Welcome to Mastering Metrics!

Econ 140 Spring 2025, Section 1

Jonathan Old

Syllabus/OH bCourses Website Feedback form (**Always open**) RStudio

Roadmap

1. Getting Started & Housekeeping
2. Time for your questions
3. Econ 140: The Big Picture
4. Correlation and Causation

Getting Started & Housekeeping

Introduce yourself ...

... with one interesting and one boring fact about you that you would like us to know!

Housekeeping

- What you can expect in this section and from me as GSI:

Housekeeping

- What you can expect in this section and from me as GSI:
  - Safe Space

Housekeeping

- What you can expect in this section and from me as GSI:
  - Safe Space
  - Diversity

Housekeeping

- What you can expect in this section and from me as GSI:
  - Safe Space
  - Diversity
  - Challenges in this course

Housekeeping

- What you can expect in this section and from me as GSI:
  - Safe Space
  - Diversity
  - Challenges in this course
  - Mental Health

Housekeeping

- What you can expect in this section and from me as GSI:
  - Safe Space
  - Diversity
  - Challenges in this course
  - Mental Health
- Core values: Transparency, Integrity, Respect

Housekeeping

- What you can expect in this section and from me as GSI:
  - Safe Space
  - Diversity
  - Challenges in this course
  - Mental Health
- Core values: Transparency, Integrity, Respect
- I am your learning companion. I see grades as feedback, not as an evaluation of your abilities or personality.

What your peers advise!

What advice would you give to another student who is considering taking this course?

Comments
Attend every section, they make the course more engaging and easier to comprehend
Review and asking a lot of questions, going to outside help (SLC)
Go to class and pay attention and you should be okay.
Be active in sessions. Cooperate with others.

What advice would you give to another student who is considering taking this course?

Comments
Ask questions and try your best not to fall behind because it often compounds!
Stay on top of your work. This class is not one where you can skip lectures and cram material right before exams. Attend lectures and sections in person even if things aren't making sense.
Make sure you attend lectures, and definitely make sure you attend discussion sections – but also take advantage of the recordings to go over and cement the concepts and materials ahead of the exams.
You should have some programming skills, even if its the slightest knowledge of R or Python. It helps.
Pay attention in lecture/review the recordings if needed and go to section.
Please ask questions, no matter how simple the question may be.
Do the section handouts if you have time.
Attend sections, approach GSIs and Professors if needed. Do assignments on time and start reviewing early.

What your peers advise (2)!

What advice would you give to another student who is considering taking this course?

Comments
Go to office hours and create study groups!
Take advantage of GSI resources!
Don't fall behind lecture content

What advice would you give to another student who is considering taking this course?

Comments
Go to section and read the textbook!
Go to section!

What your peers advise (3)!

Please identify area(s) where you think the course could be improved.

Comments
Get rid of the group work.
–more practice problems
pace
n/a

What advice would you give to another student who is considering taking this course?

Comments
attend the sections.
–this isn't your regular math or coding class, come to class and exams with a creative, story telling mindset so you can best answer exam questions and paint a causal story
relax
Go to Jonathan's section.
attend lectures and office hours
go to office hour. I understood everything correctly in office hours.
Read the textbook! Attend section! It's so easy and it significantly increases your understanding.

What your peers critiqued!

Please identify area(s) where you think the course could be improved.

Comments
Sometimes I felt like in discussion topics were simplified a little too much compared to how they were described in lecture. I felt like lectures would be more complex and harder to understand while discussion sometimes seems like I was learning common sense. Now, this could have been a result of my GSI's incredible teaching style (since he was amazing at explaining concepts clearly and answering questions) but I just wanted to be sure that we were not skimping over any of the topics in lecture in section.
No improvements necessary.
The tone could be a little bit faster maybe
n/a
Perhaps break down concepts even further, for easier understanding.

Please identify area(s) where you think the course could be improved.

Comments
N/A
I think it could be beneficial to all students if section was made optional. When the students who don't have much interest in section show up, it brings down the morale and energy of the whole room. I felt bad because Jonathan would put in a good amount of energy into his sections and it often wasn't reciprocated
Recorded discussion sections.
This class felt very conceptual, or at least lectures. I think having more examples of applying the concepts would have been useful. I know the worksheets were intended for that but I think the worksheets were sometimes very dense and I wouldn't be able to finish them.
Sometimes the material was a bit boring/dry.
Uploading slides ahead of time.
Perhaps more section assignment explanation time.

What your peers critiqued (2)!

Please identify area(s) where you think the course could be improved.

Comments
N/A
Grammer was unclear sometimes.
Could be potentially more engaging, but good practice is most important and this is not a huge issue.
NA

Please identify area(s) where you think the course could be improved.

Comments
I really liked it when summery slides where made. Like before the midterm, keep up with more of that.
Hard to follow because the concepts are so hard
The main weakness is sometimes the material was taught different to how it was in lecture or we covered material in a different way than what the instructor would want from us from the midterm and final.
Release section assignments earlier.

What your peers critiqued (3)!

What did you like LEAST about the section with Jonathan?

Comments
Not a lot of students showed up
Discussing in groups
N/a
Last minute changes, OH too early

What did you like LEAST about the section with Jonathan?

Comments
I did not like the group work
Ummmm, i feel like jonathan had a really great section which was organized and prioritized participation. Maybe something I would add is more practice problems and more in depth walk through for answers on practice problems.
Emphasis on groupwork.
nothing
OH times. I kind of wish he offered it at alternative day of the week so I could attend.
n/a
office hour is too early << this is highly subjective
N/A

Newer feedback!

Which section that Jonathan taught has helped you THE LEAST? examples: Intro,regression, IV, DiD, RDD, coding, etc)?Why?

Comments
Probably coding but I already know coding so that was never an issue
I think DiD or coding, since the coding section was super light.
Coding. We just don't spend enough time on it. This makes sense since coding is not something we are tested on in this course.
Coding but that probably because I dislike R immensely
There was not a ton of coding examples in class. This may have made the problem sets more difficult.
Anything regarding code, it tends to be that Berkeley on the whole does a poor job of meeting students where they are code–wise, this is not a dig on you at all so much as I personally believe that R is not a great language to be teaching this course in. Not a relevant comment on your teaching as I thought you did the best with what you had, but I find R is finicky and obnoxious in general.
Maybe can talk a bit more about R

Newer feedback!

What advice would you give to another student who is considering taking this course?

Comments
Go to Jonathan's section
Make sure to attend sections and watch YouTube videos on each topic! Also attend lecture for sure.
Definitely make sure you have a solid understanding of intro level statistics and economics. They are critical in understanding the more advanced topics of econometrics.
Pay attention to the content and practice and ask your GSIs for help if you need it
Take it with the other professor in the fall
I would tell them to buy the textbook and primarily use the textbook and SLC review sessions for studying. I would also recommend them to start the problem sets as soon as they are assigned.
Learn python and go to class
Well it is required for Econ majors. I wouldn't recommend this as a fun course

Newer feedback!

What advice would you give to another student who is considering taking this course?

Comments
Stay engaged and read the textbook.
Don't fall behind.
Going to discussion really helps you understand the material clearly.
Go to lecture and pay attention.
section helps
Know how to code.
Learn how to code in R

Your questions

Any questions?

...Remember – this is a safe space! Every question is useful!

Econ 140: The Big Picture

The Big Picture

Econometrics can help us

- Understand causal relationships
- Critically consume news, research, claims
- Make decisions

in

- Our private lives
- (Economic) Research
- Business
- Public Policy

The Big Picture

- Econometrics gives us amazing tools to answer economic questions, but also to answer any quantitative question based on data. For example:
  1. Will my outcomes be better if I invest my time into finding summer internships or into my coursework?
  2. What is the effect of democracy on economic growth?
  3. Will a firm make more money if it introduces pay-for-performance contracts?
  4. Do minimum wages increase unemployment?

  In short: What is the effect of X on Y?
- Difference between econometrics and data science.
  - Causality vs prediction (see here)
  - Economic models

Correlation and Causation

Dissecting Bad Causal Claims

Discuss in groups of 4: Why is this statement problematic?



Over the past 60 years, more spending on police hasn't meant less crime

The Washington Post

Intuitively, one might worry that reducing police spending would lead to a spike in crime. A review of spending on state and local police over the past 60 years, though, shows no correlation nationally between spending and crime rates.

In 1960, about $2 billion was spent by state and local governments on police. There were about 1,887 crimes per 100,000 Americans, including 161 violent crimes. By 1980, spending had increased to $14.6 billion — and crime rates had soared to 5,950 crimes per 100,000 Americans and 597 violent crimes. Over the next two decades, those rates thankfully fell, down to about 4,120 crimes per 100,000 people and 507 violent crimes. Spending spiked to more than $87 billion. Eighteen years later — by 2018, the most recent year for which full data are available — crime rates had fallen further to 2,580 crimes per 100,000, including 381 violent crimes.

Figure 1: Police spending and Crime (Source)

Dissecting Bad Causal Claims II

Discuss in groups of 4: Why is this statement problematic?

Three's a crowd: Having more than 2 kids linked to weaker brain function

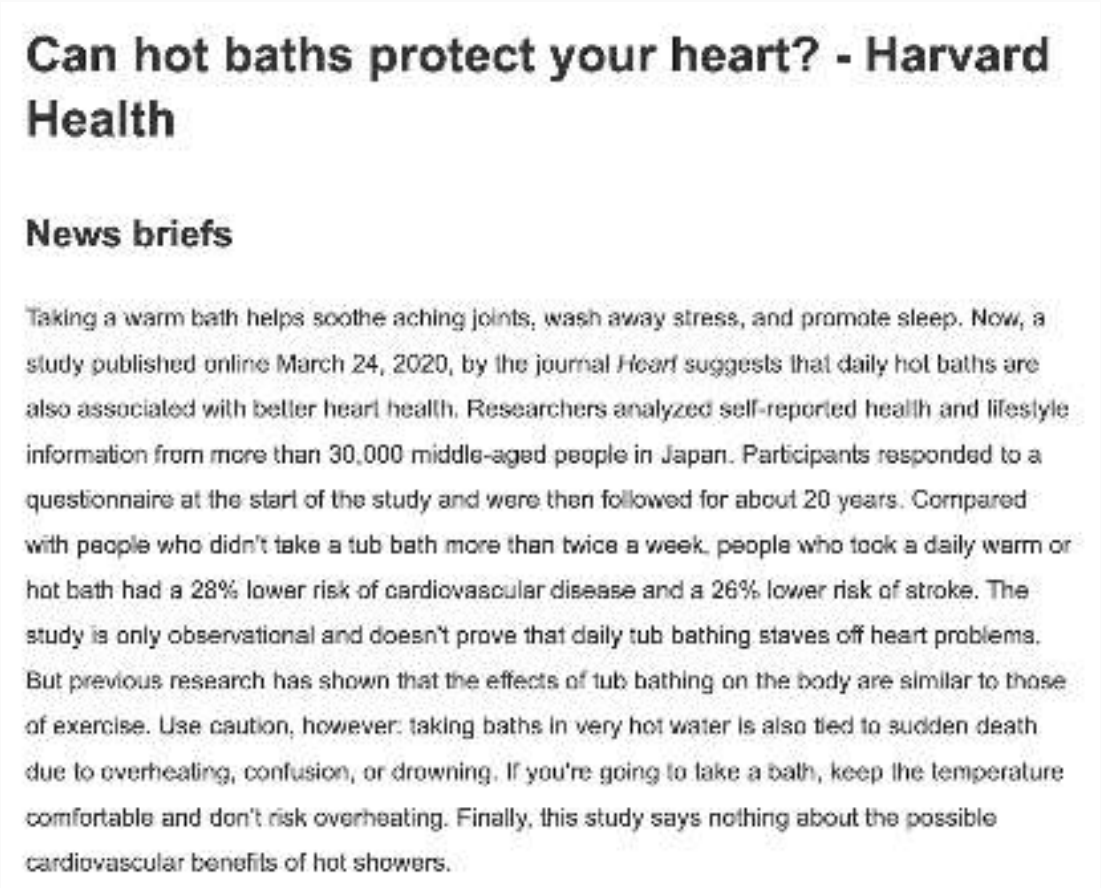


NEW YORK – Everything in moderation – even kids? Researchers from Columbia University and Université Paris-Dauphine report having more than two kids may have a negative impact on late-life cognition. The study shows that older parents with just two children appeared sharper cognitively than those with three.

Figure 2: Number of Children and Cognitive Function(Source)

Dissecting Bad Causal Claims III

Discuss in groups of 4: Why is this statement problematic?



Can hot baths protect your heart? - Harvard Health

News briefs

Taking a warm bath helps soothe aching joints, wash away stress, and promote sleep. Now, a study published online March 24, 2020, by the journal *Heart* suggests that daily hot baths are also associated with better heart health. Researchers analyzed self-reported health and lifestyle information from more than 30,000 middle-aged people in Japan. Participants responded to a questionnaire at the start of the study and were then followed for about 20 years. Compared with people who didn't take a tub bath more than twice a week, people who took a daily warm or hot bath had a 28% lower risk of cardiovascular disease and a 26% lower risk of stroke. The study is only observational and doesn't prove that daily tub bathing staves off heart problems. But previous research has shown that the effects of tub bathing on the body are similar to those of exercise. Use caution, however: taking baths in very hot water is also tied to sudden death due to overheating, confusion, or drowning. If you're going to take a bath, keep the temperature comfortable and don't risk overheating. Finally, this study says nothing about the possible cardiovascular benefits of hot showers.

Figure 3: Hot baths and health (Source)

Dissecting Bad Causal Claims IV

Discuss in groups of 4: Why is this statement problematic?



The New York Times

Another Benefit to Going to Museums? You May Live Longer

Researchers in Britain found that people who go to museums, the theater and the opera were less likely to die in the study period than those who didn't.

Leonardo da Vinci's "The Virgin and Child" at the Louvre in Paris. A British study found that people who engaged with the arts had a lower likelihood of dying during the study period than those who did not. Thibault Camus/Associated Press

Figure 4: Museums and longevity (Source)

Dissecting Bad Causal Claims IV

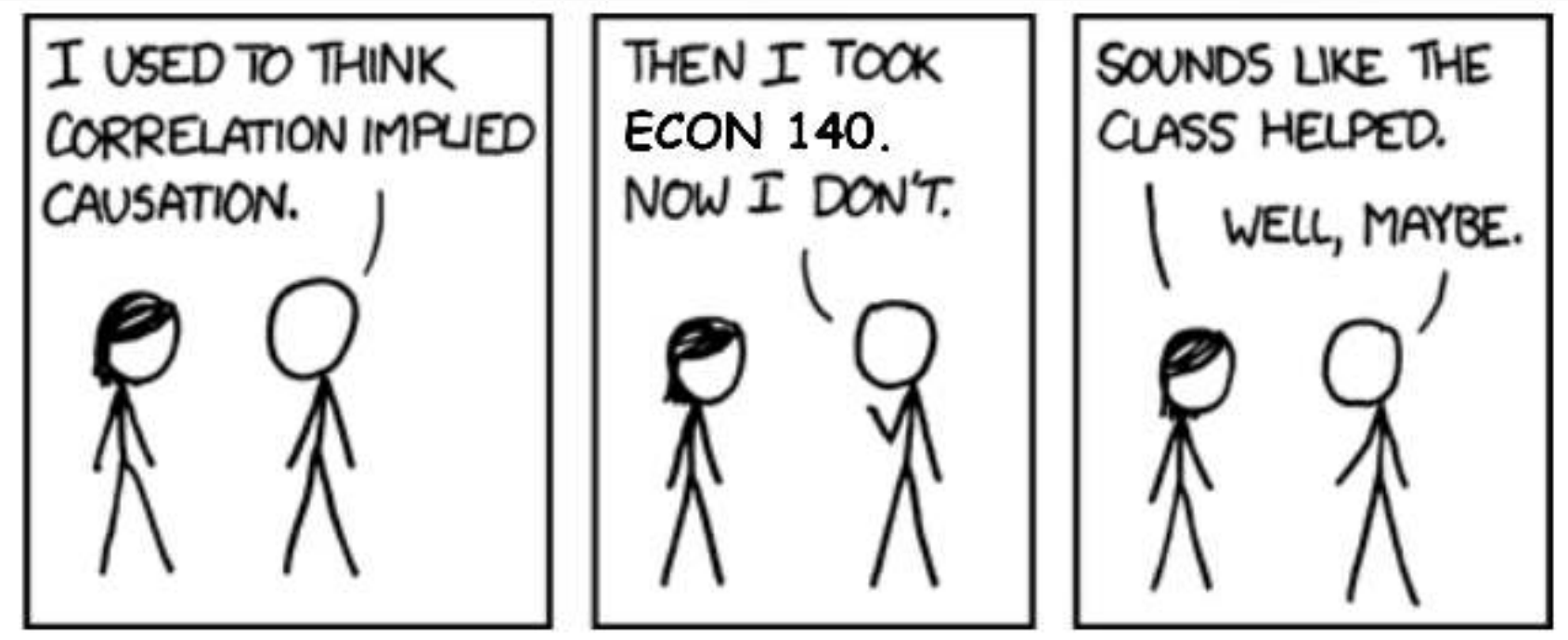


Figure 5: Correlation and causation

Quiz Time!

We have seen that
“Correlation Does Not Imply Causation”
What about:
“No Correlation Implies No Causation”?

From Counterfactuals to Linear Regression

Econ 140 Spring 2025, Section 2

Jonathan Old

Syllabus/OH bCourses Website Feedback form (**Always open**) RStudio

Roadmap

1. Time for your questions

2. Selection bias

3. Potential outcomes

4. RCTs

5. A non-mathematical intro to OLS – OR: standard errors, t-statistics, and hypothesis tests: What is that all about? Your choice

Your questions

Any questions?

...Remember – this is a safe space! Every question is useful!

Selection bias

How to think about Selection Bias

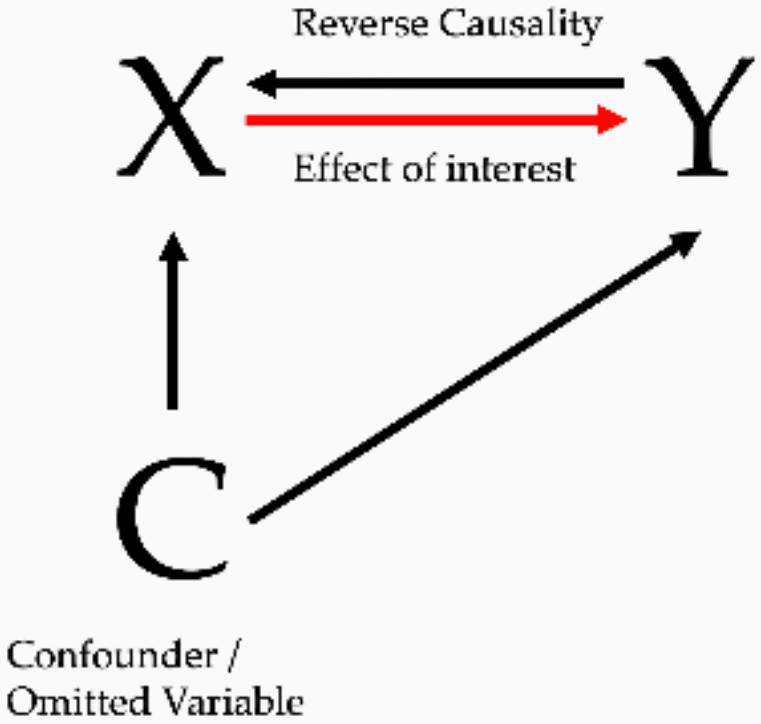


Figure 1: Selection bias

Dissecting Bad Causal Claims III

Discuss in groups of 2: Why is this statement problematic?



The New York Times

Another Benefit to Going to Museums? You May Live Longer

Researchers in Britain found that people who go to museums, the theater and the opera were less likely to die in the study period than those who didn't.

Leonardo da Vinci's "The Virgin and Child" at the Louvre in Paris. A British study found that people who engaged with the arts had a lower likelihood of dying during the study period than those who did not. Thibault Camus/Associated Press

Figure 2: Museums and longevity (Source)

Potential outcomes

Potential outcomes

- "Potential outcomes" is a framework that can help us **think through causal claims**: An alternative to math, drawing errors, or thinking things through
- We like to **write things down** in a rigorous way: Transparent, easy to verify, easy to replicate
- Potential outcomes are hypothetical outcomes
- Example: Your exam score when you go to all sections vs. when you go to no sections
- Think of potential outcomes as "parallel universes"

Our main challenge: We NEVER observe an individual at more than one status at the same time!

Potential outcomes: Notation

We write the potential outcomes as:

$$\left.\begin{aligned} Y_{i0} &= \text{Outcome of individual } i \text{ with "status" } 0 \\ Y_{i1} &= \text{Outcome of individual } i \text{ with "status" } 1 \end{aligned}\right\} \text{ Counterfactual Outcomes}$$

Alternative way of writing it: $Y_i(0)$ and $Y_i(1)$

"Status" can be anything

- Treatment assignment: 0 or 1
- Actual treatment: 0 or 1
- Drinking expensive whiskey or not
- Can also be: Multi-valued (number of children) or continuous (hours studied)

Potential outcomes: Notation

- We are often interested in the expected (think: average) potential outcome of a group of individuals with a given status.
- We write the group behind a conditional sign: $\mathrm{E}\left[\text{Score}_{i0} \mid \text{iPad}_i = 0\right]$ gives the potential outcome of a group of people that had no iPad, in the ”parallel universe” where they don’t have an iPad.
- Then, $\mathrm{E}\left[\text{Score}_{i1} \mid \text{iPad}_i = 0\right]$ gives the potental outcome of the same group (that currently have no iPad), in the ”parallel universe” where they do have an iPad.

Estimating the effect of iPads on grades

Let us start with a difference-in-means comparison:

$$\Delta = E[\text{Grade}_i|\text{iPad}_i = 1] - E[\text{Grade}_i|\text{iPad}_i = 0]$$

Add and subtract $E[\text{Grade}_i(0)|\text{iPad}_i = 1]$:

$$= E[\text{Grade}_i(1)|\text{iPad}_i = 1] - E[\text{Grade}_i(0)|\text{iPad}_i = 1] \quad +$$
$$E[\text{Grade}_i(0)|\text{iPad}_i = 1] - E[\text{Grade}_i(0)|\text{iPad}_i = 0]$$

Use properties of expectations:

$$= E[\text{Grade}_i(1) - \text{Grade}_i(0)|\text{iPad}_i = 1] \quad +$$
$$E[\text{Grade}_i(0)|\text{iPad}_i = 1] - E[\text{Grade}_i(0)|\text{iPad}_i = 0]$$
$$=$$

Estimating the effect of iPads on grades

Let us start with a difference-in-means comparison:

$$\Delta = E[\text{Grade}_i|\text{iPad}_i = 1] - E[\text{Grade}_i|\text{iPad}_i = 0]$$

Add and subtract $E[\text{Grade}_i(0)|\text{iPad}_i = 1]$:

$$= E[\text{Grade}_i(1)|\text{iPad}_i = 1] - E[\text{Grade}_i(0)|\text{iPad}_i = 1] \quad +$$
$$E[\text{Grade}_i(0)|\text{iPad}_i = 1] - E[\text{Grade}_i(0)|\text{iPad}_i = 0]$$

Use properties of expectations:

$$= E[\text{Grade}_i(1) - \text{Grade}_i(0)|\text{iPad}_i = 1] \quad +$$
$$E[\text{Grade}_i(0)|\text{iPad}_i = 1] - E[\text{Grade}_i(0)|\text{iPad}_i = 0]$$

$$= \text{Treatment effect for group with iPad} +$$

Estimating the effect of iPads on grades

Let us start with a difference-in-means comparison:

$$\Delta = E[\text{Grade}_i|\text{iPad}_i = 1] - E[\text{Grade}_i|\text{iPad}_i = 0]$$

Add and subtract $E[\text{Grade}_i(0)|\text{iPad}_i = 1]$:

$$= E[\text{Grade}_i(1)|\text{iPad}_i = 1] - E[\text{Grade}_i(0)|\text{iPad}_i = 1] \quad +$$
$$E[\text{Grade}_i(0)|\text{iPad}_i = 1] - E[\text{Grade}_i(0)|\text{iPad}_i = 0]$$

Use properties of expectations:

$$= E[\text{Grade}_i(1) - \text{Grade}_i(0)|\text{iPad}_i = 1] \quad +$$
$$E[\text{Grade}_i(0)|\text{iPad}_i = 1] - E[\text{Grade}_i(0)|\text{iPad}_i = 0]$$

$$= \text{Treatment effect for group with iPad} + \text{Selection bias}$$

Estimating the effect of iPads on grades

Let us start with a difference-in-means comparison:

$$\Delta = E[\text{Grade}_i|\text{iPad}_i = 1] - E[\text{Grade}_i|\text{iPad}_i = 0]$$

Add and subtract $E[\text{Grade}_i(0)|\text{iPad}_i = 1]$:

$$= E[\text{Grade}_i(1)|\text{iPad}_i = 1] - E[\text{Grade}_i(0)|\text{iPad}_i = 1] \quad +$$
$$E[\text{Grade}_i(0)|\text{iPad}_i = 1] - E[\text{Grade}_i(0)|\text{iPad}_i = 0]$$

Use properties of expectations:

$$= E[\text{Grade}_i(1) - \text{Grade}_i(0)|\text{iPad}_i = 1] \quad +$$
$$E[\text{Grade}_i(0)|\text{iPad}_i = 1] - E[\text{Grade}_i(0)|\text{iPad}_i = 0]$$
$$= \text{Treatment effect for group with iPad} + \text{Selection bias}$$

Selection bias: Students with and without iPad have different potential grades: **even if they both *had* iPads, they would be different**.

