

EC902/EC907: Quantitative Methods: Econometrics A

Lecture 1: Introduction

Manuel Bagues

Warwick University

October 10, 2022
Lecture Slides

Introduction

- Two groups of students
 - EC902 (MSc in Economics)
 - EC907 (MSc in Behavioural and Economic Science)
- Two tracks
 - Quantitative Methods A (EC902, more applied)
 - Quantitative Methods B (EC910, more mathematically oriented)
- Two different terms
 - 1st term: Microeconometrics (Manuel Bagues)
 - 2nd term: Macroeconometrics (Subham Kailthya) (just for EC902!)
- Main lectures (1st term)
 - Synchronous lectures: Mon: 16:00-18:00
 - Asynchronous lectures: Available on Wednesday each week
 - Watch the asynchronous lecture before the following synchronous lecture
 - In the synchronous lecture we will discuss its content

Introduction

- We will make polls frequently using [vevox.app](#)
 - ① Please, open [vevox.app](#) in your computer or download the app in your mobile
 - ② Enter session ID: 144-921-069
 - ③ Reply to the 1st question: [How are you feeling today?](#)

- Problem sets
 - Available on Wednesday each week
 - Discussed in the following week's tutorial
 - No need to hand them in...
 - ... but make sure you to prepare them!
 - Work on them in **study groups** (you can sign up in moodle)
- Tutorials
 - Tutorials: from week 3 on (next week)
 - 1st session: Stata exercise and Problem set 1
 - Bring laptop to face-to-face meetings!

- Participation in lectures and classes is highly encouraged!
 - Positive externality
 - Speak up in classes
- Feedback is very welcome (please don't be shy!)
 - Email me if something is not clear, you would like more examples about X, etc.
- My office hours:
 - Wednesdays 11:00-12:00 and 13:00-14:00
 - You can book a meeting using this link:
manuel-bagues.youcanbook.me
 - I am also available most days outside office hours (just email me in advance)

Evaluation: EC902

- Coursework (45%)
 - Introductory Mathematics and Statistics: test 1 (4%) and test 2 (6%)
 - **Midterm** (10%): 1 hour test, Wednesday Nov 16, 10:00 am
 - **Project** (25%): 3,000 word
 - Short practical exercise investigating some empirical question using the methods of the module
 - 3-member groups
 - The group sign up will be opened in week 7 of the Autumn Term
 - Confirm group membership by 09 December 2022 (end of 10th week)
 - Submission deadline: 20 March 2023, 12:00pm
 - More detailed information in this **moodle link**
- Exam (55%)
 - May

Evaluation: EC907

- Coursework (40%)
 - Introductory Mathematics and Statistics: test 1 (8%) and test 2 (12%)
 - Midterm (20%): 1 hour test, Wednesday Nov 16, 10:00am
- Exam (60%)
 - May

Course material

- Recordings of synchronous and asynchronous lecture
 - EC902 page (in echo 360)
- Slides
 - Available in **moodle** a few hours before each lecture
- Main textbooks:
 - Cunningham, Scott (2021), *Causal Inference: The Mixtape*, Yale University Press. (free online version available [here](#))
 - Angrist, J. and J.S. Pischke (2009), *Mostly Harmless Econometrics*, Princeton University Press. (ebook available following this [link](#))
 - You may also want to read first the baby version: Angrist, J. and J.S. Pischke (2014), *Mastering Metrics: The Path from Cause to Effect*, Princeton University Press.
- And if you need to refresh some basic knowledge of econometrics:
 - Wooldridge, J. (2013), *Introductory Econometrics: A Modern Approach*. South-Western College Publishing.

Statistical software: STATA

- Why STATA?
 - Easy to use!
 - Most applied researchers use it
 - Large on-line community
 - Drawback: proprietary software
- You can download it from the university:
Warwick.ac.uk/econ-stata
- Main alternatives to Stata
 - R
 - SPSS
 - Matlab
 - SAS
 - Stan
- Poll 2

Statistical software: STATA

- Very good introduction to stata available in moodle: [link](#)
- I will also introduce some STATA commands along the course.
- Useful for problem sets, project, dissertation & beyond
- STATA helpdesk for MSc students
- IT Services Training Team run STATA training courses in the Autumn Term. Courses can be booked online from the training portal: [link](#)

Structure of the course

- ① Introduction
- ② Randomized control trials (RCTs)
- ③ Identification based on observables (e.g. OLS regression controlling for observable characteristics)
- ④ Instrumental variables (IV)
- ⑤ Differences-in-differences (DID)
- ⑥ Regression discontinuity design (RDD)
- ⑦ Inference (e.g. how to cluster standard errors)
- Poll 3

Nobel prize in Economics 2021

The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2021



III. Niklas Elmehed © Nobel Prize Outreach.

David Card



III. Niklas Elmehed © Nobel Prize Outreach.

Joshua D. Angrist



III. Niklas Elmehed © Nobel Prize Outreach.

Guido W. Imbens

The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2021 was divided, one half awarded to David Card “for his empirical contributions to labour economics”, the other half jointly to Joshua D. Angrist and Guido W. Imbens “for their methodological contributions to the analysis of causal relationships.”

Readings for weeks 1-2 (available in moodle)

- Recommended readings:
 - Scientific background document for the 2021 Nobel Prize in Economics:
 - Answering causal questions using observational data
 - Causal Inference: The Mixtape (Cunningham)
 - Chapters 1 and 4
- Additional readings:
 - Harmless Econometrics (Angrist and Pischke)
 - Preface, Chapters 1-2

Introduction: types of questions

- Three types of basic questions
 - ➊ Descriptive
 - ➋ Prediction/forecasting
 - ➌ Causal
- This module focuses on causal questions...
- ... but descriptive/forecasting questions are also interesting!

1. Descriptive

There are at least three challenges for the econometrician addressing a descriptive question:

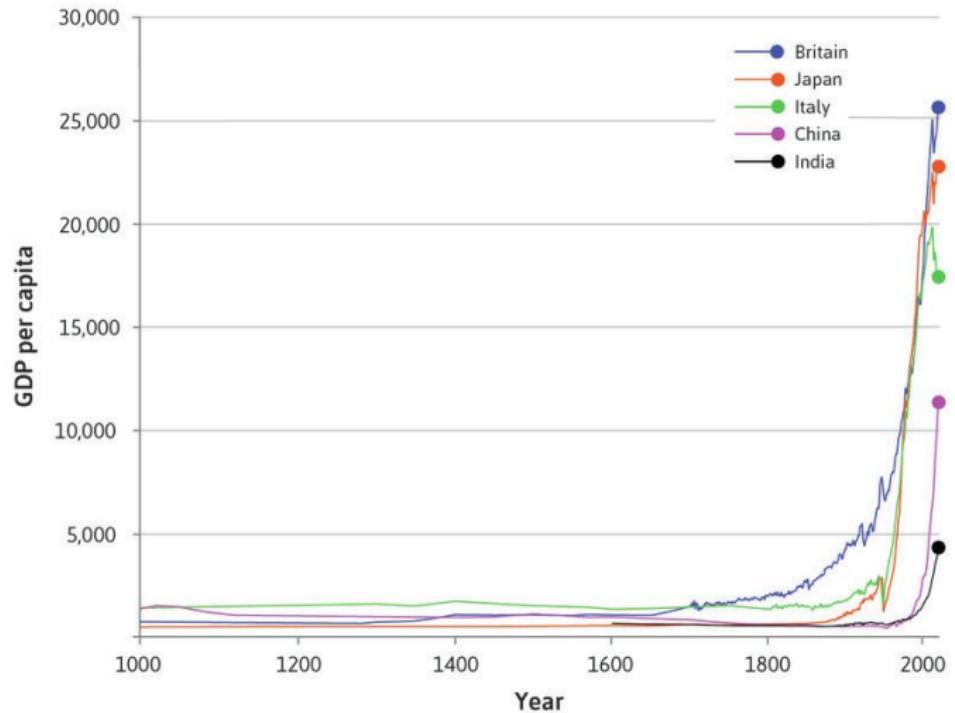
① Measurement

- Example 1: Economic growth during the last 2,000 years
- Example 2: Measure of political ‘slant’

1. Descriptive

Angus Maddison: [Maddison project](#)

History's hockey stick: Gross domestic product per capita in five countries (1000-2015).



1. Descriptive

Text-analysis

- Can we measure the ideological slant in news coverage?
 - Gentzkow and Shapiro (2010) measure the similarity of news outlet's language to that of a congressional Republican or Democrat.
 - ‘death tax’ vs. ‘estate tax’
 - ‘war on terror’ vs. ‘war in Iraq’

1. Descriptive

① Measurement

② Sampling

- We typically do not observe the full population but rather a sample. We want to make inferences about the population based on the sample.
 - Example: Survey on the labour market outcomes of Warwick University graduates

1. Descriptive

- ① Measurement
- ② Sampling
- ③ Summary Statistics

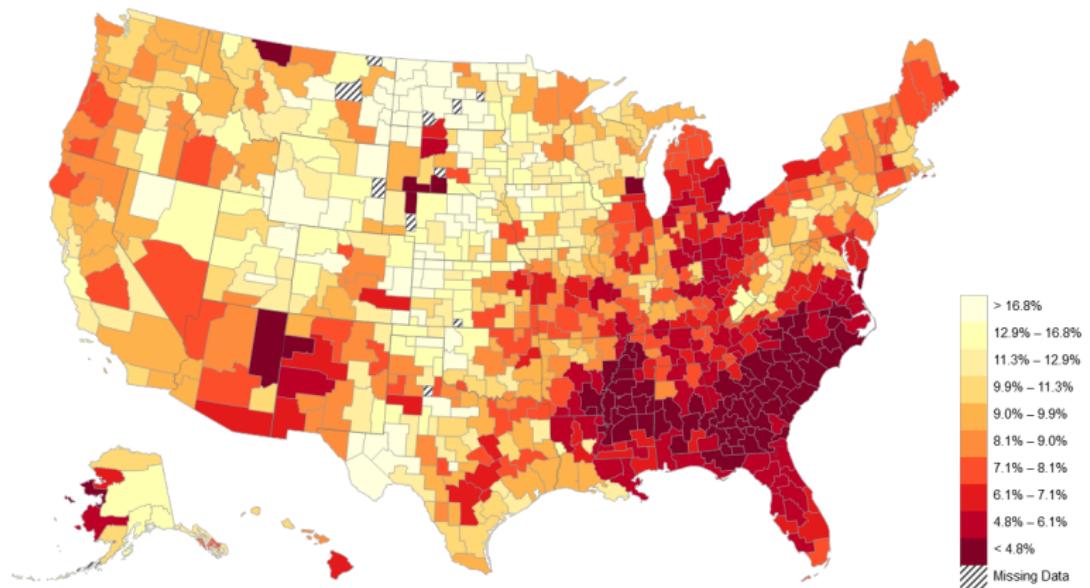
- Often the data for some of these questions is complicated and we need to find a nice way to summarize it
 - Example: Intergenerational income mobility

1. Descriptive

Chetty's Equality of Opportunity project

Intergenerational income mobility

Fraction of children who reach the top fifth of the national income distribution, conditional on having parents in the bottom fifth



2. Forecasting

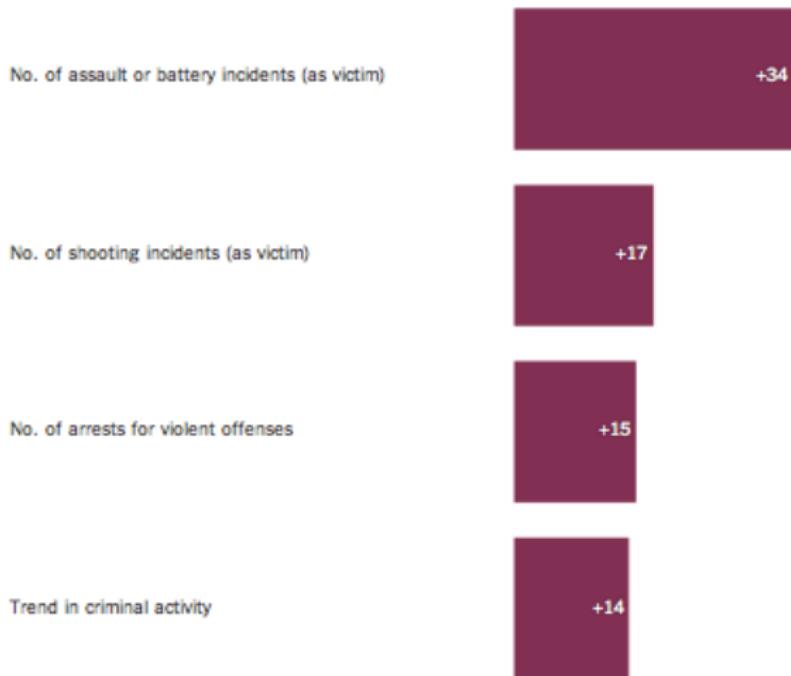
To predict future events

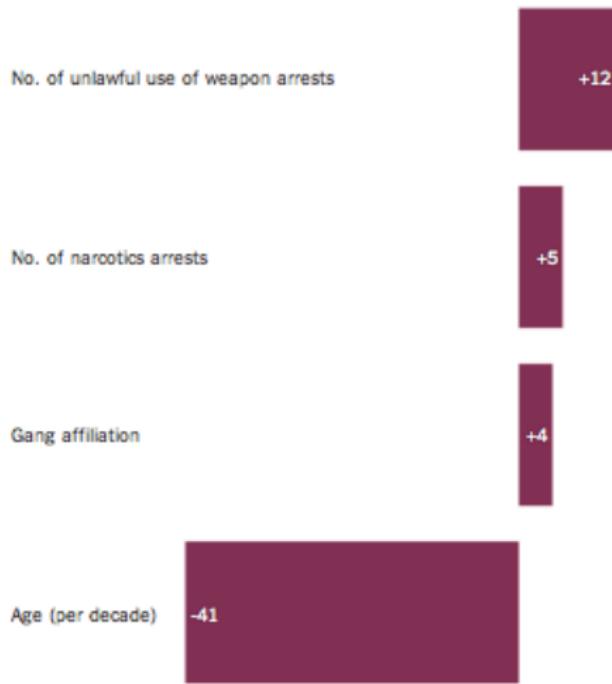
Examples:

- Future GDP growth/unemployment/inflation?
- Who will commit a crime?
 - Chicago's 'Minority report': Algorithm that tries to predict who is most likely to be involved in a shooting is used for roundups ([nytimes article](#))

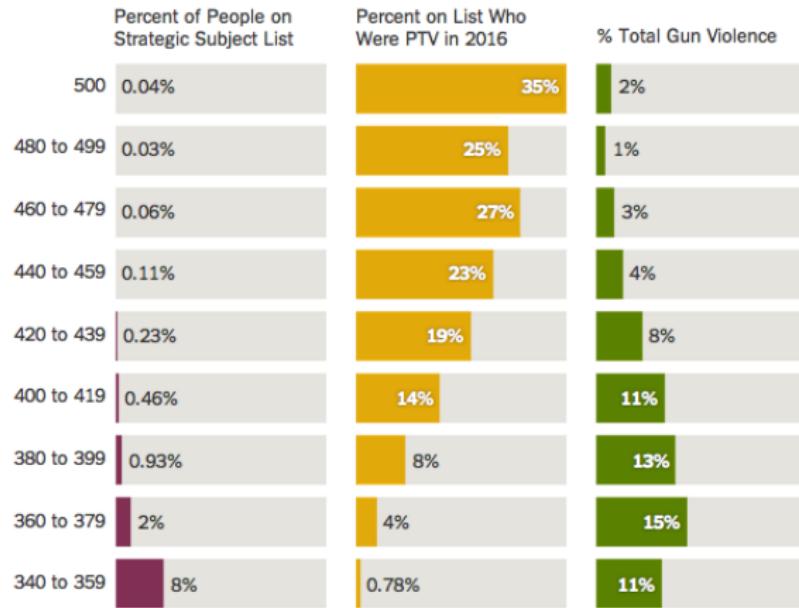
Biggest Risk Factors, and the Rewards of Age

The estimated impact of each characteristic on the final risk scores. The risk score declined by 41 points for every older 10-year age range (20 to 30, then 30 to 40, for example).





The system seems to predict crime quite accurately:



Note:

“Party to Violence” (PTV) meaning they were involved in a shooting or murder.

- These algorithms are spreading fast...
 - Already used by US courts to decide whether to release prisoners on parole
 - Firms
 - Technical and moral issues
 - Ex.: Car insurance and young men
- Big revolution linked to **machine learning** and **big data**

2. Forecasting

- Some times there are very high stakes to these questions:
 - If you can predict some small anomaly in the stock market you can potentially make a lot of money.

3. Causal effects

Two types of causal questions ([Gelman and Imbens 2013](#)):

- Reverse causal inference: search for *causes of effects* (Why?).
 - Why does Finland perform so well in standardized international education exams (e.g. PISA)?
- Forward causal questions: estimation of *effects of causes* (What if ...?).
 - Does teachers' IQ affect students performance?
 - Class size? Parents' involvement? Teachers' salaries or training?

3. Causal effects

Economists very often are motivated by *why* questions but, when they conduct their research, they tend proceed by addressing *what if* questions.

Examples:

- How does taking this module affect the grade that you will obtain in your master dissertation?
 - Note that this is different from the predictive question: "What is the grade that students taking this module will obtain with their master dissertation?"
- How does a positive or a negative facebook post affect your sentiment?
- Does death penalty decrease crime rates?
- Would it be profitable for a firm to allow employees to work from home? (**Marissa Mayer (Yahoo)** vs. Bloom)
- Are employees more satisfied if they are informed about the salaries of their colleagues? (**Card, Mas, Moretti and Saez 2012**)

Different types of questions

- One nice way to think about the difference between these three types of analysis:
 - Descriptive: If we had enough data we would know the answer.
 - Forecasting: If we had enough data and we wait long enough, we would know the answer.
 - Causality: Unless we have a plausible empirical strategy (e.g. an RCT, more on this later), we will never know the answer to this question.

Forecasting/prediction vs. causal questions

Do not confuse them!

- Example 1:

- Causal: How does taking this course affects the grade that you will obtain in your master thesis?
- Predictive: Grades in the master thesis for people taking *Quantitative Methods A* and those taking *Quantitative Methods B*?

- Example 2:

- Predictive: Do people that drink coffee have lower life expectancy?
- Causal: Would I live longer if I quit coffee?

Forecasting/prediction vs. causal questions

These are two different ‘ball games’

- Forecasting
 - Find a model with a high predictive power ($\uparrow R\text{-squared}$)
- Causal effect
 - ① Consistency of the estimate
 - we need some exogenous variation in the treatment (e.g. RCT, more on this later)
 - (in other words, the control group is comparable to the treatment group)
 - ② Coefficient of interest is precisely estimated
 - \downarrow standard error
 - Note: R-squared is ‘pretty much’ irrelevant in this case

Forecasting/prediction vs. causal questions

- Economists usually think about causal questions
 - We will spend the most time in this course thinking about causal relationships and trying to “identify” them.
- But you might also want to acquire elsewhere the skills necessary to address *prediction/forecasting* questions
 - EC994: Applications of Data Science
 - Note: elective available at the MSc in Economics (but not for the MSc in Behavioural and Economic Science)

Food for thought

- What is your question of interest? (For instance, for your master thesis research proposal)
- Is it a **descriptive, forecasting** or a **causal** question?

