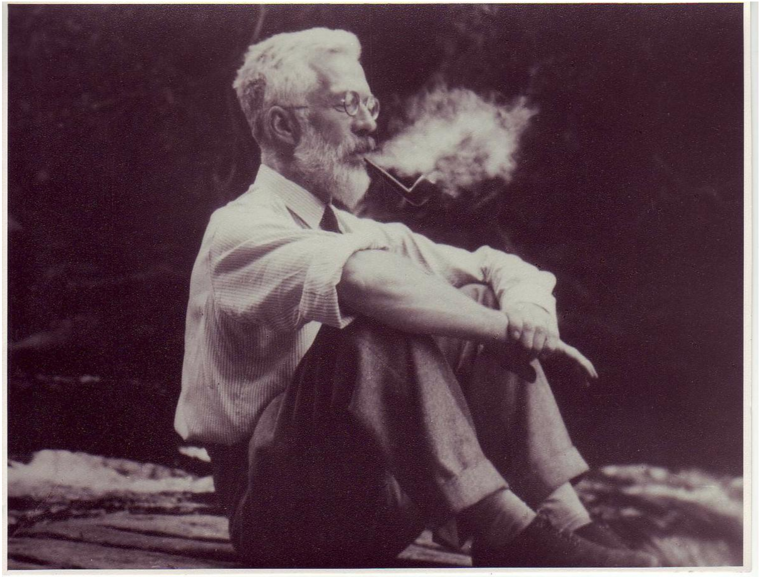
Example: Smoking and Mortality

- Cochran 1968, Biometrics

	Yearly death rates per 1,000 person
Non-smokers	13.5
Cigarettes smokers	13.5
Cigars/pipes	17.4

- How should we interpret this descriptive evidence?

Smoking and causal inference in statistics: Ronald Fisher



Example: Smoking and Mortality

- Non-smokers and smokers differ in age

	Mean age (years)
Non-smokers	57.0
Cigarettes smokers	53.2
Cigars/pipes	59.7

- Age is correlated with smoking behaviour, and probably affects also mortality

Example: Smoking and Mortality

- We could compare death rates within age groups (matching by age)
- This way, we neutralize any imbalances in the observed sample related with age

Matching:

- Divide the sample into several age groups
- Compute death rates for smokers and non-smokers by age group
- Compare smokers and non-smokers by age group:

$$E[Y_i|D_i = 1, A_i = a] - E[Y_i|D_i = 0, A_i = a]$$

and calculate the average effect using some weight.

Example: Smoking and Mortality

- Adjusted Average Death Rates

	Yearly death rates per 1,000 person
Non-smokers	13.5
Cigarettes smokers	17.7
Cigars/pipes	14.2

- cigarette smokers had relatively low death rates only because they were younger on average
- perhaps the three groups are unbalanced in another variable... (any idea?)

Regression analysis: a brief introduction

In practice, there are many details to worry about when implementing a matching strategy. This leads us to regression analysis.

- Example: How does schooling affect wages?
  $Y_i(s) = \alpha + \rho s_i + u_i$
- where
  $Y_i(s)$ is earnings (outcome)
  s_i is schooling (treatment)
  α is the intercept, level of earnings when no schooling, $(Y_i(0))$
  ρ is the slope, how wages vary with schooling?

OLS estimator

- OLS (Ordinary Least Squares) estimator minimizes the sum of squared residuals

$$\begin{aligned}\tilde{\rho} &= \frac{\sum_{i=1}^{n}(s_i - \bar{s})\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n}(s_i - \bar{s})^2} \\ &= \frac{Cov(Y_i, s_i)}{Var(s_i)} \qquad (2)\end{aligned}$$

OLS estimator

- Under some assumptions, OLS is an estimator with some desirable properties:
- Assumptions
  1. A1. Linearity (in parameters): $y_i = \alpha + x_i'\beta + \epsilon_i$
  2. A2. Exogeneity: $E(\epsilon_i|x_i) = 0$
  3. A3. No linear dependency (multicollinearity)
  4. A4. $Var(\epsilon|X) = \sigma^2$ (homoscedasticity) and $Cov(\epsilon_i, \epsilon_j|X) = 0$
- Under these assumptions OLS is unbiased and efficient (BLUE)

If schooling would be randomly assigned...