Selection Bias revisited

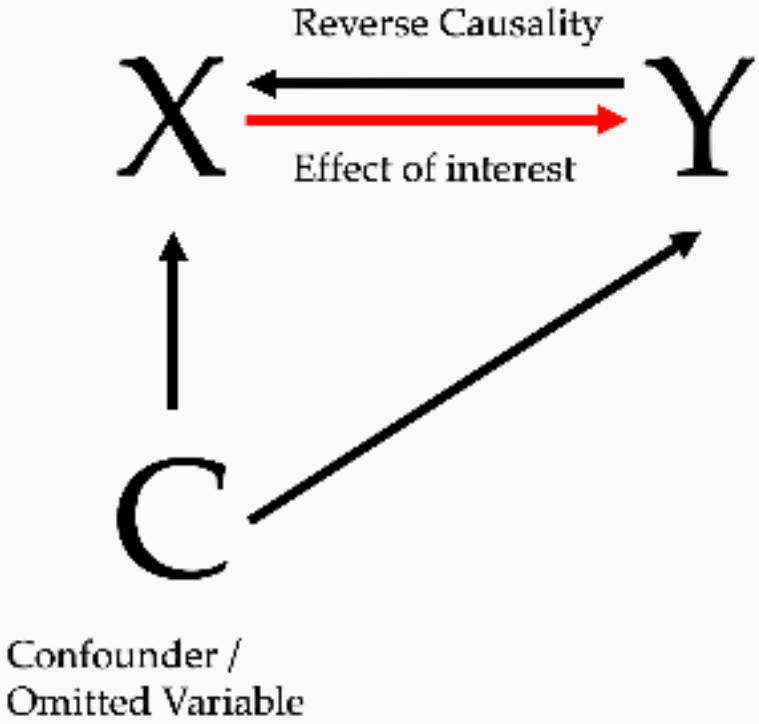


Figure 3: Selection bias

RCTs

RCTs solve selection bias

We had:

$$\Delta = E[\text{Grade}_i(1) - \text{Grade}_i(0)|\text{iPad}_i = 1] \quad + \\ E[\text{Grade}_i(0)|\text{iPad}_i = 1] - E[\text{Grade}_i(0)|\text{iPad}_i = 0]$$

- The second line was selection bias: The potential grade of individuals with and without iPad is different
- If the treatment (iPad) is **independent** of the potential outcomes, then:

$$\text{iPad}_i \perp (\text{Grade}_i(1), \text{Grade}_i(0))$$
$$\Rightarrow E[\text{Grade}_i(0)|\text{iPad}_i = 1] = E[\text{Grade}_i(0)|\text{iPad}_i = 0]$$

and selection bias will be zero.

Remaining issues with RCTs

RCTs solve the selection bias problem by **randomly assigning treatment** to different groups.

Discuss in groups:

1. Is this the same thing as saying ”We drew a random sample from the population?”
2. What are some problems with RCTs? (ethical/practical/econometric)

RCTs have revolutionized economics

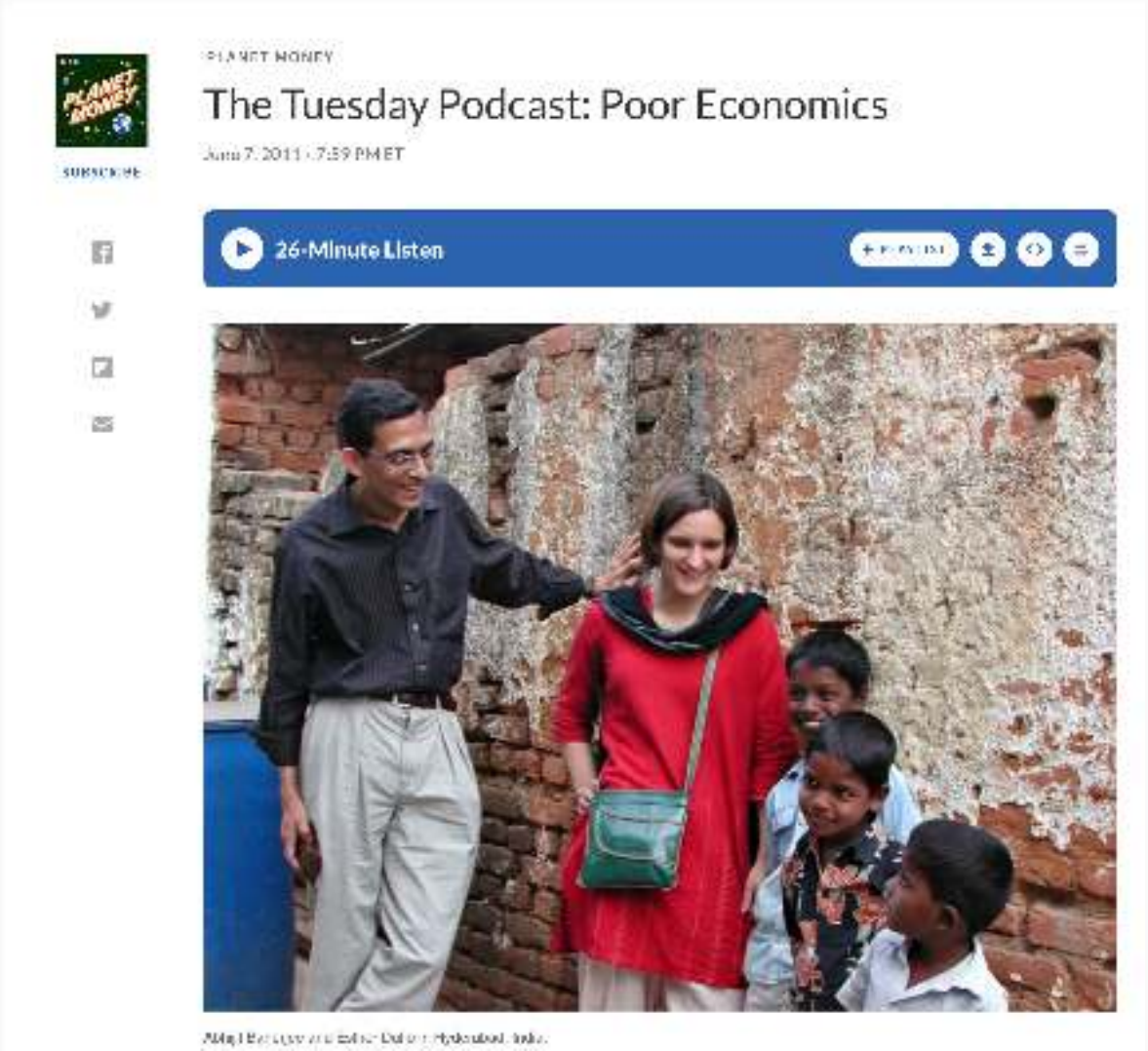


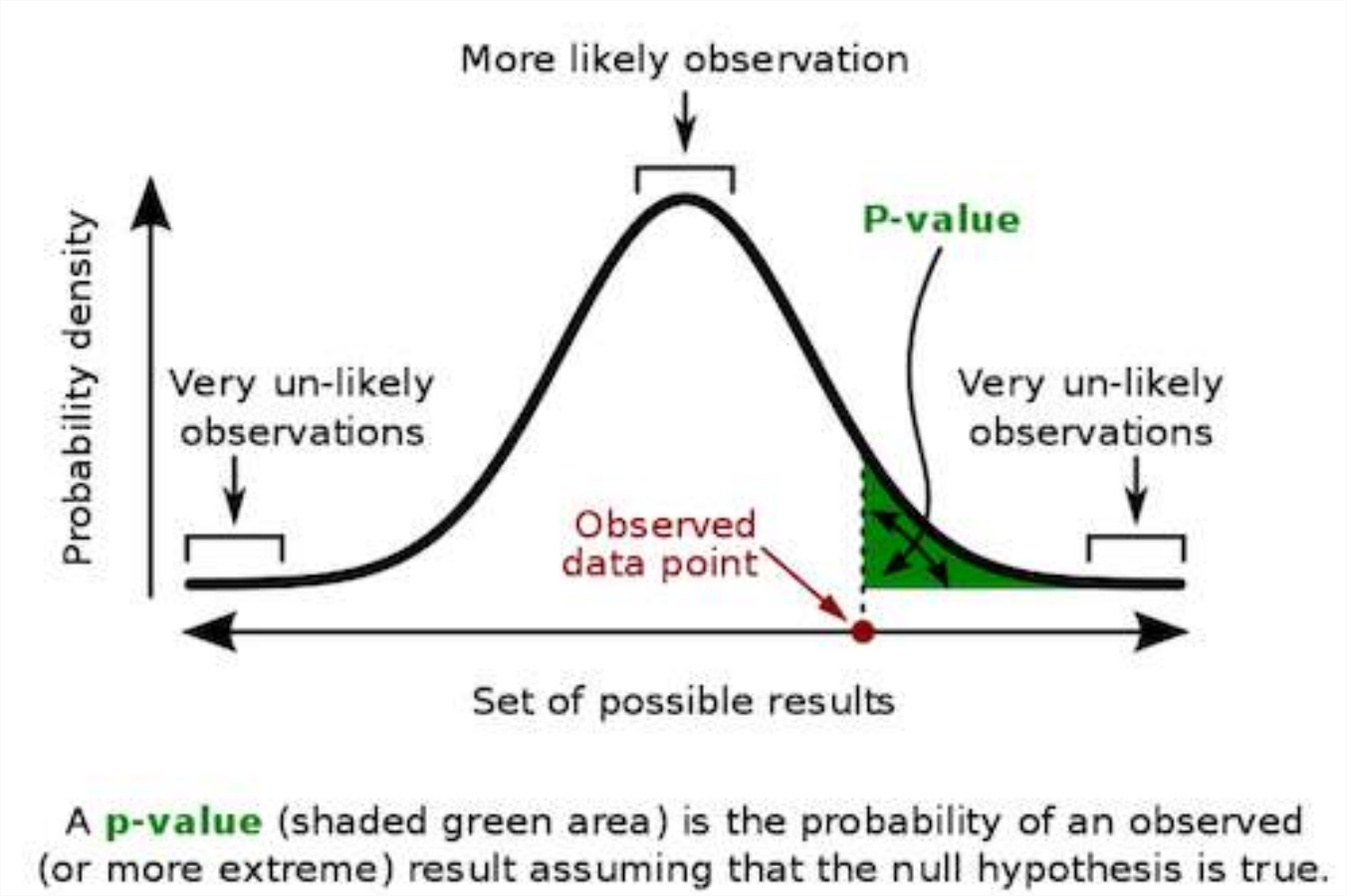
Figure 4: Abhijeet Banerjee and Ester Duflo. (Source)

OLS or hypothesis tests: Your choice

Hypothesis testing

- Random variables: Our estimator is a random variable (randomly drawn from population)
- Standard error: Random variables have a standard deviation, estimators have standard errors. This quantifies their uncertainty
- Statistics: The two keywords are the law of large numbers and the central limit theorem: The sum/mean over many draws from a random variable will be normally distributed
- For hypothesis test: null and alternative hypothesis.
- Assume the null hypothesis is true, then see how plausible results are if it is actually true.
- If they are implausible – we reject the null hypothesis! Otherwise: Fail to reject.

It’s all connected



Bias

Econ 140 Spring 2025, Section 3

Jonathan Old

Syllabus/OH bCourses Website Feedback form (**Always open**) RStudio

Roadmap

1. Time for your questions
2. Recap
3. Selection bias in action: Omitted Variable Bias
4. Time for your questions
5. Inference and hypothesis testing
6. Exam practice

Your questions

Any questions?

...Remember – this is a safe space! Every question is useful!

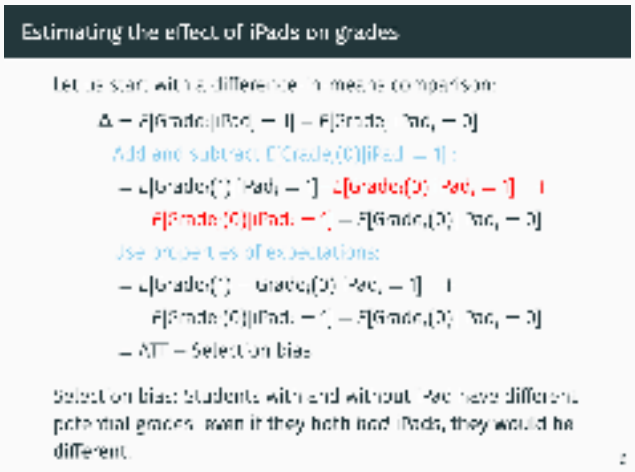
Recap

Recap: Regression

See on blackboard

Recap: Selection bias

We saw that whenever we do a difference-in-means comparison (or a regression), we get:



Causal Effect

+

Selection Bias

- We cannot observe them - so we can never be sure!
- Econometrics is all about uncertainty. You can always state that there are different possibilities and you cannot know for sure

Recap: How to think about Selection Bias

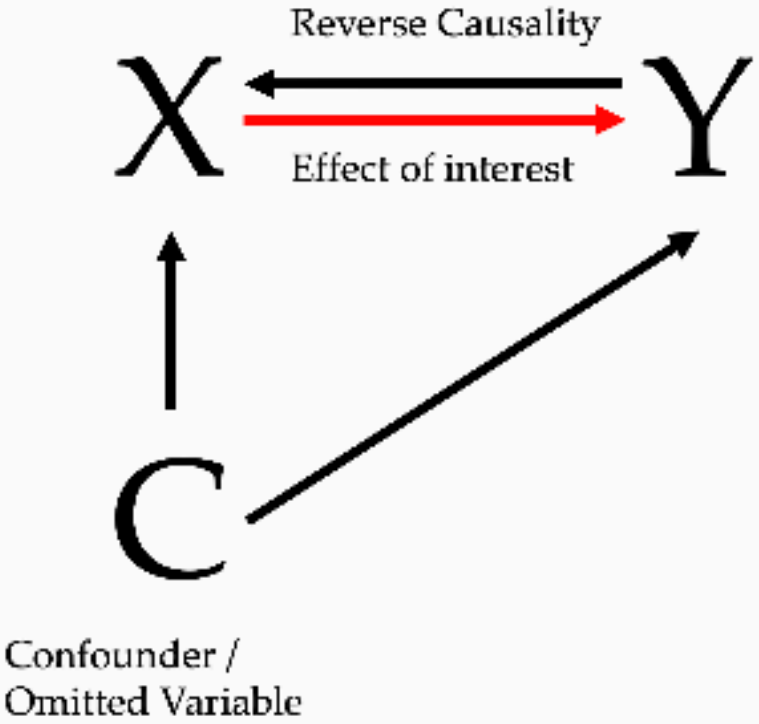


Figure 1: Selection bias

Selection bias in action: Omitted Variable Bias

The most important slide of this course (until the midterm)

Let Y_i be the outcome variable, X_i our regressor of interest, W a series of control variables, and Z_i the "omitted" variable.

$$\text{[Long regression]} \quad Y_i = \alpha_L + \beta_L X_i + \lambda Z_i + W\beta_{WL} + e_i^S$$

$$\text{[Short regression]} \quad Y_i = \alpha_S + \beta_S X_i + W\beta_{WS} + e_i^L$$

$$\text{[Auxiliary regression]} \quad Z_i = \pi_0 + \pi_1 X_i + W\pi_{WP} + v_i$$

Then, the **Omitted variable bias formula** states that:

$$\underbrace{\beta_S}_{\text{Short}} = \underbrace{\beta_L}_{\text{Long}} + \underbrace{\lambda}_{\text{Omitted}} \cdot \underbrace{\pi_1}_{\text{Included}}$$

The OVB formula describes what happens to our coefficient of interest, β, as we include one additional variable Z in the regression. We call $\lambda\pi_1$ the **omitted variable bias**. Direction of bias: multiply our guesses for the signs of λ and π_1.

Control variables

Control variables are additional variables (or covariates) **included in a regression**. We do this for **various reasons** (in decreasing order of importance):

- To remove selection bias / omitted variable bias
- To increase precision of our estimates
- To know about the (conditional/partial) correlation of other variables
- To better predict the outcome

Let's see *graphically* how control variables work!

Your questions

Any questions?

...Remember – this is a safe space! Every question is useful!

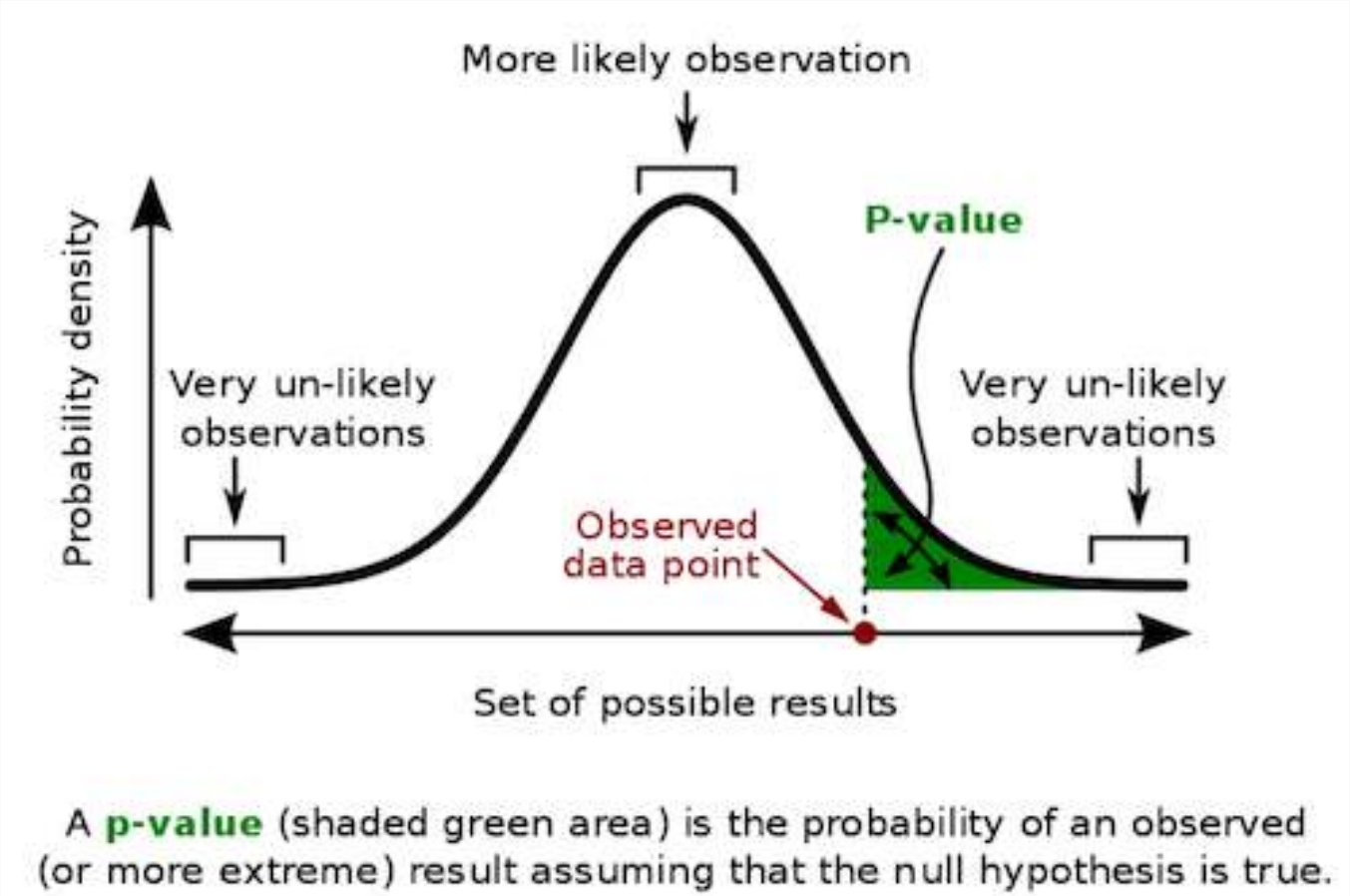
Inference and hypothesis testing

Hypothesis testing

We need a few ingredients:

- Random variables: Our estimator is a random variable (randomly drawn from population)
- Standard error: Random variables have a standard deviation, estimators have standard errors. This quantifies their uncertainty
- Statistics: The two keywords are the law of large numbers and the central limit theorem: The sum/mean of many random variables will follow normal distribution
- For hypothesis test: null and alternative hypothesis.
- We assume that the null hypothesis is true and then see how plausible results are, given the null is true.
- If they are implausible – we reject the null hypothesis! Otherwise: Fail to reject.

It’s all connected



Hypothesis testing mini-cheatsheet

$$\left| \frac{\hat{\beta}}{\text{SE}(\hat{\beta})} \right| \geq 1.96$$

$$\Leftrightarrow \left| \text{t-stat} \right| \geq 1.96$$

$$\Leftrightarrow \text{p-value} \leq 0.05$$

$$\Leftrightarrow 0 \notin \text{CI}$$

If you are testing the null hypothesis H_0: $\beta = 0$ (against the alternative hypothesis H_0: $\beta \neq 0$), then all of these are equivalent, and you can use any of these.

Exam practice

Practice exam question: 1a)

The Ministry of Truth is interested in a rumour that **air pollution could impact mental health**. One of the most harmful pollutants is fine particulate matter PM2.5, which comes from operations that involve the burning of fuels such as wood, oil, coal, etc. A research team is sent to investigate the rumour. The team **randomly** selects and surveys 19,920 people across 71 districts of the country. The key variable, Exposure E_i, is a **dummy variable** equal to 1 if the individual i is exposed to a large amount of PM2.5 in the last 24 hours, and 0 otherwise. The team also conducts a standardised questionnaire to record **depressive symptoms** in the last month, called the Kessler Psychological Distress scale (K6). The questionnaire results in a score, Depression_i, that ranges from 0 to 24; and the higher the score, the more severe the depressive symptoms for individual i. The variable has a sample average of 2.96. **Running regressions with Depression D_i as the dependent variable, you obtain the following results**:

Practice exam question: 1a)

Dependent variable: Depression $_i$

Regressor	(1)	(2)	(3)
Exposure $_i$	0.834	0.614	0.554
	(0.032)	(0.045)	(0.042)
Exposure $_i \times$ Female $_i$		0.065	
		(0.024)	
Female $_i$		−0.739	−0.825
		(0.036)	(0.066)
Age_i			0.452
			(0.132)
Age_i^2			0.524
			(0.121)

Notes: All estimations contain a constant term. Robust standard errors are in the parentheses. Age_i is the age (years old) of individual i, and Age_i^2 is the square of Age_i.

Practice Exam question: 1a)

a) Interpreting the coefficient in Column (1), a journalist, Katherine, claims: ”Since participants are randomly selected, we can infer that exposure to a large amount of PM2.5 does cause depression.”

i. Explain carefully why Katherine is wrong. Come up with two confounding variables, specifying the direction of bias(es) if there are any. Which assumption(s) would she need to impose for the causality claim to hold?

ii. What is the correct interpretation from Column (1) that Katherine should have made?

(Detailed) Suggested Answer: 1a)

i. **Random selection is not the same thing as random assignment to treatment!** Survey respondents may be systematically different from each other in ways that are correlated with depression and pollution exposure. Therefore, the results from a regression can not be interpreted causally (and are biased). A priori, it is unclear in which direction the bias would go, but we could imagine that (**Only one explanation needed for exam**):
On rainy days, pollution is lower (-) and people may be reporting more depression symptoms (+), leading to downward bias. More wealthy people choose to live in less polluted areas (-) and they may have less depression (e.g., better access to mental health resources) (-), leading to upward bias.
For the regression causality claim to hold, we need to assume that: People exposed to pollution and those not exposed to pollution would have, on average, the same depression level, had they been exposed to the same level of pollution. In other words: **Both groups would have to have the same potential depression outcomes**.

ii. On average, people that were exposed to pollution had a 0.8 points higher score on the depression scale. The difference between the two groups is significant at the 5% level.

Pratice Exam question: 1b)

b) Interpret column (2) of the regression table

i. A colleague notes the the coefficient on Female_i is significant, and states: ”The effect of being female on depression is significantly different from zero”. Do you agree with the statement? Why or why not?

ii. How is pollution exposure related to depression, for men? And how for women?