EC902/EC907: Quantitative Methods: Econometrics A

Lecture 2.1: Causal effects
(Asynchronous lecture)

Manuel Bagues

Warwick University

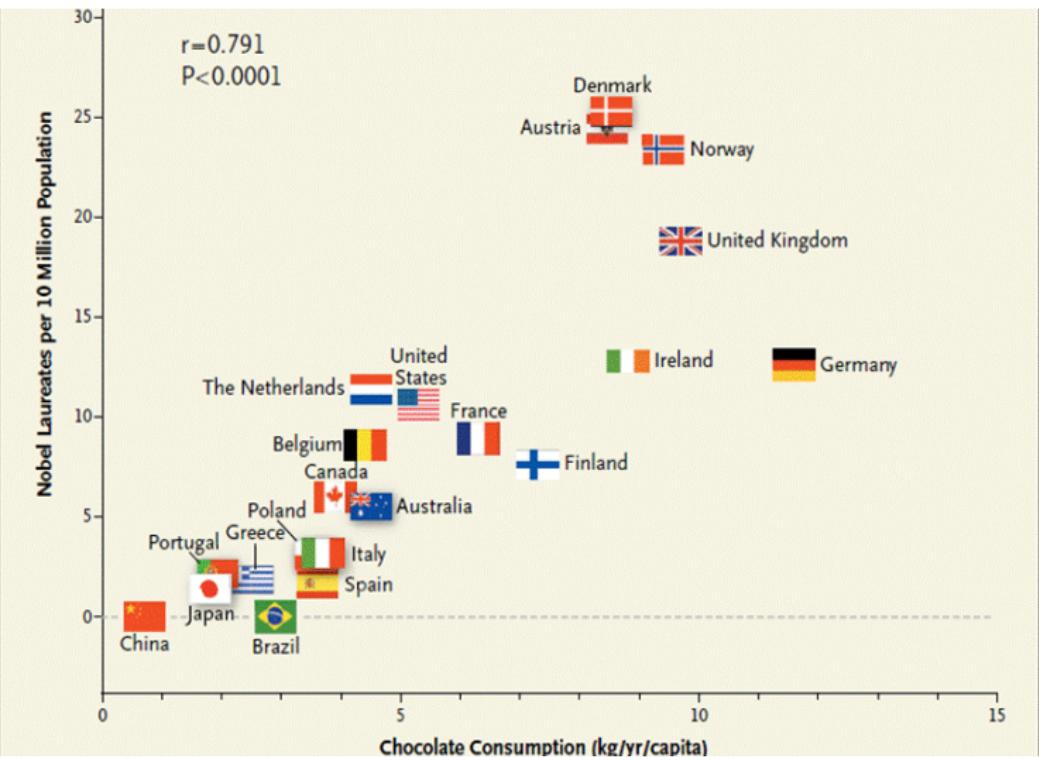
October 10, 2022
Lecture Slides

Roadmap

- Introduction
 - Types of questions → Synchronous lecture Oct. 10 (slides 1)
 - Descriptive
 - Predictive/forecasting
 - Causal
 - Correlation vs. causation → Asynchronous lecture Oct. 10(slides 2.1)
 - Potential outcomes framework → Asynchronous lecture Oct. 10 (slides 2.2)
 - Randomized control trials (RCTs) → Asynchronous lecture Oct. 10 (slides 2.3)

Correlation does not (necessarily) imply causation

- $\text{Cor}(x,y) \neq 0$
 - ① x implies y
 - ② y implies x
 - ③ z implies x and y
- This is obvious when we consider simple examples:
 - People that sleep less tend to live longer
 - Red cars are more likely to get involved in accidents
 - Countries that eat more chocolate receive more Nobel prizes
 - Messerli 2012

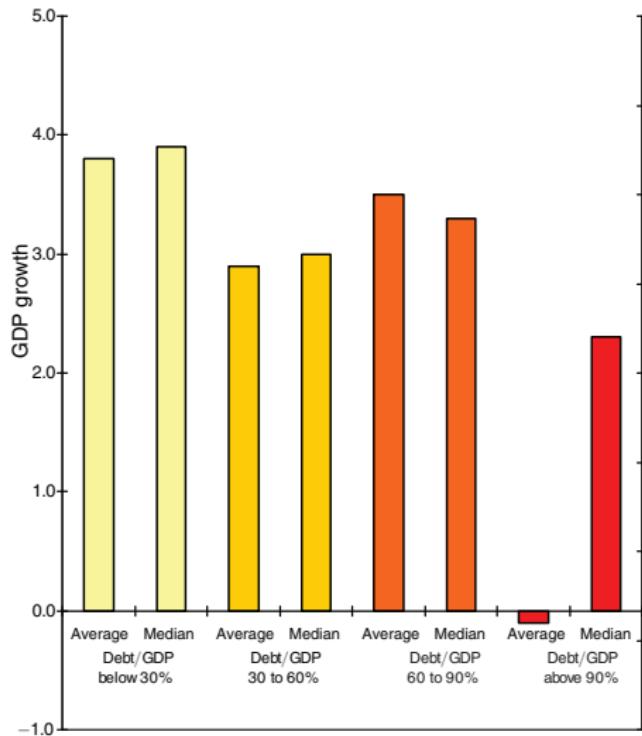


Misinterpreting descriptive evidence as causal

- Potentially confusing examples:
 - GDP growth and public debt

Does high public debt lead to lower economic growth?

Reinhart and Rogoff 2010



- This evidence has been interpreted by some observers as support for the austerity agenda .
- Paul Ryan (Path to Prosperity, 2013, p. 78):
 - *A well-known study completed by economists Ken Rogoff and Carmen Reinhart (...) found conclusive empirical evidence that gross debt exceeding 90 percent of the economy **has a significant negative effect** on economic growth.*
- What do you think?

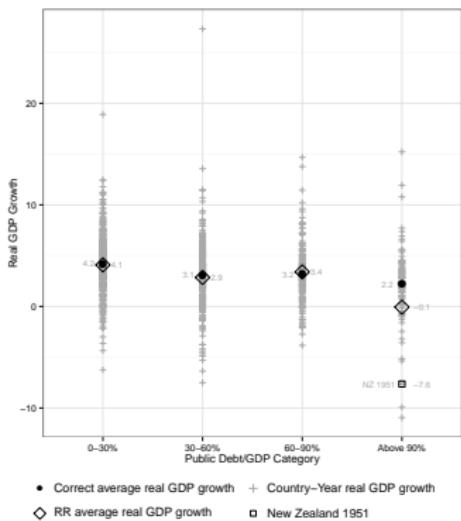
Does high public debt lead to lower economic growth?

- Two problems:
 - ① There might a reverse causality problem and/or an omitted variable problem
 - ② It turned out that the relationship was not that strong:
 - Herndon, Ash and Pollin 2013: *We replicate Reinhart and Rogoff (2010A and 2010B) and find that selective exclusion of available data, coding errors and inappropriate weighting of summary statistics lead to serious miscalculations that inaccurately represent the relationship between public debt and GDP growth among 20 advanced economies.*

Is high public debt associated to lower economic growth?

Reinhart and Rogoff 2010, Herndon, Ash and Pollin 2013

Figure 1: Real GDP growth by public debt/GDP categories, country-years, 1946–2009



Notes. The unit of observation in the scatter diagram is country-year with real GDP growth plotted against four debt/GDP categories. Our replication of RR published values for average real GDP growth within category are printed to the right. Corrected values for average real GDP growth within category are printed to the left.

Source: Authors' calculations from working spreadsheet provided by RR.

- Other potentially confusing examples:
 - Students who own a laptop tend to perform better in PISA evaluations
 - Women in boards - the glass cliff ([Ryan and Haslam 2007](#))
 - Countries where a higher share of working-age families live with their parents have a higher covid-19 fatality rate ([Bayer and Kuhn 2020](#))

y can cause x even if x takes place before y

- Note also that y can cause x even if x takes place before y
- Example (i): rain
 - When many people carry their umbrellas in the morning usually it rains in the afternoon
 - Not a bad idea to use this fact to predict rain...
 - ... however, subsidizing umbrellas is not a great policy to increase rain
- Example (ii): Impact of electoral results on the stock market
 - markets expect an extremist party to win the election → stock market experiences a negative impact before the election

THE FAMILY CIRCUS



8-5
© 1988 Bill Amend Inc.
Used by Courtesy of United Feature Syndicate, Inc.

"I wish they didn't turn on that seatbelt sign so much! Every time they do, it gets bumpy."

EC902/EC907: Quantitative Methods: Econometrics A

Lecture 2.2: Potential outcomes framework
(Asynchronous lecture)

Manuel Bagues

Warwick University

October 10, 2022
Lecture Slides

Roadmap

- Introduction
 - Types of questions → Synchronous lecture (slides 1)
 - Descriptive
 - Predictive/forecasting
 - Causal
 - Correlation vs. causation → Asynchronous Oct. 10 (slides 2.1)
 - Potential outcomes framework → Asynchronous lecture Oct. 10 (slides 2.2)
 - Random assignment → Asynchronous lecture Oct. 10 (slides 2.3)

Potential outcomes framework

Rubin causal model

- How should we interpret observational evidence?
observational evidence = causal effect + selection bias
- The impact of the treatment may be heterogeneous
→ Average treatment effect on the treated (ATET) \neq Average treatment effect (ATE)

Potential outcomes framework

- Let's think of a treatment as a binary random variable
 $D_i = \{0, 1\}$
- Each individual has two potential outcomes (counterfactuals):

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

- The causal effect of treatment for an individual i is $Y_{1i} - Y_{0i}$
- Unfortunately, for i , we only observe Y_{1i} if $D_i = 1$ and Y_{0i} if $D_i = 0$.

Observed difference in outcomes versus the causal effect of the treatment

- Let us compare the two things that we may observe: what happens to the treatment group when they are treated ($E[Y_{1i}|D_i = 1]$) vs. what happens to the control group when they are not treated ($E[Y_{0i}|D_i = 0]$)
- How should we interpret this difference? What does it tell us about the causal effect?

$$\underbrace{E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{\text{Difference in outcomes between treated and non-treated individuals}} = \underbrace{E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1]}_{\text{Average treatment effect on treated (ATET)}} + \underbrace{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{\text{Selection bias}}$$

Observed difference in outcomes versus the causal effect of the treatment

- The difference in the observed outcomes (between the treated and the non treated) is equal to the **average treatment effect on the treated (ATET)** iff there is no selection effect:
- $\underbrace{E[Y_{0i}|D_i = 1] = E[Y_{0i}|D_i = 0]}_{\text{No selection}} \Leftrightarrow \underbrace{E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{\text{Difference in observed outcomes}} = \underbrace{E[Y_{1i} - Y_{0i}|D_i = 1]}_{\text{ATET}}$

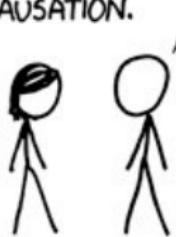
Observed difference in outcomes versus the causal effect of the treatment

- Is this a plausible assumption?
- Example:
 - Students attending Quantitative Methods A and Quantitative Methods B
 - Unfortunately, selection is likely to happen unless agents (i) are unaware of the impact of the treatment or (ii) they are irrational.

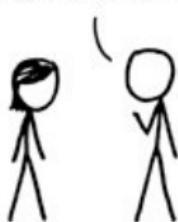
Note: ATET vs. ATE

- Average treatment effect on the treated (ATET):
 $E[Y_{1i} - Y_{0i}|D_i = 1]$
- will differ from the population average treatment effect (ATE):
 $E[Y_{1i} - Y_{0i}]$
- iff $E[Y_{1i} - Y_{0i}|D_i = 1] \neq E[Y_{1i} - Y_{0i}|D_i = 0]$
- The effect of the treatment may be heterogeneous and the distinction is often important (also for policy makers!)
 - Example: effect of taking this course for the *treated, non-treated* and *average individual*
- Different empirical methods identify the impact of the treatment on different groups of individuals
 - E.g.: RCT → ATE, DID → ATET, IV → LATE etc
 - We should always be aware of the type of effect that we are identifying

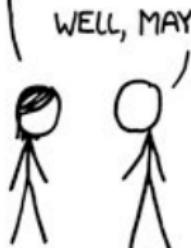
I USED TO THINK
CORRELATION IMPLIED
CAUSATION.



THEN I TOOK A
STATISTICS CLASS.
NOW I DON'T.



SOUNDS LIKE THE
CLASS HELPED.
WELL, MAYBE.



EC902/EC907: Quantitative Methods: Econometrics A

Lecture 2.3: Randomized control trials (RCTs)
(Asynchronous lecture)

Manuel Bagues

Warwick University

October 10, 2022
Lecture Slides

Roadmap

- Introduction
 - Types of questions → Synchronous lecture Oct. 10 (slides 1)
 - Descriptive
 - Predictive/forecasting
 - Causal
 - Correlation vs. causation → Asynchronous lecture Oct. 10 (slides 2.1)
 - Potential outcomes framework → Asynchronous lecture Oct. 10 (slides 2.2)
 - Randomized control trials (RCTs) → Asynchronous lecture Oct. 10 (slides 2.3)

Randomized control trials (RCTs)

- How random assignment overcomes the selection effect
- Drawbacks
 - ① Feasibility
 - ② Internal validity
 - ③ External validity

How can we estimate a causal relationship?

(*How can we overcome the selection effect?*)

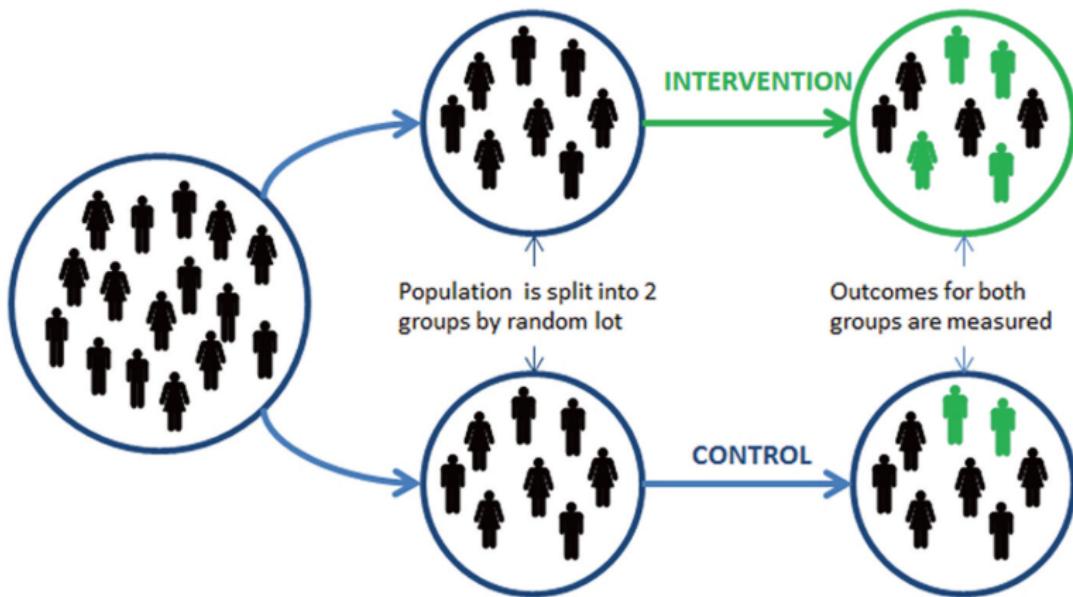
- We want to understand what would have happened to treated individuals if they were not exposed to the treatment
- The “Scientific” solution vs. the “Statistical” solution
 - In the physical sciences one can often answer this type of question by running a ‘scientific’ experiment.
- Example: **Galileo’s Leaning Tower of Pisa experiment**
 - Aristotle’s theory of gravity: objects fall at speed relative to their mass
 - Throw two balls of different weight from the tower

Leaning Tower of Pisa



- Properties:
 - Temporal stability: the response does not change if the time when a treatment is applied is varied slightly.
 - Causal transience: the response of one treatment is not affected by prior exposure of the unit to the other treatment.
 - Unit homogeneity: units are homogeneous with respect to treatments and responses.
- Unfortunately in Social Sciences none of these assumptions is plausible. Instead we use the statistical solution.
- Randomized experiment: Setting where the assignment mechanism does not depend on characteristics of the units, either observed or unobserved, and the researcher has control over the assignments

Random assignment



Random assignment

- Random assignment solves the selection problem since it makes D_i independent of potential outcomes, hence:

$$\begin{aligned}E[Y_{1i}|D_i = 1] &= E[Y_{1i}|D_i = 0] \\E[Y_{0i}|D_i = 1] &= E[Y_{0i}|D_i = 0]\end{aligned}$$

- Therefore, a comparison of the treatment and control group outcomes provides information about the causal impact of the **average treatment effect on the treated (ATET)**:

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

Random assignment

- Moreover, since the composition of the treatment and the control group is similar, the ATET is equal to the population average treatment effect (ATE):

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

Note: With other empirical methods very often the ATET will not be equal to the ATE.

Stable unit treatment value assumption (SUTVA)

- Note that, in the above discussion, we have made an important implicit assumption called the **Stable Unit Treatment Value Assumption (SUTVA)**
- We have assumed ‘the [potential outcome] observation on one unit should be unaffected by the particular assignment of treatments to the other units’.
- This assumption may not always be plausible (e.g. there are general equilibrium effects), more on this below.
- Textbook example: vaccines

Types of randomized experiments

Lab vs. field

- Lab vs. field experiments
 - Lab: The effect of feedback on relative performance in the lab ([Azmat and Iriberry 2010](#))
 - Field: The effect of feedback on relative performance in a university ([Azmat, Bagues, Cabrales and Iriberry 2019](#))

Example: face masks

- N95 respirator face masks are designed to protect the wearer by filtering out 95% of airborne particles that measure 0.3 micrometers and larger.
- Why do we need an experiment?
 - Christine Benn and co-authors experiment in Guinea Bissau ([link to Nature article](#))

Types of randomized experiments

Stratified randomization

- Completely randomized vs stratified randomized experiments
 - To ensure that the control and treatment groups are very similar, you may partition the covariate space and assign treatment randomly to units within a cluster
- Example
 - We could randomly assign 50% of the class to the treatment group, or we could first classify individuals according to their nationality and then assign 50% of individuals within each nationality to the treatment (the latter will ensure that treatment and control are perfectly balanced in terms of nationality)

Potential drawbacks of RCTs

① Feasibility

② Internal validity

- Refers to the ability of a study to estimate causal effects within the study population.

③ External validity

- It is concerned with generalizing causal inferences, drawn for a particular population and setting, to others, where these alternative settings could involve different populations, different outcomes, or different contexts

Potential drawbacks of RCTs

Feasibility

- Problems of implementation
 - Not possible to randomize many relevant treatments (e.g. monetary policy)
 - Cost (e.g. impact of mediterranean diet)
 - Political issues (policy makers need to acknowledge ignorance),...
 - Compliance and attrition
- Ethical issues
 - Examples: STAR, babies...
 - The ethical argument is not obvious when (i) the treatment cannot be applied to everybody (maybe due to some budget constraints) and (ii) the optimal assignment rule is unknown.
 - Note: Universities typically have a Research Ethics Committee (also known as Institutional Review Board or IRB) that ensures that all research involving participants and/or their data addresses relevant ethical considerations.

Potential drawbacks of RCTs

Internal validity: Hawthorne effect

- **Hawthorne effect**

- The Illumination Experiment (Landsberger 1950, Levitt and List 2011)
- Audit study in France (Behaghel et al. 2015)

Potential drawbacks of RCTs

Internal validity: SUTVA

- Potential violation of the *Stable Unit Treatment Value Assumption (SUTVA)*
 - This framework is generally not well suited to the evaluation of system-wide reforms which are intended to have substantial general equilibrium effects.
- Example: Crepon, Duflo, Gurgand, Rathelot and Zamora 2013
Program of job placement assistance to young unemployed workers in France to improve their labor market outcomes
 - Can you think of potential ways in which the treatment may affect the control group? (hint: displacement effects)

Potential drawbacks of RCTs

- Solution → Clustered Randomized Experiment (or two-step randomized design)
 - ① 1st step: Partition of the covariate space, and randomize the number of units that are assign to treatment in each cluster
 - ② 2nd step: Randomized which units receive the treatment within each cluster
- Example: Crepon, Duflo, Gurgand, Rathelot and Zamora 2013
 - ① each of 235 local employment areas are randomly assigned a proportion P of job seekers to be assigned to treatment: either 0%, 25%, 50%, 75%, or 100%.
 - ② a fraction P of all the eligible job seekers is randomly selected to participate in the job placement program
- Interpretation of results?
 - ① Within each cluster, the treatment group has better labor market outcomes
 - ② However, on average job seekers are as likely to find a job in different areas

Potential drawbacks of RCTs

External validity

- Many authors are concerned with the potential lack of external validity of RCTs (e.g. Deaton, Manski...)
 - Treatment heterogeneity
- Problem also applies to non-experimental empirical strategies
- Possible solution:
 - Estimate heterogeneity of treatment effects and combine RCTs with structural models
 - Example: Duflo, Hanna and Ryan (AER 2012), ‘Incentives Work: Getting Teachers to Come to School’

EC902/EC907: Quantitative Methods: Econometrics A

Lecture 3 (synchronous)

Manuel Bagues

Warwick University

October 17, 2022
Lecture Slides

Roadmap so far:

- Synchronous lecture week 2:
 - Types of questions (slides 1)
- Reading first week:
 - Scientific background for the 2021 Nobel Prize in Economics
 - Introduction ‘Causal Inference: The Mixtape’
 - Introduction ‘Almost harmless Econometrics’ (*complementary reading*)
- Asynchronous lecture week 2:
 - 2.1. Correlation vs. causation (slides 2.1)
 - 2.2. Potential outcomes framework (slides 2.2)
 - 2.3. Randomized control trials (slides 2.3)
 - How they overcome the selection effect
 - Drawbacks
- Problem set 1
 - Available in moodle
 - Discussed in this week’s tutorial

Roadmap this week

- Structure:
 - Monday: synchronous lecture
 - Tuesday: asynchronous lectures
 - Wednesday: problem set 2
- Any questions so far?

Poll questions

- We will make polls frequently using [vevox.app](#)
 - ① Please, open [vevox.app](#) in your computer, in Teams or download the app in your mobile
 - ② Enter session ID: 144-921-069

Poll questions (i)

- Types of questions:
 - Have you watched the asynchronous lectures?
 - A. No, unfortunately I was too busy
 - B. I watched part of it, but not everything.
 - C. Yes, I watched everything!