- However, it is not necessarily an estimate of the causal effect of s_i on Y_i
- Only when we have random exposure of subjects to the treatment in the population, conditional on observables, we can be sure that regression analysis provides a causal estimate

Endogeneity

- When the CIA is not satisfied we say that s is endogenous
- More generally, an explanatory variable s_{ji} is said to be *endogenous* if it is correlated with unobservable factors that affect the outcome variable (error term)
- Three main cases:
    1. **Omitted variable**
    2. **Measurement error**
    3. **Simultaneity**

Omitted variable bias

- Let us assume that the "true" model states that wages are affected by schooling and ability

  $$y_i = \alpha + \rho s_i + \gamma a_i + e_i$$

  where e_i is uncorrelated with s_i and a_i
- Unfortunately, we do not have a good measure for ability, and thus can only estimate the following short regression

  $$y_i = \tilde{\alpha} + \tilde{\rho} s_i + u_i$$

  where $u_i = \gamma a_i + e_i$
- Generally $\tilde{\rho}$ and ρ are different, unless:
  1. $\gamma = 0$
  2. s_i and a_i are uncorrelated in the sample
- Let us see in which sense they are different

What happens if we omit a variable

- Let us calculate the OLS estimator of $\tilde{\rho}$:

$$y_i = \tilde{\alpha} + \tilde{\rho} s_i + u_i$$

- OLS estimate

$$\tilde{\rho}_{ols} = \frac{Cov(y_i, s_i)}{Var(s_i)}$$

- Plug in the long regression

$$= \frac{Cov(\alpha + \rho s_i + \gamma a_i + e_i, s_i)}{Var(s_i)}$$

- Recall $Cov(e_i, s_i) = 0$

$$= \rho + \gamma \frac{Cov(a_i, s_i)}{Var(s_i)}$$

What happens if we omit a variable

$$\tilde{\rho}_{ols} = \rho + \underbrace{\gamma \frac{Cov(a_i, s_i)}{Var(s_i)}}_{\text{omitted variable bias}}$$

- $\Rightarrow \tilde{\rho}$ is generally biased for ρ
- Two cases in which there is **no** omitted variable bias:
  1. $\gamma = 0$ (a is not in the true model!)
  2. s and a are uncorrelated

What happens if we omit a variable

	$Corr(s, a) > 0$	$Corr(s, a) < 0$
$\gamma > 0$	POSITIVE BIAS	NEGATIVE BIAS
$\gamma < 0$	NEGATIVE BIAS	POSITIVE BIAS

Is adding controls always a good idea?

- CIA suggests that one way to deal with the omitted variable bias would be to include additional controls so that we are able to control for all the omitted variables
- However, adding controls may not always be a good idea
- **Bad controls** are variables that are themselves potential outcome variables in the notional experiment at hand
    1. Controlling for occupation in college-earnings regression
    2. IQ after schooling as proxy for ability in schooling-earnings regression (late proxy)

Is adding controls always a good idea?

- Let's see an example: controlling for occupation
- Occupation is affected by college. Does it make sense to look at the effect of college on earnings conditional on occupation?
- W_i is a dummy for white collar jobs, C_i a dummy for colleges, and Y_i earnings
- Counterfactual outcomes: $Y_{0i}, Y_{1i}, W_{0i}, W_{1i}$
- As usual we observe:

$$Y_i = C_i Y_{1i} + (1 - C_i) Y_{0i}$$
$$W_i = C_i W_{1i} + (1 - C_i) W_{0i}$$

- Let's assume that C_i is randomly assigned $\Rightarrow$ no trouble estimating its causal effect on both Y_i and W_i
- Let us assume that we want to see the impact of C_i on Y_i for white collar workers

Bad controls

- Given the assumptions we can easily estimate:

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

and

$$E[W_i|C_i = 1] - E[W_i|C_i = 0] = E[W_{1i} - W_{0i}|C_i = 1]$$

- But we want to know

$$E[Y_{1i} - Y_{0i}|C_i = 1, W_i = 1]$$

Bad controls

- We can either control for W_i in a regression or regress Y_i on C_i in the sample where $W_i = 1$:

$$E[Y_i|W_i = 1, C_i = 1] - E[Y_i|W_i = 1, C_i = 0] =$$
$$E[Y_{1i}|W_{1i} = 1, C_i = 1] - E[Y_{0i}|W_{0i} = 1, C_i = 0]$$

- By the joint independence of $\{Y_{1i}, W_{1i}, Y_{0i}, W_{0i}\}$ and C_i:

$$E[Y_{1i}|W_{1i} = 1, C_i = 1] - E[Y_{0i}|W_{0i} = 1, C_i = 0] =$$
$$E[Y_{1i}|W_{1i} = 1] - E[Y_{0i}|W_{0i} = 1]$$

Bad controls

- Calculating the above we see the problem:

$$E[Y_{1i}|W_{1i}=1]-E[Y_{0i}|W_{0i}=1]$$
$$=E[Y_{1i}-Y_{0i}|W_{1i}=1]+\{E[Y_{0i}|W_{1i}=1]-E[Y_{0i}|W_{0i}=1]\}$$

- The bias is due to the fact that college is likely to change the composition of the pool of white collars
- You need an explicit model of the links between college, occupation, and earning

Example

EXAMPLE: Case with a bad control

		Osku	Mia	Heikki	Maija
Potential grade without the course	Y_{oi}	3	5	3	5
Potential grade with the course	Y_{1i}	4	4	4	5
Seminar attendance without the course	W_{oi}	0	1	0	1
Seminar attendance with the course	W_{1i}	1	1	1	1
Treatment (took the course)	D_i	1	0	0	1
Seminar attendance	W_i	1	1	0	1
Realized thesis grade	Y_i	4	5	3	5
Treatment effect on grades	$Y_{1i} - Y_{0i}$	1	-1	1	0
Treatment effect on seminar attendance	$W_{1i} - W_{0i}$	1	0	1	0

Check that the observed differences between treated and the non-treated are same as the effect of treatment on treated for both Y and W!

What is the observed difference of Y between treated and the non-treated when W=1?

Is this equal to the effect of treatment on the treated when $W_{1i} = 1$?

Is $E[Y_{0i}|W_{1i} = 1] = E[Y_{0i}|W_{0i} = 1]$ in this case?

OLS estimator

- The exogeneity assumption, $E(\epsilon_i|x_i) = 0$, implies that $Cov(x_i, \epsilon_i) = 0$
- Then, the OLS estimator of β:

$$\begin{aligned} \hat{\beta}_{OLS} &= \frac{Cov(y,x)}{Var(x)} \\ &= \frac{Cov(\alpha+\beta x+\epsilon,x)}{Var(x)} \\ &= \beta\frac{Var(x)}{Var(x)} + \frac{Cov(x,\epsilon)}{Var(x)} \\ &= \beta \end{aligned}$$

Bad controls

$$
\begin{aligned}
E[Y_i|W_i = 1, C_i = 1] - E[Y_i|W_i = 1, C_i = 0] \\
&= E[Y_{1i}|W_{1i} = 1, C_i = 1] - E[Y_{0i}|W_{0i} = 1, C_i = 0] \\
&= E[Y_{1i}|W_{1i} = 1] - E[Y_{0i}|W_{0i} = 1] \\
&= E[Y_{1i}|W_{1i} = 1] - E[Y_{0i}|W_{1i} = 1] \\
&+ E[Y_{0i}|W_{1i} = 1] - E[Y_{0i}|W_{0i} = 1] \\
&= E[Y_{1i} - Y_{0i}|W_{1i} = 1] + E[Y_{0i}|W_{1i} = 1] - E[Y_{0i}|W_{0i} = 1]
\end{aligned}
$$