(Detailed) Suggested Answer: 1b)

i. It is difficult to make such interpretations when interaction terms are involved. Taking partial derivatives, the "effect" of being female is:

$$\frac{\partial \text{Depression}_i}{\partial \text{Female}_i} = -0.739 + 0.065 \cdot \text{Exposure}_i$$

We can do inference (and test significance) at $\text{Exposure}_i = 0$ (just looking at the coefficient for female, -0.739 is significant). We can also do it for any other level of exposure, but for that we also need to take the other coefficient into account and cannot just use the table.

Comment 1: Less relevant for the exam, but important to be aware of!
Comment 2: We can always do inference on the interaction term, which is significant here!

ii. Taking partial derivatives, the "effect" of pollution exposure is:

$$\frac{\partial \text{Depression}_i}{\partial \text{Exposure}_i} = 0.614 + 0.065 \cdot \text{Female}_i$$

Hence, the "effect" for Males is 0.614 and the "effect" for Females is larger ($0.614 + 0.065 = 0.779$). The effect of pollution is also **signficantly** larger than for men, because the coefficient on the interaction term is significantly different from zero.

If time permits: Practice exam question, midterm 2022.

Practicing regression

Econ 140 Spring 2025, Section 4

Jonathan Old

Syllabus/OH bCourses Website Feedback form (**Always open**) RStudio

Roadmap

1. Time for your questions
2. Omitted Variable Bias
3. Bad Controls
4. Quadratic Terms
5. Interaction terms
6. Logs
7. Topics we've glossed over so far
8. Exam practice

Your questions

Any questions?

...Remember – this is a safe space! Every question is useful!

Omitted Variable Bias

The most important slide of this course (until the midterm)

Let Y_i be the outcome variable, X_i our regressor of interest, W a series of control variables, and Z_i the "omitted" variable.

$$\text{[Long regression]} \quad Y_i = \alpha_L + \beta_L X_i + \lambda Z_i + W\beta_{WL} + e_i^S$$

$$\text{[Short regression]} \quad Y_i = \alpha_S + \beta_S X_i + W\beta_{WS} + e_i^L$$

$$\text{[Auxiliary regression]} \quad Z_i = \pi_0 + \pi_1 X_i + W\pi_{WP} + v_i$$

Then, the **Omitted variable bias formula** states that:

$$\underbrace{\beta_S}_{\text{Short}} = \underbrace{\beta_L}_{\text{Long}} + \underbrace{\lambda}_{\text{Omitted}} \cdot \underbrace{\pi_1}_{\text{Included}}$$

The OVB formula describes what happens to our coefficient of interest, β, as we include one additional variable Z in the regression. We call $\lambda\pi_1$ the **omitted variable bias**. Direction of bias: multiply our guesses for the signs of λ and π_1.

Bad Controls

Control variables

Control variables are additional variables (or covariates) **included in a regression**. We do this for **various reasons** (in decreasing order of importance):

- To remove selection bias / omitted variable bias
- To increase precision of our estimates
- To know about the (conditional/partial) correlation of other variables
- To better predict the outcome

Bad controls

- Some controls are called ”bad controls”. These are:
  1. Variables that are themselves outcomes of a treatment: What happens if you control for the change in English test scores in the regression below?
  2. Variables that moderate the treatment effect, e.g. controlling for occupation choice in gender wage gap regression . . .
- Rule of Thumb: Good controls are either pre-determined or immutable characteristics.
- Another way to think about it: Controls help us make ”apples to apples” comparisons. We should think before what exactly we want to compare to each other.

Mathematically, good and bad controls are the same thing.

We need to use our  to distinguish them!

Quadratic Terms

Making OLS more interesting

- We can extend the simple OLS framework

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

  to something richer:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + e_i$$

- All questions of the type *"how is Y_i expected to change if we change X_i", keeping all other variables in the regression fixed* can be solved with **partial derivatives** – in this case:

$$\frac{\partial Y_i}{\partial X_i} =$$

Making OLS more interesting

- We can extend the simple OLS framework

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

  to something richer:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + e_i$$

- All questions of the type *"how is Y_i expected to change if we change X_i", keeping all other variables in the regression fixed* can be solved with **partial derivatives** – in this case:

$$\frac{\partial Y_i}{\partial X_i} = \beta_1 + 2 \cdot \beta_2 \cdot X_i$$

Example for Quadratic Terms

Let us see how to use quadratic terms on `Datahub`

Interaction terms

How to think about interaction terms

We will cover this as an exercise. My most important advice to you is that you should use partial derivatives!
The slides after this will be skipped in class, but feel free to use them for your own learning.

Interaction terms: Making OLS more interesting

Let us consider the model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$$

where Y_i is a country's GDP per capita, X_{1i} the value of its natural resources, and X_{2i} a measure of how democratic it is.

1. How do we interpret β_1?

Interaction terms: Making OLS more interesting

Let us consider the model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$$

where Y_i is a country's GDP per capita, X_{1i} the value of its natural resources, and X_{2i} a measure of how democratic it is.

1. How do we interpret β_1?
   Keeping democracy fixed, increasing the value of a country's natural resources by one unit is associated with β_1 higher GDP per capita.
2. How do we interpret β_2?

Interaction terms: Making OLS more interesting

Let us consider the model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$$

where Y_i is a country's GDP per capita, X_{1i} the value of its natural resources, and X_{2i} a measure of how democratic it is.

1. How do we interpret β_1?
   Keeping democracy fixed, increasing the value of a country's natural resources by one unit is associated with β_1 higher GDP per capita.
2. How do we interpret β_2?
   Keeping natural resources fixed, increasing a country's democracy score by one unit is associated with β_2 higher GDP per capita.

Interaction Terms (ii)

Now, let us extend the model to:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + e_i$$

1. What is the ”effect” of X_{1i} on Y_i?

Interaction Terms (ii)

Now, let us extend the model to:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + e_i$$

1. What is the "effect" of X_{1i} on Y_i? $\beta_1 + \beta_3 X_{2i}$

Interaction Terms (ii)

Now, let us extend the model to:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + e_i$$

1. What is the "effect" of X_{1i} on Y_i? $\beta_1 + \beta_3 X_{2i}$
2. How do we interpret β_1?

Interaction Terms (ii)

Now, let us extend the model to:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + e_i$$

1. What is the "effect" of X_{1i} on Y_i? $\beta_1 + \beta_3 X_{2i}$
2. How do we interpret β_1? **The effect of an additional unit of X_{1i} , if X_{2i} is equal to 0.**

Interaction Terms (ii)

Now, let us extend the model to:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + e_i$$

1. What is the "effect" of X_{1i} on Y_i? $\beta_1 + \beta_3 X_{2i}$
2. How do we interpret β_1? **The effect of an additional unit of X_{1i} , if X_{2i} is equal to 0.**
3. How do we interpret β_2?

Interaction Terms (ii)

Now, let us extend the model to:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + e_i$$

1. What is the "effect" of X_{1i} on Y_i? $\beta_1 + \beta_3 X_{2i}$
2. How do we interpret β_1? **The effect of an additional unit of X_{1i} , if X_{2i} is equal to 0.**
3. How do we interpret β_2? **The effect of an additional unit of X_{2i} , if X_{1i} is equal to 0.**

Interaction Terms (ii)

Now, let us extend the model to:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + e_i$$

1. What is the "effect" of X_{1i} on Y_i? $\beta_1 + \beta_3 X_{2i}$
2. How do we interpret β_1? **The effect of an additional unit of X_{1i} , if X_{2i} is equal to 0.**
3. How do we interpret β_2? **The effect of an additional unit of X_{2i} , if X_{1i} is equal to 0.**
4. How do we interpret $\beta_1 + \beta_3$?

Interaction Terms (ii)

Now, let us extend the model to:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + e_i$$

1. What is the "effect" of X_{1i} on Y_i? $\beta_1 + \beta_3 X_{2i}$
2. How do we interpret β_1? **The effect of an additional unit of X_{1i} , if X_{2i} is equal to 0.**
3. How do we interpret β_2? **The effect of an additional unit of X_{2i} , if X_{1i} is equal to 0.**
4. How do we interpret $\beta_1 + \beta_3$? **The effect of an additional unit of X_{1i} , if X_{2i} is equal to 1.**

Rule of thumb: Always use partial derivatives to make sure that you are right!

Interactions with only dummy variables

- Take any two binary variables, for example: Studies at UC Berkeley or not (UCB_i), and is from California or not (Cali_i).
- The regression with interactions looks like this:

$$Y_i = \underbrace{\beta_0}_{17.77} + \underbrace{\beta_1}_{2.28} \text{UCB}_i + \underbrace{\beta_2}_{0.95} \text{Cali}_i + \underbrace{\beta_3}_{-0.97} \text{UCB}_i \times \text{Cali}_i$$

- We can write this in a table, and get all group averages

	Cali = 1	Cali=0
UCB = 1	$\beta_0 + \beta_1 + \beta_2 + \beta_3$ =17.77+2.28+0.95-0.97	$\beta_0 + \beta_1$ = 17.77 + 2.28
UCB = 0	$\beta_0 + \beta_2$ = 17.77 + 0.95	β_0 =17.77

- Take differences between cells to get different effects!

Logs

Notes on logarithms

- We can take logs of whole equations to get linear models (problem set)
- We can also take logs of specific variables, especially when they have long tails (wealth in the US, GDP per capita, etc.)
- We can get to the right interpretation of log-specifications with just math

Notes on logarithms

- We can take logs of whole equations to get linear models (problem set)
- We can also take logs of specific variables, especially when they have long tails (wealth in the US, GDP per capita, etc.)
- We can get to the right interpretation of log-specifications with just math
- But I will make your life easier with a cheat sheet.

Logs: Cheatsheet

Model	LHS	RHS	A change in x by ...	is associated with a change in y by ...
Level-Level	y	x	1 unit	β_1 units
Level-Log	y	$\log(x)$	1%	$\beta_1/100$ units
Log-Level	$\log(y)$	x	1 unit	$100\beta_1$%
Log-Log	$\log(y)$	$\log(x)$	1%	β_1%

If you want to get a bonus star from me, write "approximately" in log-interpretations.

Topics we've glossed over so far

What if the outcome variable is binary (a dummy variable)?

Let's run the regression

$$\text{Defaulted}_i = \alpha + \beta \tilde{\text{Credit Score}}_i + e_i$$

where Defaulted_i is equal to 1 if individual i has ever defaulted on a loan (mortgage, credit card, auto loan, etc.), and $\tilde{\text{Credit Score}}_i$ is i's credit score, **minus the average credit score in the sample** (Note: US credit scores range from 300 to 850 points).

1. You run a regression and get $\hat{\alpha}$=0.1. How do you interpret this? Does this number make sense here?
2. Your estimate for β is $\hat{\beta} = 0.001$. Interpret.

What if the outcome variable is binary (a dummy variable)?

Let's run the regression

$$\text{Defaulted}_i = \alpha + \beta \tilde{\text{Credit Score}}_i + e_i$$

where Defaulted_i is equal to 1 if individual i has ever defaulted on a loan (mortgage, credit card, auto loan, etc.), and $\tilde{\text{Credit Score}}_i$ is i's credit score, **minus the average credit score in the sample** (Note: US credit scores range from 300 to 850 points).

1. You run a regression and get $\hat{\alpha}$=0.1. How do you interpret this? Does this number make sense here?
2. Your estimate for β is $\hat{\beta} = 0.001$. Interpret.

With a dummy dependent variable, changing X_i by one unit increases the probability of $Y_i = 1$ by $\hat{\beta} \cdot 100$ percentage points.

Exam practice

Practice exam question: 1a)

The Ministry of Truth is interested in a rumour that **air pollution could impact mental health**. One of the most harmful pollutants is fine particulate matter PM2.5, which comes from operations that involve the burning of fuels such as wood, oil, coal, etc. A research team is sent to investigate the rumour. The team **randomly** selects and surveys 19,920 people across 71 districts of the country. The key variable, Exposure E_i, is a **dummy variable** equal to 1 if the individual i is exposed to a large amount of PM2.5 in the last 24 hours, and 0 otherwise. The team also conducts a standardised questionnaire to record **depressive symptoms** in the last month, called the Kessler Psychological Distress scale (K6). The questionnaire results in a score, Depression_i, that ranges from 0 to 24; and the higher the score, the more severe the depressive symptoms for individual i. The variable has a sample average of 2.96. **Running regressions with Depression D_i as the dependent variable, you obtain the following results**:

Practice exam question: 1a)

Dependent variable: Depression $_i$

Regressor	(1)	(2)	(3)
Exposure $_i$	0.834	0.614	0.554
	(0.032)	(0.045)	(0.042)
Exposure $_i \times$ Female $_i$		0.065	
		(0.024)	
Female $_i$		−0.739	−0.825
		(0.036)	(0.066)
Age_i			0.452
			(0.132)
Age_i^2			0.524
			(0.121)

Notes: All estimations contain a constant term. Robust standard errors are in the parentheses. Age_i is the age (years old) of individual i, and Age_i^2 is the square of Age_i.

Pratice Exam question: 1a)

a) Interpreting the coefficient in Column (1), a journalist, Katherine, claims: ”Since participants are randomly selected, we can infer that exposure to a large amount of PM2.5 does cause depression.”

i. Explain carefully why Katherine is wrong, specifying the direction of bias(es) if there is any. Which assumption(s) would she need to impose for the causality claim to hold?

ii. What is the correct interpretation from Column (1) that Katherine should have made?

(Detailed) Suggested Answer: 1a)

i. **Random selection is not the same thing as random assignment to treatment!** Survey respondents may be systematically different from each other in ways that are correlated with depression and pollution exposure. Therefore, the results from a regression can not be interpreted causally (and are biased). A priori, it is unclear in which direction the bias would go, but we could imagine that (**Only one explanation needed for exam**):
On rainy days, pollution is lower (-) and people may be reporting more depression symptoms (+), leading to downward bias. More wealthy people choose to live in less polluted areas (-) and they may have less depression (e.g., better access to mental health resources) (-), leading to upward bias.
For the regression causality claim to hold, we need to assume that: People exposed to pollution and those not exposed to pollution would have, on average, the same depression level, had they been exposed to the same level of pollution. In other words: **Both groups would have to have the same potential depression outcomes**.

ii. On average, people that were exposed to pollution had a 0.8 points higher score on the depression scale. The difference between the two groups is significant at the 5% level.

Pratice Exam question: 1b)

b) Interpret column (2) of the regression table

i. A colleague notes the the coefficient on Female_i is significant, and states: ”The effect of being female on depression is significantly different from zero”. Do you agree with the statement? Why or why not?

ii. How is pollution exposure related to depression, for men? And how for women?

(Detailed) Suggested Answer: 1b)

i. It is difficult to make such interpretations when interaction terms are involved. Taking partial derivatives, the "effect" of being female is:

$$\frac{\partial \text{Depression}_i}{\partial \text{Female}_i} = -0.739 + 0.065 \cdot \text{Exposure}_i$$

We can do inference (and test significance) at $\text{Exposure}_i = 0$ (just looking at the coefficient for female, -0.739 is significant). We can also do it for any other level of exposure, but for that we also need to take the other coefficient into account and cannot just use the table.

Comment 1: Less relevant for the exam, but important to be aware of!
Comment 2: We can always do inference on the interaction term, which is significant here!

ii. Taking partial derivatives, the "effect" of pollution exposure is:

$$\frac{\partial \text{Depression}_i}{\partial \text{Exposure}_i} = 0.614 + 0.065 \cdot \text{Female}_i$$

Hence, the "effect" for Males is 0.614 and the "effect" for Females is larger ($0.614 + 0.065 = 0.779$). The effect of pollution is also **signficantly** larger than for men, because the coefficient on the interaction term is significantly different from zero.

If time permits: Practice exam question, midterm 2022.

PGetting fit for the midterm

Econ 140 Spring 2025, Section 5

Jonathan Old

Syllabus/OH bCourses Website Feedback form (**Always open**) RStudio

Roadmap

1. Time for your questions
2. Quadratic Terms
3. Interaction terms
4. Exam practice
5. Logs
6. Topics we've glossed over so far
7. Time for your questions

Your questions

Any questions?

...Remember – this is a safe space! Every question is useful!

Quadratic Terms

Making OLS more interesting

- We can extend the simple OLS framework

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

  to something richer:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + e_i$$

- All questions of the type *"how is Y_i expected to change if we change X_i", keeping all other variables in the regression fixed* can be solved with **partial derivatives** – in this case:

$$\frac{\partial Y_i}{\partial X_i} =$$

Making OLS more interesting

- We can extend the simple OLS framework

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

  to something richer:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + e_i$$

- All questions of the type *"how is Y_i expected to change if we change X_i", keeping all other variables in the regression fixed* can be solved with **partial derivatives** – in this case:

$$\frac{\partial Y_i}{\partial X_i} = \beta_1 + 2 \cdot \beta_2 \cdot X_i$$

Interaction terms

Interaction Terms (ii)

Now, let us extend the model to:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + e_i$$

1. What is the ”effect” of X_{1i} on Y_i?

Interaction Terms (ii)

Now, let us extend the model to:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + e_i$$

1. What is the "effect" of X_{1i} on Y_i? $\beta_1 + \beta_3 X_{2i}$

Interaction Terms (ii)

Now, let us extend the model to:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + e_i$$

1. What is the "effect" of X_{1i} on Y_i? $\beta_1 + \beta_3 X_{2i}$
2. How do we interpret β_1?

Interaction Terms (ii)

Now, let us extend the model to:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + e_i$$

1. What is the "effect" of X_{1i} on Y_i? $\beta_1 + \beta_3 X_{2i}$
2. How do we interpret β_1? **The effect of an additional unit of X_{1i} , if X_{2i} is equal to 0.**

Interaction Terms (ii)

Now, let us extend the model to:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + e_i$$

1. What is the "effect" of X_{1i} on Y_i? $\beta_1 + \beta_3 X_{2i}$
2. How do we interpret β_1? **The effect of an additional unit of X_{1i} , if X_{2i} is equal to 0.**
3. How do we interpret β_2?

Interaction Terms (ii)

Now, let us extend the model to:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + e_i$$

1. What is the "effect" of X_{1i} on Y_i? $\beta_1 + \beta_3 X_{2i}$
2. How do we interpret β_1? **The effect of an additional unit of X_{1i} , if X_{2i} is equal to 0.**
3. How do we interpret β_2? **The effect of an additional unit of X_{2i} , if X_{1i} is equal to 0.**

Interaction Terms (ii)

Now, let us extend the model to:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + e_i$$

1. What is the "effect" of X_{1i} on Y_i? $\beta_1 + \beta_3 X_{2i}$
2. How do we interpret β_1? **The effect of an additional unit of X_{1i} , if X_{2i} is equal to 0.**
3. How do we interpret β_2? **The effect of an additional unit of X_{2i} , if X_{1i} is equal to 0.**
4. How do we interpret $\beta_1 + \beta_3$?

Interaction Terms (ii)

Now, let us extend the model to:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + e_i$$

1. What is the "effect" of X_{1i} on Y_i? $\beta_1 + \beta_3 X_{2i}$
2. How do we interpret β_1? **The effect of an additional unit of X_{1i} , if X_{2i} is equal to 0.**
3. How do we interpret β_2? **The effect of an additional unit of X_{2i} , if X_{1i} is equal to 0.**
4. How do we interpret $\beta_1 + \beta_3$? **The effect of an additional unit of X_{1i} , if X_{2i} is equal to 1.**

Rule of thumb: Always use partial derivatives to make sure that you are right!

Exam practice

Practice exam question: 1a)

The Ministry of Truth is interested in a rumour that **air pollution could impact mental health**. One of the most harmful pollutants is fine particulate matter PM2.5, which comes from operations that involve the burning of fuels such as wood, oil, coal, etc. A research team is sent to investigate the rumour. The team **randomly** selects and surveys 19,920 people across 71 districts of the country. The key variable, Exposure E_i, is a **dummy variable** equal to 1 if the individual i is exposed to a large amount of PM2.5 in the last 24 hours, and 0 otherwise. The team also conducts a standardised questionnaire to record **depressive symptoms** in the last month, called the Kessler Psychological Distress scale (K6). The questionnaire results in a score, Depression_i, that ranges from 0 to 24; and the higher the score, the more severe the depressive symptoms for individual i. The variable has a sample average of 2.96. **Running regressions with Depression D_i as the dependent variable, you obtain the following results**:

Practice exam question: 1a)

Dependent variable: Depression $_i$

Regressor	(1)	(2)	(3)
Exposure $_i$	0.834	0.614	0.554
	(0.032)	(0.045)	(0.042)
Exposure $_i \times$ Female $_i$		0.065	
		(0.024)	
Female $_i$		−0.739	−0.825
		(0.036)	(0.066)
Age_i			0.452
			(0.132)
Age_i^2			0.524
			(0.121)

Notes: All estimations contain a constant term. Robust standard errors are in the parentheses. Age_i is the age (years old) of individual i, and Age_i^2 is the square of Age_i.

Pratice Exam question: 1a)

a) Interpreting the coefficient in Column (1), a journalist, Katherine, claims: ”Since participants are randomly selected, we can infer that exposure to a large amount of PM2.5 does cause depression.”

i. Explain carefully why Katherine is wrong, specifying the direction of bias(es) if there is any. Which assumption(s) would she need to impose for the causality claim to hold?

ii. What is the correct interpretation from Column (1) that Katherine should have made?

(Detailed) Suggested Answer: 1a)

i. **Random selection is not the same thing as random assignment to treatment!** Survey respondents may be systematically different from each other in ways that are correlated with depression and pollution exposure. Therefore, the results from a regression can not be interpreted causally (and are biased). A priori, it is unclear in which direction the bias would go, but we could imagine that (**Only one explanation needed for exam**):
On rainy days, pollution is lower (-) and people may be reporting more depression symptoms (+), leading to downward bias. More wealthy people choose to live in less polluted areas (-) and they may have less depression (e.g., better access to mental health resources) (-), leading to upward bias.
For the regression causality claim to hold, we need to assume that: People exposed to pollution and those not exposed to pollution would have, on average, the same depression level, had they been exposed to the same level of pollution. In other words: **Both groups would have to have the same potential depression outcomes**.

ii. On average, people that were exposed to pollution had a 0.8 points higher score on the depression scale. The difference between the two groups is significant at the 5% level.

Pratice Exam question: 1b)

b) Interpret column (2) of the regression table

i. A colleague notes the the coefficient on Female_i is significant, and states: ”The effect of being female on depression is significantly different from zero”. Do you agree with the statement? Why or why not?

ii. How is pollution exposure related to depression, for men? And how for women?

(Detailed) Suggested Answer: 1b)

i. It is difficult to make such interpretations when interaction terms are involved. Taking partial derivatives, the "effect" of being female is:

$$\frac{\partial \text{Depression}_i}{\partial \text{Female}_i} = -0.739 + 0.065 \cdot \text{Exposure}_i$$

We can do inference (and test significance) at $\text{Exposure}_i = 0$ (just looking at the coefficient for female, -0.739 is significant). We can also do it for any other level of exposure, but for that we also need to take the other coefficient into account and cannot just use the table.

Comment 1: Less relevant for the exam, but important to be aware of!
Comment 2: We can always do inference on the interaction term, which is significant here!

ii. Taking partial derivatives, the "effect" of pollution exposure is:

$$\frac{\partial \text{Depression}_i}{\partial \text{Exposure}_i} = 0.614 + 0.065 \cdot \text{Female}_i$$

Hence, the "effect" for Males is 0.614 and the "effect" for Females is larger ($0.614 + 0.065 = 0.779$). The effect of pollution is also **signficantly** larger than for men, because the coefficient on the interaction term is significantly different from zero.

Logs

Logs: Cheatsheet

Model	LHS	RHS	A change in x by ...	is associated with a change in y by ...
Level-Level	y	x	1 unit	β_1 units
Level-Log	y	$\log(x)$	1%	$\beta_1/100$ units
Log-Level	$\log(y)$	x	1 unit	$100\beta_1$%
Log-Log	$\log(y)$	$\log(x)$	1%	β_1%

If you want to get a bonus star from me, write "approximately" in log-interpretations.

Topics we've glossed over so far

Your questions

Any questions?

...Remember – this is a safe space! Every question is useful!

What if the outcome variable is binary (a dummy variable)?

Let's run the regression

$$\text{Defaulted}_i = \alpha + \beta \tilde{\text{Credit Score}}_i + e_i$$

where Defaulted_i is equal to 1 if individual i has ever defaulted on a loan (mortgage, credit card, auto loan, etc.), and $\tilde{\text{Credit Score}}_i$ is i's credit score, **minus the average credit score in the sample** (Note: US credit scores range from 300 to 850 points).

1. You run a regression and get $\hat{\alpha}$=0.1. How do you interpret this? Does this number make sense here?
2. Your estimate for β is $\hat{\beta} = 0.001$. Interpret.

What if the outcome variable is binary (a dummy variable)?

Let's run the regression

$$\text{Defaulted}_i = \alpha + \beta \widetilde{\text{Credit Score}}_i + e_i$$

where Defaulted_i is equal to 1 if individual i has ever defaulted on a loan (mortgage, credit card, auto loan, etc.), and $\widetilde{\text{Credit Score}}_i$ is i's credit score, **minus the average credit score in the sample** (Note: US credit scores range from 300 to 850 points).