Poll questions (ii)

- Types of questions:
 - In a recent paper titled “Historical Analysis of National Subjective Wellbeing using Millions of Digitized Books”, Daniel Sgroi from the University of Warwick and co-authors conduct a quantitative analysis of digitized text from millions of books published over the past 200 years to measure the evolution of national subjective well-being over the past two centuries in a number of countries.
 - Would you say that the authors are addressing a:
 - A. Descriptive question
 - B. Predictive question
 - C. Causal question

Poll questions (ii)

- Types of questions:
 - In a recent paper titled “Historical Analysis of National Subjective Wellbeing using Millions of Digitized Books”, Daniel Sgroi from the University of Warwick and co-authors conduct a quantitative analysis of digitized text from millions of books published over the past 200 years to measure the evolution of national subjective well-being over the past two centuries in a number of countries.
 - Would you say that the authors are addressing a:
 - A. Descriptive question
 - B. Predictive question
 - C. Causal question
 - D. Stupid question

Correlation and causation

Slides 2.1, see also Mixtape chapter 1)

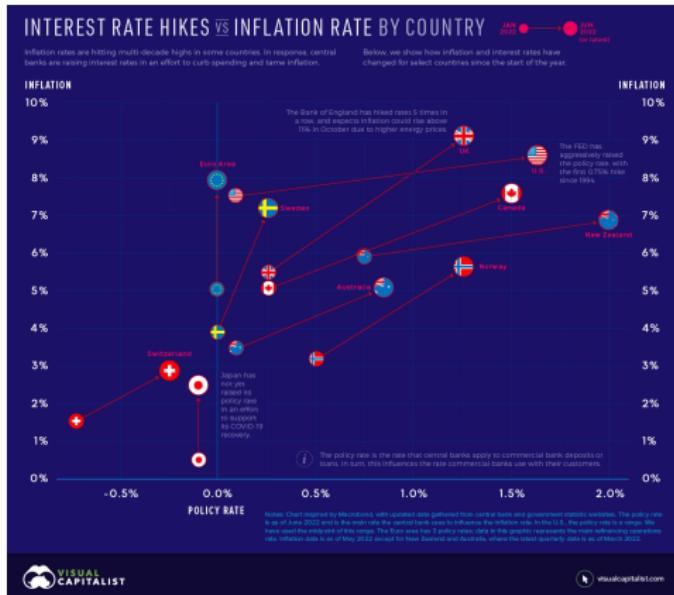
- When we rely on observational data (i.e. non-experimental), correlation is unlikely to mean causation
- Example: Erdogan and the relationship between interest rates and inflation

Erdogan's views on interest rates and inflation

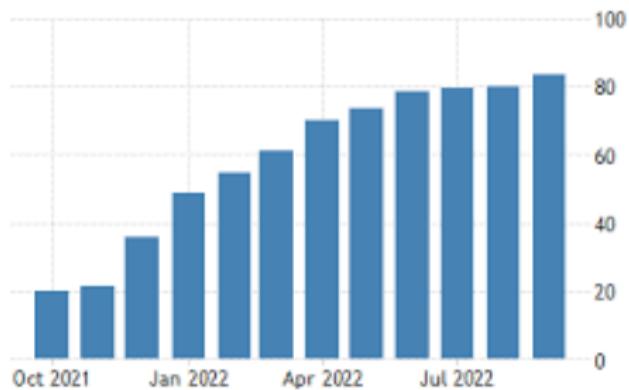
Newspaper article on 29 Nov 2021

- *President Recep Tayyip Erdogan has once again reiterated his unconventional hypothesis that reducing interest rates will lead to lower inflation.*
- *"Inflation is the result, high interest rates are the cause," Erdogan said*
- *Speaking to reporters on his flight back from Turkmenistan, Erdogan said that he studied economics and that he will never give up on his stance against high interest rates.*
- *Erdogan gave the United States and Israel as examples to prove his point.*
 - "When we take a look at these countries, our thesis' correctness can be seen," he said.

Interest rates and inflation: cross-country correlation



Inflation in Turkey



Potential outcomes framework

slides 2.2, see also Mixtape chapter 4

$$\underbrace{E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{\text{Difference in outcomes between treated and non-treated individuals}} =$$

$$\underbrace{E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1]}_{\text{Average treatment effect on treated (ATET)}} + \underbrace{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{\text{Selection bias}}$$

Potential outcomes framework

ATET vs. ATE

- Average treatment effect on the treated (ATET):
 $E[Y_{1i} - Y_{0i}|D_i = 1]$
- may differ from the population average treatment effect (ATE):
 $E[Y_{1i} - Y_{0i}]$
- The effect of the treatment may be heterogeneous and the distinction is often important (also for policy makers!)

Potential outcomes framework

Numerical hypothetical example

Imagine a population with 4 units:

i	Y_{1i}	Y_{0i}	Y_i	D_i	$Y_{1i} - Y_{0i}$
1	3	0	3	1	3
2	1	1	1	1	0
3	1	0	0	0	1
4	1	1	1	0	0

Note: numbers in red reflect magnitudes that are not observable (e.g. what would have happened to a treated individual in the absence of the treatment) but which, for the sake of the example, we pretend to know.

Question 1: $E[Y_{i1}/D_i = 1] - E[Y_{i0}/D_i = 0]$?

Potential outcomes framework

Numerical hypothetical example

Imagine a population with 4 units:

i	Y_{1i}	Y_{0i}	Y_i	D_i	$Y_{1i} - Y_{0i}$
1	3	0	3	1	3
2	1	1	1	1	0
3	1	0	0	0	1
4	1	1	1	0	0

Note: numbers in red reflect magnitudes that are not observable (e.g. what would have happened to a treated individual in the absence of the treatment) but which, for the sake of the example, we pretend to know.

Question 1: $E[Y_{i1}/D_i = 1] - E[Y_{i0}/D_i = 0]$?

$$E[Y_{i1}/D_i = 1] - E[Y_{i0}/D_i = 0] = \frac{3+1}{2} - \frac{0+1}{2} = 1.5$$

i	Y_{1i}	Y_{0i}	Y_i	D_i	$Y_{1i} - Y_{0i}$
1	3	0	3	1	3
2	1	1	1	1	0
3	1	0	0	0	1
4	1	1	1	0	0

Question 2: $\alpha_{ATE} = E[Y_1 - Y_0]$?

i	Y_{1i}	Y_{0i}	Y_i	D_i	$Y_{1i} - Y_{0i}$
1	3	0	3	1	3
2	1	1	1	1	0
3	1	0	0	0	1
4	1	1	1	0	0

Question 2: $\alpha_{ATE} = E[Y_1 - Y_0]$?

$$E[Y_1 - Y_0] = \frac{3 + 0 + 1 + 0}{4} = 1$$

i	Y_{1i}	Y_{0i}	Y_i	D_i	$Y_{1i} - Y_{0i}$
1	3	0	3	1	3
2	1	1	1	1	0
3	1	0	0	0	1
4	1	1	1	0	0

Question 3: $\alpha_{ATE} = E[Y_1 - Y_0 | D_i = 1]$?

i	Y_{1i}	Y_{0i}	Y_i	D_i	$Y_{1i} - Y_{0i}$
1	3	0	3	1	3
2	1	1	1	1	0
3	1	0	0	0	1
4	1	1	1	0	0

Question 3: $\alpha_{ATE} = E[Y_1 - Y_0 | D_i = 1]$?

$$E[Y_1 - Y_0 | D_i = 1] = \frac{3 + 0}{2} = 1.5$$

Randomized control trials

Slide 2.3

- Random assignment solves the selection problem since it makes D_i independent of potential outcomes \rightarrow no selection effect

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

- Drawbacks
 - Feasibility
 - Internal validity
 - External validity

Poll questions (iii)

- Randomized control trials identify:
 - A. average treatment effect on the non-treated
 - B. average treatment effect on the treated
 - C. all of the above
 - D. none of the above

Poll questions (iii)

- Randomized control trials identify:
 - A. average treatment effect on the non-treated
 - B. average treatment effect on the treated
 - **C. all of the above**
 - D. none of the above

Random assignment

- Note also that, since the composition of the treatment and the control group is similar, the **ATET** is equal to the **population average treatment effect (ATE)**:

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

Poll questions (iv)

- The Stable Unit Treatment Value Assumption (SUTVA) is violated:
 - A. when individuals in the treatment group are aware that they are part of an experiment
 - B. when individuals in the control group are also somehow affected by the treatment
 - C. when not everybody in the treatment group actually takes the treatment

Poll questions (iv)

- The Stable Unit Treatment Value Assumption (SUTVA) is violated:
 - A. when individuals in the treatment group are aware that they are part of an experiment
 - **B. when individuals in the control group are also somehow affected by the treatment**
 - C. when not everybody in the treatment group actually takes the treatment

Poll questions (v)

- Randomized control trials tend to have:
 - A. high internal validity
 - B. high external validity
 - C. all of the above

Poll questions (v)

- Randomized control trials tend to have:
 - A. **high internal validity**
 - B. high external validity
 - C. all of the above

Poll questions (vi)

- The Hawthorne effect
 - A. implies a violation of the Stable Unit Treatment Value Assumption (SUTVA)
 - B. is a potential concern when subjects are aware that they are part of an experiment
 - C. all of the above

Poll questions (vi)

- The Hawthorne effect
 - A. implies a violation of the Stable Unit Treatment Value Assumption (SUTVA)
 - **B. is a potential concern when subjects are aware that they are part of an experiment**
 - C. all of the above

Random assignment - Inference

- How do we estimate the ATE/ATET in an unbiased/consistent way?
 - We simply compare sample means
- Inference: precision of our estimates?
 - ① Large samples: Sampling-based approach → Two sample t-test
 - ② Small samples: Randomization inference → Fisher's exact test

EC902/EC907: Econometrics A

Lecture 4.1

Manuel Bagues

Warwick University

Roadmap

- Randomized controlled trials (RCT)
 - How random assignment overcomes the selection effect (slides 2.3)
 - Main drawbacks (slides 2.3)
 - Estimation (slides 4.1)
 - Inference
 - N is large → sampling-based inference (slides 4.1)
 - N is small → randomization-based inference (slides 4.2)
 - Examples of RCTs (slides 4.3)

- How do we estimate the causal impact of a treatment in an RCT in an unbiased and consistent way?
 - We compare the difference in outcomes between the treatment and the control group
 - We calculate the st. error of this estimate, to determine how likely it is that the observed difference reflects the existence of a causal effect

Estimation in randomized experiments

- Let us consider a randomized trial with N individuals. Our estimand of interest is ATE.

$$\alpha_{ATE} = E[Y_1 - Y_0] = E[Y|D = 1] - E[Y|D = 0]$$

- Using the analogy principle, we construct an estimator:

$$\hat{\alpha} = \bar{Y}_1 - \bar{Y}_0$$

where

$$\bar{Y}_1 = \frac{\sum Y_i \cdot D_i}{\sum D_i} = \frac{1}{N_1} \sum_{D_i=1} Y_i$$

$$\bar{Y}_0 = \frac{\sum Y_i \cdot (1 - D_i)}{\sum (1 - D_i)} = \frac{1}{N_0} \sum_{D_i=0} Y_i$$

where $N_1 = \sum_i D_i$ and $N_0 = N - N_1$

- $\hat{\alpha}$ is an unbiased and consistent estimator of α_{ATE}

Unbiasedness & Consistency

- Unbiasedness:
 - An estimator is unbiased if, on average, it is correct.
 - $E(\hat{\alpha}) = \alpha$
- Consistency:
 - An estimator is consistent if, as sample size gets larger, the estimate converges to the ‘true’ value
 - As $n \rightarrow \infty, \hat{\alpha} \rightarrow \alpha$

Example

- We want to know the impact of a certain treatment, for instance a drug that reduces the mortality rate for people with COVID-19
- We run an experiment with N subjects, and we assigned half to the treatment group and half to the control group.
- As an estimator, we use the difference between the average outcome value in the treatment and control groups.

$$\hat{\alpha} = \bar{Y}_1 - \bar{Y}_0$$

- It is **unbiased** because, if we run multiple times this experiment (always with the same sample size), our estimate will be on average equal to the true effect
- It is **consistent** because if we increase the sample size, the estimate will tend to become closer to the true effect

Inference

- Generally the difference in outcomes between the treatment and control group ($\hat{\alpha}$) will not be exactly equal to zero
- This observed difference may reflect (i) chance or (ii) the causal impact of the treatment
- We need to calculate how likely would it be to observe such a difference, if the treatment had no effect.
- Different approaches depending on the sample size:
 - ① Large samples: Sampling-based approach → Two sample t-test
 - ② Small samples: Randomization inference → Fisher's exact test

Outline

1 Estimation

2 Inference

- Sampling-based inference

Sampling-based inference

- In this case inference is based on the idea that the subjects are a random sample from a much larger population.
- It is common in empirical analyses to view the sample analyzed as a random sample drawn randomly from a large, essentially infinite super-population.
- Uncertainty is viewed as arising from this sampling, with knowledge of the full population leading to full knowledge of the estimands.
- Traditional sampling based approach considers the treatment assignments to be fixed, while the outcomes are random.

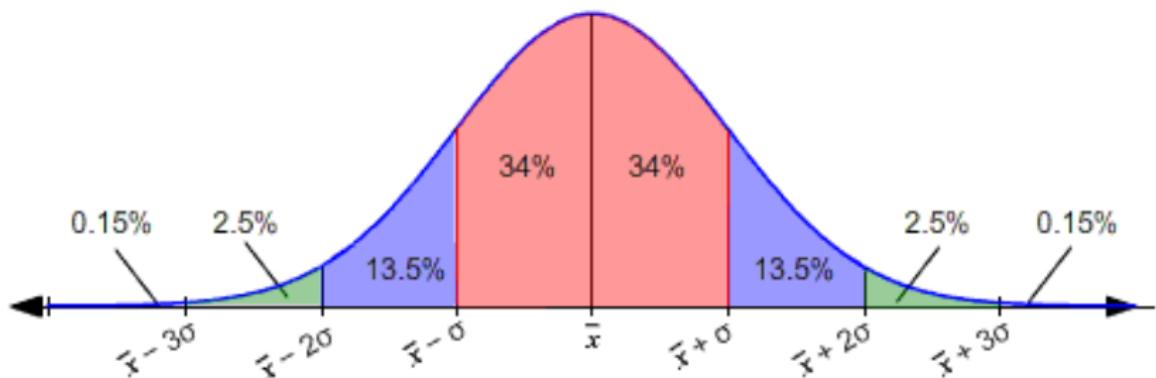
Central limit theorem

- Luckily enough, when sample size if large enough, we can use the central limit theorem (CLT) for independent and identically distributed random variables.
- From the CLT we know that, given an underlying population with mean μ and standard deviation σ , if we draw a sample of size n , the sample mean ($\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$) is normally distributed as follows:
 - $E[\bar{y}] = \mu$
 - $\text{variance}[\bar{y}] = \frac{\sigma^2}{n}$
 - $\text{st.error} = \text{st.dev.}[\bar{y}] = \sqrt{\text{variance}[\bar{y}]} = \frac{\sigma}{\sqrt{n}}$
- Or, similarly:

$$\bar{y} \xrightarrow{d} N\left(\mu, \frac{\sigma^2}{n}\right)$$

Example of the Central limit theorem

- Let us consider the height of the British male population
- The average height in the overall population is equal to 175 cm. with standard deviation equal to 10 cm.
- If start taking random samples of 100 British men, what would be the distribution of their sample mean?
→ Normally distributed with expected value equal to 175 cm.
and st. error equal to one $[\frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{100}}]$
- What is the 95% confidence interval (CI)?
 - In a normal distribution, to calculate the 95% CI we can simply add and subtract 2*st.error (1.96, if you want to be more precise)
 - The sample mean will be 95% of the time between 173 and 177



Testing in large samples: Two sample t-test

- Our object of interest is the difference between two sample means (the average outcomes of the treatment group minus the control group):

$$\hat{\alpha} = \bar{Y}_1 - \bar{Y}_0$$

- The sum of two (independent) normally distributed variables is also normally distributed, with its mean being the sum of the two means, and its variance being the sum of the two variances.

Therefore:

$$\hat{\alpha} \xrightarrow{d} N(\alpha_{ATE}, \frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_0^2}{n_0}) \quad (1)$$

- This expression provides information about the expected distribution of our estimator, $\hat{\alpha}$. Most importantly, it tells us how much variability we should expect in the difference between the mean of the treatment and the control group.

- Note that in equation (1), given that the actual variance of Y_1 and Y_0 is unknown, we have used the estimated one:

$$\hat{\sigma}_1^2 = \frac{1}{N_1 - 1} \sum_{D_i=1} (Y_i - \bar{Y})^2$$

- and $\hat{\sigma}_0^2$ is analogously defined.

- Let us consider the hypothesis that the treatment has no effect ($\alpha_{ATE} = 0$)
- Then $\hat{\alpha}$ would be normally distributed as follows:

$$\hat{\alpha} \xrightarrow{d} N\left(0, \frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_0^2}{n_0}\right)$$

- and its standard error (the st. dev. of $\hat{\alpha}$) is $\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_0^2}{n_0}}$
- We know that, because of its normality, with 95% probability the realizations of $\hat{\alpha}$ should lie within 1.96 standard errors of zero.
- Therefore, we would reject the null hypothesis $H_0 : \alpha_{ATE} = 0$ against the alternative hypothesis $H_1 : \alpha_{ATE} \neq 0$ whenever $|\hat{\alpha}| > 1.96 * st.error(\hat{\alpha})$

- Imagine that we run an RCT to estimate the impact of taking this module on students' future lifetime income.
- (Normal) probability distribution of $\hat{\alpha}$ is summarized by two moments of its distribution:
 - Point estimate: difference between treatment and control group
 - Standard error: provides information about the accuracy of the estimate
- Indicate whether following estimates (in pounds) indicate that the impact is significantly different from zero.
 - $\hat{\alpha}$ (*st. error*)
 - i 100,000 (100,000)
 - ii 100 (100,000)
 - iii 100 (30)
 - iv 100,000 (20,000)
- You just need to check whether the point estimate is larger (in absolute value) than twice the standard error

- Imagine that we run an RCT to estimate the impact of taking this module on students' future lifetime income.
- (Normal) probability distribution of $\hat{\alpha}$ is summarized by two moments of its distribution:
 - Point estimate: difference between treatment and control group
 - Standard error: provides information about the accuracy of the estimate
- Indicate whether following estimates (in pounds) indicate that the impact is significantly different from zero.
 - $\hat{\alpha}$ (*st. error*)
 - i 100,000 (100,000)
 - ii 100 (100,000)
 - iii 100 (30)
 - iv 100,000 (20,000)
- where **blue** indicates statistical significance; **red** lack of it

- To characterize whether an effect is statistically significant we also use:
 - P-values: How likely it is that such a large difference would have been observed if there was no effect (e.g. 5%)
 - Stars are often used to report significance levels (e.g. ** to indicate p-value<5%)
- Another useful way to summarize this distribution:
 - 95% confidence interval : point estimate $\pm 2 \times$ standard error
- Note: 95% confidence intervals are helpful to interpret estimates. Just add and subtract the st. error times 2 to the point estimate
 - E.g.: 100,000 (100,000) \rightarrow CI=(-100,000, 300,000)
- $\hat{\alpha}$ not statistically different from zero $\not\Rightarrow \alpha$ is equal to zero
 - Precisely estimated zeros vs. uninformative estimates
- Corollary: Estimates are useful when they are precise

Statistical significance vs. economic significance

- Which ones of the following estimates are significant in ‘economic’ terms (independently of whether they are statistically significant)
 - $\hat{\alpha}$ (*st. error*)
 - iii 100 (30)
 - iv 100,000 (20,000)

Statistical significance vs. economic significance

- Which ones of the following estimates are significant in ‘economic’ terms (independently of whether they are statistically significant)
 - $\hat{\alpha}$ (*st. error*)
 - iii 100 (30)
 - iv 100,000 (20,000)
- where blue indicates ‘economic’ significance; red lack of it

EC902/EC907: Econometrics A

Lecture 4.2

Manuel Bagues

Warwick University

Roadmap

- Randomized controlled trials (RCT)
 - How random assignment overcomes the selection effect (slides 2.3)
 - Main drawbacks (slides 2.3)
 - Estimation (slides 4.1)
 - Inference
 - N is large → sampling-based inference (slides 4.1)
 - N is small → randomization-based inference (slides 4.2)
 - Examples of RCTs (slides 4.3)

Testing in small samples: Randomization-based inference

Athey and Imbens (2016), ‘The Econometrics of Randomized Experiments’

- Randomization-based inference (or permutation tests)
 - Uncertainty in estimates arises naturally from the random assignment of the treatments, rather than from hypothesized sampling from a large population.
 - Randomization-based approach takes the subject’s potential outcomes as fixed, and considers the assignment of subjects to treatments as random.
- Two possible ways to implement the randomization-based approach
 - ① Check all possible assignments → Fisher’s exact test
 - ② Check a large random sample of possible assignments

Fisher's exact test

- Fisher's exact test with small N:

$$H_0 : Y_1 = Y_0$$

$$H_1 : Y_1 \neq Y_0$$

- Let Ω be the set of possible randomization realizations
- We only observe the outcomes for one possible realization:

$$\hat{\alpha} = \bar{Y}_1 - \bar{Y}_0$$

- Under the sharp null hypothesis, we can calculate the value that the difference in means would have taken under any other realizations of the assignment of observations to the treatment and control group, $\hat{\alpha}(\omega)$, for $\omega \in \Omega$

Testing in small samples: Fisher's exact test

- Straightforward implementation:
 - ① Calculate the difference between the treatment and the control group ($\hat{\alpha}$)
 - ② Hypothetical scenarios: all possible different ways in which the assignment could have happened and calculate the difference in outcomes in each one of these hypothetical realizations [$\hat{\alpha}(\omega)$]
 - ③ Check how exceptional is the actual difference in outcomes ($\hat{\alpha}$) within the set of hypothetical assignments [$\hat{\alpha}(\omega)$]