1. You run a regression and get $\hat{\alpha}$=0.1. How do you interpret this? Does this number make sense here?
2. Your estimate for β is $\hat{\beta} = 0.001$. Interpret.

With a dummy dependent variable, changing X_i by one unit increases the probability of $Y_i = 1$ by $\hat{\beta} \cdot 100$ percentage points.

Hypothesis testing mini-cheatsheet

$$\left| \frac{\hat{\beta}}{\text{SE}(\hat{\beta})} \right| \geq 1.96$$

$$\Leftrightarrow \left| \text{t-stat} \right| \geq 1.96$$

$$\Leftrightarrow \text{p-value} \leq 0.05$$

$$\Leftrightarrow 0 \notin \text{CI}$$

If you are testing the null hypothesis H_0: $\beta = 0$ (against the alternative hypothesis H_0: $\beta \neq 0$), then all of these are equivalent, and you can use any of these.

If time permits: Practice exam question, midterm 2024.

Instrumental Variables

Econ 140 Spring 2025, Section 7

Jonathan Old

Syllabus/OH bCourses Website Feedback form (**Always open**) RStudio

Roadmap

1. Time for your questions
2. (Group) data projects
3. Introduction to Instrumental Variables
4. IV Conditions
5. IV Summary
6. Time for your questions
7. Group work
8. Time for your questions

Your questions

Any questions?

...Remember – this is a safe space! Every question is useful!

(Group) data projects

Your time for questions

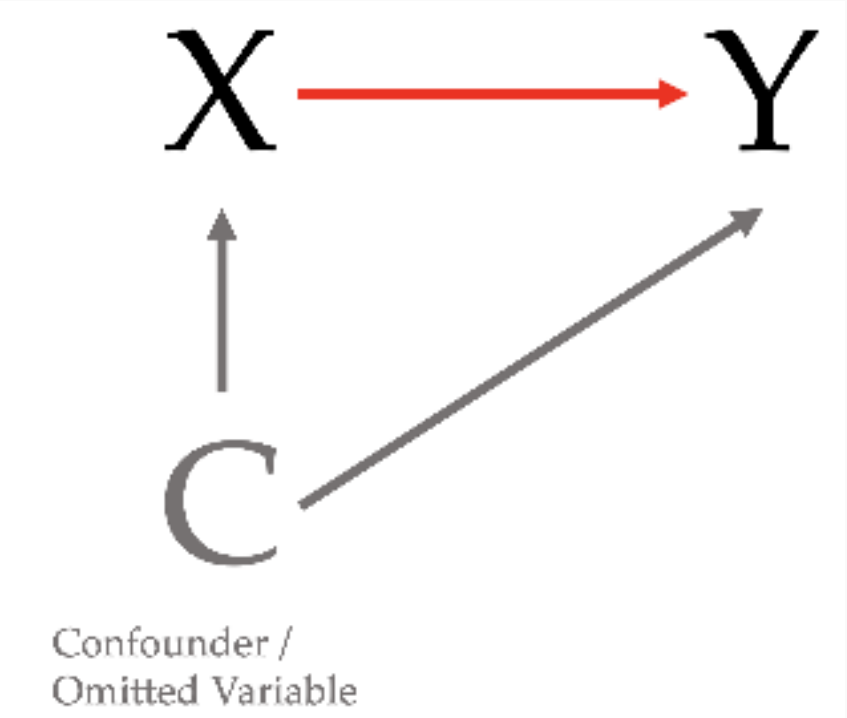
I prepared a document that can help you find data if you are lost. See it on bCourses.

The first deadline is on April 5.

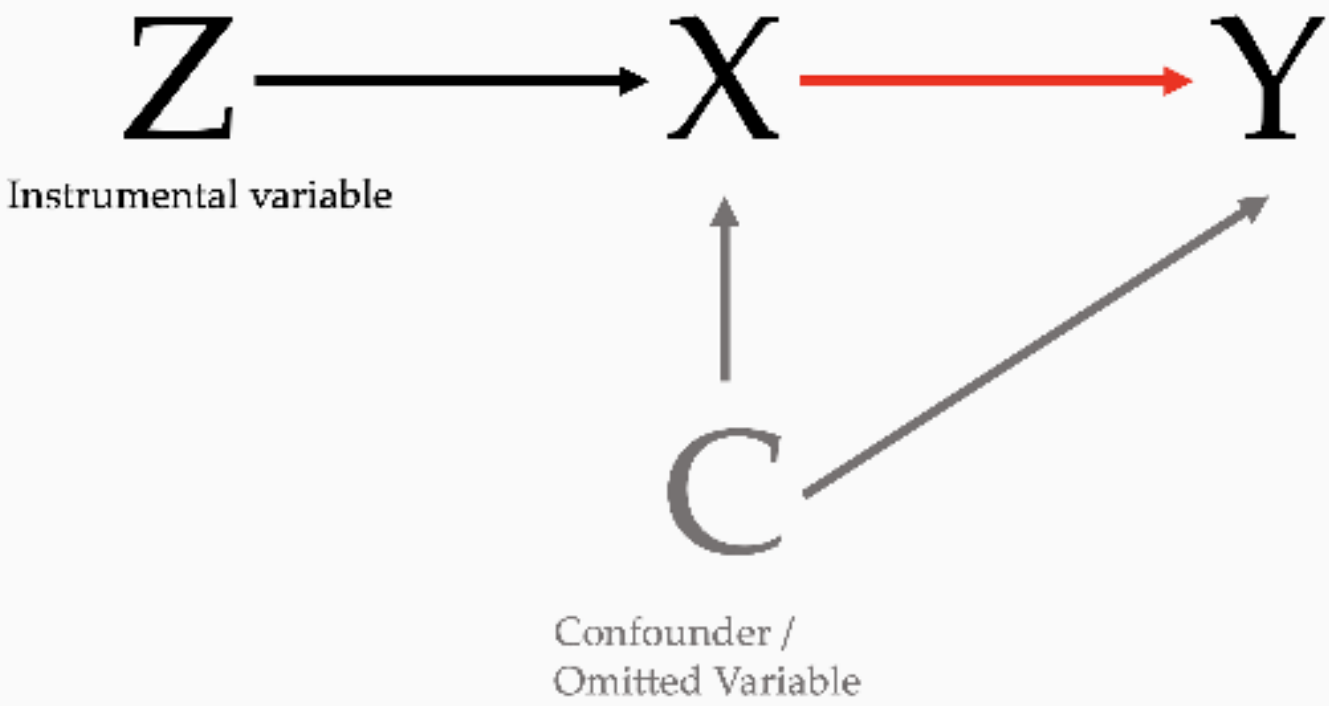
Group work encouraged!

Introduction to Instrumental Variables

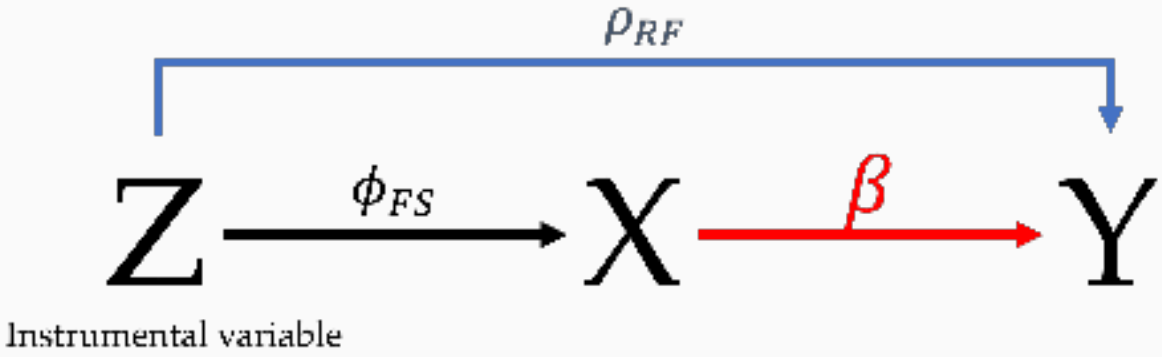
Recap: Omitted Variable Bias



Instrumental variables: The setup



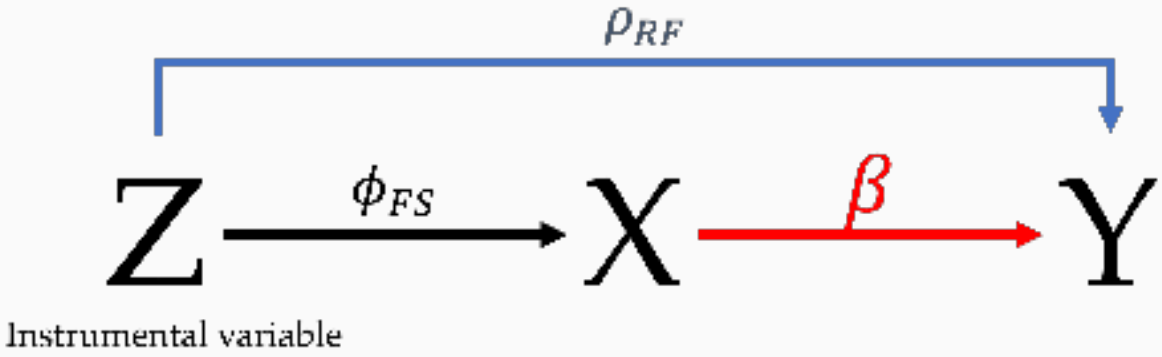
Recap: IV ”rescales” the effect



A simple example:

- We want to know the effect of chocolate (X) on happiness (Y), using a randomized voucher as instrument (Z).
- We find: people with voucher were 3 points more happy ($\rho = 3$), and ate 0.5 more chocolates ($\phi = 0.5$).
- Then, the effect of eating one more chocolate is:

Recap: IV ”rescales” the effect



A simple example:

- We want to know the effect of chocolate (X) on happiness (Y), using a randomized voucher as instrument (Z).
- We find: people with voucher were 3 points more happy ($\rho = 3$), and ate 0.5 more chocolates ($\phi = 0.5$).
- Then, the effect of eating one more chocolate is: $\beta = \rho/\phi = 3/0.5 = 6$.

Calculating the IV coefficient

What is the effect of **eating chocolate** (D) on happiness (Y).

- Why not estimate: $Y_i = \alpha + \beta D_i + \varepsilon_i$?

Randomly give voucher to buy chocolate at 90% discount (Z).

- Why not estimate: $Y_i = \alpha + \beta Z_i + \varepsilon_i$?

Let us set up some regressions:

$$\text{Regression of interest: } Y_i = \alpha + \beta D_i + e_i$$

$$\text{First stage: } D_i = \alpha_1 + \phi_{FS} Z_i + u_i$$

$$\text{Reduced Form: } Y_i = \alpha_2 + \rho_{RF} Z_i + v_i$$

$$\text{Plug in regression of interest: } Y_i = \alpha + \beta(\alpha_1 + \phi_{FS} \cdot Z_i + u_i) + e_i$$

$$\text{Get back reduced form: } = \underbrace{(\alpha + \beta\alpha_1)}_{\alpha_2} + \underbrace{(\beta\gamma)}_{\rho_{RF}} Z_i + \underbrace{(\beta u_i + e_i)}_{v_i}$$

So we see that $\rho_{RF} = \beta\phi_{FS} \Leftrightarrow \beta = \rho_{RF}/\phi_{FS}$

Interpretation of the IV coefficient

- How do we interpret ϕ_{FS} ?

Interpretation of the IV coefficient

- How do we interpret ϕ_{FS} ? The average difference in chocolate consumption between those who got a voucher and those who didn't

Interpretation of the IV coefficient

- How do we interpret ϕ_{FS} ? The average difference in chocolate consumption between those who got a voucher and those who didn't
- How do we interpret ρ_{RF} ?

Interpretation of the IV coefficient

- How do we interpret ϕ_{FS} ? The average difference in chocolate consumption between those who got a voucher and those who didn't
- How do we interpret ρ_{RF} ? The average difference in happiness between those who got a voucher and those who didn't

Interpretation of the IV coefficient

- How do we interpret ϕ_{FS} ? The average difference in chocolate consumption between those who got a voucher and those who didn't
- How do we interpret ρ_{RF} ? The average difference in happiness between those who got a voucher and those who didn't

$$\beta = \frac{\rho_{RF}}{\phi_{FS}} = \frac{E\left[Y_i \mid Z_i = 1\right] - E\left[Y_i \mid Z_i = 0\right]}{E\left[D_i \mid Z_i = 1\right] - E\left[D_i \mid Z_i = 0\right]}$$

IV gives us the treatment effect for the compliers

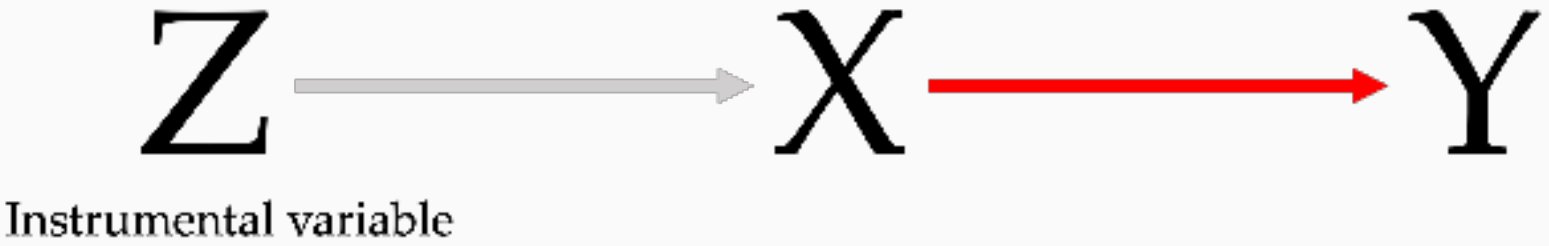
Potential outcomes! (unobserved)		***Does not get voucher (Z=0)***	
Gets voucher (Z=1)		*Eats chocolate (D=1)*	*Does not eat chocolate (D=0)*
	Eats chocolate (D=1)	Always-takers: E(D|Z=1)=E(D|Z=0)=1 ➔ E(Y|Z=1)=E(Y|Z=0)	Compliers
	Does not eat chocolate (D=0)	Defiers	Never-takers: E(D|Z=1)=E(D|Z=0)=0 ➔ E(Y|Z=1)=E(Y|Z=0)

IV Conditions

Does IV always work?

- No! It only works if we have a valid instrument
- For this, we need three conditions:

1. Relevance: *Z* must truly affect *X*



Does IV always work?

- No! It only works if we have a valid instrument
- For this, we need three conditions:

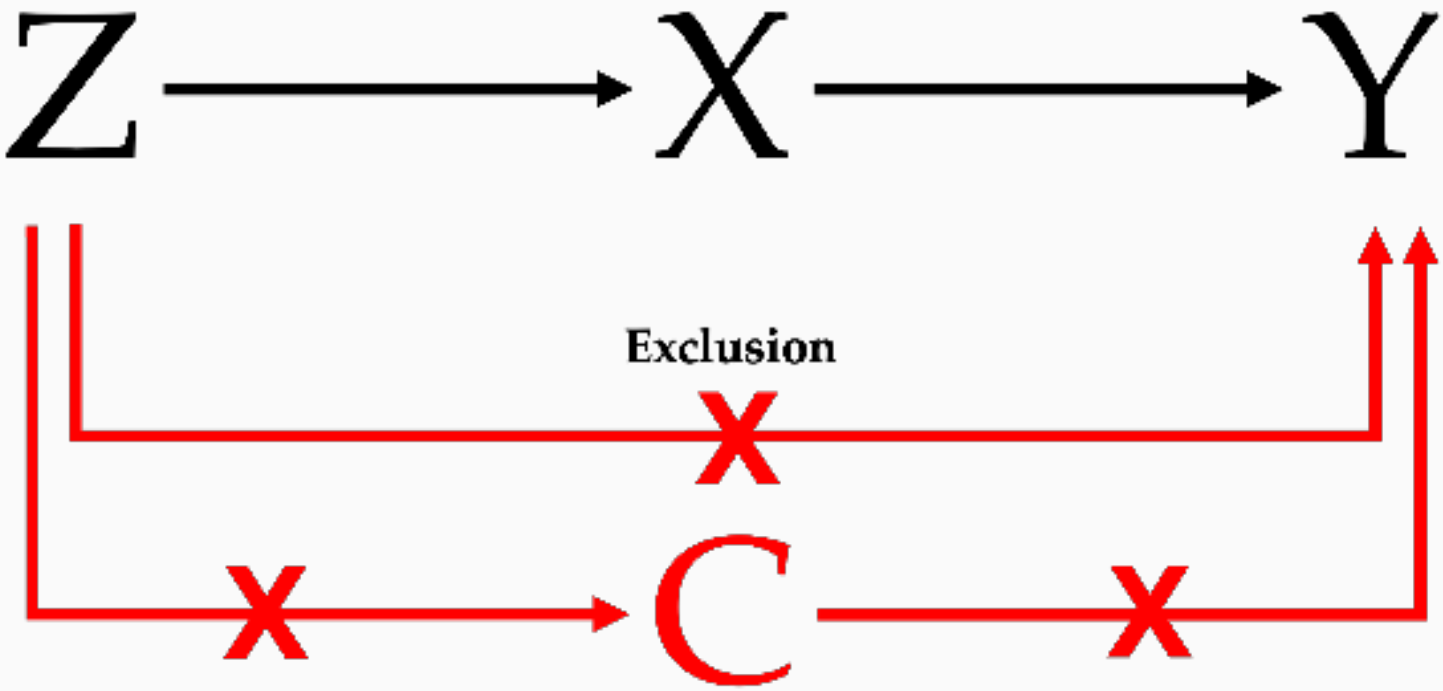
2. Independence/Exogeneity: Z is as good as randomly assigned



Does IV always work?

- No! It only works if we have a valid instrument
- For this, we need three conditions:

3. Exclusion: The **ONLY** way that *Z* affects *Y* is via *X*!

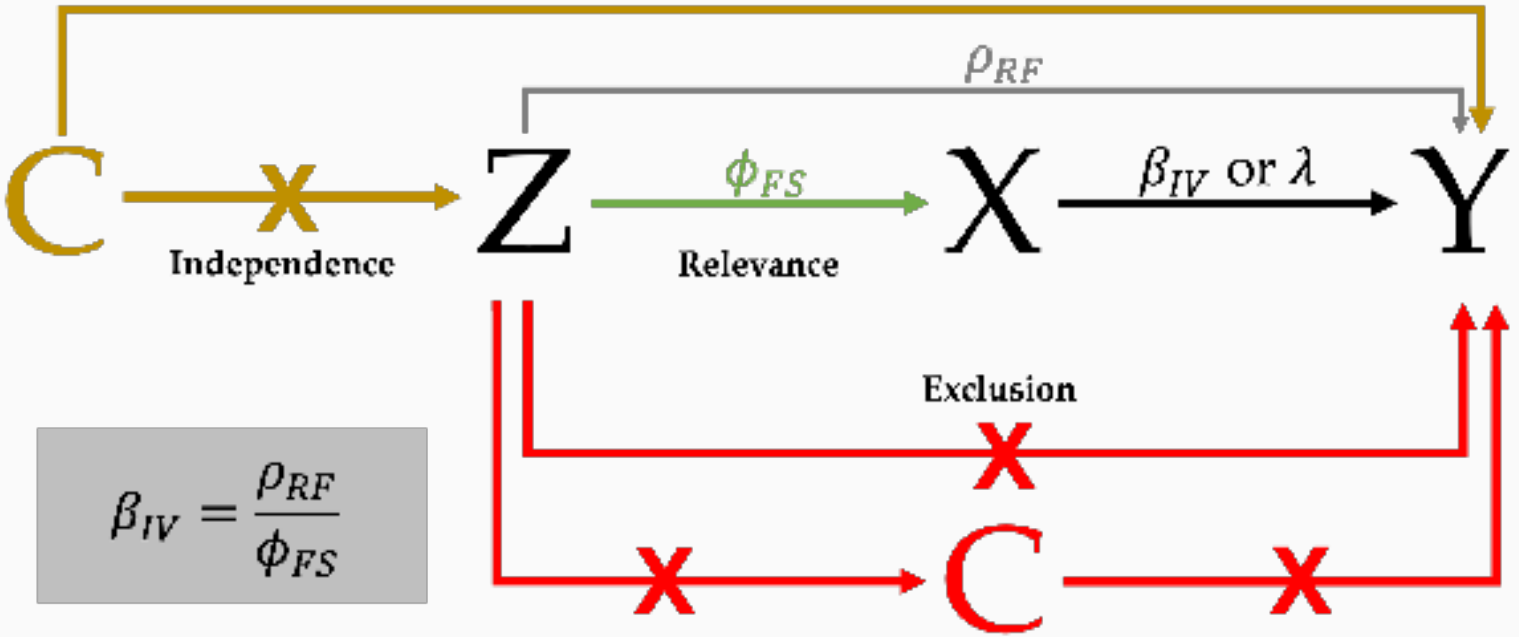


IV Summary

IV summary

We need the following three assumptions for IV to work:

1. **Relevance**: *Z* must truly affect *X*
2. **Independence/Exogeneity**: *Z* is as good as randomly assigned
3. **Exclusion Restriction**: The **only** way that *Z* affects *Y* is via *X*.



$$\beta_{IV} = \frac{\rho_{RF}}{\phi_{FS}}$$

How IV estimates can be different from OLS estimates

- **OVB:** If OLS had omitted variable bias, and our instrument is valid, then the IV estimate should be different. We can check whether this is plausible with the OVB formula
- **Measurement error:** If we have random measurement error in the independent variable (X), then we can use IV to overcome this. In that case, the IV coefficient will be larger (in absolute value) than the OLS coefficient
- **LATE:** OLS gives us ATT + Selection Bias, while IV gives us the treatment effect on the compliers (LATE). The ATT may be different from LATE, even without selection bias.
- **Invalid IV:** Hard to determine what exactly is going on
- **Sampling variation**: This can just happen by chance

IV and controls

- There is a catch in IV: independence (exogeneity) and exclusion restriction only need to hold **conditionally**
- This means: If we **know** C, we can control for it in the IV regression, and the concerns go away
- But then, we have the old OLS problems again (we maybe need to control for lots of things)
- So we should only do this in cases where we really know all the C's.

Your questions

Any questions?

...Remember – this is a safe space! Every question is useful!

Group work

Group work

Group 1: We are interested in the effect of being in the army on crime. We instrument being in the army with a lottery (paper)

Group 2: We are interested in the effect of income on conflict. We instrument income with rainfall (paper)

Group 3: We are interested in the effect of air pollution on mortality. We instrument local air pollution with wind direction (paper)

1. Relevance: Z must truly affect X
2. Independence/Exogeneity: Z is as good as randomly assigned
3. Exclusion restriction: The **only** way that Z affects Y is via X

Your job: Discuss whether these assumptions hold!

Your questions

Any questions?

...Remember – this is a safe space! Every question is useful!

Instrumental Variables: Practice

Econ 140 Spring 2025, Section 8

Jonathan Old

Syllabus/OH bCourses Website Feedback form (**Always open**) RStudio

Roadmap

1. Time for your questions
2. (Group) data projects
3. Introduction to Instrumental Variables
4. IV Summary
5. Time for your questions
6. Rainfall IV
7. Group work
8. Time for your questions

Your questions

Any questions?

...Remember – this is a safe space! Every question is useful!

(Group) data projects

Your time for questions

I prepared a document that can help you find data if you are lost. See it on bCourses.

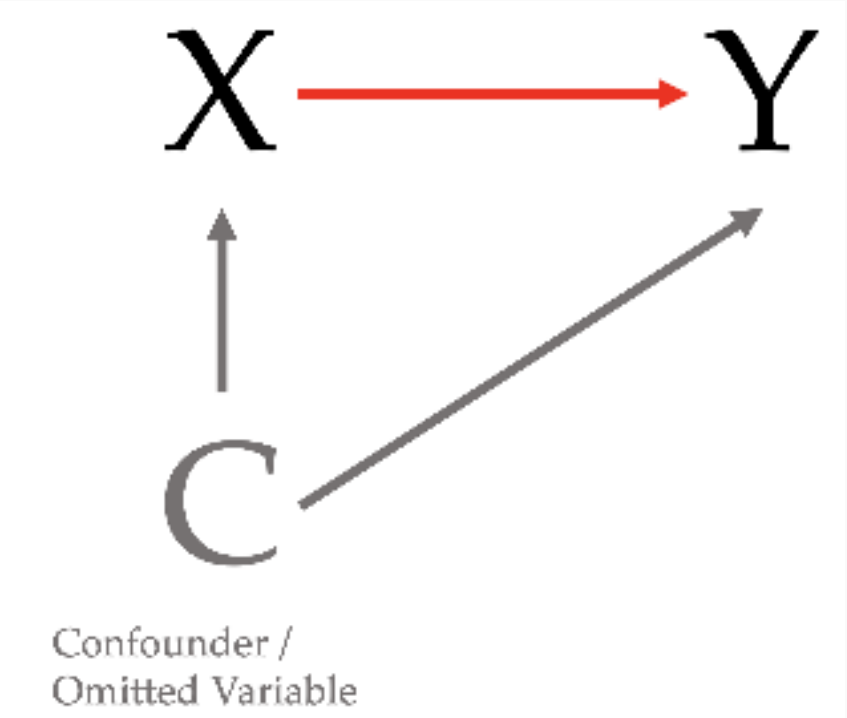
The first deadline is on April 5 (day after tomorrow).

Group work encouraged!

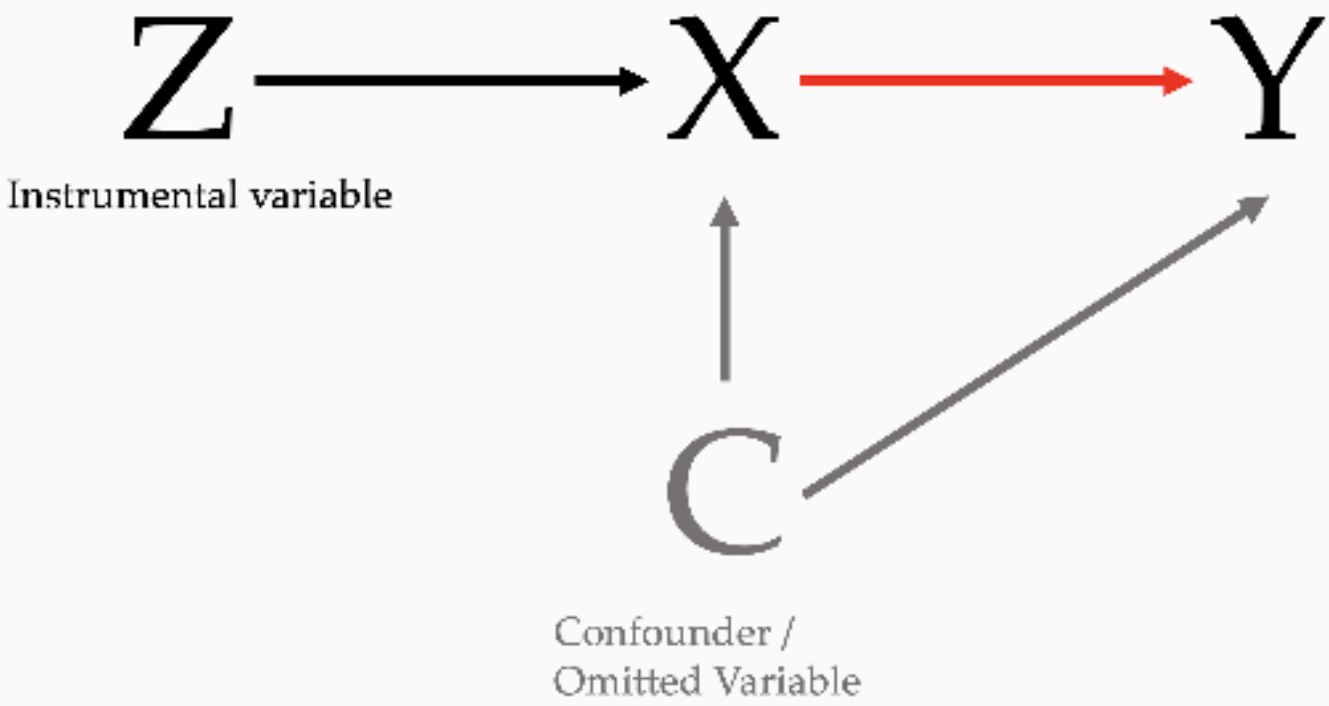
Most important tip: All good questions in this course are of the form: What is the effect of X on Y?

Introduction to Instrumental Variables

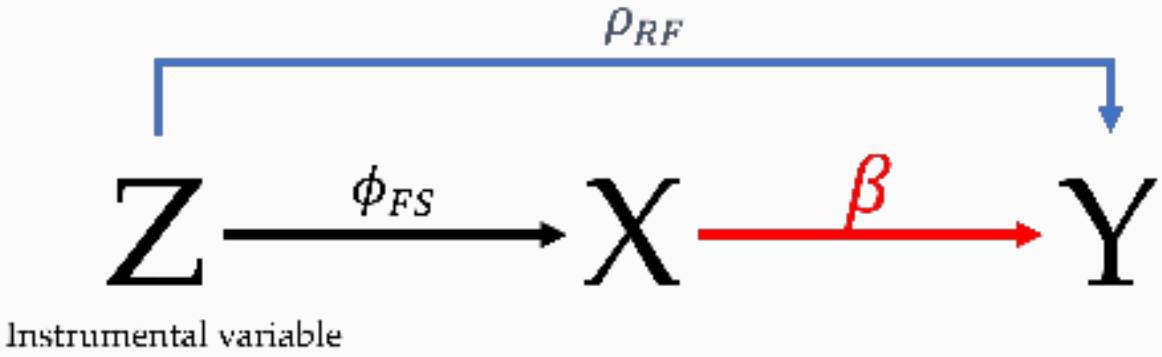
Recap: Omitted Variable Bias



Instrumental variables: The setup



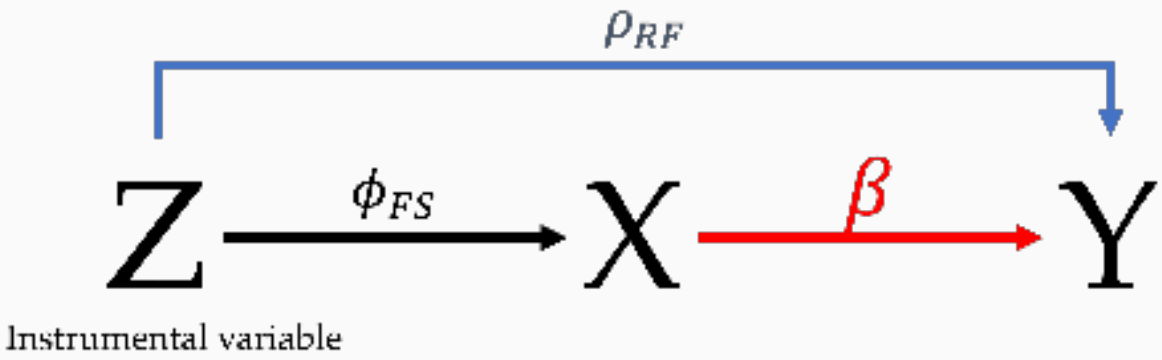
Recap: IV ”rescales” the effect



A simple example:

- We want to know the effect of chocolate (X) on happiness (Y), using a randomized voucher as instrument (Z).
- We find: people with voucher were 3 points more happy ($\rho = 3$), and ate 0.5 more chocolates ($\phi = 0.5$).
- Then, the effect of eating one more chocolate is:

Recap: IV ”rescales” the effect



A simple example:

- We want to know the effect of chocolate (X) on happiness (Y), using a randomized voucher as instrument (Z).
- We find: people with voucher were 3 points more happy ($\rho = 3$), and ate 0.5 more chocolates ($\phi = 0.5$).
- Then, the effect of eating one more chocolate is: $\beta = \rho/\phi = 3/0.5 = 6$.

IV in R

Let's see how IV and 2SLS works on `Datahub`

Calculating the IV coefficient

What is the effect of **eating chocolate** (D) on happiness (Y).

- Why not estimate: $Y_i = \alpha + \beta D_i + \varepsilon_i$?

Randomly give voucher to buy chocolate at 90% discount (Z).

- Why not estimate: $Y_i = \alpha + \beta Z_i + \varepsilon_i$?

Let us set up some regressions:

$$\text{Regression of interest: } Y_i = \alpha + \beta D_i + e_i$$

$$\text{First stage: } D_i = \alpha_1 + \gamma Z_i + u_i$$

$$\text{Reduced Form: } Y_i = \alpha_2 + \delta Z_i + v_i$$

$$\text{Plug in regression of interest: } Y_i = \alpha + \beta(\alpha_1 + \gamma \cdot Z_i + u_i) + e_i$$

$$\text{Get back reduced form: } = \underbrace{(\alpha + \beta\alpha_1)}_{\alpha_2} + \underbrace{(\beta\gamma)}_{\delta} Z_i + \underbrace{(\beta u_i + e_i)}_{v_i}$$

$$\text{So we see that } \quad \delta = \beta\gamma \Leftrightarrow \beta = \delta/\gamma$$

IV gives us the treatment effect for the compliers

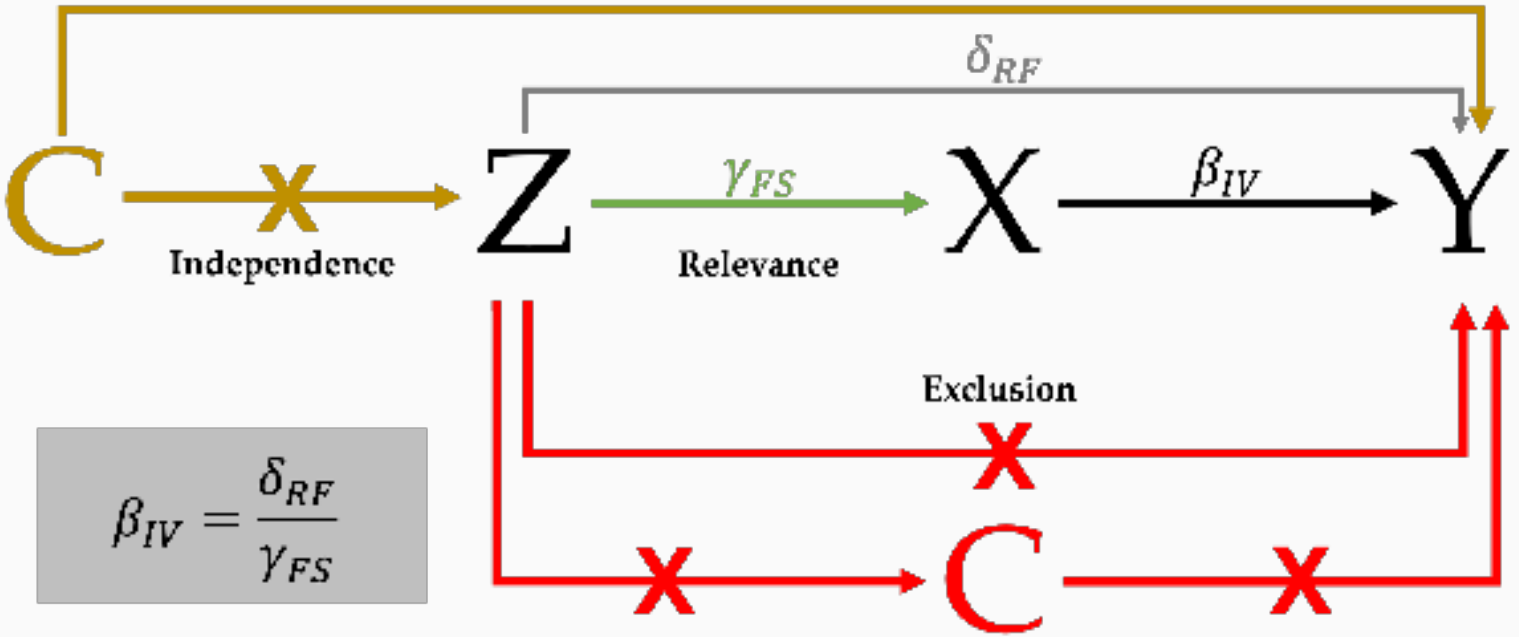
Potential outcomes! (unobserved)		***Does not get voucher (Z=0)***	
Gets voucher (Z=1)		*Eats chocolate (D=1)*	*Does not eat chocolate (D=0)*
	Eats chocolate (D=1)	Always-takers: E(D|Z=1)=E(D|Z=0)=1 ➔ E(Y|Z=1)=E(Y|Z=0)	Compliers
	Does not eat chocolate (D=0)	Defiers	Never-takers: E(D|Z=1)=E(D|Z=0)=0 ➔ E(Y|Z=1)=E(Y|Z=0)

IV Summary

IV summary

We need the following three assumptions for IV to work:

1. **Relevance**: *Z* must truly affect *X*
2. **Independence/Exogeneity**: *Z* is as good as randomly assigned
3. **Exclusion Restriction**: The **only** way that *Z* affects *Y* is via *X*.



$$\beta_{IV} = \frac{\delta_{RF}}{\gamma_{FS}}$$

Your questions

Any questions?

...Remember – this is a safe space! Every question is useful!

Rainfall IV

Intro

NEXT ARTICLE >

Economic Shocks and Civil Conflict: An Instrumental Variables Approach

Edward Miguel, Shanker Satyanath, and Ernest Sergenti

University of California, Berkeley and National Bureau of Economic ResearchNew York University

PDF | PDF PLUS | Abstract | Full Text

Abstract

Estimating the impact of economic conditions on the likelihood of civil conflict is difficult because of endogeneity and omitted variable bias. We use rainfall variation as an instrumental variable for economic growth in 41 African countries during 1981–99. Growth is strongly negatively related to civil conflict: a negative growth shock of five percentage points increases the likelihood of conflict by one-half the following year. We attempt to rule out other channels through which rainfall may affect conflict. Surprisingly, the impact of growth shocks on conflict is *not* significantly different in richer, more democratic, or more ethnically diverse countries.

I. Introduction

Details | Figures | References | Cited by



Journal of Political Economy
Volume 112, Number 4
August 2004

Article DOI
https://doi-org.libproxy.berkeley.edu/10.1086/421174

Why IV?

There is a growing body of research that highlights the association between economic conditions and civil conflict (see Sambanis [2001] for a review). However, the existing literature does not adequately address the endogeneity of economic variables to civil war and thus does not convincingly establish a causal relationship. In addition to endogeneity, omitted variables—for example, government institutional quality—may drive both economic outcomes and conflict, producing misleading cross-country estimates.

Relevance

In this paper we use exogenous variation in rainfall as an instrumental variable for income growth in order to estimate the impact of economic growth on civil conflict.[1] Weather shocks are plausible instruments for growth in gross domestic product in economies that largely rely on rain-fed agriculture, that is, neither have extensive irrigation systems nor are heavily industrialized. The instrumental variable method makes it credible to assert that the association between economic conditions and civil war is a causal relationship rather than simply a correlation. As such,

Sub-Saharan Africa is the ideal region for this identification strategy: the World Development Indicator database indicates that only 1 percent of cropland is irrigated in the median African country, and the agricultural sector remains large. We find that weather shocks are in fact closely related to income growth in sub-Saharan Africa (in the first-stage regression). However, our identification strategy is inappropriate for other regions of the world, since weather is not sufficiently closely linked to income growth.[2] Although the analysis is not global, it is likely to be of exceptional interest from both the research and policy perspectives, since the incidence of civil wars in Africa is high and has increased in the past two decades.

Exclusion

following year. We attempt to rule out other channels through which rainfall may affect conflict. Surprisingly, the impact of growth shocks on conflict is *not* significantly different in richer, more democratic, or more ethnically diverse countries.

Exclusion

While it is intuitively plausible that the rainfall instruments are exogenous, they must also satisfy the exclusion restriction: weather shocks should affect civil conflict only through economic growth. In the Introduction above, we acknowledge the possibility that economic channels other than per capita economic growth per se (i.e., income inequality or rural poverty rates) may be key underlying causes of civil conflict in the aftermath of adverse rainfall shocks; unfortunately we do not have reliable cross-country data on these other intermediate channels. There are, however, central government budget figures for ap-

Exclusion

A more serious violation of the exclusion restriction is the possibility that high levels of rainfall might directly affect civil conflict independently of economic conditions. For instance, floods may destroy the road network and thus make it more costly for government troops to contain rebel groups. Note that this first possibility is not a serious threat to our estimation strategy, since higher levels of rainfall are empirically associated with significantly *less* conflict in the reduced-form regressions.

Exclusion

because of more difficult transportation conditions. To explore this possibility, we estimated the impact of rainfall shocks on the extent of the usable road network using World Bank data, and we did not find a statistically significant relationship. In fact, the point estimates on current and lagged rainfall are both positive (e.g., the coefficient on current rainfall growth is 192, standard error 1,025; regression not shown), which argues against the theory above. Another potential violation of

Interested in these topics?



Group work

Past exam question

We will tackle the difficult Fall 2023 exam question together:
Link

Group work

Group 1: We are interested in the effect of being in the army on crime. We instrument being in the army with a lottery (paper)

Group 2: We are interested in the effect of protestant religion on economic growth. We instrument protestantism in a region with the distance to Wittenberg (paper)

Group 3: We are interested in the effect of air pollution on mortality. We instrument local air pollution with wind direction (paper)

1. Relevance: Z must truly affect X
2. Independence/Exogeneity: Z is as good as randomly assigned
3. Exclusion restriction: The **only** way that Z affects Y is via X

Your job: Discuss whether these assumptions hold!

Your questions

Any questions?

...Remember – this is a safe space! Every question is useful!

Differences in Differences

Econ 140 Spring 2025, Section 9

Jonathan Old

Syllabus/OH bCourses Website Feedback form (**Always open**) RStudio

Roadmap

1. Time Series
2. Diff-in-Diff
3. Making DiD more general: Panel Data estimation
4. Practice

Any questions?

...Remember – this is a safe space! Every question is useful!

Time Series

Why empirical microeconomists don't use time series

Spurious correlations here.

Unsolicited financial advice here.

A great idea to lose your tuition fees is here.

Time series topics

- Autoregressive and moving-average processes
- Serial correlation (of the errors)
- Unit roots
- Trends
- Detecting trend breaks
- ...

My prediction: Not the most important topic on the exam. But come to OH if you have questions or want to know more!

Practice

See here

Diff-in-Diff

Motivation for Differences in Differences

- In the first part of the course, we talked about comparison between groups or other units (cross-sectional data)
- But we have also seen some comparisons over time: time-series data
- Both of these have very big and obvious problems, but we can use them **together** using (panel data) in a powerful tool: Differences-in-differences!

How to get the causal effect of a treatment

In 2021, UC Berkeley offered free mental health coachings to students with pre-existing mental health issues. We want to evaluate the effect of this policy and we collect a depression score (0-10) for all students at UC Berkeley.

	2020	2022
Free Mental Health: Treated	6	6
No Free Mental Health: Untreated	4	5

We can do several comparisons:

How to get the causal effect of a treatment II

	2020	2022
Free Mental Health: Treated		6
No Free Mental Health: Untreated		5

We can do several comparisons:

- Comparison 1: Compare treated and non-treated group after the intervention

How to get the causal effect of a treatment II

	2020	2022
Free Mental Health: Treated		6
No Free Mental Health: Untreated		5

We can do several comparisons:

- Comparison 1: Compare treated and non-treated group after the intervention
- Estimated treatment effect? $6 - 5 = 1$

How to get the causal effect of a treatment II

	2020	2022
Free Mental Health: Treated		6
No Free Mental Health: Untreated		5

We can do several comparisons:

- Comparison 1: Compare treated and non-treated group after the intervention
- Estimated treatment effect? $6 - 5 = 1$
- Problems?

How to get the causal effect of a treatment II

	2020	2022
Free Mental Health: Treated		6
No Free Mental Health: Untreated		5

We can do several comparisons:

- Comparison 1: Compare treated and non-treated group after the intervention
- Estimated treatment effect? $6 - 5 = 1$
- Problems?
- **Identifying assumption**: The average difference between groups is due to the treatment only. Without the treatment, the average outcome of the treated group would have been equal to the average outcome of the control group.

How to get the causal effect of a treatment III

	2020	2022
Free Mental Health: Treated	6	6
No Free Mental Health: Untreated		

We can do several comparisons:

- Comparison 2: Compare treated group before and after the intervention

How to get the causal effect of a treatment III

	2020	2022
Free Mental Health: Treated	6	6
No Free Mental Health: Untreated		

We can do several comparisons:

- Comparison 2: Compare treated group before and after the intervention
- Estimated treatment effect? $6 - 6 = 0$

How to get the causal effect of a treatment III

	2020	2022
Free Mental Health: Treated	6	6
No Free Mental Health: Untreated		

We can do several comparisons:

- Comparison 2: Compare treated group before and after the intervention
- Estimated treatment effect? $6 - 6 = 0$
- Problems?

How to get the causal effect of a treatment III

	2020	2022
Free Mental Health: Treated	6	6
No Free Mental Health: Untreated		

We can do several comparisons:

- Comparison 2: Compare treated group before and after the intervention
- Estimated treatment effect? $6 - 6 = 0$
- Problems?
- Identifying assumption: The average difference across time is due to the treatment only. Without the treatment, the average outcome of the treated group would not have changed.

How to get the causal effect of a treatment: DiD

	2020	2022
Free Mental Health: Treated	6	6
No Free Mental Health: Untreated	4	5

We can do several comparisons:

- Comparison 3: Compare treated to untreated group, before and after the intervention. **Differences in differences!**

How to get the causal effect of a treatment: DiD

	2020	2022
Free Mental Health: Treated	6	6
No Free Mental Health: Untreated	4	5

We can do several comparisons:

- Comparison 3: Compare treated to untreated group, before and after the intervention. **Differences in differences!**
- Estimated treatment effect?
  $(6 - 6) - (5 - 4) = (6 - 5) - (6 - 4) = -1$

How to get the causal effect of a treatment: DiD

	2020	2022
Free Mental Health: Treated	6	6
No Free Mental Health: Untreated	4	5

We can do several comparisons:

- Comparison 3: Compare treated to untreated group, before and after the intervention. **Differences in differences!**
- Estimated treatment effect?
  $(6 - 6) - (5 - 4) = (6 - 5) - (6 - 4) = -1$
- Identifying assumption: Parallel trends: Without the treatment, the average increase in the outcome of the treated would have been the same as the average increase in the outcome of the untreated.

Sidenote

... We can NOT observe the identifying assumption! We can find evidence for or against it, but we can never be sure!