Example

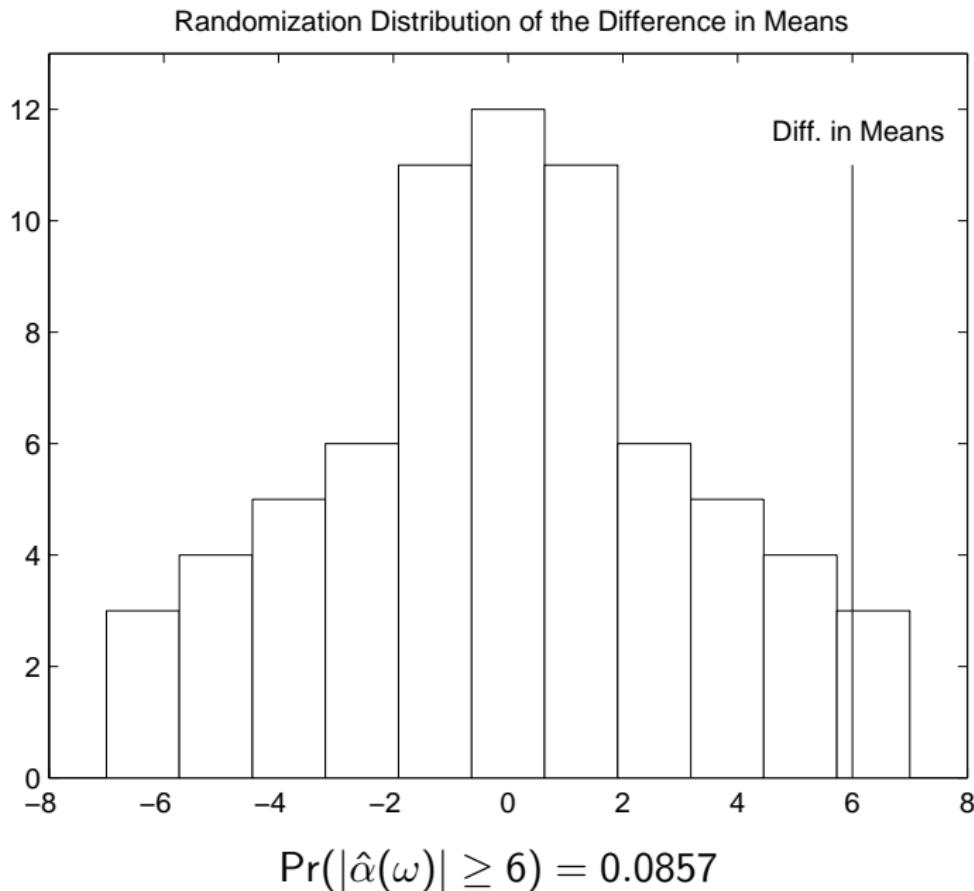
Imagine a population with 8 units, and suppose we assign 4 individuals to the treatment

i	1	2	3	4	5	6	7	8			
Y_i	12	4	6	10	6	0	1	1	\bar{Y}_1	\bar{Y}_0	$\hat{\alpha}$
D_i	1	1	1	1	0	0	0	0	8	2	6

Now let us use Fisher's exact test to quantify how likely would be to observe such a large $\hat{\alpha}$ if the treatment had no effect

i	1	2	3	4	5	6	7	8	\bar{Y}_1	\bar{Y}_0	$\hat{\alpha}(\omega)$
$\omega=1$	1	1	1	1	0	0	0	0	8	2	6
$\omega=2$	1	1	1	0	1	0	0	0	7	3	4
$\omega=3$	1	1	1	0	0	1	0	0	5.5	4.5	1
$\omega=4$	1	1	1	0	0	0	1	0	5.75	4.25	1.5
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
$\omega=70$	0	0	0	0	1	1	1	1	2	8	-6

Testing in Small Samples: Fisher's Exact Test



Example

- In the actual experiment, the difference in outcomes between the treatment and control group was equal to 6
- There are 70 possible ways in which we can assign 4 individuals to the treatment group and 4 to the control group.
- In 8.6% (6 out 70) of these possible hypothetical assignments the difference between the (hypothetical) treatment and control groups was equal or larger than 6
- 8.6% is our p-value - it reflects the probability that we are observing such a large $\hat{\alpha}$ just by chance, even if the treatment in reality had no effect

EC902/EC907: Econometrics A

Lecture 4.3

Manuel Bagues

Warwick University

Roadmap

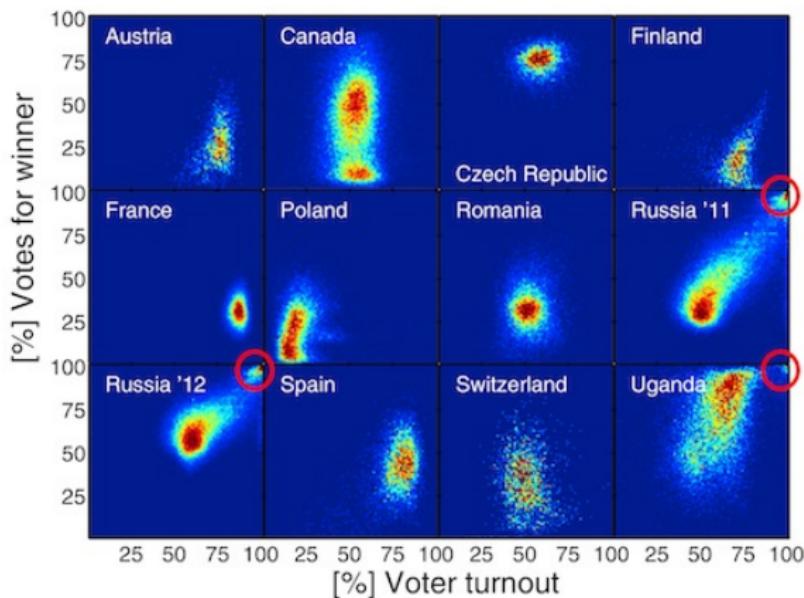
- Randomized controlled trials (RCT)
 - How random assignment overcomes the selection effect (slides 2.3)
 - Main drawbacks (slides 2.3)
 - Estimation (slides 4.1)
 - Inference
 - N is large → sampling-based inference (slides 4.1)
 - N is small → randomization-based inference (slides 4.2)
 - Examples of RCTs (slides 4.3)

Example: Electoral fraud in Russia

- Motivation:
 - Is there any electoral fraud in Russia?
 - How much?
- Available evidence:
 - Anecdotal evidence
 - Statistical evidence

Circumstancial evidence (i)

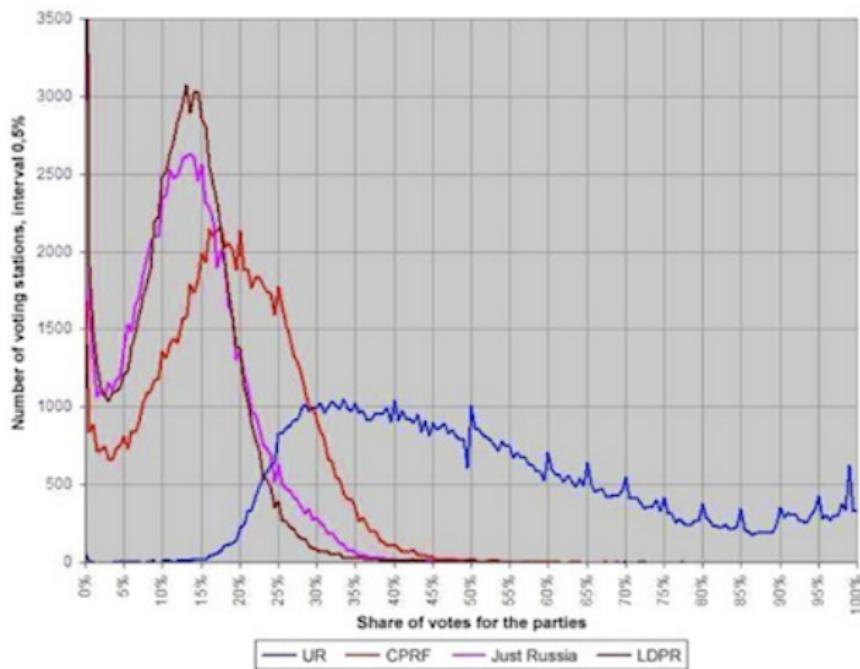
Bimodal distribution of votes



Circumstancial evidence (ii)

Spikes in the distribution of votes for United Russia

Kobak, Shpilkin and Pschenichnikov (2016)



Circumstancial evidence (ii)

"We do not believe Churov [the head of the electoral committee], we believe Gauss!"

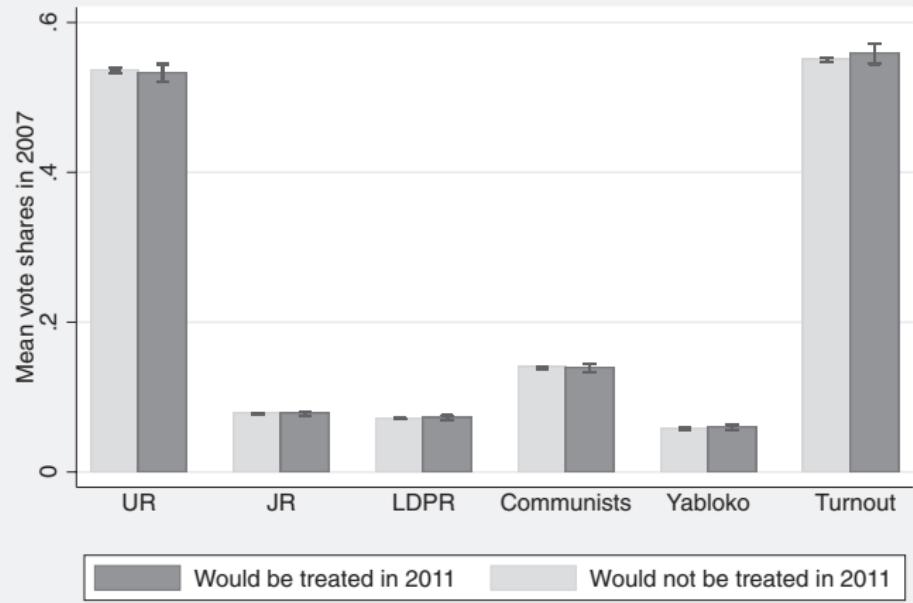


Let us rephrase slightly our question in a *treatment effects* fashion:

- Would electoral results change if there were independent observers in the polling stations?
- To address this question, we can try to send observers to some (non-randomly selected) polling stations, as some NGOs and international organizations do.
 - How informative would this be?
- Can you propose a better approach?
 - Enikolopov, Korovkina, Petrova, Sonin and Zakharov 2013

- Random assignment of independent observers to 156 of 3,164 polling stations in the city of Moscow
- Treatment:
 - Observers can only prevent the most obvious types of fraud
 - Intention-to-treat (not full compliance): Some of these observers were removed before the vote counting process was finished
- Within each district, polling stations were sorted according to their official number assigned by Central Election Committee. Every 25th polling station within an electoral district, starting from the first, was assigned for observation
- How can we verify whether the randomization worked?

Placebo test



```

. clear

. use "PNAS_data_2011_2007.dta"

.

. ttest er_share if year==2007, by(treat)

```

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]
0	949	.5231358	.0027977	.0861861	.5176453 .5286262
1	56	.5286899	.0139387	.1043077	.5007561 .5566237
combined	1,005	.5234452	.002752	.0872432	.5180449 .5288456
diff		-.0055541	.0120021		-.0291062 .017998

diff = mean(0) - mean(1) t = -0.4628
Ho: diff = 0 degrees of freedom = 1003

Ha: diff < 0 Pr(T < t) = 0.3218	Ha: diff != 0 Pr(T > t) = 0.6436	Ha: diff > 0 Pr(T > t) = 0.6782
------------------------------------	---	------------------------------------

- How does Stata calculate the standard error of $\hat{\alpha}$?
- From the Central Limit Theorem we know the distribution of $\hat{\alpha}$

$$st.error(\hat{\alpha}) = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_0^2}{n_0}}$$

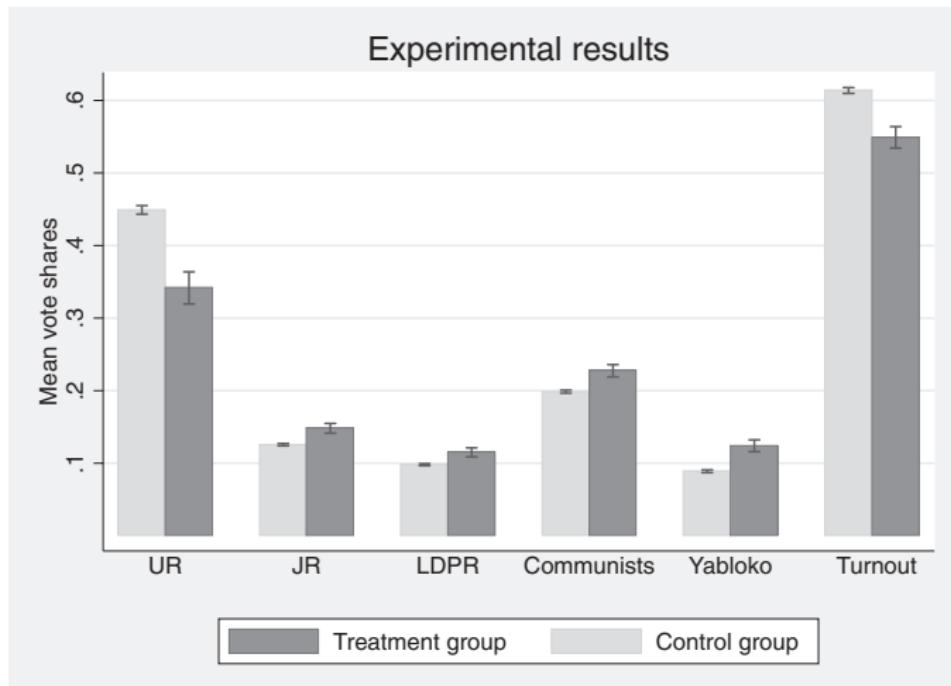
- If we assume that the variance is similar in both samples:

$$st.error(\hat{\alpha}) = \sqrt{\frac{\hat{\sigma}^2}{n_1} + \frac{\hat{\sigma}^2}{n_0}} = \sqrt{\frac{0.08724^2}{949} + \frac{0.08724^2}{56}} = 0.012$$

- Let us also calculate the 95% confidence interval for $\hat{\alpha}$

$$\begin{aligned} 95\%CI &= (\hat{\alpha} - 1.96 * st.error(\hat{\alpha}), \hat{\alpha} + 1.96 * st.error(\hat{\alpha})) = \\ &= (-.0056 - 1.96 * 0.012, -.0056 + 1.96 * 0.012) = (-.029, 0.018) \end{aligned}$$

Main results:



```
. ttest er_share if year==2011, by(treat)
```

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]
0	3,008	.4491337	.0030033	.1647148	.443245 .4550223
1	156	.3416313	.0112328	.1402975	.3194422 .3638205
combined	3,164	.4438333	.0029374	.165225	.438074 .4495926
diff		.1075023	.0134341		.081162 .1338427

diff = mean(0) - mean(1) t = 8.0022
Ho: diff = 0 degrees of freedom = 3162

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0
Pr(T < t) = 1.0000 Pr(|T| > |t|) = 0.0000 Pr(T > t) = 0.0000

- Main results
 - The actual share of votes for the incumbent United Russia party is 11 percentage points lower in treatment areas (34% instead of 45%). P-value is
 - The turnout at the polling stations with observers was lower by 6.5 percentage points
- Interpretation?

Field experiment estimate of electoral fraud in Russian parliamentary elections

Randomization inference

- Let us now use randomization inference to calculate p-values
- Procedure:
 - Assign randomly the treatment to 156 observations (out of 3,064)
 - Compare these (fake) treatment and control groups: calculate the difference in outcomes
 - Repeat this 10,000 times
 - Check how often the difference in outcomes is larger (in absolute terms) than the difference observed taking into account the actual assignment to the treatment (>10.8 in this case)
- Let us start with just one placebo:

```
. ***Randomization inference  
. *Let us generate N placebos where we assign 156 randomly chosen observations to the treatment  
. *example with N=1  
. *for simplicity let us keep only the relevant sample  
. keep if year==2011 & treat!=.  
5,062 observations deleted)  
  
. generate random = runiform()  
. sort random  
. gen placebo=0  
. replace placebo=1 if _n<=156  
156 real changes made)  
. drop random  
  
. browse id treat placebo er_share
```

Stata/SE 15.0 File Edit View Data Graphics Statistics User Window Help

Data Editor (Browse) — PNAS_data_2011_2007.dta

Edit Browse

Filter Variables Properties Snapshots

placebo[19]

	id	treat	placebo	er_share					
1	2231	0	1	.3807947					
2	1130	0	1	.6591376					
3	2171	0	1	.245628					
4	950	0	1	.4953387					
5	1685	0	1	.2659288					
6	2586	0	1	.6816667					
7	315	1	1	.3507194					
8	1907	1	1	.5067638					
9	2712	0	1	.7369353					
10	2381	0	1	.5057938					
11	3062	0	1	.2210788					
12	369	0	1	.3537549					
13	2733	0	1	.5354331					
14	1508	0	1	.4640934					
15	2179	0	1	.6697149					
16	640	0	1	.4657763					
17	2465	0	1	.415056					
18	2405	0	1	.4694015					
19	534	0	1	.3438155					
20	1735	0	1	.5004143					
21	2467	0	1	.4166239					
22	1592	1	1	.4694486					
23	337	1	1	.3526531					
24	2045	0	1	.5000654					

```
. bys placebo: sum er_share
```

```
-> placebo = 0
```

Variable	Obs	Mean	Std. Dev.	Min	Max
er_share	3,008	.4450423	.1651021	.1273234	.9400212

```
-> placebo = 1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
er_share	156	.4205222	.1663976	.1631623	.8626778

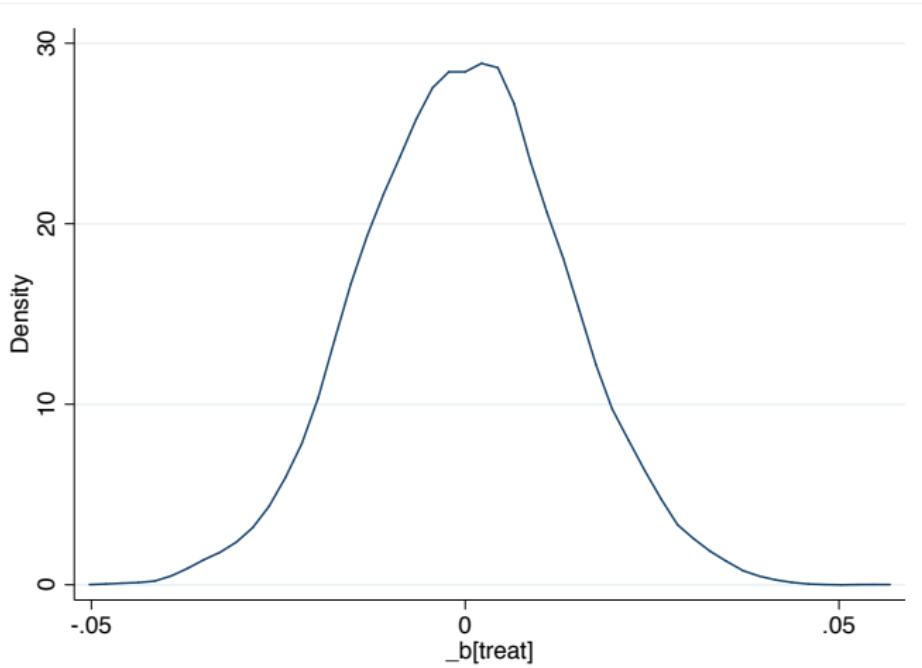
```
.  
end of do-file
```