How to get the causal effect of a treatment: DiD

	2020	2022
Free Mental Health: Treated	6	6
No Free Mental Health: Untreated	4	5

We can calculate the DiD estimate in two ways (here: superscripts stand for treatment/control group, subscripts denote unit i and time $t \in \{0, 1\}$)

$$DiD = E[\ \underbrace{\left(Y_{i1}^{T} - Y_{i0}^{T}\right)}_{\text{Change for treated}} - \underbrace{\left(Y_{i1}^{C} - Y_{i0}^{C}\right)}_{\text{Change for untreated}}\ ]$$

$$= E[\ \underbrace{\left(Y_{i1}^{T} - Y_{i1}^{C}\right)}_{\text{After-difference}} - \underbrace{\left(Y_{i0}^{T} - Y_{i0}^{C}\right)}_{\text{Before-difference}}\ ]$$

DiD and Potential Outcomes

Let us look at the second one in more detail:

$$
\begin{aligned}
DiD &= E[\left(Y_{i1}^T - Y_{i1}^C\right) - \left(Y_{i0}^T - Y_{i0}^C\right)] \\
&= \underbrace{\left(E[Y_{i1}^T] - E[Y_{i1}^C]\right)}_{\text{After-difference}} - \underbrace{\left(E[Y_{i0}^T] - E[Y_{i0}^C]\right)}_{\text{Before-difference}}
\end{aligned}
$$

Compare this with out formula for selection bias (in time period 1):

$$
\begin{aligned}
\Delta &= E[Y_{i1}^T] - E[Y_{i1}^C] \\
&= \underbrace{E[\mathrm{Y}_{i1}^T(1) - \mathrm{Y}_{i1}^T(0)]}_{\text{ATT}} + \underbrace{\left(E[\mathrm{Y}_{i1}^T(0)] - E[\mathrm{Y}_{i1}^C(0)]\right)}_{\text{Selection Bias}}
\end{aligned}
$$

DiD and Potential Outcomes

Let us look at the second one in more detail:

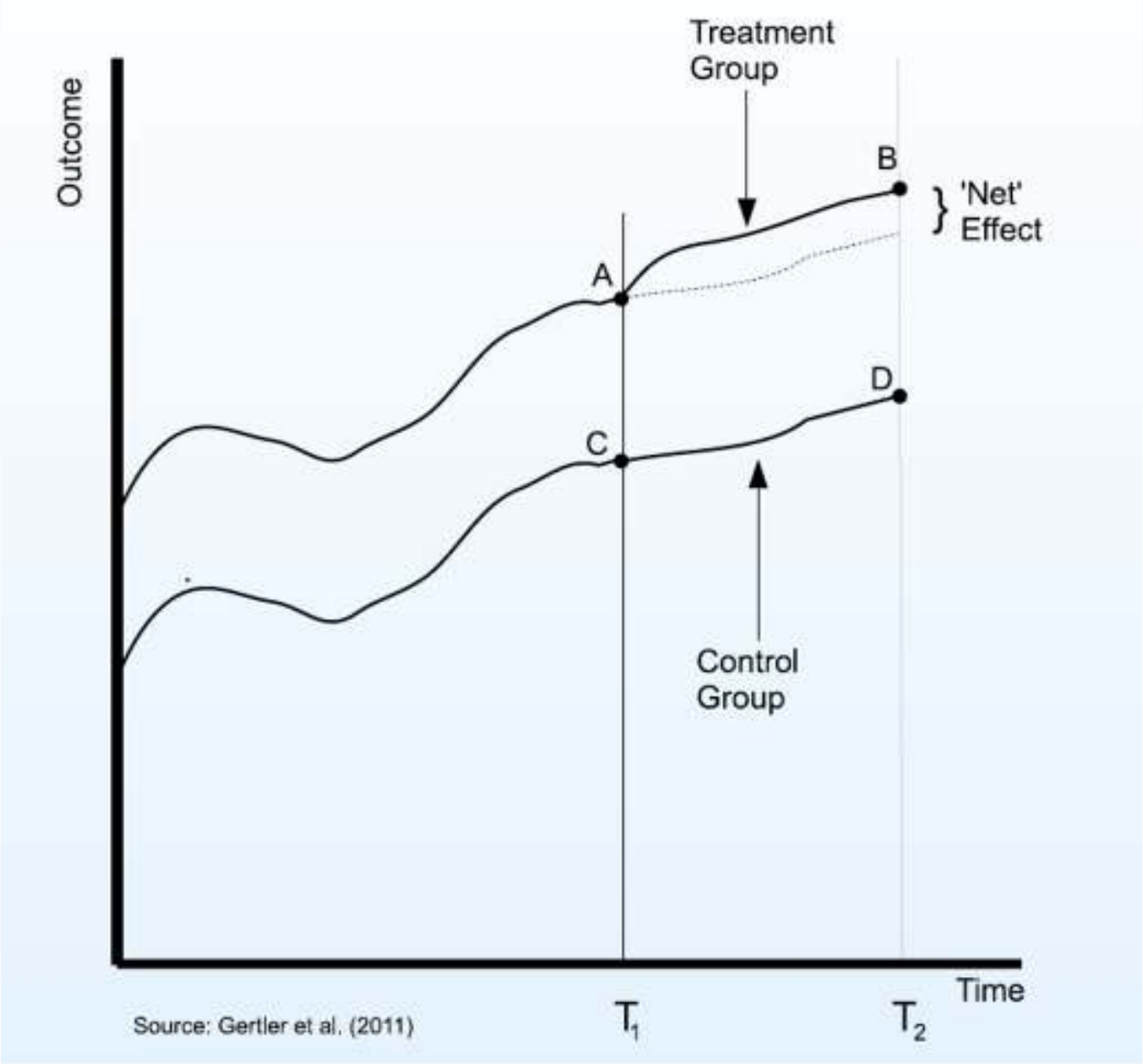
$$DiD = E[\left(Y_{i1}^T - Y_{i1}^C\right) - \left(Y_{i0}^T - Y_{i0}^C\right)]$$
$$= \underbrace{\left(E[Y_{i1}^T] - E[Y_{i1}^C]\right)}_{\text{After-difference}} - \underbrace{\left(E[Y_{i0}^T] - E[Y_{i0}^C]\right)}_{\text{Before-difference}}$$

Compare this with out formula for selection bias (at $t = 1$):

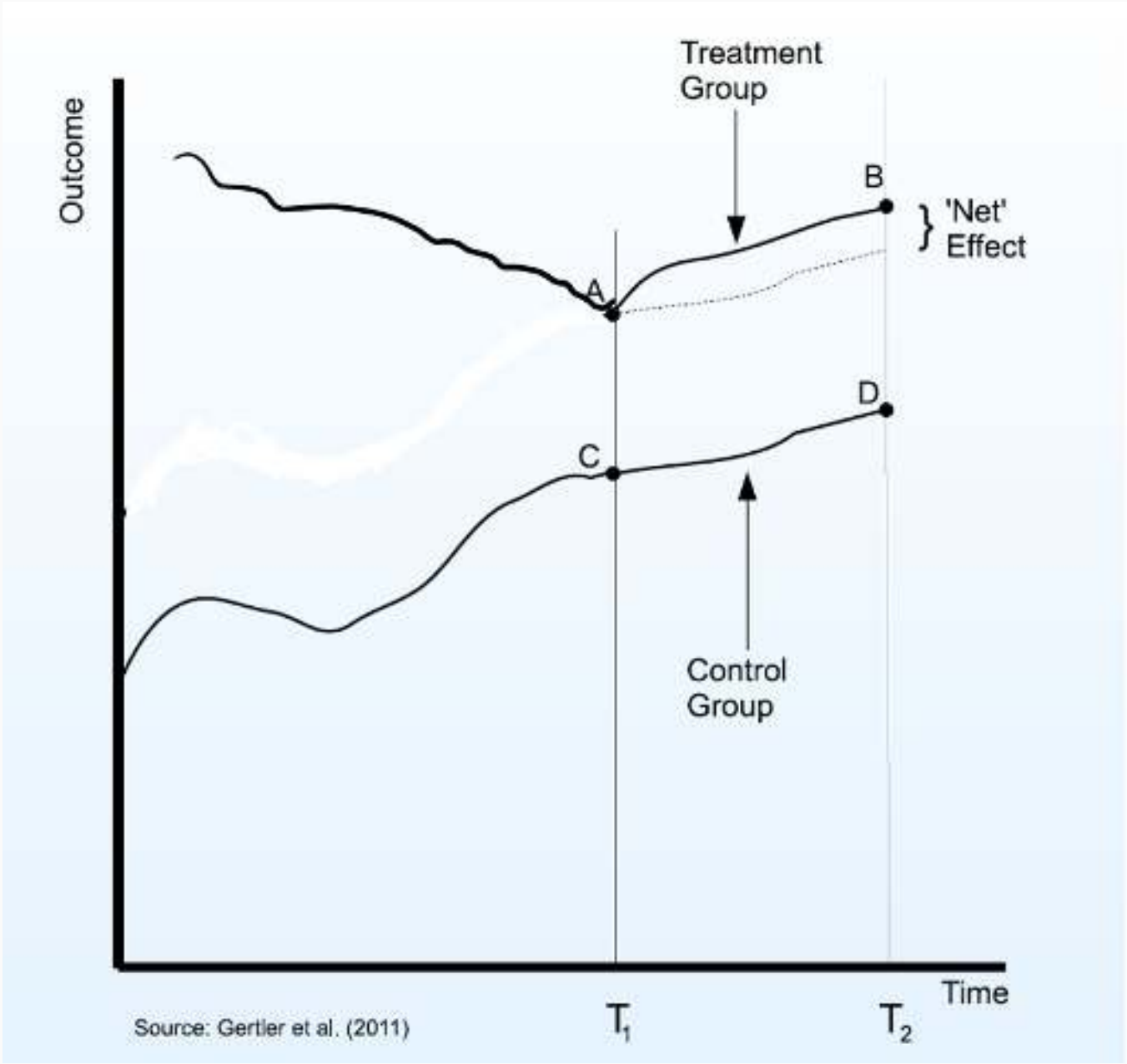
$$\Delta = E[Y_{i1}^T] - E[Y_{i1}^C]$$
$$= \underbrace{E[Y_{i1}^T(1) - Y_{i1}^T(0)]}_{\text{ATT}} + \underbrace{\left(E[Y_{i1}^T(0)] - E[Y_{i1}^C(0)]\right)}_{\text{Selection Bias}}$$

So DiD=ATT if the before-difference is equal to the selection bias in period 1: If the treatment group hadn't gotten the treatment (potential outcome = 0), the difference to the control group would have been the same as in period 0.

Parallel trends assumption



(Potential) violation of a parallel trends assumption



Estimating DiD with regressions

We can set up a simple linear regression to estimate a DiD model:

$$Y_{it} = \alpha + \beta \text{ Treated }_i + \gamma \text{ Post }_t + \delta \text{ Treated }_i \cdot \text{ Post }_t + u_{it}$$

	2020	2022
Free Mental Health: Treated		
No Free Mental Health: Untreated	α	

Estimating DiD with regressions

We can set up a simple linear regression to estimate a DiD model:

$$Y_{it} = \alpha + \beta \text{ Treated }_i + \gamma \text{ Post }_t + \delta \text{ Treated }_i \cdot \text{ Post }_t + u_{it}$$

	2020	2022
Free Mental Health: Treated		
No Free Mental Health: Untreated	α	$\alpha + \gamma$

Estimating DiD with regressions

We can set up a simple linear regression to estimate a DiD model:

$$Y_{it} = \alpha + \beta \text{ Treated }_i + \gamma \text{ Post }_t + \delta \text{ Treated }_i \cdot \text{ Post }_t + u_{it}$$

	2020	2022
Free Mental Health: Treated	$\alpha + \beta$	
No Free Mental Health: Untreated	α	$\alpha + \gamma$

Estimating DiD with regressions

We can set up a simple linear regression to estimate a DiD model:

$$Y_{it} = \alpha + \beta \text{ Treated }_i + \gamma \text{ Post }_t + \delta \text{ Treated }_i \cdot \text{ Post }_t + u_{it}$$

	2020	2022
Free Mental Health: Treated	$\alpha + \beta$	$\alpha + \beta + \gamma + \delta$
No Free Mental Health: Untreated	α	$\alpha + \gamma$

Estimating DiD with regressions

We can set up a simple linear regression model to get the DiD estimate:

$$Y_{it} = \alpha + \beta \text{ Treated }_i + \gamma \text{ Post }_t + \delta \text{ Treated }_i \cdot \text{ Post }_t + u_{it}$$

	2020	2022
Free Mental Health: Treated	6	6
No Free Mental Health: Untreated	4	5

With our data, we would get:

- $\alpha = 4$
- $\beta = 2$
- $\gamma = 1$
- $\delta = -1$

Making DiD more general: Panel Data estimation

Limitations of the 2x2 DiD framework

In the simple 2x2 framework, we estimated:

$$Y_{it} = \alpha + \beta \text{ Treated }_i + \gamma \text{ Post }_t + \delta \text{ Treated }_i \cdot \text{ Post }_t + u_{it}$$

How can we extend this framework?

Intro to Panel Data i

A panel dataset may look like this:

	country	iso2c	iso3c	year	oda_per_cap	perc_below_pov_line	gdp_per_cap	inflation
316	Chile	CL	CHL	2015	3.0402700	0.4	13569.9478	4.34877353
317	Chile	CL	CHL	2016	9.8673527	NA	13785.6883	3.78619356
318	Chile	CL	CHL	2017	3.8277324	0.3	15045.5277	2.18271847
319	Chile	CL	CHL	2018	NA	NA	15795.7085	2.43488981
320	Chile	CL	CHL	2019	NA	NA	14631.9469	2.55754476
321	Chile	CL	CHL	2020	NA	0.7	13094.4595	3.04549085
322	Chile	CL	CHL	2021	NA	NA	16265.0960	4.52456838
323	China	CN	CHN	2015	-0.2219718	1.2	8016.4314	1.43702361
324	China	CN	CHN	2016	-0.5702808	0.8	8094.3634	2.00000162
325	China	CN	CHN	2017	-0.7090169	0.7	8816.9869	1.59313600
326	China	CN	CHN	2018	-0.5027731	0.4	9905.5420	2.07479040
327	China	CN	CHN	2019	0.4323084	0.1	10143.8382	2.89923416
328	China	CN	CHN	2020	0.4059415	NA	10408.6698	2.41942189
329	China	CN	CHN	2021	-0.4002945	NA	12556.3331	0.98101514
330	Colombia	CO	COL	2015	28.7760584	4.9	6228.4263	4.98983116
331	Colombia	CO	COL	2016	23.2331299	4.9	5938.4639	7.51346025
332	Colombia	CO	COL	2017	17.5871894	4.3	6450.3196	4.31431326
333	Colombia	CO	COL	2018	36.1416353	4.5	6782.0379	3.24056933
334	Colombia	CO	COL	2019	17.4085507	5.3	6438.0602	3.52301933

Intro to Panel Data ii

It may also look like this:

	Crime_2006	Crime_2007	Crime_2008	Crime_2009	Crime_2010	Crime_2011	Crime_2012	Crime_2013	Crime_2014	Crime_2015	Crime_2016	Crime_2017
New York	0,74075276	0,80312483	0,57854938	0,58480987	0,20097276	0,80826596	0,97070048	0,02437202	0,07831943	0,84572914	0,75749919	0,98948363
Los Angeles	0,81840795	0,63440595	0,75200179	0,45218138	0,54532883	0,0015679	0,39512876	0,41847691	0,66467241	0,01946605	0,55885996	0,27521867
Chicago	0,81682982	0,00089416	0,74575534	0,5521186	0,60461758	0,34694645	0,47874246	0,15925374	0,34667588	0,83944993	0,44825934	0,58746291
Houston	0,05885616	0,79349661	0,18033028	0,51317119	0,79642654	0,50269029	0,56157808	0,77173283	0,60797558	0,40461519	0,785332	0,6477959
Phoenix	0,2568483	0,28097691	0,82722175	0,11994755	0,884851	0,83800072	0,08978678	0,21613254	0,97306225	0,57291156	0,27341507	0,20641572
Philadelphia	0,86364857	0,30478317	0,07818559	0,82930161	0,28324875	0,60204952	0,4467311	0,1165021	0,15597767	0,36430802	0,38034388	0,37091781
San Antonio	0,98256294	0,13215208	0,55220195	0,25274349	0,22117177	0,63049228	0,74109619	0,17937274	0,50914785	0,68773119	0,17854372	0,50634586
San Diego	0,26052609	0,90551179	0,96145068	0,04932108	0,55859179	0,57753748	0,38254621	0,36728436	0,39429268	0,40312699	0,55506545	0,51665312
Dallas	0,85045327	0,33482019	0,15361924	0,74231968	0,55762722	0,05057305	0,57960707	0,71774087	0,63565644	0,04305477	0,06451316	0,27719973
San Jose	0,75452737	0,21462214	0,24820183	0,2727016	0,00127307	0,36822105	0,32304996	0,9116018	0,86006177	0,05904204	0,1124391	0,83247218
Austin	0,45398495	0,27082977	0,89920786	0,82480029	0,52877813	0,9452614	0,82578514	0,828743	0,20606871	0,4439492	0,38857348	0,8154967
Jacksonville	0,18638302	0,66744344	0,38392491	0,83800597	0,30374371	0,16785163	0,53614039	0,15558203	0,14260844	0,51748284	0,07054862	0,70962828
Fort Worth	0,13254183	0,78355507	0,83321159	0,33460775	0,00772913	0,78127289	0,61026409	0,25552213	0,27373821	0,78588972	0,80784366	0,1307398
Columbus	0,21457262	0,59542339	0,65347292	0,16532587	0,35580394	0,7113844	0,72036005	0,39323372	0,73022336	0,32649945	0,95442814	0,44923343
Indianapolis	0,97993863	0,2017115	0,58883837	0,54914541	0,22495244	0,58047281	0,01984661	0,88891756	0,34920826	0,326581	0,11407985	0,76592867
Charlotte	0,13043704	0,44544929	0,13581143	0,32176217	0,06451777	0,60708331	0,24805805	0,91489751	0,39053247	0,36409354	0,94249222	0,12812588

Interpreting FE regressions

Suppose we collect yearly data on firms' profits and the gender of the firms' CEO, for a sample of 1,000 firms over 20 years.

We can estimate this model with firm and year fixed effects:

$$\text{Profits}_{it} = \beta_0 + \beta_1 \text{Female CEO}_{it} + \alpha_i + \delta_t + \varepsilon_{it}$$

How do we interpret β_1?

Interpreting FE regressions

Suppose we collect yearly data on firms' profits and the gender of the firms' CEO, for a sample of 1,000 firms over 20 years.

We can estimate this model with firm and year fixed effects:

$$\text{Profits}_{it} = \beta_0 + \beta_1 \text{Female CEO}_{it} + \alpha_i + \delta_t + \varepsilon_{it}$$

How do we interpret β_1?

The same way as in an OLS regression!

β_1 describes how having a female CEO is associated with profits, holding constant all firm-specific factors that do not change over time, and all time-specific factors that affect all firms equally. β_1 is the effect within each firm (and year!), controlling for the firm-specific and year-specific fixed effects.

Visualization

See separate gif!

Limitations of fixed effects

- Imagine we want to estimate the effect of annual income on happiness and we have a panel following 1,000 individuals over 10 years. You think there are many omitted variables and include individual-level fixed effects. You are **also** interested in the effect of parental income on happiness. What's the problem?

Limitations of fixed effects

- Imagine we want to estimate the effect of annual income on happiness and we have a panel following 1,000 individuals over 10 years. You think there are many omitted variables and include individual-level fixed effects. You are **also** interested in the effect of parental income on happiness. What's the problem?
- We want to estimate the effect of democracy on economic growth in a panel of 180 countries over 60 years. We can include country-fixed effects to account for any effect coming from different countries having a different climate, terrain, access to the sea, culture, etc. Are you satisfied with this approach?

Limitations of fixed effects

- Imagine we want to estimate the effect of annual income on happiness and we have a panel following 1,000 individuals over 10 years. You think there are many omitted variables and include individual-level fixed effects. You are **also** interested in the effect of parental income on happiness. What's the problem?
- We want to estimate the effect of democracy on economic growth in a panel of 180 countries over 60 years. We can include country-fixed effects to account for any effect coming from different countries having a different climate, terrain, access to the sea, culture, etc. Are you satisfied with this approach?
- We want to estimate the effect of education on wages and think of using individual-level fixed effects. What can go wrong?

Any questions?

...Remember – this is a safe space! Every question is useful!

Practice

Practice

See here

Exercises and Questions around IV

Econ 140, Section 10

Jonathan Old

Roadmap

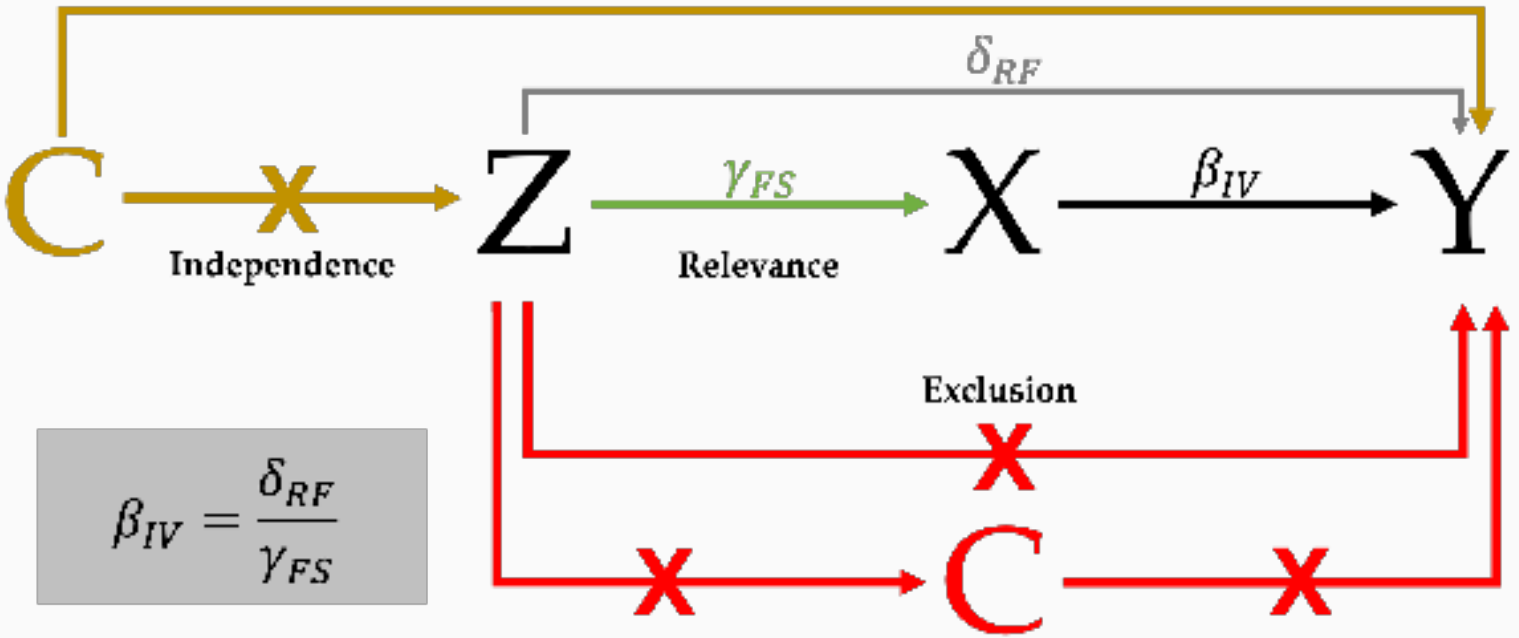
1. Recap: IV
2. Group work
3. IV Exam Exercise
4. Questions

Recap: IV

Recap: IV summary

We need the following three assumptions for IV to work:

1. **Relevance**: *Z* must truly affect *X*
2. **Independence/Exogeneity**: *Z* is as good as randomly assigned
3. **Exclusion Restriction**: The **only** way that *Z* affects *Y* is via *X*.



$$\beta_{IV} = \frac{\delta_{RF}}{\gamma_{FS}}$$

How IV estimates can be different from OLS estimates

- **OVB:** If OLS had omitted variable bias, and our instrument is valid, then the IV estimate should be different. We can check whether this is plausible with the OVB formula
- **Measurement error:** If we have random measurement error in the independent variable (X), then we can use IV to overcome this. In that case, the IV coefficient will be larger (in absolute value) than the OLS coefficient
- **LATE:** OLS gives us ATT + Selection Bias, while IV gives us the treatment effect on the compliers (LATE). The ATT may be different from LATE, even without selection bias.
- **Invalid IV:** Hard to determine what exactly is going on
- **Sampling variation**: This can just happen by chance

IV and controls

- There is a catch in IV: independence (exogeneity) and exclusion restriction only need to hold **conditionally**
- This means: If we **know** C, we can control for it in the IV regression, and the concerns go away
- But then, we have the old OLS problems again
- So we should only do this in cases where we really know all the C's.

Group work

Group work

Group 1: We are interested in the effect of being in the army on crime. We instrument being in the army with a lottery (paper)

Group 2: We are interested in the effect of income on conflict. We instrument income with rainfall (paper)

Group 3: We are interested in the effect of air pollution on mortality. We instrument local air pollution with wind direction (paper)

1. Relevance: Z must truly affect X
2. Independence/Exogeneity: Z is as good as randomly assigned
3. Exclusion restriction: The **only** way that Z affects Y is via X

Your job: Discuss whether these assumptions hold!

IV Exam Exercise

IV Exam exercise i

The land of Econometrea consists of 34 islands. Each island has their own university and students attend uni on their home island. Students in Econometrea either have to attend lectures in person or watch them online. Priyanka has obtained data for the students from all the universities and would like to study the effect of watching lectures online on the students' exam scores. For each student she has the exam score (0 - 100, 70+ is a first, 40 and below a fail), which she uses as her left hand side variable, the fraction of lectures watched online, and how on how many days the student visited the library each week. Priyanka obtains the following regression results:

IV Exam exercise ii

	Estimation Method				
	OLS	IV	IV	IV	IV
Regressor	(1)	(2)	(3)	(4)	(5)
Fraction online	−4.91	−2.08	−0.56	−0.75	6.09
	(0.11)	(0.22)	(0.40)	(0.90)	(3.22)
Library visits			0.91		2.09
			(0.08)		(0.43)
Sample	All	All	All	Unis with lottery	Unis with lottery

All regressions also contain a constant term and a dummy variable for each but one of the universities. Robust standard errors are reported in parentheses.

IV Exam exercise iii

(a) What is the interpretation of the coefficient in column (1). If this were a causal effect, would it be big or small? Explain whether this estimate is likely to have a causal interpretation.
(b) Wei Min observes that some halls of residence are close to the lecture rooms while others are further away and suggests to use the distance of halls from lecture halls as an instrumental variable for the fraction of lectures watched online. Results are in column (2). Explain which assumptions need to be satisfied for the IV to deliver better causal estimates than OLS here. Discuss the validity of the

IV Exam exercise iv

assumptions in this case.

	Estimation Method				
	OLS	IV	IV	IV	IV
Regressor	(1)	(2)	(3)	(4)	(5)
Fraction online	−1.91	−2.08	−0.56	−0.75	6.09
	(0.11)	(0.22)	(0.40)	(0.90)	(3.22)
Library visits			0.91		2.09
			(0.08)		(0.43)
Sample	All	All	All	Unis with lottery	Unis with lottery

All regressions also contain a constant term and a dummy variable for each but one of the universities. Robust standard errors are reported in parentheses.

IV Exam exercise v

Solution

- Students who watch all lectures online, on average, have around 4.9 fewer points on the exam than students who went to all lecturs in person. This is a moderately strong effect, but likely not causal (think of omitted variables, such as student motivation)

IV Exam exercise vi

Solution

- **Relevance:** Proximity to lecture halls must correlate strongly with lecture attendance. This is plausible from our experience, and we could test this using an F-test.
- **Independence/Exogeneity:** Proximity must be as good as randomly allocated. This might not hold here, for example, if more motivated students live closer to campus.
- **Exclusion restruction:** Proximity to lecture halls only influences exam scores because it has an effect on lecture attendance. This can be violated if proximity also affects the probability to go and study in the library, or if residence halls closer to campus are also closer to bars

IV Exam exercise vii

(c) Alma points out to Wei Min that students can pick which hall they want to live in and that some halls are located in Study Village, close to lecture halls and the library while others are located further away in Party Town, surrounded by pubs. How does this information affect your assessment of the IV strategy?
(d) Manuel realises that the data also include a variable for the number of times a student has checked into the library per week. He suggests to rerun the instrumental variables regression adding this variable as a control. Results for this regression are displayed in column (3). Assess Manuel's strategy.