```
. dis .4450423-.4205222  
.0245201
```

Field experiment estimate of electoral fraud in Russian parliamentary elections

Randomization inference

- In this case, the difference was equal to 0.0245
- Now we have to do it 9,999 times more
- We can do it ourselves writing a loop
- Or we can simply google to see if there is any available command (e.g. "randomization inference stata command")
 - Bingo - there is an author-written command: 'ritest'
 - we type "net install ritest" to install it
- We can execute it and we can plot the results of these 10,000 permutations



Field experiment estimate of electoral fraud in Russian parliamentary elections

Randomization inference

- In none of the 10,000 permutations the ‘simulated’ difference is larger than the actual one (11p.p.) → p-value=0.0000 (0/10,000)
- Note that in this case sampling- and randomized-based inference provide similar results.
- Question: More generally, when would you think that they might lead to different results?

Field experiment estimate of electoral fraud in Russian parliamentary elections

Randomization inference

- In none of the 10,000 permutations the ‘simulated’ difference is larger than the actual one (11p.p.) → p-value=0.0000 (0/10,000)
- Note that in this case sampling- and randomized-based inference provide similar results.
- Question: More generally, when would you think that they might lead to different results?
 - when N is small! (because the CLT does not apply)

Examples from the business world

Which **experiment** could be used to capture the causal effect?

- AB testing
- John List and Uber -

<https://www.bbc.co.uk/news/stories-54613947>

Other examples

Which **experiment** could be used to capture the causal effect?

- Would it be profitable for the call center of a travel agency to allow their employees to work from home?
 - Bloom et al 2012
- Are employees more satisfied if they are informed about the salaries of their colleagues?
 - Card et al 2011
- Does the gender composition of hiring committees matter?
 - Bagues and Esteve-Volart 2010
- Do arrests decrease domestic violence?
 - National Institute of Justice's Spouse Assault Replication Program
- Do monetary incentives crowd out intrinsic motivation?
 - Gneezy and Rustichini 2000
 - Lacetera et al. 2012

- Do “modern managerial” practices increase firms’ productivity?
(lean manufacturing principles)
 - Bloom et al. 2010
- An increase in the salaries offered in the public sector attracts candidates that are less committed to public service
 - Dal Bo, Finan and Rossi 2012
- How can we decrease the impact of AIDS in Subsaharan Africa?
 - Dupas 2011
- Do Indian teachers react to incentives?
 - Duflo et al 2012

EC902/EC907: Econometrics A

Lecture 5

Manuel Bagues

Warwick University

October 24, 2022
Lecture Slides

Roadmap

- Randomized controlled trials (RCT)
 - How random assignment overcomes the selection effect (slides 2.3)
 - Main drawbacks (slides 2.3)
 - Estimation (slides 4.1)
 - Inference
 - N is large \rightarrow sampling-based inference (slides 4.1)
 - N is small \rightarrow randomization-based inference (slides 4.2)
 - Examples of RCTs (slides 4.3)

Our estimator

Let us consider the simple case where we have a **dychotomic treatment** (e.g. enrolling in this module)

- ① To estimate the average treatment effect (ATE) we conduct an RCT...
- ② ... and we use a very simple estimator:

$$\hat{\alpha} = \bar{Y}_1 - \bar{Y}_0$$

which has two very desirable properties: it is **unbiased** and **consistent**

Unbiasedness & Consistency

- Unbiasedness:
 - An estimator is unbiased if, on average, it is correct.
 - $E(\hat{\alpha}) = \alpha$
- Consistency:
 - An estimator is consistent if, as sample size gets larger, the estimate converges to the ‘true’ value
 - As $n \rightarrow \infty, \hat{\alpha} \rightarrow \alpha$
- Typically we will use estimators that satisfy these two properties

Poll questions

- We will make some polls using [vevox.app](#)
 - ➊ Please, open [vevox.app](#) in your computer, in Teams or download the app in your mobile
 - ➋ Enter session ID: 144-921-069

Poll question

- Types of questions:
 - Have you watched the asynchronous lectures?
 - A. No, unfortunately I was too busy
 - B. I watched part of it, but not everything.
 - C. Yes, I watched everything!

Poll question 1

- Imagine the following (fictitious and absurd) estimator.

$$\hat{\alpha} = \bar{Y}_1 - \bar{Y}_0 + \lambda$$

- where λ is a random variable uniformly distributed between -1 and +1
- Which of the following statements is true. This estimator is:
 - ① unbiased but inconsistent
 - ② biased but consistent
 - ③ unbiased and consistent
 - ④ biased and inconsistent

Poll question 1

- Imagine the following (fictitious and absurd) estimator.

$$\hat{\alpha} = \bar{Y}_1 - \bar{Y}_0 + \lambda$$

- where λ is a random variable uniformly distributed between -1 and +1
- Which of the following statements is true. This estimator is:
 - ① unbiased but inconsistent
 - ② biased but consistent
 - ③ unbiased and consistent
 - ④ biased and inconsistent

Our estimator

- ➊ To estimate the average treatment effect of a treatment (ATE) we conduct an RCT...
- ➋ We use a very simple estimator: $\hat{\alpha} = \bar{Y}_1 - \bar{Y}_0$
- ➌ Once we have a value for $\hat{\alpha}$, we need to calculate how likely it is that it reflects the outcome of random sampling (and not the impact of the treatment)
 - Large N: t-test
 - Small N: randomization inference

Central limit theorem

Testing in large samples: Sampling-based inference

- Luckily enough, when sample size if large enough, we can use the central limit theorem (CLT) for independent and identically distributed random variables.
- From the CLT we know that, given an underlying population with mean μ and standard deviation σ , if we draw a (sufficiently large) sample of size n , the sample mean ($\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$) is normally distributed as follows:
 - $E[\bar{y}] = \mu$
 - $\text{variance}[\bar{y}] = \frac{\sigma^2}{n}$
 - $\text{st.error} = \text{st.dev.}[\bar{y}] = \sqrt{\text{variance}[\bar{y}]} = \frac{\sigma}{\sqrt{n}}$
- Or, similarly:

$$\bar{y} \xrightarrow{d} N\left(\mu, \frac{\sigma^2}{n}\right)$$

Example: Application of central limit theorem

- For simplicity let us (provisionally) assume that we know that, in a given population of students, mean performance was distributed with mean equal to 65 and standard deviation equal to 10
 - Question: intuitive explanation of st. dev.? [click here](#)
- Let us consider that we run an experiment in this population with a sample of 100 students
 - For instance, the experiment could imply not allowing this sample to use lecture capture
 - Theoretically, it is unclear what to expect: more resources vs. time-consistency problems

Poll question 2

- Under the null hypothesis that the treatment has no impact, if we take a sample of 100 students, which of the following statements is true? (note: assume that in the underlying population has mean equal to 65 and standard deviation equal to 10)
 - ① 95% of the time the sample mean will be between 64 and 66
 - ② 95% of the time the sample mean will be between 63 and 67
 - ③ 95% of the time the sample mean will be between 62 and 68
 - ④ 95% of the time the sample mean will be between 60 and 70
 - ⑤ It is not possible to answer this question with the available information

Poll question 2

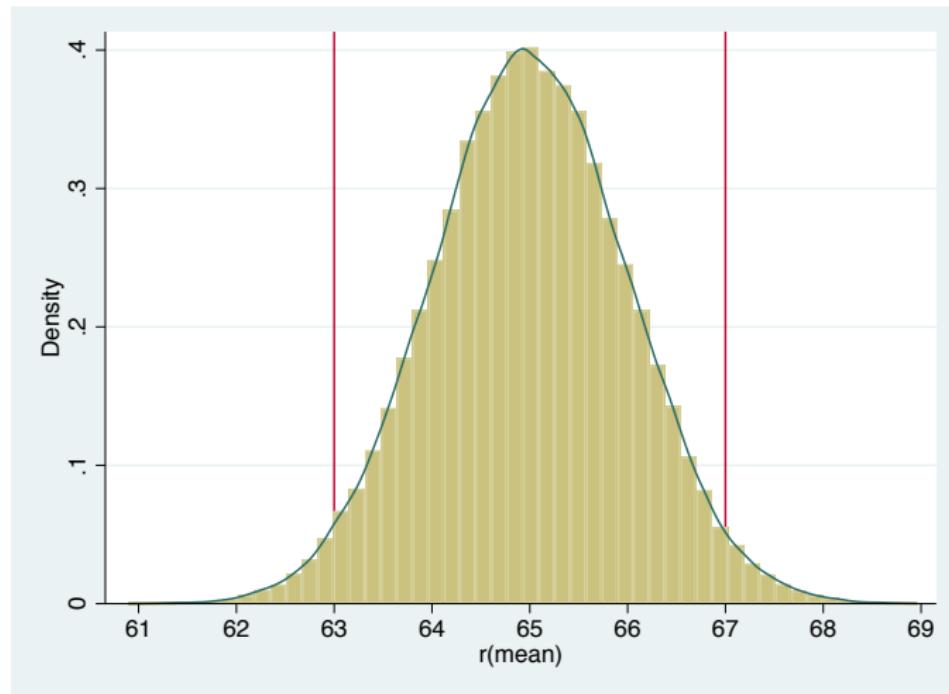
- Under the null hypothesis that the treatment has no impact, if we take a sample of 100 students, which of the following statements is true? (note: assume that in the underlying population has mean equal to 65 and standard deviation equal to 10)
 - ① 95% of the time the sample mean will be between 64 and 66
 - ② 95% of the time the sample mean will be between 63 and 67
 - ③ 95% of the time the sample mean will be between 62 and 68
 - ④ 95% of the time the sample mean will be between 60 and 70
 - ⑤ It is not possible to answer this question with the available information

Simulation: distribution of the sample mean (N=100)

Population distribution: uniform between 47.7 and 83.3 (mean=65, st. dev.=10)

Simulation: We draw 100,000 samples with N=100, and calculate the mean for each sample.

As predicted by the CLT, average sample mean=65.0, st. dev. of sample mean=1.00, mass outside the 95% interval=5.0%



- Going on with the previous experiment, imagine that we observe that the average performance in the sample of 100 students participating in the experiment was equal to 63, what do we make out of this?

- Going on with the previous experiment, imagine that we observe that the average performance in the sample of 100 students participating in the experiment was equal to 63, what do we make out of this?
- The CLT indicates that there is **less than 5% probability** that this happened by chance!
- If we consider a 5% threshold, we would conclude that we cannot reject the possibility that preventing access to lecture capture affected negatively the performance of students
- Probability of a false positive?

- Going on with the previous experiment, imagine that we observe that the average performance in the sample of 100 students participating in the experiment was equal to 63, what do we make out of this?
- The CLT indicates that there is **less than 5% probability** that this happened by chance!
- If we consider a 5% threshold, we would conclude that we cannot reject the possibility that preventing access to lecture capture affected negatively the performance of students
- Probability of a false positive?
 - 5 % (by definition)

- Going on with the previous experiment, imagine that we observe that the average performance in the sample of 100 students participating in the experiment was equal to 63, what do we make out of this?
- The CLT indicates that there is **less than 5% probability** that this happened by chance!
- If we consider a 5% threshold, we would conclude that we cannot reject the possibility that preventing access to lecture capture affected negatively the performance of students
- Probability of a false positive?
 - 5 % (by definition)
- Probability of a false negative?

- Going on with the previous experiment, imagine that we observe that the average performance in the sample of 100 students participating in the experiment was equal to 63, what do we make out of this?
- The CLT indicates that there is **less than 5% probability** that this happened by chance!
- If we consider a 5% threshold, we would conclude that we cannot reject the possibility that preventing access to lecture capture affected negatively the performance of students
- Probability of a false positive?
 - 5 % (by definition)
- Probability of a false negative?
 - We would need to make some assumption about the potential impact of the treatment

Poll question 3

- Let us consider again a population with mean 65 and standard deviation equal to 10, but let us change the sample size.
- If we conduct our experiment on a sample of **4 students**. Which of the following statements is true?
 - ① 95% of the time the sample mean will be between 55 and 75
 - ② 95% of the time the sample mean will be between 60 and 70
 - ③ 95% of the time the sample mean will be between 63 and 67
 - ④ 95% of the time the sample mean will be between 64 and 66
 - ⑤ It is not possible to answer this question with the available information

Poll question 3

- We take a sample of 4 students. Which of the following statements is true?
 - ① 95% of the time the sample mean will be between 55 and 75
 - ② 95% of the time the sample mean will be between 60 and 70
 - ③ 95% of the time the sample mean will be between 63 and 67
 - ④ 95% of the time the sample mean will be between 64 and 66
 - ⑤ It is not possible to answer this question with the available information

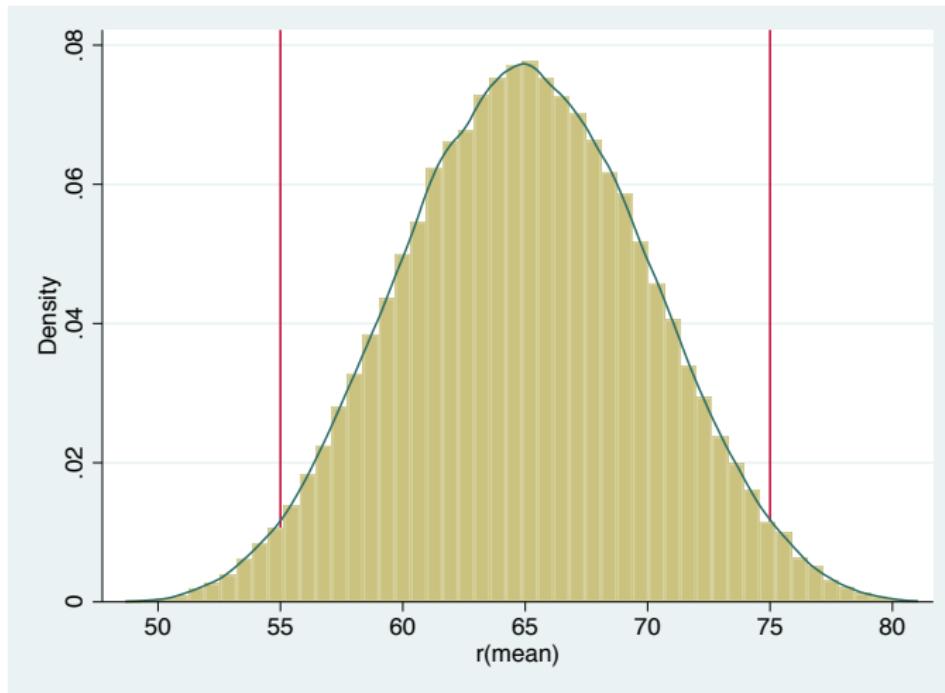
- To answer the previous question we would need more information on the distribution of the underlying population (e.g. normality, uniform, binomial...)
 - Remember, the CLT says that, if sample size is large, the distribution of the sample mean is normal
 - What ‘large’ means depends on the underlying distribution
 - Rule of thumb: $N>30$
- Let us consider a couple of examples, using samples of size 4
 - ① Normal distribution (mean=65, st. dev.=10)
 - ② Uniform distribution (mean=65, st. dev.=10)

Simulation: distribution of the sample mean

Population distribution: **normal distribution**, mean=65, st. dev.=10

Simulation: We draw 100,000 samples with **N=4**

Results: Average sample mean=65.0, st. dev. of sample mean=5.00, **mass outside the 95% interval=5.0%**

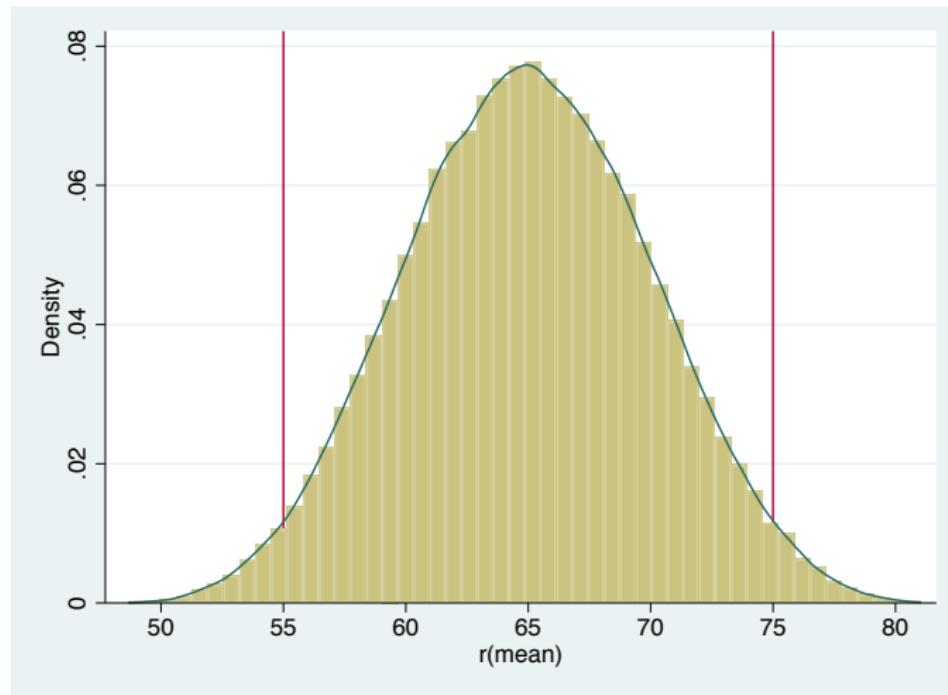


Simulation: distribution of the sample mean (N=4)

Population distribution: uniform between 47.7 and 83.3 (mean=65, st. dev.=10)

Simulation: We draw 100,000 samples with N=4, and calculate the mean for each sample.

Average sample mean=65.0, st. dev. of sample mean=5.00, **mass outside the 95% interval=4.7%**



Our estimator (continued)

Initially we said:

- ① To estimate the average treatment effect of a treatment (ATE) we conduct an RCT...
- ② We use a very simple estimator: $\hat{\alpha} = \bar{Y}_1 - \bar{Y}_0$
- ③ Once we have a value for $\hat{\alpha}$, we need to calculate how likely it is that it reflects the outcome of random sampling (and not the impact of the treatment)
 - Large N: t-test
 - Small N: randomization inference

Let us now assume that we are in a context where N is large, and let us use the CLT to estimate the st.dev.($\hat{\alpha}$), a.k.a. the standard error

Testing in large samples: Equality of means test

- The sum of two (independent) normally distributed variables is also normally distributed, with its mean being the sum of the two means, and its variance being the sum of the two variances.

Therefore:

$$\hat{\alpha} \xrightarrow{d} N(\alpha_{ATE}, \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}) \quad (1)$$

- If we assume similar variance in both samples ($\sigma = \sigma_1 = \sigma_2$) and the sample of the treatment and control group is similar ($\frac{n}{2} = n_1 = n_0$)

$$\hat{\alpha} \xrightarrow{d} N(\alpha_{ATE}, \frac{\sigma^2}{\frac{n}{2}} + \frac{\sigma^2}{\frac{n}{2}}) = N(\alpha_{ATE}, \frac{4 * \sigma^2}{n}) \quad (2)$$

- $st.error(\hat{\alpha}) = st.dev.\hat{\alpha} = \sqrt{\frac{4 * \sigma^2}{n}} = \frac{2 * \sigma}{\sqrt{n}}$
- This expression provides information about the expected distribution of our estimator, $\hat{\alpha}$. It tells us how much variability we should expect in $\hat{\alpha}$ just by chance.

Testing in large samples: Equality of means test

- Note that so far we have implicitly assumed that the population standard deviation (σ) was known. However, in nearly all practical statistical work, the population standard deviation of these errors is unknown and has to be estimated from the data.
- In this case the t-student distribution is often used to account for the extra uncertainty that results from this estimation.
- However, in practice for N large the t-distribution converges to a Normal.

Can you manually calculate the standard error and confidence intervals using the previous formula?