IV Exam exercise viii

Solution

- This would be a violation of independence/exogeneity: The students who live closer to the campus are just inherently different than students who live far away
- Library visits are a bad control: They are another outcome of the instrument, and so controlling for them introduces additional bias.

IV Exam exercise ix

	Estimation Method				
	OLS	IV	IV	IV	IV
Regressor	(1)	(2)	(3)	(4)	(5)
Fraction online	−4.91	−2.08	−0.56	−0.75	6.09
	(0.11)	(0.22)	(0.40)	(0.90)	(3.22)
Library visits			0.91		2.09
			(0.08)		(0.43)
Sample	All	All	All	Unis with lottery	Unis with lottery

All regressions also contain a constant term and a dummy variable for each but one of the universities. Robust standard errors are reported in parentheses.

IV Exam exercise x

(e) Jenny notices that there are five universities which assign students to their halls of residence by a lottery. She suggests to run the IV model from columns (2) and (3) for the subsample of students from these universities only. Results are displayed in columns (4) and (5). Assess Jenny's regressions.
(f) Drawing on the results in the table above, what have you learned from this exercise about the causal effect of watching lectures online on students' exam results?

IV Exam exercise xi

Solution

- For these universities, the distance is as good as randomly assigned, and so we have no violations of the independence/exogeneity restriction. However, the exclusion restriction may still not hold.
- In a simply OLS regression, it may look as if watching lectures online causes worse grades. However, we can correct for this using a valid IV, and find that the effect is only very small and statistically insignificant.

Questions

From Diff in Diff to Panel Data

Econ 140, Section 11

Jonathan Old

Roadmap

1. Recap: DiD

2. Making DiD more general: Panel Data estimation

3. Practice question

Recap: DiD

How to get the causal effect of a treatment: DiD

	2020	2022
Free Mental Health: Treated	6	6
No Free Mental Health: Untreated	4	5

We can do several comparisons:

- Comparison 3: Compare treated to untreated group, before and after the intervention. **Differences in differences!**

How to get the causal effect of a treatment: DiD

	2020	2022
Free Mental Health: Treated	6	6
No Free Mental Health: Untreated	4	5

We can do several comparisons:

- Comparison 3: Compare treated to untreated group, before and after the intervention. **Differences in differences!**
- Estimated treatment effect?
  $(6 - 6) - (5 - 4) = (6 - 5) - (6 - 4) = -1$

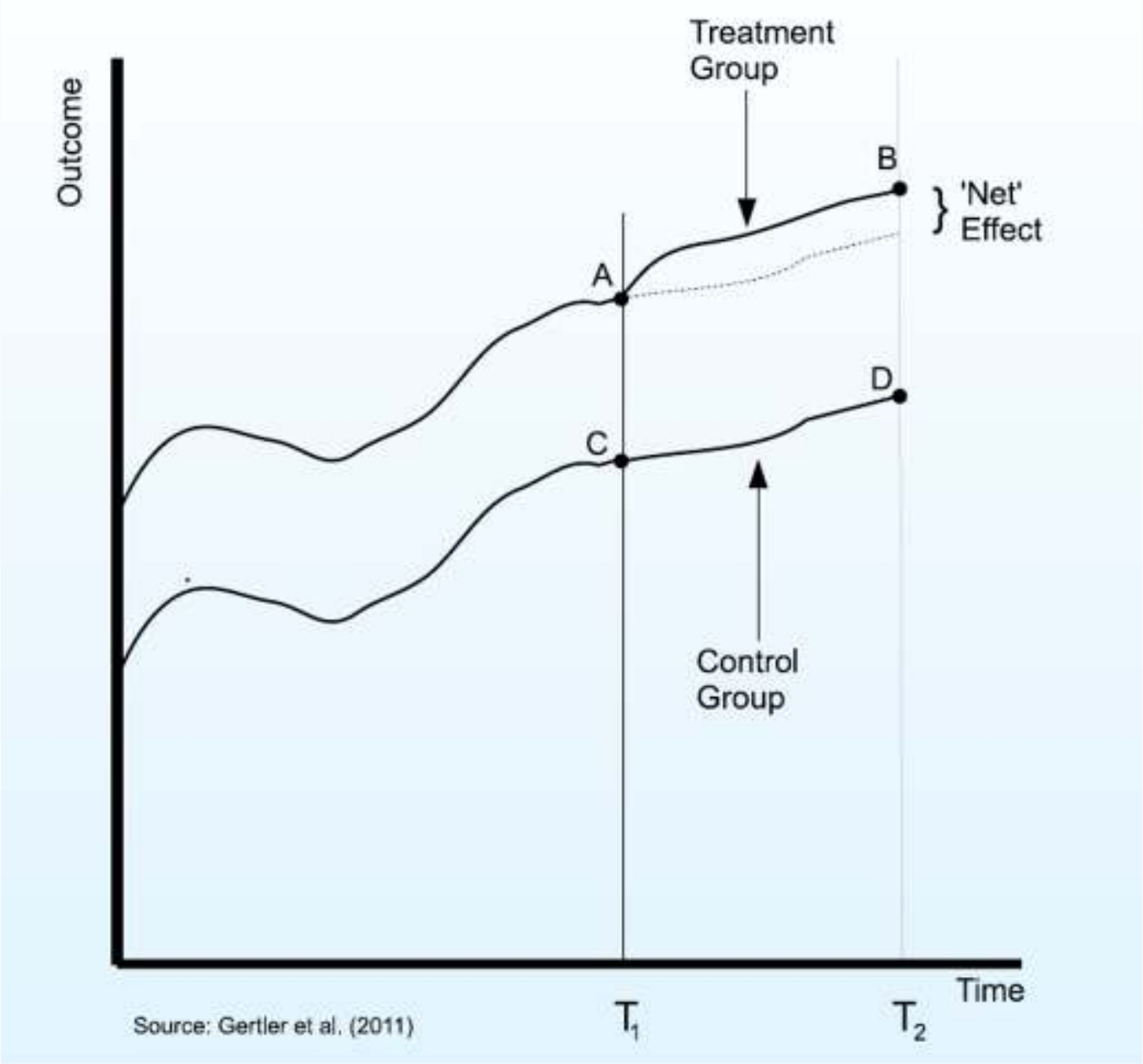
How to get the causal effect of a treatment: DiD

	2020	2022
Free Mental Health: Treated	6	6
No Free Mental Health: Untreated	4	5

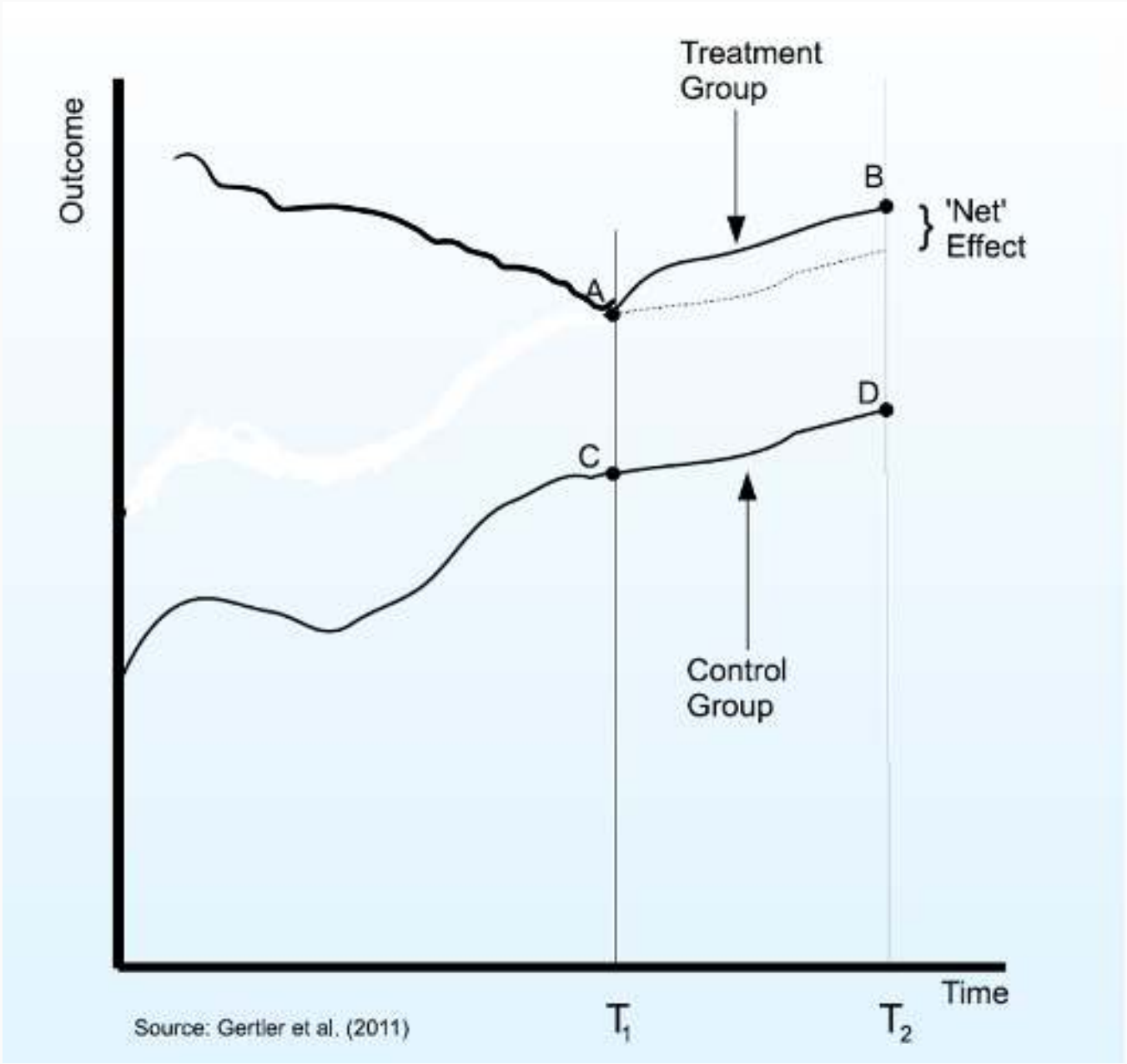
We can do several comparisons:

- Comparison 3: Compare treated to untreated group, before and after the intervention. **Differences in differences!**
- Estimated treatment effect?
  $(6 - 6) - (5 - 4) = (6 - 5) - (6 - 4) = -1$
- Identifying assumption: Parallel trends: Without the treatment, the average increase in the outcome of the treated would have been the same as the average increase in the outcome of the untreated.

Parallel trends assumption



(Potential) violation of a parallel trends assumption



Estimating DiD with regressions

We can set up a simple linear regression to estimate a DiD model:

$$Y_{it} = \alpha + \beta \text{ Treated }_i + \gamma \text{ Post }_t + \delta \text{ Treated }_i \cdot \text{ Post }_t + u_{it}$$

	2020	2022
Free Mental Health: Treated	$\alpha + \beta$	$\alpha + \beta + \gamma + \delta$
No Free Mental Health: Untreated	α	$\alpha + \gamma$

Any questions?

… Remember – Every question is useful!

Making DiD more general: Panel Data estimation

Limitations of the 2x2 DiD framework

In the simple 2x2 framework, we estimated:

$$Y_{it} = \alpha + \beta \text{ Treated }_i + \gamma \text{ Post }_t + \delta \text{ Treated }_i \cdot \text{ Post }_t + u_{it}$$

How can we extend this framework?

Intro to Panel Data i

A panel dataset may look like this:

	country	iso2c	iso3c	year	oda_per_cap	perc_below_pov_line	gdp_per_cap	inflation
316	Chile	CL	CHL	2015	3.0402700	0.4	13569.9478	4.34877353
317	Chile	CL	CHL	2016	9.8673527	NA	13785.6883	3.78619356
318	Chile	CL	CHL	2017	3.8277424	0.3	15045.5277	2.18271847
319	Chile	CL	CHL	2018	NA	NA	15795.7085	2.43488981
320	Chile	CL	CHL	2019	NA	NA	14631.9469	2.55754478
321	Chile	CL	CHL	2020	NA	0.7	13094.4595	3.04549085
322	Chile	CL	CHL	2021	NA	NA	16265.0960	4.52456838
323	China	CN	CHN	2015	-0.2219718	1.2	8016.4314	1.43702361
324	China	CN	CHN	2016	-0.5702808	0.8	8094.3634	2.00000162
325	China	CN	CHN	2017	-0.7090169	0.7	8816.9869	1.59313600
326	China	CN	CHN	2018	-0.5027741	0.4	9905.3420	2.07479040
327	China	CN	CHN	2019	0.4323084	0.1	10143.8382	2.89923416
328	China	CN	CHN	2020	0.4059415	NA	10408.6698	2.41942189
329	China	CN	CHN	2021	-0.4002945	NA	12556.3331	0.98101514
330	Colombia	CO	COL	2015	28.7760584	4.9	6228.4263	4.98983116
331	Colombia	CO	COL	2016	23.2331299	4.9	5938.4639	7.51346025
332	Colombia	CO	COL	2017	17.5871894	4.3	6450.3196	4.31431326
333	Colombia	CO	COL	2018	36.1416353	4.5	6782.0379	3.24056933
334	Colombia	CO	COL	2019	17.4085507	5.3	6458.0602	3.52301933

Intro to Panel Data ii

It may also look like this:

	Crime_2006	Crime_2007	Crime_2008	Crime_2009	Crime_2010	Crime_2011	Crime_2012	Crime_2013	Crime_2014	Crime_2015	Crime_2016	Crime_2017
New York	0,74075276	0,80312483	0,57854938	0,58480987	0,20097276	0,80826596	0,97070048	0,02437202	0,07831943	0,84572914	0,75749919	0,98948363
Los Angeles	0,81840795	0,63440595	0,75200179	0,45218138	0,54532883	0,0015679	0,39512876	0,41847691	0,66467241	0,01946605	0,55885996	0,27521867
Chicago	0,81682982	0,00089416	0,74575534	0,5521186	0,60461758	0,34694645	0,47874246	0,15925374	0,34667588	0,83944993	0,44825934	0,58746291
Houston	0,05885616	0,79349661	0,18033028	0,51317119	0,79642654	0,50269029	0,56157808	0,77173283	0,60797558	0,40461519	0,785332	0,6477959
Phoenix	0,2568483	0,28097691	0,82722175	0,11994755	0,884851	0,83800072	0,08978678	0,21613254	0,97306225	0,57291156	0,27341507	0,20641572
Philadelphia	0,86364857	0,30478317	0,07818559	0,82930161	0,28324875	0,60204952	0,4467311	0,1165021	0,15597767	0,36430802	0,38034388	0,37091781
San Antonio	0,98256294	0,13215208	0,55220195	0,25274349	0,22117177	0,63049228	0,74109619	0,17937274	0,50914785	0,68773119	0,17854372	0,50634586
San Diego	0,26052609	0,90551179	0,96145068	0,04932108	0,55859179	0,57753748	0,38254621	0,36728436	0,39429268	0,40312699	0,55506545	0,51665312
Dallas	0,85045327	0,33482019	0,15361924	0,74231968	0,55762722	0,05057305	0,57960707	0,71774087	0,63565644	0,04305477	0,06451316	0,27719973
San Jose	0,75452737	0,21462214	0,24820183	0,2727016	0,00127307	0,36822105	0,32304996	0,9116018	0,86006177	0,05904204	0,1124391	0,83247218
Austin	0,45398495	0,27082977	0,89920786	0,82480029	0,52877813	0,9452614	0,82578514	0,828743	0,20606871	0,4439492	0,38857348	0,8154967
Jacksonville	0,18638302	0,66744344	0,38392491	0,83800597	0,30374371	0,16785163	0,53614039	0,15558203	0,14260844	0,51748284	0,07054862	0,70962828
Fort Worth	0,13254183	0,78355507	0,83321159	0,33460775	0,00772913	0,78127289	0,61026409	0,25552213	0,27373821	0,78588972	0,80784366	0,1307398
Columbus	0,21457262	0,59542339	0,65347292	0,16532587	0,35580394	0,7113844	0,72036005	0,39323372	0,73022336	0,32649945	0,95442814	0,44923343
Indianapolis	0,97993863	0,2017115	0,58883837	0,54914541	0,22495244	0,58047281	0,01984661	0,88891756	0,34920826	0,326581	0,11407985	0,76592867
Charlotte	0,13043704	0,44544929	0,13581143	0,32176217	0,06451777	0,60708331	0,24805805	0,91489751	0,39053247	0,36409354	0,94249222	0,12812588

Interpreting FE regressions

Suppose we collect yearly data on firms' profits and the gender of the firms' CEO, for a sample of 1,000 firms over 20 years.

We can estimate this model with firm and year fixed effects:

$$\text{Profits}_{it} = \beta_0 + \beta_1 \text{Female CEO}_{it} + \alpha_i + \delta_t + \varepsilon_{it}$$

How do we interpret β_1?

Interpreting FE regressions

Suppose we collect yearly data on firms' profits and the gender of the firms' CEO, for a sample of 1,000 firms over 20 years.

We can estimate this model with firm and year fixed effects:

$$\text{Profits}_{it} = \beta_0 + \beta_1 \text{Female CEO}_{it} + \alpha_i + \delta_t + \varepsilon_{it}$$

How do we interpret β_1?

The same way as in an OLS regression!

β_1 describes how having a female CEO is associated with profits, holding constant all firm-specific factors that do not change over time, and all time-specific factors that affect all firms equally. β_1 is the effect within each firm (and year!), controlling for the firm-specific and year-specific fixed effects.

Visualization

See separate gif!

Limitations of fixed effects

- Imagine we want to estimate the effect of annual income on happiness and we have a panel following 1,000 individuals over 10 years. You think there are many omitted variables and include individual-level fixed effects. You are **also** interested in the effect of parental income on happiness. What's the problem?

Limitations of fixed effects

- Imagine we want to estimate the effect of annual income on happiness and we have a panel following 1,000 individuals over 10 years. You think there are many omitted variables and include individual-level fixed effects. You are **also** interested in the effect of parental income on happiness. What's the problem?
- We want to estimate the effect of democracy on economic growth in a panel of 180 countries over 60 years. We can include country-fixed effects to account for any effect coming from different countries having a different climate, terrain, access to the sea, culture, etc. Are you satisfied with this approach?

Limitations of fixed effects

- Imagine we want to estimate the effect of annual income on happiness and we have a panel following 1,000 individuals over 10 years. You think there are many omitted variables and include individual-level fixed effects. You are **also** interested in the effect of parental income on happiness. What's the problem?
- We want to estimate the effect of democracy on economic growth in a panel of 180 countries over 60 years. We can include country-fixed effects to account for any effect coming from different countries having a different climate, terrain, access to the sea, culture, etc. Are you satisfied with this approach?
- We want to estimate the effect of education on wages and think of using individual-level fixed effects. What can go wrong?

Any questions?

… Remember – Every question is useful!

Practice question

Did Welfare Reform Increase Employment? In the 1990s, in an effort known as 'Welfare Reform' the United States implemented a series of policies that aimed to reduce welfare dependency by incentivizing work and raising the hassle of being on welfare. These reforms were implemented at the state level between 1993 and 1997. In this exercise, you will investigate whether welfare reform successfully raised employment rates.

The file welfare.csv contains information on individuals in the U.S. between 1990 and 2005. The key variables are

- employed: dummy equal to one if individual is employed
- reform: dummy equal to one if the individual is living in a state that has welfare reform in place in that year
- hsgrad: dummy equal to one if individual has a HS degree
- white: dummy equal to one if individual is white
- woman: dummy equal to one if individual identifies as a woman
- avg_benefit: the average welfare benefit (between 1990 and 2005) for a family of three in that individual's state

Estimate a regression of (1) employed against reform (2) employed against reform, hsgrad, white, and woman and interpret the coefficient on reform in both regressions.

Suggest a variable that varies across states but plausibly varies little-or not at all-over time and that could cause omitted variable bias in regression (2).

Estimate regression (2) using state fixed effects. How does the coefficient change? What does this suggest about the ommitted variable bias in part (b)? Which omitted variables do we account for with state FE?

Estimate regression (2) using time fixed effects. Interpret the interpret the coefficient on reform. How does the coefficient change? What does this suggest about the ommitted variable bias in part (b)? Which omitted variables do we account for with time FE?

Estimate regression (2) using time AND state fixed effects. Interpret the interpret the coefficient on reform. How does the coefficient change? What does this suggest about the ommitted variable bias in part (b)? Which omitted variables do we account for with state and time FE?

Which specification (no FE, state FE, time FE, or state + time FE) is most credible and why?

See you on R!

Semester Review, Questions, Exam Practice

Econ 140, Section 12

Jonathan Old

Roadmap

1. Regression Discontinuity
2. Broad topics for the exam
3. IV Exam Exercise
4. Other exam practice

Course evaluation



Regression Discontinuity

The world is full of arbitrary rules

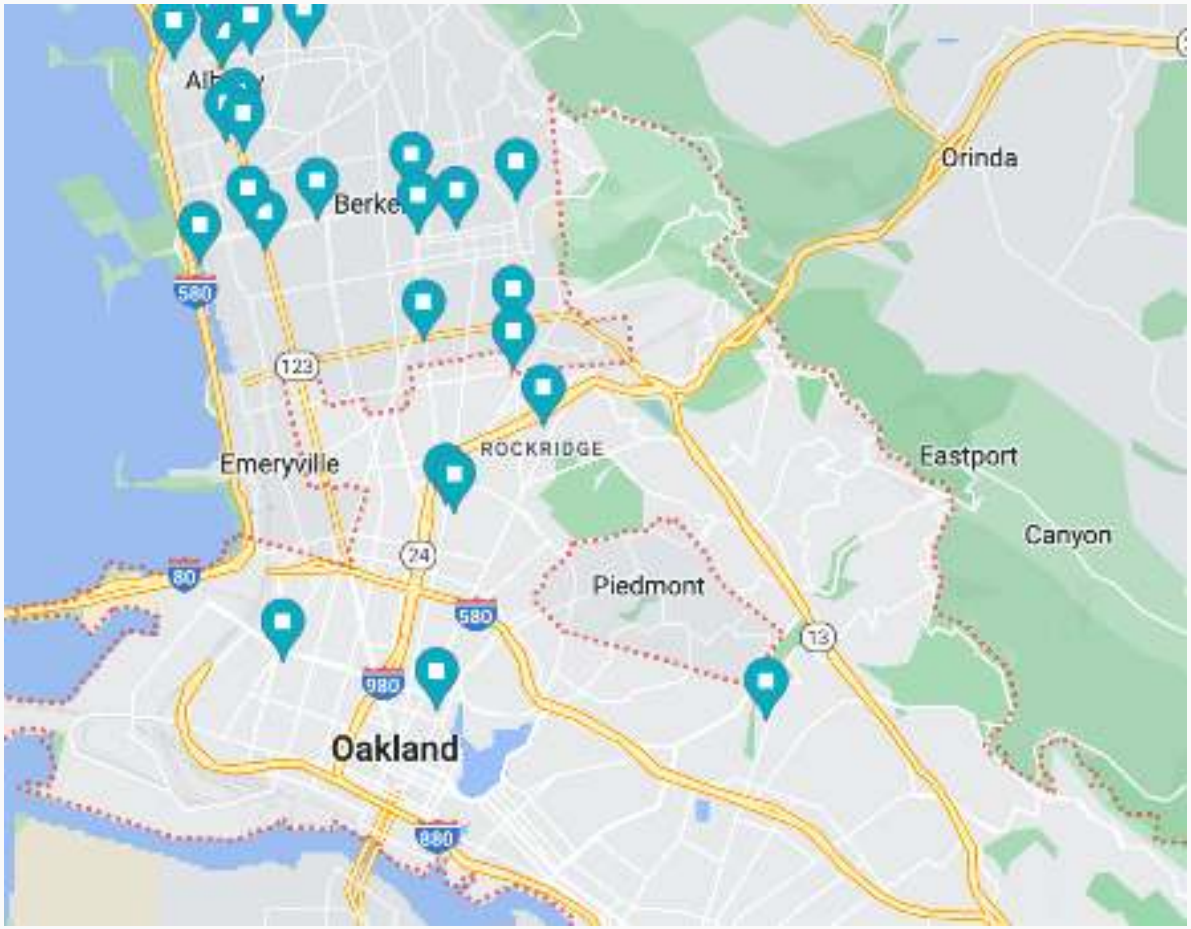
- Students receive a scholarship if their GPA is above 3.0
- Children are allowed to start school if they are five years old before January 31 of that year
- Individuals are eligible for a microfinance loan if they own less than 0.5 acres of land
- Legislators are elected if they receive over 50% of the vote

This creates wonderful discontinuities in the data that we can exploit for econometric analysis!