```
. ttest email , by(race)
```

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]
b	2,435	.4796715	.0101263	.4996892	.4598144 .4995285
w	2,435	.4788501	.0101256	.4996551	.4589944 .4987058
combined	4,870	.4792608	.0071594	.499621	.4652251 .4932964
diff		.0008214	.0143203		-.0272528 .0288955

diff = mean(b) - mean(w) t = 0.0574

Ho: diff = 0 degrees of freedom = 4868

Ha: diff < 0

Pr(T < t) = 0.5229

Ha: diff != 0

Pr(|T| > |t|) = 0.9543

Ha: diff > 0

Pr(T > t) = 0.4771

- Moreover, we know that, because of its normality, with 95% probability the realizations of $\hat{\alpha}$ should lie within 1.96 standard errors of zero.
- Therefore, we would reject the null hypothesis $H_0 : \alpha_{ATE} = 0$ whenever $|\hat{\alpha}| > 1.96 * st.error(\hat{\alpha})$
- Note: failure to reject $H_0 \not\Rightarrow \alpha = 0$

Interpreting results

Example: RCT conducted in China to estimate the impact of the drug remdesivir on mortality
Wang et al., Lancet (2020)

- Randomised, double-blind, placebo-controlled, multicentre trial at ten hospitals in Hubei, China.
- This trial is registered with ClinicalTrials.gov, NCT04257656.
- Between Feb 6, 2020, and March 12, 2020, 237 patients were enrolled and randomly assigned to a treatment group (158 to remdesivir and 79 to placebo)
- Patients were randomly assigned in a 2:1 ratio to intravenous remdesivir (200 mg on day 1 followed by 100 mg on days 2-10 in single daily infusions) or the same volume of placebo infusions for 10 days.
- Outcome variable: mortality rate by day 28
 - Treatment group: 14%
 - Control group: 13%
 - $\hat{\alpha}=1.1\%$ (st.error=4.7%)

Poll question 4:

- Are the authors' estimates [$\hat{\alpha}=1.1\%$, $\text{st.error}(\hat{\alpha})=4.7\%$] significantly different from zero at the 5% level?
 - ① Yes
 - ② No
 - ③ It is not possible to tell based on the information provided

Poll question 4:

- Are the authors' estimates [$\hat{\alpha}=1.1\%$, $\text{st.error}(\hat{\alpha})=4.7\%$] significantly different from zero at the 5% level?
 - ① Yes
 - ② No
 - ③ It is not possible to tell based on the information provided

Poll question 5:

- What is the 95% confidence interval for the estimate?
[$\hat{\alpha} = 1.1\%$ ($st.error = 4.7\%$)]
 - ① (-8.1%,10.3%)
 - ② (-3.6%,5.8%)
 - ③ (-9.4%,9.4%)

Poll question 5:

- What is the 95% confidence interval for the estimate?
[$\hat{\alpha} = 1.1\%$ ($st.error = 4.7\%$)]
 - ① (-8.1%,10.3%)
 - ② (-3.6%,5.8%)
 - ③ (-9.4%,9.4%)

Poll question 6:

- Based on the above results, what would you conclude:
 - ① we cannot reject the possibility that the drug is ineffective
 - ② we cannot reject the possibility that the drug increases mortality by 10 p.p.
 - ③ we cannot reject the possibility that the drug decreases mortality by 5 p.p.
 - ④ all of the above

Poll question 6:

- Based on the above results, what would you conclude:
 - ① we cannot reject the possibility that the drug is ineffective
 - ② we cannot reject the possibility that the drug increases mortality by 10 p.p.
 - ③ we cannot reject the possibility that the drug decreases mortality by 5 p.p.
 - ④ all of the above

Poll question 7:

- Based on the above results, would you suggest that:
 - ① the use of remdesivir should be immediately stopped, given that it might increase mortality rate
 - ② the use of remdesivir should be immediately stopped, given that it is ineffective
 - ③ more research is needed to establish whether remdesivir is a useful treatment

Poll question 7:

- Based on the above results, would you suggest that:
 - ① the use of remdesivir should be immediately stopped, given that it might increase mortality rate
 - ② the use of remdesivir should be immediately stopped, given that it is ineffective
 - ③ more research is needed to establish whether remdesivir is a useful treatment

Randomization inference: Problem set 2, question C

- Suppose that we randomly assign 3 individuals out of 6 to the treatment group and we observe the following outcomes:

i	1	2	3	4	5	6
D_i	1	1	1	0	0	0
Y_i	3	2	1	2	1	0

What is the difference in outcomes between the treatment and control groups?

Randomization inference: Problem set 2, question C

- Suppose that we randomly assign 3 individuals out of 6 to the treatment group and we observe the following outcomes:

i	1	2	3	4	5	6
D_i	1	1	1	0	0	0
Y_i	3	2	1	2	1	0

What is the difference in outcomes between the treatment and control groups?

$$\hat{\alpha} = E[Y_{i1}/D_i = 1] - E[Y_{i0}/D_i = 0] = (3+2+1)*\frac{1}{3} - (2+1+0)*\frac{1}{3} = 1$$

- Does the treatment have a statistically significant impact? In other words, would you say that, if the impact had no impact, there is less than a 5% probability of observing such a large difference between the treatment and control groups?
- Before we make a formal calculation, what would be your educated guess based on the observed magnitude of the difference ($\hat{\alpha}=1$)?
 - ➊ p-value<0.05
 - ➋ $0.05 < \text{p-value} < 0.10$
 - ➌ $0.10 < \text{p-value}$

- Let us now use Fisher's Exact Test to calculate the p-value (i.e. how likely it is to observe such a large difference under the null hypothesis that the treatment has no effect). How should we proceed?

- Let us now use Fisher's Exact Test to calculate the p-value (i.e. how likely it is to observe such a large difference under the null hypothesis that the treatment has no effect). How should we proceed?
 - we consider all possible assignments that might potentially have happened (such that 3 individuals are assigned to the treatment group and 3 to the control) (\rightarrow 20 possible assignments)
 - we calculate the difference in outcomes between treatment and control in each potential assignment
 - finally we check how often we observe a difference that is as large in magnitude to the one observed between the actual treatment and control group?

Poll question 8

i	1	2	3	4	5	6	alpha:
Y	3	2	1	2	1	0	
D	1	1	1	0	0	0	
w=1	1	1	1	0	0	0	1.00
w=2	1	1	0	1	0	0	1.67
w=3	1	1	0	0	1	0	1.00
w=4	1	1	0	0	0	1	0.33
w=5	1	0	1	1	0	0	1.00
w=6	1	0	1	0	1	0	0.33
w=7	1	0	1	0	0	1	-0.33
w=8	1	0	0	1	1	0	1.00
w=9	1	0	0	1	0	1	0.33
w=10	1	0	0	0	1	1	-0.33
w=11	0	1	1	1	0	0	0.33
w=12	0	1	1	0	1	0	-0.33
w=13	0	1	1	0	0	1	-1.00
w=14	0	1	0	1	1	0	0.33
w=15	0	1	0	1	0	1	-0.33
w=16	0	1	0	0	1	1	-1.00
w=17	0	0	1	1	1	0	-0.33
w=18	0	0	1	1	0	1	-1.00
w=19	0	0	1	0	1	1	-1.67
w=20	0	0	0	1	1	1	-1.00

Poll question 6: what is the p-value? (i) 10%, (ii) 5%, (iii) 50%, (iv) 40%

i	1	2	3	4	5	6	
Y	3	2	1	2	1	0	
D	1	1	1	0	0	0	
w=1	1	1	1	0	0	0	1.00
w=2	1	1	0	1	0	0	1.67
w=3	1	1	0	0	1	0	1.00
w=4	1	1	0	0	0	1	0.33
w=5	1	0	1	1	0	0	1.00
w=6	1	0	1	0	1	0	0.33
w=7	1	0	1	0	0	1	-0.33
w=8	1	0	0	1	1	0	1.00
w=9	1	0	0	1	0	1	0.33
w=10	1	0	0	0	1	1	-0.33
w=11	0	1	1	1	0	0	0.33
w=12	0	1	1	0	1	0	-0.33
w=13	0	1	1	0	0	1	-1.00
w=14	0	1	0	1	1	0	0.33
w=15	0	1	0	1	0	1	-0.33
w=16	0	1	0	0	1	1	-1.00
w=17	0	0	1	1	1	0	-0.33
w=18	0	0	1	1	0	1	-1.00
w=19	0	0	1	0	1	1	-1.67
w=20	0	0	0	1	1	1	-1.00

alpha:

- Out of 20 possible combinations, the effect was as large 10 times
→ p-value=0.50
- If the treatment had no effect, there would be a 50% probability that we observe such a large difference.
- Note: a large p-value does not necessarily mean that the treatment has no effect, it means simply that we cannot reject such possibility.

Thank you for your attention!

Short aside: standard deviation

- Measure of dispersion: the square root of the mean squared deviation
 - $\sigma = \sqrt{\frac{\sum_{i=1}^N (y_i - \mu)^2}{N}}$
- When the population is **normally distributed**, its interpretation is quite straightforward:
 - 95% of observations that lie within 2 st. dev. of the mean, and 99.7% within 3 st. dev..
 - Ex: if grades are standardized with mean=65 and st.dev.=10, you know that around 95% obtained between 45 and 85.

back

EC902/EC907: Econometrics A

Lecture 6.1

Manuel Bagues

Warwick University

Road map

- So far: RCTs with a dychotomic treatment
- This lecture: RCTs with a continuous treatment
 - Linear regression model
 - Estimation: OLS

Linear regression model: a brief introduction

- So far, we had considered a simple case where the treatment was a dychotomic variable:

$$\alpha = E(y/D_i = 1) - E(y/D_i = 0) = E(y_1) - E(y_0)$$

- Estimation just required a comparison of sample means:

$$\hat{\alpha} = \bar{y}_1 - \bar{y}_0$$

- Let us now consider a more general situation where we are interested in the impact of a continuous treatment (e.g. amount of fertilizer, drug dose, subsidy, class size ...) .

- Let us assume the following data generating process:

$$y_i = \beta_0 + \beta_1 * x_i + u_i \quad (1)$$

where y_i is the outcome variable,

x_i is the amount of treatment received,

β_0 is the intercept,

β_1 is the slope (how the outcome varies with the treatment),

u_i represents other (unobservable) factors that affect the outcome.

- A few implicit assumptions in this equation:
 - the role of other factors: additive error term
 - homogeneous effect (similar β_0 and β_1 for all individuals)
 - Functional relationship between x and y (eg. linear vs quadratic)
- Without loss of generality, let us also assume that:

$$E(u_i) = 0 \quad (2)$$

- Moreover, due to random assignment:

$$E(u_i/x_i) = E(u_i) \quad (3)$$

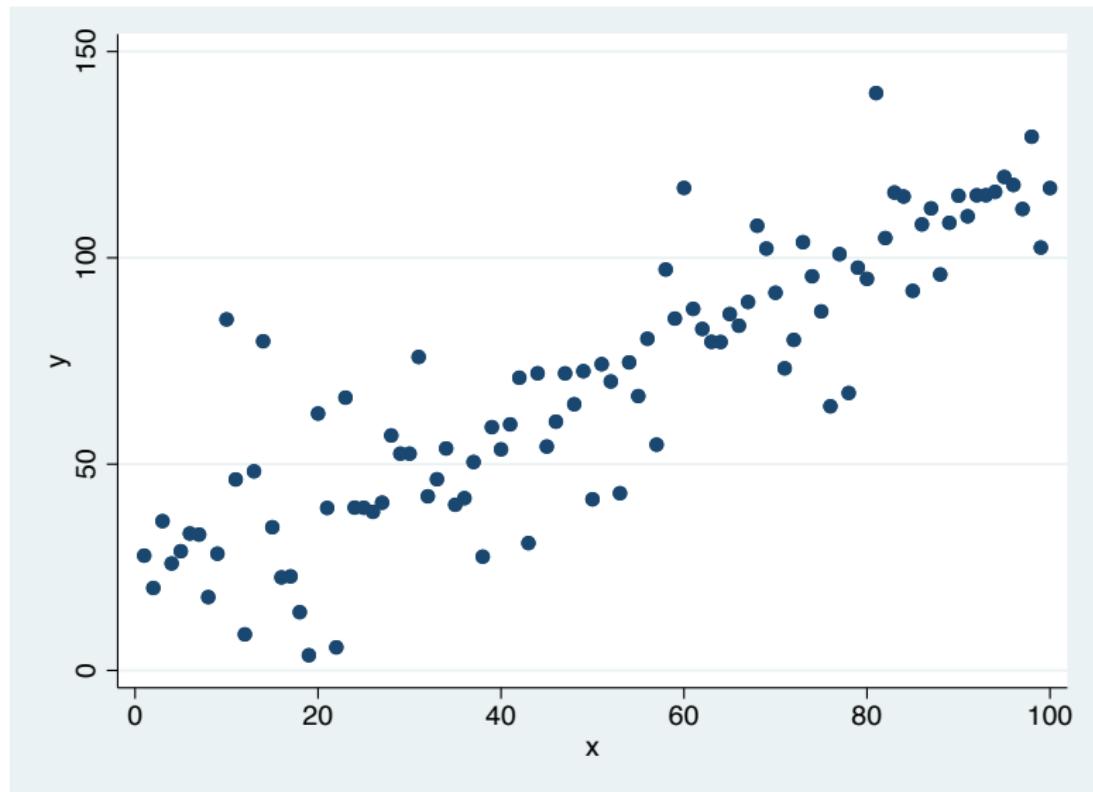
- Combining the above assumptions, it follows that:

$$E(y_i|x_i) = \beta_0 + \beta_1 * x_i + \underbrace{E(u_i/x_i)}_{=0} = \beta_0 + \beta_1 * x_i$$

and $E(\Delta y_i|\Delta x_i) = \beta_1$

- Now, imagine that we conduct an RCT where we randomize the amount of x received by each individual, and we collect information on y

Graphical representation:



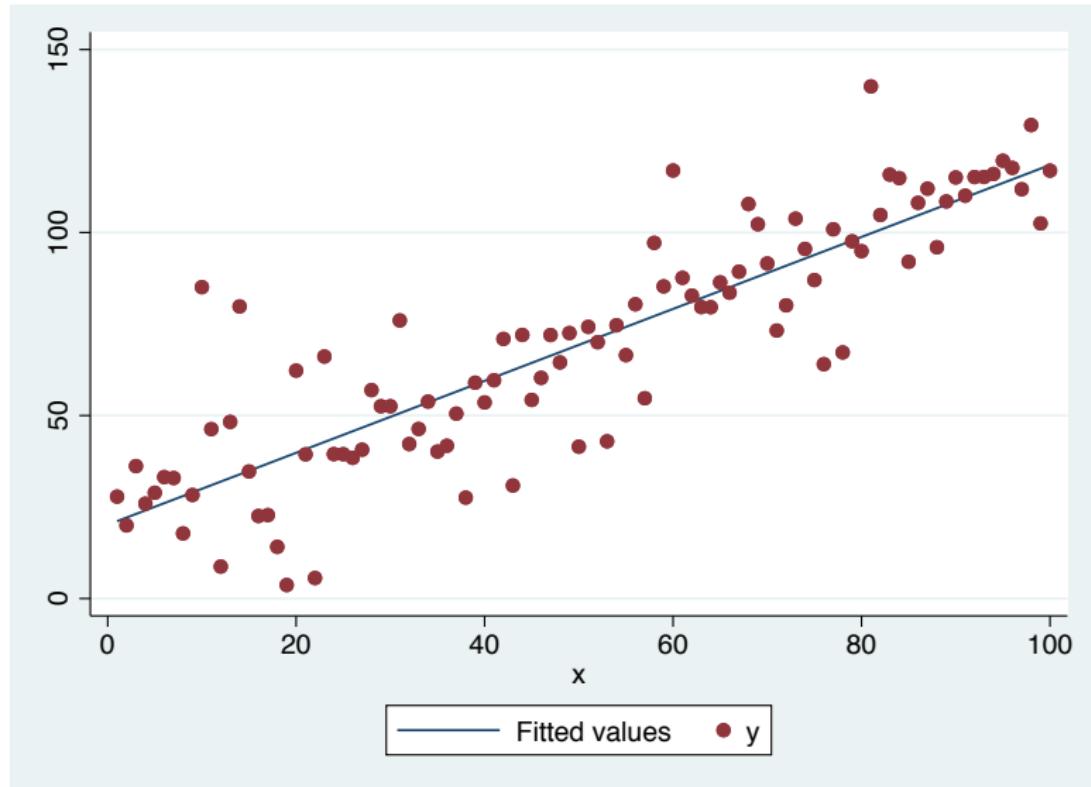
OLS estimator

- Based on the information about the data generation process and the observed data, how do we estimate the value of $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$?
- An intuitive answer is that we need to find what is the ‘best’ possible straight line fitting the data plot.
- For instance, we may use the Ordinary Least Squares (OLS) estimates, i.e. the line $(\hat{\beta}_0 + \hat{\beta}_1 * x)$ that minimizes the (square) distance between a proposed prediction line and the observed data points $(\sum(y - \hat{\beta}_0 - \hat{\beta}_1 * x)^2)$

$$\hat{\beta}_1^{OLS} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{Cov(y_i, x_i)}{V(x_i)}$$

$$\hat{\beta}_0^{OLS} = \bar{y} - \hat{\beta} * \bar{x}$$

Graphical representation:



OLS estimator in matrix form

- We can also express the model in matrix notation (with n observations and $k=1$ variables):

- $$\underbrace{\mathbf{y}}_{nx1} = \underbrace{\mathbf{X}}_{nx2} \underbrace{\boldsymbol{\beta}}_{2x1} + \underbrace{\mathcal{U}}_{nx1}$$

where:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \mathcal{U} = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}$$

OLS estimator in matrix form

- The OLS estimator minimizes the sum of squared residuals
 - $\hat{\beta}_{OLS} = \arg \min_{\beta} (y - X\beta)'(y - X\beta)$
- Solving this problem we find that the best fit is given by:
$$\hat{\beta}_{OLS} = (X'X)^{-1}(X'Y)$$
- Note: above we use the symbol ' to denote a transposed matrix (e.g. X' is equivalent to X^T)

Other ways of motivating the OLS estimator

- So far we have vaguely justified the OLS estimator because it provides a ‘good’ fit.
- Let me provide three better reasons why we will use OLS to estimate linear regression models
 - ① method of moments approach

Method of moments approach to OLS

- The unbiasedness of the OLS estimator can be derived directly from the exogeneity assumption using a method of moments approach
- Let us consider the following model:
 - i. $Y = X\beta + U$
- where the exogeneity assumption holds:
 - ii. $E(X' \cdot U) = 0$
- If we substitute (ii) in (i):
 - $E(X' \cdot (Y - X\beta)) = 0$
 - $E(X'Y - X'X\beta) = 0$
 - $E[\underbrace{(X'X)^{-1}X'Y}_{\hat{\beta}_{OLS}}] = \beta$

Other ways of motivating the OLS estimator

- Let me provide three better reasons why we may want to use OLS to estimate a linear regression model
 - Method of moments approach
 - Best linear unbiased estimator (BLUE)

OLS is BLUE

Gauss-Markov theorem

Under some conditions, OLS provides the ‘best’ linear unbiased estimator (BLUE), where ‘best’ means that it has the smallest variance ($E(\hat{\beta}^{OLS} - \beta)^2$) of all possible linear estimators.

- ① Correct specification: the data-generating process is **linear**
- ② No perfect **multicollinearity** between regressors
 - Otherwise $\mathbf{X}'\mathbf{X}$ is not invertible and the OLS estimator cannot be computed
- ③ Spherical errors: (i) **Homoscedasticity** and (ii) **no autocorrelation**
(more on this later on)
- ④ Strict **exogeneity**: $E(\epsilon|\mathbf{X}) = 0$
 - a. $E(\epsilon) = 0$
 - b. $E(\mathbf{X}' \cdot \epsilon) = 0$
 - Remember that condition (4) is required to give OLS estimates a causal interpretation!

Other ways of motivating the OLS estimator

- Let me provide three better reasons why we may want to use OLS to estimate a linear regression model
 - ① Method of moments approach
 - ② Best linear unbiased estimator (BLUE)
 - ③ If the error term is normally distributed $\Rightarrow \beta^{MLE} = \beta^{OLS}$
 - Error term would be normally distributed if it is the composite of a sufficiently large number of minor influences (Central Limit Theorem).
 - Maximum Likelihood Estimation provides the estimates that are more likely to have generated the data
 - Note: alternative distributional assumptions might lead to other estimators (e.g. Laplacian distribution $\Rightarrow \beta^{MLE}$ = least absolute errors estimator)

EC902/EC907: Econometrics A

Lecture 6.2

Manuel Bagues

Warwick University

Roadmap

- Randomized controlled trials (RCT)
 - Linear regression model
 - Estimation: OLS
- Identification based on observables
 - Conditional independence assumption
 - Ideal experiment
- OLS
 - Threats to validity

How can we estimate causal estimates without running an RCT?

- RCTs solve the selection problem
- Unfortunately, it is not possible to run a controlled experiment with every research question
- Often we will need to rely on observational data using some **identification strategy**.

Causality without experiments

The **identification strategy** refers to the manner in which a researcher uses observational data (i.e. data not generated by a randomized trial) to approximate a real experiment and identify causal effects. In other words, it is the way in which we try to find a control group that is comparable to the treatment group.

- Main empirical strategies:
 - ① *Random assignment*
 - ② *Selection based on observables* ([today](#) and [next week!](#))
 - ③ *Instrumental variables*
 - ④ *Difference-in-differences*
 - ⑤ *Regression discontinuity design*

Selection based on observables

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

The Conditional Independence Assumption (CIA)

- To understand the CIA let's begin with an example.
- We would like to estimate the impact of taking this module on students' grade in the master dissertation (Y_i).
- Let $D_i = 1$ if student i takes this module and $D_i = 0$ otherwise.
- A naive comparison of observed averages may not be very informative:

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

- Students in the treatment and control group are likely to differ in many dimensions that affect the outcome, leading to a selection bias ($E[Y_{0i}|D_i = 1] \neq E[Y_{0i}|D_i = 0]$)

Causality and the CIA

- But we can try to control for some of these differences (e.g. whether they have previously taken an econometrics module)
- Let us compare the treatment and control group, taking into account observable characteristics:

$$E[Y_{1i}|\mathbf{X}_i, D_i = 1] - E[Y_{0i}|\mathbf{X}_i, D_i = 0] = E[Y_{1i} - Y_{0i}|\mathbf{X}_i, D_i = 1] + \\ + E[Y_{0i}|\mathbf{X}_i, D_i = 1] - E[Y_{0i}|\mathbf{X}_i, D_i = 0]$$

- The CIA is valid when, conditioning on a set of observed characteristics X_i , the bias disappears

$$E[Y_{0i}|\mathbf{X}_i, D_i = 1] = E[Y_{0i}|\mathbf{X}_i, D_i = 0]$$

- Hence,

$$E[Y_{1i}|\mathbf{X}_i, D_i = 1] - E[Y_{0i}|\mathbf{X}_i, D_i = 0] = E[Y_{1i} - Y_{0i}|\mathbf{X}_i, D_i = 1]$$

Numerical example

- Let us consider the following hypothetical example. We have information on a random sample of 6 students, three of them took Econometrics A and the other three Econometrics B. For the sake of the exercise, let us pretend that we can observe all potential outcomes, which we present in the following table:

i	Y_i	D_i	Y_{1i}	Y_{0i}	$Y_{1i} - Y_{0i}$
1	75	1	75	70	5
2	70	1	70	65	5
3	65	1	65	60	5
4	75	0	80	75	5
5	70	0	75	70	5
6	65	0	70	65	5

- Let us suppose that we are interested in the average treatment effect on the treated (ATET)

$$E[Y_{i1} - Y_{i0}/D_i = 1] = (5 + 5 + 5) * \frac{1}{3} = 5$$

- But, in real life, we can typically observe only the difference in performance between the treatment and the control group.

$$\begin{aligned} E[Y_{i1}/D_i = 1] - E[Y_{i0}/D_i = 0] &= \\ (75 + 70 + 65) * \frac{1}{3} - (75 + 70 + 65) * \frac{1}{3} &= 0 \end{aligned}$$

- In this example, the two do not coincide, reflecting the existence of a selection bias.

- We would like to control for variables that explain selection. For instance, let us assume that we observe whether the individual has taken previously an econometrics module ($X_i = 1$ if yes, otherwise equal to zero)
- Note that individuals with an econometrics background are more likely to select into the control group (Econometrics B)

i	Y_i	D_i	Y_{1i}	Y_{0i}	$Y_{1i} - Y_{0i}$	X_i
1	75	1	75	70	5	0
2	70	1	70	65	5	0
3	65	1	65	60	5	0
4	75	0	80	75	5	1
5	70	0	75	70	5	1
6	65	0	70	65	5	0

How would you implement here an identification strategy based on observables?

- Identification based on observables: Let us compare individuals who have similar observable characteristics.
- Since none of the individuals in the treatment group has previously taken econometrics, we only consider in the control group those individuals who did not take previously an econometrics module ($X_i = 0$). We can do a simple comparison of means:

$$E[Y_{i1}/\mathbf{X}_i = 0, D_i = 1] - E[Y_{i0}/\mathbf{X}_i = 0, D_i = 0] =$$

$$= \frac{1}{3} * (75 + 70 + 65) - 65 = 70 - 65 = 5$$

- which happens to be equal to the ATET! (we managed to get rid of the selection bias)

Causality and the CIA

- How representative was this numerical example?
- In general, would you expect individuals in the control and treatment group with similar observable characteristics to be similar in all other relevant unobserved characteristics? In other words, how relevant is the selection problem in practice?
- Selection would not be an issue if agents were irrational or fully uninformed. However, in general we rarely randomize our choices (even conditional on observables)