The most important rules of regression discontinuity

- The essence of RD is to compare people (units) just above the cut-off to others just below
- Crucial assumptions: There are no jumps but our jump at the jump // The effect of X (our running variable) and any other variable C on Y (our outcome) is smooth around the discontinuity
- Identifying assumption: In the absence of the ”treatment” (which was allocated by a discontinuous rule), the outcomes of the treated would have been essentially the same as the outcomes of the untreated
- Limitations:
  1. We only estimate a **local** effect!
  2. There may be strategic behavior at the cutoff (example: school district Oakland/Piedmont)

Oakland and Piedmont



Implementing regression discontinuity designs

With a running variable X, an outcome Y, and a treatment D allocated at the discontinuity, we estimate:

$$Y_i = \alpha + f(X_i) + \beta D_i + \varepsilon_i$$

- $f(X_i)$ is a smooth function of the running variable - e.g. just linear, or quadratic
- We could also estimate whether the **slope** changes at the discontinuity
- We are eventually interested in β: Does the outcome change at the discontinuity?

Broad topics for the exam

Introduction

- Interpreting regression tables
- p-values: How surprised we should be to observe the world as it is if your hypothesis about how it works were true?
- Omitted variable bias
- Interpretation of control variables and their role
- Interaction terms (dummy or continuous) – dummyXdummy: get means for every group

Instrumental variables

- Calculation
- Three assumptions: Relevance, Independence (Exogeneity), Exclusion restriction
- Think of three regressions: First stage, reduced form, ”target regression” (Second stage)
- Connection to RCTs (imperfect compliance)

Differences in Differences

- Parallel trends assumption
- Calculate with tables
- Calculate using regression tables

Panel Data and Fixed effects

- Fixed effects: Calculation
- Useful way to think about it: We add a lot of controls, but key problem remains (time-varying confounders: Things that change over time that affect different units differently, e.g.: climate change ...)
- Connection of two-way fixed effects with DiD

Regression Discontinuities

- ”No jump but our jump at the jump”
- Fuzzy DiD is just IV

General and overarching concepts

- Comparison of OLS and other estimates: What changed and why? (Hint: Always go back to OVB)
- Be creative but also clear: Most thinking questions can be solved with reasonable economic intuition, your knowledge of the world, and what you learned on the course.
- Keep in mind the big picture: We are just trying to find a valid counterfactual, to estimate causal effects!

IV Exam Exercise

IV Exam exercise i

The land of Econometrea consists of 34 islands. Each island has their own university and students attend uni on their home island. Students in Econometrea either have to attend lectures in person or watch them online. Priyanka has obtained data for the students from all the universities and would like to study the effect of watching lectures online on the students' exam scores. For each student she has the exam score (0 - 100, 70+ is a first, 40 and below a fail), which she uses as her left hand side variable, the fraction of lectures watched online, and how on how many days the student visited the library each week. Priyanka obtains the following regression results:

IV Exam exercise ii

	Estimation Method				
	OLS	IV	IV	IV	IV
Regressor	(1)	(2)	(3)	(4)	(5)
Fraction online	−4.91	−2.08	−0.56	−0.75	6.09
	(0.11)	(0.22)	(0.40)	(0.90)	(3.22)
Library visits			0.91		2.09
			(0.08)		(0.43)
Sample	All	All	All	Unis with lottery	Unis with lottery

All regressions also contain a constant term and a dummy variable for each but one of the universities. Robust standard errors are reported in parentheses.

IV Exam exercise iii

(a) What is the interpretation of the coefficient in column (1). If this were a causal effect, would it be big or small? Explain whether this estimate is likely to have a causal interpretation.
(b) Your friend realizes that you also have data from before Covid, when all students had to attend all lectures in person. She suggests to construct a panel dataset and to run a regression and to include time and individual fixed effects. What assumption would be required for this regression to deliver a causal effect?

IV Exam exercise iv

Solution

- Students who watch all lectures online, on average, have around 4.9 fewer points on the exam than students who went to all lecturs in person. This is a moderately strong effect, but likely not causal (think of omitted variables, such as student motivation)

IV Exam exercise v

Solution

- For this to work, we need the parallel assumption to hold:
  In the absence of Covid and the introduction of the online policy, individuals with more online attendance would have had the same change in exam scores as individuals with less online attendance.
  In this case, this might not be very plausible, because Covid probably affected individuals who are likely to stay at home differently than individuals who attend class in person.

IV Exam exercise vi

(c) Wei Min observes that some halls of residence are close to the lecture rooms while others are further away and suggests to use the distance of halls from lecture halls as an instrumental variable for the fraction of lectures watched online. Results are in column (2). Explain which assumptions need to be satisfied for the IV to deliver better causal estimates than OLS here. Discuss the validity of the assumptions in this case.

IV Exam exercise vii

	Estimation Method				
	OLS	IV	IV	IV	IV
Regressor	(1)	(2)	(3)	(4)	(5)
Fraction online	−1.91	−2.08	−0.56	−0.75	6.09
	(0.11)	(0.22)	(0.40)	(0.90)	(3.22)
Library visits			0.91		2.09
			(0.08)		(0.43)
Sample	All	All	All	Unis with lottery	Unis with lottery

All regressions also contain a constant term and a dummy variable for each but one of the universities. Robust standard errors are reported in parentheses.

IV Exam exercise viii

Solution

- **Relevance:** Proximity to lecture halls must correlate strongly with lecture attendance. This is plausible from our experience, and we could test this using an F-test.
- **Independence/Exogeneity:** Proximity must be as good as randomly allocated. This might not hold here, for example, if more motivated students live closer to campus.
- **Exclusion restruction:** Proximity to lecture halls only influences exam scores because it has an effect on lecture attendance. This can be violated if proximity also affects the probability to go and study in the library, or if residence halls closer to campus are also closer to bars

IV Exam exercise ix

(d) Alma points out to Wei Min that students can pick which hall they want to live in and that some halls are located in Study Village, close to lecture halls and the library while others are located further away in Party Town, surrounded by pubs. How does this information affect your assessment of the IV strategy?
(e) Manuel realises that the data also include a variable for the number of times a student has checked into the library per week. He suggests to rerun the instrumental variables regression adding this variable as a control. Results for this regression are displayed in column (3). Assess Manuel's strategy.

IV Exam exercise x

Solution

- This would be a violation of independence/exogeneity: The students who live closer to the campus are just inherently different than students who live far away
- Library visits are a bad control: They are another outcome of the instrument, and so controlling for them introduces additional bias.

IV Exam exercise xi

	Estimation Method				
	OLS	IV	IV	IV	IV
Regressor	(1)	(2)	(3)	(4)	(5)
Fraction online	−4.91	−2.08	−0.56	−0.75	6.09
	(0.11)	(0.22)	(0.40)	(0.90)	(3.22)
Library visits			0.91		2.09
			(0.08)		(0.43)
Sample	All	All	All	Unis with lottery	Unis with lottery

All regressions also contain a constant term and a dummy variable for each but one of the universities. Robust standard errors are reported in parentheses.

IV Exam exercise xii

(f) Jenny notices that there are five universities which assign students to their halls of residence by a lottery. She suggests to run the IV model from columns (2) and (3) for the subsample of students from these universities only. Results are displayed in columns (4) and (5). Assess Jenny's regressions.
(g) Drawing on the results in the table above, what have you learned from this exercise about the causal effect of watching lectures online on students' exam results?

IV Exam exercise xiii

Solution

- For these universities, the distance is as good as randomly assigned, and so we have no violations of the independence/exogeneity restriction. However, the exclusion restriction may still not hold.
- In a simply OLS regression, it may look as if watching lectures online causes worse grades. However, we can correct for this using a valid IV, and find that the effect is only very small and statistically insignificant.

Other exam practice

RDD question

Question 2

The local government wants to estimate the impact on future earnings of a job training program that it operated in 2016 and 2017. Access to the program is governed by an eligibility rule: only individuals whose income in the prior tax year was less than £12,000 can participate.

A. **(20 marks)** Explain how you could use this eligibility rule to estimate the causal effect of the program. Describe any data you would need, write down the regression equation(s) you would estimate, define all variables precisely, and explain how you would interpret the regression results. If you make any additional assumptions, state them clearly.

B. **(10 marks)** A colleague worries that because the eligibility rule was public, your estimates of the program's causal effect may be biased. Why might this pose a problem for identification? How could you use the data to assess the validity of this concern?

DiD question

Question 3

(10 marks) On 28 August 2017, a new state law forced the city of St. Louis, Missouri, United States to reduce its minimum wage from $10 per hour to $7.70 per hour. You have monthly employment data for at the municipal level for the entire state in the months of May 2017 and November 2017. How would you construct the difference-in-differences estimate for the effect of reducing the minimum wage? What assumption would have to hold for this to be a valid causal estimate of the effect? Precisely describe one concern that you may have with the validity of this assumption. Describe two things you could do to assess the validity of this assumption.

Method	Estimation details	Identifying assumption	Formal assumption	Threats to identification	How to verify identifying assumption?	Which effect is estimated?
OLS with controls	$Y_i = \beta_0 + \boldsymbol{\beta_1} X_i + \gamma C_i + \epsilon_i$ The slope coefficient, $\boldsymbol{\beta_1}$, tells us how much we expect Y to change for a one-unit change in X, conditional on controls C.	Conditional on controls, in the absence of the treatment, the outcome of the treated group would have been the same as the outcome of the control group	$E[\epsilon_i \mid X_i] = Cov(X_{i_i} \epsilon_i) = 0$ $E[Y_i(0) \mid T_i = 1] = E[Y_i(0) \mid T_i = 0]$	OVB, classical measurement error, reverse causality, misspecfication of functional form, systematic measurement error, sample selection: Anything that causes the error to be correlated with X!	• Add set of sensible controls, see how coefficient changes • Come up with plausible stories	ATT
RCT with perfect compliance	Estimate with OLS	Treatment is, conditionally, randomly assigned – same as in OLS.	Same as in OLS	Same as in OLS	Balance table: At baseline, check whether there are significant differences between treatment and control.	ATT = ATE
IV	• First stage: $X_i = \alpha_0 + \alpha_1 Z_i + \pi C_i + u_i$ • Reduced form: $Y_i = \gamma_0 + \gamma_1 Z_i + \eta C_i + v_i$ • Target regression: $Y_i = \beta_0 + \boldsymbol{\beta_1} X_i + \delta C_i + \epsilon_i$ IV coefficient is first stage divided by reduced form. Estimated with 2SLS, esp. for multiple instruments.	We are comparing units with different levels of the treatment X, where the only difference we use is coming from differences in the instrument Z. 1. No defiers / monotonicity 2. Instrument relevance 3. Instrument exogeneity 4. Exclusion restriction	1. $X_i(1) \geq X_i(0)$ 2. $Cov(X_i, Z_i) \neq 0$ 3. $Cov(Z_i, \epsilon_i) = 0$ 4. $Cov(Z_i, \epsilon_i) = 0$ Mathematically speaking, exogeneity and exclusion restriction are the same.	Violations of IV assumptions.	1. Has to be assumed. 2. F-test>10, with one instrument for one endogenous variable: square of t-statistic 3. Balance table 4. Use brain	LATE LATE = ATT if there are no always-takers
RCT with imperfect compliance	Use assignment to treatment group as instrument for actual treatment status. Estimate with IV. The Reduced form estimates ITT.	Same as IV	Same as IV	Same as in IV & RCT	Same as IV & RCT	ITT (reduced form) or: LATE (= ATT if there are no always-takers)
DiD	In 2x2 case, take differences in table. Regression implementation: $Y_{it} = \alpha_0 + \alpha_1 T_i + \delta A_t + \boldsymbol{\beta} T_i \times A_t + \epsilon_{it}$, i.e. use an indicator for treatment group, an indicator for after time period, and interaction. $\boldsymbol{\beta}$ is the DiD coefficient.	Parallel Trends: In the absence of the treatment, the outcome of the treatment group would have changed by the same amount as the outcome of the control group. DiD removes any bias from the fact that: Individuals are inherently different from each other, or specific time periods are inherently different from each other.	$E[Y_{i0}^T] - E[Y_{i0}^C] = E[Y_{i0}^T(0)] - E[Y_{i0}^C(0)]$ $Cov(X_{it}, \epsilon_{it}) = 0$	Violations of parallel trends: 1. Something changed over time that affected treated and control group differently 2. The effect of X on Y changed over time, in way that is correlated with X	• Look at pre-trends: If they are not parallel, there is a problem • But we cannot prove this, ergo: Use brain	ATT, but estimated within groups, not between
Fixed Effects	Essentially the same as in DiD (for two-way fixed effects): $Y_{it} = \alpha_0 + \alpha_i + \delta_t + \boldsymbol{\beta} T_{it} + \epsilon_{it}$ α_i are unit fixed effects (dummy variables), δ_t are time fixed effects. Estimate with a seriers of dummy variables, or by demeaning. Can only estimate effects for time-varying treatments.	No time-changing confounders, which is the same thing as DiD, or as OLS with controls, but only using time-changing confounders.	$Cov(X_{it}, \epsilon_{it}) = 0$ which is the same assumption as in DiD.	Time-changing confounders: Did anything happen over time that affected T and C differently? Did the effect of X on Y change over time in a way that differs across units?	• Same as DiD • Also the same as OLS with controls, but look at time-changing confounders	ATT, But estimated within groups, not between
Sharp RD	$Y_{it} = \alpha_0 + f(X_i) + \boldsymbol{\beta} T_i + \epsilon_i$ Where $f(X_i)$ is a flexible functional form – linear or polynomial – of the running variable that determines the treatment. T_i is an indicator for whether i is above the threshold. Need to choose bandwitdth, the "window" we use to estimate the size of the jump.	There is no jump, expect our jump, at the jump, i.e., at the threshold, the only thing that jumps is the probability of treatment, i.e. 1. Probability of getting treatment jumps at threshold 2. No other pre-determined variable jumps at the threshold 3. The only way that being above the threshold affects the outcome is by changing the probability of getting the treatment. No strategic behavior / bunching around threshold!		• Other baseline variables jump at threshold • Strategic behavior and bunching around threshold (information) • Threshold is also used for other purposes	• Run RD (or make plot) with other baseline variables • Use brain: If eligibility for program X and for program Y use the same threshold, then not identified • Look at plots to identify bunching/strategic behavior around threshold	ATT
Fuzzy RD is just IV	Estimate with IV/2SLS. Use "being above the threshold" as instrument for the treatment. In all specifications, control for running variable.	Same as Sharp RD, which is same as IV (verify that the two sets of conditions are exactly the same!)		• Same as sharp RD = same as IV	• Same as sharp RD • Same as IV	LATE (= ATT if there are no always-takers, i.e. if no one below the threshold gets the treatment)

Terminology

OLS: Ordinary Least Squares
RCT: Randomized controlled Trial
DiD: Differences-in-Differences
IV: Instrumental variables
RD: Regression discontinuity

ATE: Average Treatment Effect
ATT = ToT: Average Treatment Effect for the Treated
LATE: Local Average Treatment Effect = Treatment Effect for the Compliers

Numbers in parentheses denote **potential outcomes.**
Subscripts with numbers denote **time periods**.

© Jonathan Old

Reading a regression table

Outcome: Happiness	(1)	(2)
Gender $_i$	0.834	0.614
	(0.032)	(0.045)
Education $_i$		−0.739
		(0.036)

Notes: Estimations contain a constant term. Standard errors in parentheses.

- Gender is a dummy variable. Average happiness of individuals with gender=1 is 0.834 higher than of individuals with gender=0. This is significant at the 5% level, because the coefficient is more than twice the standard error
- Conditional on education (keeping education fixed/comparing people with the same level of education), gender=1 is associated with 0.614 higher happiness.
- Conditional on gender, each additional year of education is associated with 0.739 less points on the happiness scale.

Omitted Variable Bias

Let Y_i be the outcome variable, X_i our regressor of interest, W a series of control variables, and Z_i the "omitted" variable.

$$\text{[Long regression]} \quad Y_i = c_1 + \beta_L X_i + \lambda Z_i + W\pi + e_i$$

$$\text{[Short regression]} \quad Y_i = c_2 + \beta_S X_i + W\pi + u_i$$

$$\text{[Auxiliary regression]} \quad Z_i = c_3 + \pi_1 X_i + W\pi + v_i$$

Then, the **Omitted variable bias formula** states that:

$$\underbrace{\beta_S}_{\text{Short}} = \underbrace{\beta_L}_{\text{Long}} + \underbrace{\lambda}_{\text{Omitted}} \cdot \underbrace{\pi_1}_{\text{Included}}$$

The OVB formula describes what happens to our coefficient of interest, β, as we include one additional variable Z in the regression. We call $\lambda\pi_1$ the **omitted variable bias**. Direction of bias: multiply our guesses for the signs of δ and γ.

Good and Bad controls

- Some controls are called ”bad controls”. These are:
  1. Variables that are themselves outcomes of a treatment: Treatment $\rightarrow$ Bad Control
  2. Variables that moderate the treatment effect: Treatment $\rightarrow$ Bad Control $\rightarrow$ Outcome
- Rule of Thumb: Good controls are either pre-determined or immutable characteristics.
- Another way to think about it: Controls help us make ”apples to apples” comparisons. We want to compare units that, in the absence of the treatment, would have the same outcomes, and differ only because they have different levels of the treatment.

Logs: Cheatsheet

Model	LHS	RHS	A change in x by ...	is associated with a change in y by ...
Level-Level	y	x	1 unit	β_1 units
Level-Log	y	$\log(x)$	1%	$\beta_1/100$ units
Log-Level	$\log(y)$	x	1 unit	$100\beta_1$%
Log-Log	$\log(y)$	$\log(x)$	1%	β_1%

If you want to get a bonus star from me, write ”approximately” in log-interpretations.

Hypothesis testing

$$\left| \frac{\hat{\beta}}{\text{SE}(\hat{\beta})} \right| \geq 1.96$$

$$\Leftrightarrow \left| \text{t-stat} \right| \geq 1.96$$

$$\Leftrightarrow \text{p-value} \leq 0.05$$

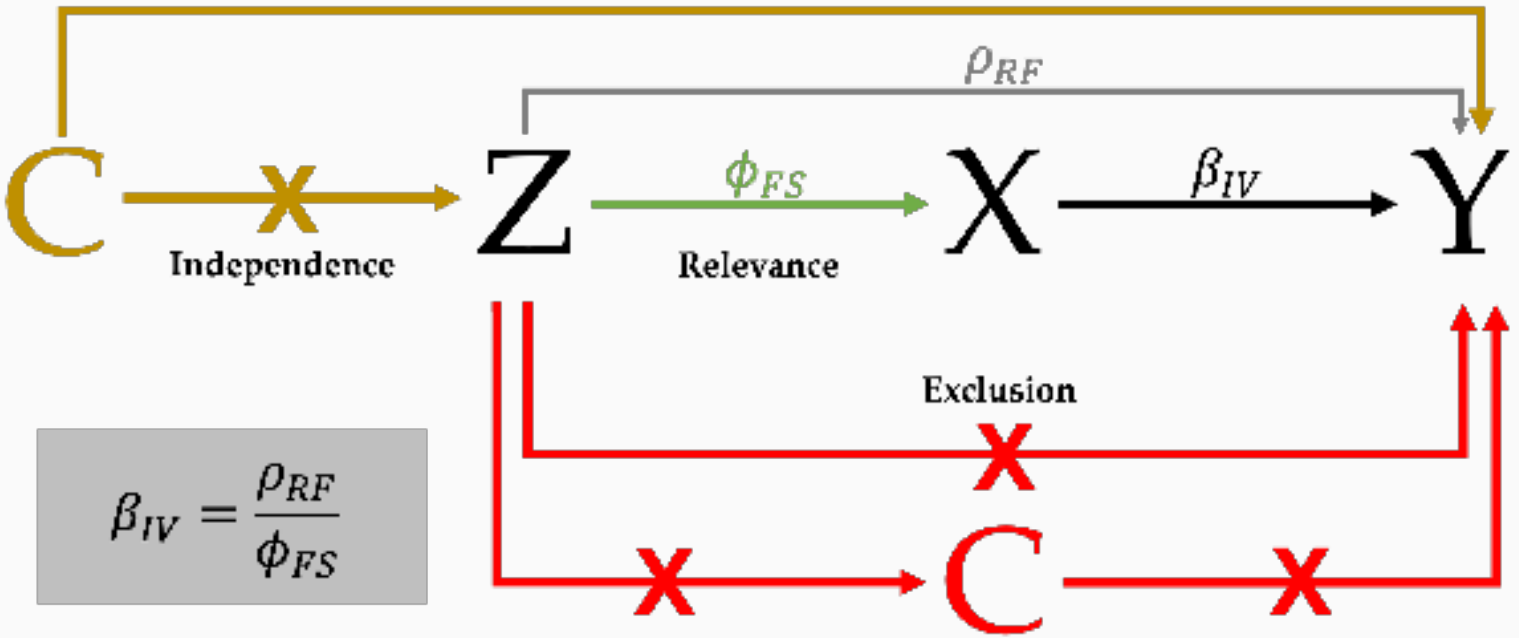
$$\Leftrightarrow 0 \notin \text{CI}$$

When testing the null hypothesis H_0: $\beta = 0$ (against the alternative hypothesis H_0: $\beta \neq 0$), then if any of these conditions holds, we reject the null.

IV summary

We need the following three assumptions for IV to work:

1. **Relevance**: *Z* must truly affect *X*
2. **Independence/Exogeneity**: *Z* is as good as randomly assigned
3. **Exclusion Restriction**: The **only** way that *Z* affects *Y* is via *X*.



$$\beta_{IV} = \frac{\rho_{RF}}{\phi_{FS}}$$

IV: The LATE is the treatment effect for the compliers

Potential outcomes! (unobserved)		***Does not get voucher (Z=0)***	
Gets voucher (Z=1)		*Eats chocolate (D=1)*	*Does not eat chocolate (D=0)*
	Eats chocolate (D=1)	Always-takers: $E(D\|Z=1)=E(D\|Z=0)=1$ ➔ $E(Y\|Z=1)=E(Y\|Z=0)$	Compliers
	Does not eat chocolate (D=0)	Defiers	Never-takers: $E(D\|Z=1)=E(D\|Z=0)=0$ ➔ $E(Y\|Z=1)=E(Y\|Z=0)$

Differences in Differences

	Before	After
Treated	$\alpha + \beta$	$\alpha + \beta + \gamma + \delta$
Untreated	α	$\alpha + \gamma$

We can calculate the DiD estimate in two ways:

$$DiD = E[\underbrace{\left(Y_{i1}^T - Y_{i0}^T\right)}_{\text{Change for treated}} - \underbrace{\left(Y_{i1}^C - Y_{i0}^C\right)}_{\text{Change for untreated}}] = [(\alpha + \beta + \gamma + \delta) - (\alpha + \beta)] - [(\alpha + \gamma) - (\alpha)]$$

$$= E[\underbrace{\left(Y_{i1}^T - Y_{i1}^C\right)}_{\text{After-difference}} - \underbrace{\left(Y_{i0}^T - Y_{i0}^C\right)}_{\text{Before-difference}}] = [(\alpha + \beta + \gamma + \delta) - (\alpha + \gamma)] - [(\alpha + \beta) - (\alpha)]$$

We need to assume parallel trends **after the treatment** for causality. Verify using data **before the treatment**. Estimate DiD with regression:

$$Y_{it} = \alpha + \beta \text{ Treated }_i + \gamma \text{ Post }_t + \delta \text{ Treated }_i \cdot \text{ Post }_t + u_{it}$$

To generalize, estimate model with unit fixed effects α_i and time FE δ_t:

$$Y_{it} = \alpha_i + \delta_t + \beta^{FE} \text{ X }_{it} + u_{it}$$

Regression discontinuity designs

Setup: We need a running variable X, impacting a treatment D discontinuously at a threshold, and an outcome Y.

Assumption: The effect of X and any other variable C on outcome Y is smooth around the discontinuity. In particular: No strategic behavior around cutoff. Can do a "balance check" around threshold.

Sharp RD: Probability of D jumps from 0 to 1 at threshold. To estimate:

$$Y_i = \alpha + \beta_1 X_i + \beta_2 X_i \cdot D_i + \beta^{RD} D_i + \varepsilon_i$$

Fuzzy RD: Probability jumps at threshold. Then, we simply use IV and instrument for D with being just above the threshold.

How to approach an exam question

1. Think: About the question, about the real world
2. Start with the numbers you see
    - One-sentence summary
    - Direction (positive or negative?)
    - Statistical significance (significant or insignificant, level?)
    - (Economic) magnitude (big or small?)
3. Then: Establish whether estimated relationship is causal or not
    - What do the results mean? Correlation (interesting) or causality (policy-relevant)
    - Is X-variable randomized? Do we have valid counterfactuals?
    - If not: Do you expect bias? Of which sort (OVB, reverse causality, bad controls, . . .)?
    - Find a plausible story for bias (using the OVB formula)

There are no traps!