- What may drive selection?
 - Information, differences in preferences...
 - These (unobserved) selection factors may affect the outcome variable

Causality and the CIA

- For relevant ‘treatments’, selection is usually a relevant problem. Note that there is crucial paradox in empirical studies that rely on an identification strategy based on observables. In order to estimate the effect of a certain treatment, we need to assume that, conditional on certain observables, this treatment was assigned in a random way. However, those who benefit more from the treatment probably may have tried to get more "treatment".
- Corollary: we should be always be very cautious about the interpretation of estimates when the identification relies on observables

Reminder: why selection is not a problem in an RCT?

- Random assignment: by construction, we know that the control and treatment groups are comparable
- Identification based on observables: we **hope** that, once we control for observables, the control and treatment groups are comparable. The crucial assumption of the identification based on observables strategy is that, conditional on observables, the assignment of the treatment is as good as random (*Conditional independence assumption*)
- However, if the treatment and the control group differ in terms of the variables that we observe, why should we expect them to be similar in the dimensions that we do not observe?

Identification based on observables

Ideal experiment

- To assess the reliability of an empirical strategy that relies on observables, it is useful to reflect about the ‘ideal experiment’ that the author is hoping to capture. Is it plausible?
- Let us consider a particular example: The effect of having a distinct black name (**Fryer and Levitt 2004**)
- How would you address this question using a identification based on observables approach? What would you control for?
- What is the ideal RCT that the authors are hoping to capture with their empirical strategy?
- Note: if you have a paper which is based on observables, make sure always to think about your implicit RCT

Identification based on observables

Ideal experiment

Other examples:

- The impact of paternity leave on infant-parent attachment, using a survey and comparing fathers who took the paternity leave and those who did not (Krueger 1993)
- The impact of hours slept on life expectancy
- The impact of R&D expenditure on countries growth rate
- The impact of computers on students' performance

EC902/EC907: Econometrics A

Lecture 6.3

Manuel Bagues

Warwick University

Roadmap

- Identification based on observables
 - Conditional independence assumption (lecture 6.2)
 - Ideal experiment (lecture 6.2)
 - Estimation methods ([lecture 6.3](#))
 - Main threats to the validity ([lecture 6.3](#))

Estimation

- Main methods
 - ① Statistical matching techniques (e.g. propensity score matching)
 - ② Ordinary least squares (OLS)

Main threats to the validity

- Main threats to validity
 - ① Omitted variables (be able to calculate the omitted variable bias!)
 - ② Bad controls
 - ③ Measurement error in the independent variable
 - ④ Measurement error in the dependent variable
 - ⑤ Reverse causality

Omitted variable bias

- Let us consider that this is the true model:
 - $Y = X\beta + Z\gamma + \epsilon$
 - (where for instance Y is earnings, X is education, and Z is ability)
- but unfortunately we do not observe Z (ability)
- We compute the OLS estimator using the observable information on earnings (Y) and education (X)
 - $\hat{\beta}_{OLS} = (X'X)^{-1}(X'Y)$
- Let us now compute the bias involved in this estimation

$$\begin{aligned} E[\hat{\beta}_{OLS}] &= E[(X'X)^{-1}(X'Y)] = E[(X'X)^{-1}(X'(X\beta + Z\gamma + \epsilon))] = \\ &= E[\cancel{(X'X)^{-1}X'}X\beta + \cancel{(X'X)^{-1}X'}Z\gamma + \underbrace{\cancel{(X'X)^{-1}X'}}_{E(X'\epsilon)=0}(\epsilon))] = \\ &= \beta + \underbrace{E[(X'X)^{-1}X'Z\gamma]}_{\text{omitted variable bias}} \end{aligned}$$

What happens if we omit a variable

	$E(X'Z) > 0$	$E(X'Z) < 0$
$\gamma > 0$	POSITIVE BIAS	NEGATIVE BIAS
$\gamma < 0$	NEGATIVE BIAS	POSITIVE BIAS

Bad controls

Is adding controls always a good idea?

- To avoid the existence of an omitted variable bias, you may be tempted to control for as many variables as possible
- However, it may not always be a good idea to add controls in a regression
- **Bad controls** are variables that are themselves potential outcome variables in the notional experiment at hand
- The problem is that, even if the treatment and the control group were initially similar, if we condition our analysis on a variable that is affected by the treatment, then we create a sample bias.
- Another way to think about it is that if you introduce bad controls you are shutting down one of the possible channels through which the treatment may have an effect

Bad controls

- Let's see an example: controlling for occupation in college-earnings regression
- W_i is a dummy for white collar jobs, C_i a dummy for colleges, and Y_i earnings
- As usual we have:

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

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

Bad controls

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

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

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

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

Bad controls

- Calculating the above we see the problem:

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

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

Measurement error in the independent variable

- Suppose that we are interested in estimating the causal impact of taking this module ($X_i = 1$) on performance in the Master dissertationc(Y_i).

$$Y_i = X_i\beta + \epsilon_i$$

- However, our information on who took this course \tilde{X} is imperfect and subject to a random error η ($\tilde{X} = X_i + \eta_i$)
- and $Y_i = \tilde{X}_i\beta - \underbrace{\beta\eta_i}_{e_i} + \epsilon_i = \tilde{X}_i\beta + e_i$

Measurement error in the independent variable

- The OLS estimator would be equal to:

$$E[\beta_{OLS}] = E[(\tilde{X}' \tilde{X})^{-1} \tilde{X}' Y] = E[(\tilde{X}' \tilde{X})^{-1} \tilde{X}' (\tilde{X}\beta + e)]$$
$$E[\beta_{OLS}] = \beta + E[(\tilde{X}' \tilde{X})^{-1} \tilde{X}' e]$$

- If the measurement error is random ($E[X \cdot \eta] = 0$), then it follows that:
 - $E[\beta_{OLS}] = \beta - \frac{\beta * Var(\eta)}{Var(\tilde{X})}$
- We have a problem of **attenuation bias**:
 - If $\beta > 0$, then $0 < \beta_{OLS} < \beta$
 - If $\beta < 0$, then $0 > \beta_{OLS} > \beta$

Measurement error in the dependent variable

- Assume now that the only imperfect measure we are dealing with is that of the dependent variable:

$$Y = X\beta + \epsilon$$

- We can only observe an imperfect measure Y^* , such that $Y^* = Y + e$
- For instance, going on with the previous example, we do not observe the precise grade obtained in the Master dissertation

$$Y^* - e = X\beta + \epsilon$$

$$Y^* = X\beta + \epsilon + e$$

- $E[\beta_{OLS}] = E[(X'X)^{-1}X'Y^*] = E[(X'X)^{-1}X'(X\beta + \epsilon + e)]$

Measurement error in the dependent variable

- What is relevant is how e is correlated with other regressors.
- If e is uncorrelated with X (random measurement error), then OLS is unbiased:
- $E[\beta_{OLS}] = E[(X'X)^{-1}X'X\beta + \underbrace{(X'X)^{-1}X'\epsilon}_{=0} + \underbrace{(X'X)^{-1}X'e}_{=0}] = \beta$
- However, if the measurement error is correlated with the treatment [$E(X \cdot e) \neq 0$], then β_{OLS} would be biased

Reverse causality

- Sometimes the problem is that the direction of the causal effect goes from Y to X.
- For example, consider the relationship between the stock market and electoral results.

EC902/EC907: Econometrics A

Lecture 7

Manuel Bagues

Warwick University

October 31, 2022
Lecture Slides

Today

- ① Stata helpdesk
- ② Economic vs statistical significant impact
- ③ Asynchronous lectures

- **STATA Helpdesk**

- Cholwoo Kim is available on Fridays at 13.00 - 14.00 in room FAB5.07
- Drop-in one-to-one session: Drop by if you have any questions, first-come first-served basis
- Please check first the 'Introduction to Stata' on Moodle:
<https://moodle.warwick.ac.uk/course/view.php?id=31050>

Economic vs statistical significant impact

- An estimate is *statistically significant* if we can statistically reject that it is equal to zero at the 95% level
- Several ways to verify it:
 - p-value < 0.05
 - $0 \notin 95\% \text{ C.I.}(\hat{\beta})$
 - $\hat{\beta} > 1.96 * \text{s.e.}(\hat{\beta})$ (note: assuming that $\hat{\beta}$ is normally distributed)
- Instead, *economic significance* is about the magnitude of the estimate: is $\hat{\beta}$ ‘large’?
- Sometimes the answer is obvious:
 - MSc at Warwick $\Rightarrow \uparrow$ permanent income by 1M pounds \rightarrow large effect
 - MSc at Warwick $\Rightarrow \uparrow$ permanent income by 1 pound \rightarrow small effect
- But sometimes the answer might be ambiguous:
 - E.g.: MSc at Warwick $\Rightarrow \uparrow$ permanent income by 10,000 pounds (around 1% of permanent income in the UK context)
- Note: to discuss economic significance you may want sometimes to use the bounds of the 95% C.I., particularly when the estimate

Asynchronous lectures:

- ① RCTs with a continuous treatment (slides 6.1)
 - Linear regression model
 - OLS estimation
- ② Identification based on observables (slides 6.2)
 - Empirical strategies relying on observational data
 - Conditional independence assumption
 - Ideal experiment
- ③ Identification based on observables (slides 6.3)
 - Estimation methods
 - Threats to validity

Poll questions

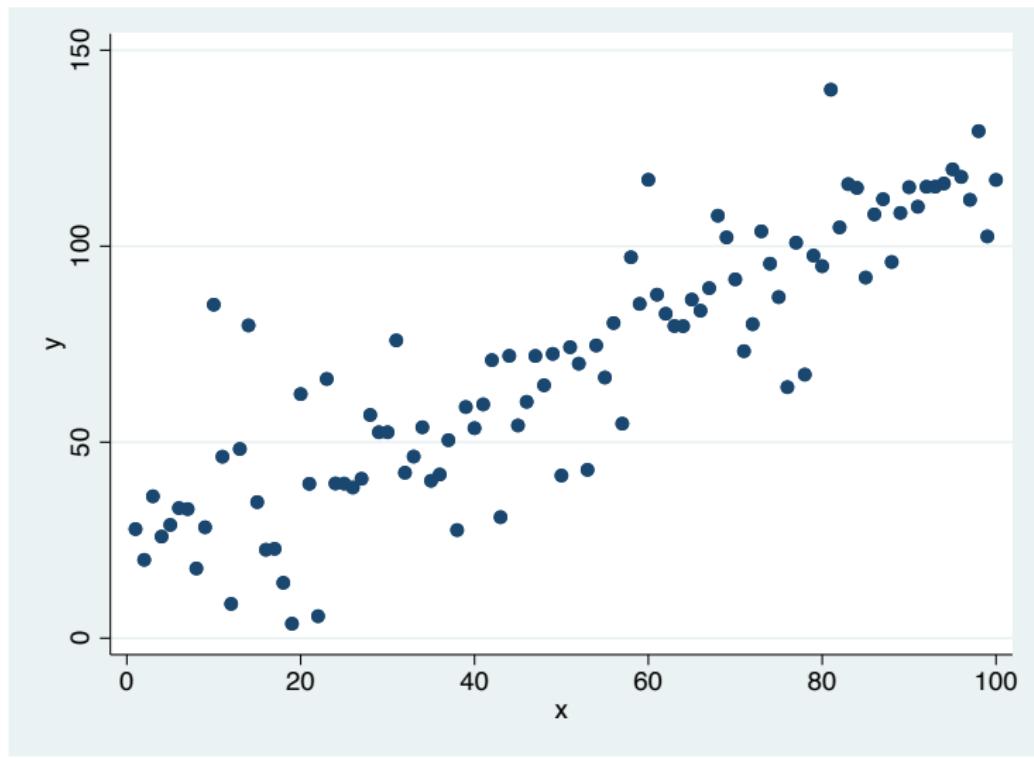
- We will make some polls using [vevox.app](#)
 - ① Please, open [vevox.app](#) in your computer, in Teams or download the app in your mobile
 - ② Enter session ID: 144-921-069

Poll question

- Have you watched the asynchronous lectures?
 - A. No, unfortunately I was too busy
 - B. I watched part of it, but not everything.
 - C. Yes, I watched everything!

We run an RCT with a continuous treatment

E.g. we provide a random quantity of drug x to a patient, and measure its impact on outcome variable y .



- Let us assume the following data generating process:

$$y_i = \beta_0 + \beta_1 * x_i + u_i$$

where y_i is the outcome variable,

x_i is the amount of treatment received,

β_0 is the intercept,

β_1 is the slope (how the outcome varies with the treatment),

u_i represents other (unobservable) factors that affect the outcome.

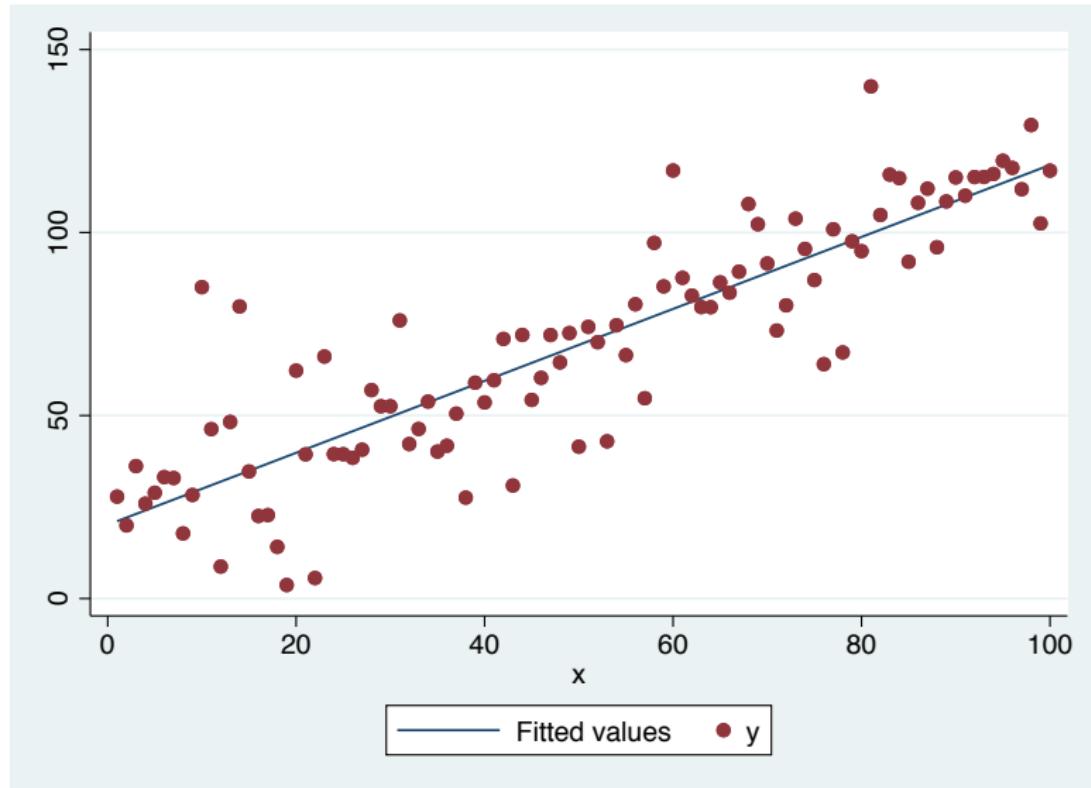
- where, due to random assignment:

$$E(u_i/x_i) = E(u_i) = 0$$

- and, as a result:

$$E(y_i|x_i) = \beta_0 + \beta_1 * x_i$$

Possible solution: OLS



Poll question 1

- ① Based on the information about the data generation process and the observed data, how can we estimate the value of $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$ in an **unbiased** way?
- Least squares: Calculate the $\hat{\beta}$ that minimizes the **squared distance** between a proposed prediction line and the observed data points
 - Least absolute deviations: Calculate the $\hat{\beta}$ that minimizes the **absolute distance** between a proposed prediction line and the observed data points
 - Least quartic: Calculate the $\hat{\beta}$ that minimizes the **quartic distance** (d^4) between a proposed prediction line and the observed data points
 - All of the above

Poll question 1

- ① Based on the information about the data generation process and the observed data, how can we estimate the value of $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$ in an **unbiased** way?
- Least squares: Calculate the $\hat{\beta}$ that minimizes the **squared distance** between a proposed prediction line and the observed data points
 - Least absolute deviations: Calculate the $\hat{\beta}$ that minimizes the **absolute distance** between a proposed prediction line and the observed data points
 - Least quartic: Calculate the $\hat{\beta}$ that minimizes the **quartic distance** (d^4) between a proposed prediction line and the observed data points
 - *All of the above*

Ways of motivating the OLS estimator

- Why we may want to use OLS to estimate a linear regression model?
 - ➊ It minimizes the squared distance
 - ➋ Method of moments approach
 - ➌ Best linear unbiased estimator (BLUE)
 - ➍ If the error term is normally distributed $\Rightarrow \beta^{MLE} = \beta^{OLS}$

Slides 6.2

- Empirical strategy
- Identification based on observables
 - Conditional independence assumption
 - Ideal experiment

Causality without experiments

The **identification strategy** refers to the manner in which a researcher uses observational data (i.e. data not generated by a randomized trial) to approximate a real experiment and identify causal effects. In other words, it is the way in which we try to find a control group that is comparable to the treatment group.

- Main empirical strategies:
 - ① *Random assignment*
 - ② *Selection based on observables*
 - ③ *Instrumental variables*
 - ④ *Difference-in-differences*
 - ⑤ *Regression discontinuity design*

Selection based on observables

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

Causality and the CIA

- In general, would you expect individuals in the control and treatment group with similar observable characteristics to be similar in all other relevant unobserved characteristics? In other words, how relevant is the selection problem in practice?
- Selection would not be an issue if agents were fully irrational or fully uninformed. Unfortunately, individuals are unlikely to randomize their choices (even conditional on observables)
- What may drive selection?
 - Information, differences in preferences...
 - These (unobserved) selection factors may affect the outcome variable

Identification based on observables

Ideal experiment

- To assess the reliability of an empirical strategy that relies on observables, it is useful to reflect about the ‘ideal experiment’ that the author is hoping to capture. Is it plausible?
- Let us consider a particular example: The effect of having a distinct black name ([Fryer and Levitt 2004](#))
- How would you address this question using a identification based on observables approach? What would you control for?
- Another example:
 - The impact of books at home on students’ performance

Main threats to the validity

- Main threats to validity
 - ① Omitted variables bias
 - ② Measurement error in the independent variable
 - ③ Measurement error in the dependent variable
 - ④ Bad controls
 - ⑤ Reverse causality

Omitted variable bias

- Let us consider that this is the true model:
 - $Y = X\beta + Z\gamma + \epsilon$
 - (where for instance Y is earnings, X is education, and Z is ability)
- but unfortunately we do not observe Z (ability)
- We compute the OLS estimator using the observable information on earnings (Y) and education (X)
 - $\hat{\beta}_{OLS} = (X'X)^{-1}(X'Y)$
- Let us now compute the bias involved in this estimation

$$\begin{aligned}E[\hat{\beta}_{OLS}] &= E[(X'X)^{-1}(X'Y)] = E[(X'X)^{-1}(X'(X\beta + Z\gamma + \epsilon))] = \\&= E[\cancel{(X'X)^{-1}} \cancel{X'X}\beta + \cancel{(X'X)^{-1}} X'Z\gamma + \underbrace{\cancel{(X'X)^{-1}} \cancel{(X'\epsilon)}}_{E(X'\epsilon)=0}] = \\&= \beta + \underbrace{E[(X'X)^{-1} X'Z\gamma]}_{\text{omitted variable bias}}\end{aligned}$$

What happens if we omit a relevant variable?

	$E(X'Z) > 0$	$E(X'Z) < 0$
$\gamma > 0$	POSITIVE BIAS	NEGATIVE BIAS
$\gamma < 0$	NEGATIVE BIAS	POSITIVE BIAS

The effect of children on the gender wage gap

- Marianne Bertrand, Claudia Goldin and Lawrence Katz (2010) study the careers of MBAs who graduated between 1990 and 2006 from a top US business school -the Booth School of Business of the University of Chicago- and how career dynamics differ by gender.
- They explore the evolution of the gender gap in earnings and labor supply for young professionals employed primarily in corporate, consulting, and financial services jobs.
- In the following Table, column (5) describes the relationship between the (log) number of hours worked and gender. Furthermore, column (6) provides also information about the difference in the number of hours worked for women with and without children.

Possible solution: identification based on observables

TABLE 5—DETERMINANTS OF THE GENDER GAP IN LABOR SUPPLY: THE ROLE OF CHILDREN

Dependent variable	Not working		Actual post-MBA experience		Log (weekly hours worked)	
	(1)	(2)	(3)	(4)	(5)	(6)
Female	0.084 [0.009]***		-0.286 [0.039]***		-0.089 [0.013]***	
Female with child		0.200 [0.024]***		-0.660 [0.094]***		-0.238 [0.031]***
Female without child		0.034 [0.007]***		-0.126 [0.031]***		-0.033 [0.012]***
Pre-MBA characteristics	Yes	Yes	Yes	Yes	Yes	Yes
MBA performance	Yes	Yes	Yes	Yes	Yes	Yes
Cohort × year dummies	Yes	Yes	Yes	Yes	Yes	Yes
Constant	-0.175 [0.145]	-0.111 [0.126]	5.929 [0.618]***	5.757 [0.550]***	3.951 [0.462]***	3.914 [0.426]***
Observations	19,366	19,286	19,366	19,286	18,611	18,535
R ²	0.07	0.11	0.98	0.98	0.14	0.16

Notes: The unit of observation is a survey respondent in a given post-MBA year. “Female with child” (“Female without child”) is a dummy variable that equals 1 if the respondent is a female and has at least one child (no child) in that year. Pre-MBA characteristics include: a dummy for US citizen, a white dummy, an Asian dummy, a dummy for “top 10” undergraduate institution, a dummy for “top 10–20” undergraduate institution, undergraduate GPA, a dummy for missing undergraduate GPA, a quadratic in age, verbal GMAT score, quantitative GMAT score, a dummy for pre-MBA industry and a dummy for pre-MBA job function. MBA performance includes overall MBA GPA and fraction of finance classes. Standard errors (in brackets) are clustered at the individual level.

***Significant at the 1 percent level.

Poll question 2

- Can we give a causal interpretation to the observed difference in the number of hours worked by women with and without children? Discuss potential selection issues and try to interpret results in terms of bounds.
- Which of the following statements is correct (more than one might be correct):
 - ① Women with children work 23.8% fewer hours than women without children
 - ② Women with children work 20.5% fewer hours than women without children
 - ③ Children decrease working hours by 20.5%
 - ④ Children decrease working hours by at most 20.5% (assuming ‘negative’ selection into motherhood)

Poll question 2

- Can we give a causal interpretation to the observed difference in the number of hours worked by women with and without children? Discuss potential selection issues and try to interpret results in terms of bounds.
- Which of the following statements is correct (more than one might be correct):
 - ① Women with children work 23.8% fewer hours than women without children
 - ② *Women with children work 20.5% fewer hours than women without children*
 - ③ Children decrease working hours by 20.5%
 - ④ *Children decrease working hours by at most 20.5% (assuming ‘negative’ selection into motherhood)*

Main threats to the validity

- Main threats to validity
 - ➊ Omitted variable bias
 - ➋ Measurement error in the independent variable
 - ➌ Measurement error in the dependent variable
 - ➍ Bad controls
 - ➎ Reverse causality

Poll question 3

- Imagine that the variable ‘number of hours worked’ was subject to some type of random measurement error such as rounding. How would this affect the interpretation of point estimates?
- Which of the following statements is correct (more than one might be correct):
 - ① It would bias upwards the OLS estimate
 - ② It would introduce attenuation bias
 - ③ It would decrease the accuracy of the estimation
 - ④ None of the above

Poll question 3

- Imagine that the variable ‘number of hours worked’ was subject to some type of random measurement error such as rounding. How would this affect the interpretation of point estimates?
- Which of the following statements is correct (more than one might be correct):
 - ① It would bias upwards the OLS estimate
 - ② It would introduce attenuation bias
 - ③ *It would decrease the accuracy of the estimation*
 - ④ None of the above

Poll question 4

- Imagine that the variable ‘children’ was subject to some type of random measurement error. How would this affect the interpretation of point estimates?
- Which of the following statements is correct (more than one might be correct):
 - ① It would bias upwards the OLS estimate
 - ② It would introduce attenuation bias
 - ③ It would decrease the accuracy of the estimation
 - ④ None of the above

Poll question 4

- Imagine that the variable ‘children’ was subject to some type of random measurement error. How would this affect the interpretation of point estimates?
- Which of the following statements is correct (more than one might be correct):
 - ① It would bias upwards the OLS estimate
 - ② *It would introduce attenuation bias*
 - ③ It would decrease the accuracy of the estimation,
 - ④ None of the above

Main threats to the validity

- Main threats to validity
 - ❶ Omitted variables bias
 - ❷ Measurement error in the independent variable
 - ❸ Measurement error in the dependent variable
 - ❹ Bad controls
 - ❺ Reverse causality

Poll question 5

- The authors control in their regressions for Pre-MBA characteristics and MBA performance. In addition to these controls, imagine you also have information about their current individual weight and height, two variables that are known to be correlated with labor market outcomes. Discuss whether it might be a good idea to take into account any of these two variables in the regression if we want to estimate how having kids affects female labor force participation.
- You would control for:
 - ① Both variables
 - ② Just height
 - ③ Just weight
 - ④ None of the above

Poll question 5

- The authors control in their regressions for Pre-MBA characteristics and MBA performance. In addition to these controls, imagine you also have information about their current individual weight and height, two variables that are known to be correlated with labor market outcomes. Discuss whether it might be a good idea to take into account any of these two variables in the regression if we want to estimate how having kids affects female labor force participation.
- You would control for:
 - ① Both variables
 - ② *Just height*
 - ③ Just weight
 - ④ None of the above

Additional example: Does class size affect students' performance?

Tennessee STAR experiment

- How can we improve students' performance?
- Should we devote more resources to reduce **class size**?
 - Example: Should we split this course in two separate groups?
- A large number of observational studies tend to find that class size is not generally associated to better student performance
 - Hanushek (1997): “No strong or systematic relationship between school inputs and student achievement”

Percentage distribution of estimated effect of key resources on student performance, based on 376 studies

Resources	Number of estimates	Statistically significant		Statistically insignificant
		Positive	Negative	
Real classroom resources				
Teacher–pupil ratio	276	14%	14%	72%
Teacher education	170	9	5	86
Teacher experience	206	29	5	66
Financial aggregates				
Teacher salary	118	20	7	73%
Expenditure per pupil	163	27	7	66
Other				
Facilities	91	9	5	86
Administration	75	12	5	83

Example: Does class size affect students' performance?

Tennessee STAR experiment

Tennessee STAR experiment

- Cost: \$12 million
- A cohort of kindergartners in 1985/86: 11,600 children in 80 schools
- The study ran for four years
- Three treatments:
 - ① small classes with 13-17 children
 - ② regular classes with 22-25 children without a teacher's aide.
 - ③ regular classes with 22-25 children with a teacher's aide.
- Within each school, students are randomly assigned to one of these groups

Example: Does class size affect students' performance?

Tennessee STAR experiment

Krueger 1999 provides an econometric analysis of the short-run effects of the experiment

- Main findings:
 - ❶ performance on standardized tests increases by four percentile points the first year students attend small classes
 - ❷ teacher aides and measured teacher characteristics have little effect
- Note: Hanushek has written also a critical article about the pitfalls of the STAR experiment

TABLE I
 COMPARISON OF MEAN CHARACTERISTICS OF TREATMENTS AND CONTROLS:
 UNADJUSTED DATA

Variable	Small	Regular	Regular/Aide	Joint P-Value ^a
1. Free lunch ^c	.47	.48	.50	.09
2. White/Asian	.68	.67	.66	.26
3. Age in 1985	5.44	5.43	5.42	.32
4. Attrition rate ^d	.49	.52	.53	.02
5. Class size in kindergarten	15.1	22.4	22.8	.00
6. Percentile score in kindergarten	54.7	49.9	50.0	.00

Table 2.2.2: Experimental estimates of the effect of class-size assignment on test scores

Explanatory variable	(1)	(2)	(3)	(4)
Small class	4.82 (2.19)	5.37 (1.26)	5.36 (1.21)	5.37 (1.19)
Regular/aide class	.12 (2.23)	.29 (1.13)	.53 (1.09)	.31 (1.07)
White/Asian (1 = yes)	—	—	8.35 (1.35)	8.44 (1.36)
Girl (1 = yes)	—	—	4.48 (.63)	4.39 (.63)
Free lunch (1 = yes)	—	—	-13.15 (.77)	-13.07 (.77)
White teacher	—	—	—	-.57 (2.10)
Teacher experience	—	—	—	.26 (.10)
Master's degree	—	—	—	-0.51 (1.06)
School fixed effects	No	Yes	Yes	Yes
R ²	.01	.25	.31	.31

Example: Does class size affect students' performance?

Tennessee STAR experiment

- Several ‘tricky’ questions:
 - ① Would you prefer that your children are taught by a teacher with a Master’s degree?
 - ② Does taking a Master’s degree improve teachers’ teaching skills?
- R-square - do we care for causal questions?

Thank you!

EC902/EC907: Econometrics A

Lecture 8.1

Manuel Bagues

Warwick University

Roadmap

- Main threats to validity of OLS:
 - ❶ Omitted variables
 - ❷ Bad controls
 - ❸ Measurement error in the independent variable
 - ❹ Measurement error in the dependent variable
 - ❺ Reverse causality
- Examples:
 - ❶ Example: The Impact of Social Networks on Remuneration in Top Executive Jobs ([slides 8.1](#))
 - ❷ Example: The Returns to Computer Use ([slides 8.2](#))
- Standard errors ([slides 8.3](#))

Example

The Impact of Social Networks on Remuneration in Top Executive Jobs
Lalanne and Seabright 2011

- This paper investigates the impact of social networks on earnings. The authors use a large data set of more than 80,000 individuals working in high positions in almost 4,000 US and UK firms over a 12 year period (from 1997 to 2009).
- The dataset contains information about individuals' demographic characteristics such as age, nationality and gender, about individuals' employment history such as earnings and position, about individuals' education characteristics such as degree obtained, field and university, and about firms' characteristics such as market capitalization, sector or number of employees.

Example

The Impact of Social Networks on Remuneration in Top Executive Jobs

- The authors use this original dataset to create a new variable, called **links**, which measures the number of individuals with whom an individual in the dataset has worked in the same firm at some point of time.
- The authors argue that this variable measures professional networks. On average **links** is equal to 200. The authors find that, controlling for other factors, individuals who have overlapped professionally with a larger number of people have higher salaries.

Example

The Impact of Social Networks on Remuneration in Top Executive Jobs

- In particular, they run the following regression:

$$\text{Log}(\text{salar} y_i) =$$

$$\beta_1 + \beta_2 * \text{links}_i + \beta_3 * \text{female}_i + \beta_4 * \text{age}_i + \beta_5 * \text{education}_i + \epsilon_i$$

- The OLS estimate of β_2 is positive and this coefficient is significantly different from zero at the standard levels. The authors conclude that links have a causal positive effect on the remuneration of top executives.

Example

The Impact of Social Networks on Remuneration in Top Executive Jobs

Reply to the following questions:

- ① Discuss the possible existence of measurement error in the variable **links** and how this would affect results.
- ② Discuss the possible existence of measurement error in the variable **salary** and how this would affect results.
- ③ Let us assume for simplicity that there is no measurement error. Which additional assumptions do we need to make in order to give a **causal interpretation** to the statistical correlation observed by the authors?
 - Explain verbally why the variable **links** might not be orthogonal to the error term, providing some examples.
 - Explain verbally which kind of variation you would need to exploit in order to interpret results in a causal way

Example

The Impact of Social Networks on Remuneration in Top Executive Jobs

- ④ Would it be a good idea to control for occupation? And for firm fixed effects?
- ⑤ Can you think about some alternative identification strategy to identify the effect of networks?

EC902/EC907: Econometrics A

Lecture 8.2

Manuel Bagues

Warwick University

Roadmap

- Main threats to validity of OLS:
- Examples:
 - ① Example: The Impact of Social Networks on Remuneration in Top Executive Jobs (slides 8.1)
 - ② Example: The Returns to Computer Use (slides 8.2)
 - Krueger 1993: How computers have changed the wage structure: evidence from microdata, 1984-1989
 - DiNardo and Pischke (1997): “The Returns to Computer Use Revisited: Have Pencils Changed the Wage Structure Too?”
- Standard errors (slides 8.3)

Example: The returns to computer use

- Next, let us analyze a classical paper that illustrates the perils of identification based on observables:
 - Alan B. Krueger (1993)
'How computers have changed the wage structure: evidence from microdata, 1984-1989'
The Quarterly Journal of Economics, Vol. 108., No. 1, 33-60.

Motivation I

- During the 80's and the beginning of the 90's there was a big change in the wage structure in the US and other countries. The relative advantage of college graduates over high-school graduates has increased from 34% in 1979 to 56% in 1991.
- What is the root for these changes?
 - International competition
 - Skill-biased technological change

Motivation II

- The reported use of computers has more than tripled during the analyzed period.
- The paper analyzes whether employees who use computers at work earn more as a result of applying their computer skills.
- Microdata allow the author to explore the robustness of the relation between computer use and wages to many control variables.

Theory

- The effect of returns to education: The new computer technology could be complement or substitute for skilled labor.
- The effect on earnings: Presumably positive, but its magnitude should depend on how difficult it is to acquire computer skills (ie: email)

Data

- Main data source: Current Population Surveys (CPS) in the US, conducted in 1984 and 1989.
- Individuals were asked if they had “direct or hands on use of computers” at work. Computer use includes programming, word processing, e-mail, etc.

TABLE I
 PERCENT OF WORKERS IN VARIOUS CATEGORIES WHO DIRECTLY
 USE A COMPUTER AT WORK

Group	1984	1989
All workers	24.6	37.4
<u>Gender</u>		
Men	21.2	32.3
Women	29.0	43.4
<u>Education</u>		
Less than high school	5.0	7.8
High school	19.3	29.3
Some college	30.6	45.3
College	41.6	58.2
Postcollege	42.8	59.7
<u>Race</u>		
White	25.3	38.5
Black	19.4	27.7
<u>Age</u>		
Age 18–24	19.7	29.4
Age 25–39	29.2	41.5
Age 40–54	23.6	39.1
Age 55–65	16.9	26.3
<u>Occupation</u>		
Blue-collar	7.1	11.6
White-collar	33.0	48.4

Methodology

- First, Krueger estimates the standard earnings function augmenting it with the dummy variable indicating whether an individual uses a computer at work, C :

$$\ln W_i = X_i \beta + C \alpha + \epsilon_i$$

where X includes various individual characteristics: education, experience, race, gender, marital status, veteran status, union status, occupation dummies (8 dummies [48 dummies]), regional dummies.

- Second, he disaggregates C for specific computer uses.

TABLE II
OLS REGRESSION ESTIMATES OF THE EFFECT OF COMPUTER USE ON PAY
(DEPENDENT VARIABLE: ln (HOURLY WAGE))

Independent variable	October 1984			October 1989		
	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	1.937 (0.005)	0.750 (0.023)	0.928 (0.026)	2.086 (0.006)	0.905 (0.024)	1.094 (0.026)
Uses computer at work (1 = yes)	0.276 (0.010)	0.170 (0.008)	0.140 (0.008)	0.325 (0.009)	0.188 (0.008)	0.162 (0.008)
Years of education	— (0.001)	0.069 (0.002)	0.048 (0.002)	— (0.002)	0.075 (0.002)	0.055 (0.002)
Experience	— (0.001)	0.027 (0.001)	0.025 (0.001)	— (0.001)	0.027 (0.001)	0.025 (0.001)
Experience-squared ÷ 100	— (0.002)	-0.041 (0.002)	-0.040 (0.002)	— (0.002)	-0.041 (0.002)	-0.040 (0.002)
Black (1 = yes)	— (0.013)	-0.098 (0.012)	-0.066 (0.012)	— (0.013)	-0.121 (0.013)	-0.092 (0.012)
Other race (1 = yes)	— (0.020)	-0.105 (0.019)	-0.079 (0.019)	— (0.020)	-0.029 (0.020)	-0.015 (0.020)
Part-time (1 = yes)	— (0.010)	-0.256 (0.010)	-0.216 (0.010)	— (0.010)	-0.221 (0.010)	-0.183 (0.010)
Lives in SMSA (1 = yes)	— (0.007)	0.111 (0.007)	0.105 (0.007)	— (0.007)	0.138 (0.007)	0.130 (0.007)
Veteran (1 = yes)	— (0.011)	0.038 (0.011)	0.041 (0.011)	— (0.012)	0.025 (0.012)	0.031 (0.011)
Female (1 = yes)	— (0.012)	-0.162 (0.012)	-0.135 (0.012)	— (0.012)	-0.172 (0.012)	-0.151 (0.012)
Married (1 = yes)	— (0.011)	0.156 (0.011)	0.129 (0.011)	— (0.011)	0.159 (0.011)	0.143 (0.011)
Married*Female	— (0.015)	-0.168 (0.015)	-0.151 (0.015)	— (0.015)	-0.141 (0.015)	-0.131 (0.015)
Union member (1 = yes)	— (0.009)	0.181 (0.009)	0.194 (0.009)	— (0.010)	0.182 (0.010)	0.189 (0.010)
8 Occupation dummies	No 0.051	No 0.446	Yes 0.491	No 0.082	No 0.451	Yes 0.486
<i>R</i> ²						

Notes. Standard errors are shown in parentheses. Sample size is 13,335 for 1984 and 13,379 for 1989. Columns (2), (3), (5), and (6) also include three region dummy variables.

Results

- Unconditionally, the computer use is associated with 27.6 % higher wages in 1984 and with 32.5 % higher wages in 1989.
- (Few) observable individual characteristics account for about a half of this effect.

TABLE III
THE RETURN TO VARIOUS USES OF COMPUTERS, OCTOBER 1989^a
(DEPENDENT VARIABLE: ln (HOURLY WAGE))

Use of computer at work	Proportion	Coefficient (std. error)
Uses computer at work for any task ^b	0.398	0.145 (0.010)
<u>Specific Task^c</u>		
Word processing	0.165	0.017 (0.012)
Bookkeeping	0.100	-0.058 (0.013)
Computer-assisted design	0.039	0.026 (0.020)
Electronic mail	0.063	0.149 (0.016)
Inventory control	0.102	-0.056 (0.013)
Programming	0.077	0.052 (0.031)
Desktop publishing or newsletters	0.036	-0.047 (0.021)
Spread sheets	0.094	0.079 (0.015)
Sales	0.060	-0.002 (0.016)
Computer games	0.019	-0.109 (0.026)
<i>R</i> ²		0.495

a. The sample and other explanatory variables are the same as in column (6) of Table II.

b. The computer use dummy variable equals one if the worker uses computers for any of the ten enumerated tasks or for any other task.

- The highest returns could be observed to electronic mail use!
Computer games show no positive effect (too bad...)
- Why email effect? Should we be concerned about selection issues?

- The author also controls for computer use at home in the following way:

$$\ln W_i = X_i \beta + C_w \alpha_1 + C_h \alpha_2 + C_x * C_h \alpha_3 + \epsilon_i$$

The rationale: Controlling for unobserved individual characteristics, which are related, first, to an individual propensity to acquire computer skills, and, second, to the probability that an employee selects this individual for jobs requiring computer use.

TABLE IV
 THE RETURN TO COMPUTER USE AT WORK, HOME, AND WORK AND HOME
 (STANDARD ERRORS ARE SHOWN IN PARENTHESES.)

Type of computer use	October 1984 (1)	October 1989 (2)	Percent of sample, 1989 (3)
Uses computer at work	0.165 (0.009)	0.177 (0.009)	39.8
Uses computer at home	0.056 (0.021)	0.070 (0.019)	12.5
Uses computer at home and work	0.006 (0.029)	0.017 (0.023)	8.6
Sample size	13,335	13,379	

Notes. The table reports coefficients for three dummy variables estimated from log hourly wage regressions. The other explanatory variables in the regressions are education, experience and its square, two race dummies, three region dummies, dummy variables indicating part-time status, residence in an SMSA, veteran status, gender, marital status, union membership, and an interaction between marital status and gender. Covariates are the same as in columns (2) and (5) of Table II.

- The effect of computer at work is still high even after controlling for the computer use at home.
- The estimated effect of the computer use at work is higher than the effect of computer use generally.

Conclusions

- The paper analyzes the effect of computer use at work on earnings.
- Various estimation models suggest that computer use at work implies around 10-15 % increase in earnings.
- Given that more educated people are more likely to use computers, Krueger calculates that computer use might account for more than a third of the increase in the rate of return to education in the US.

Some criticisms

The validity of the study depends on several strong assumptions:

- Conditional on 8 [48] occupational categories and few observable individual characteristics, the assignment of workers to computer use is (as good as) random.
- In other words, computer use is not associated to any other characteristics of the job or employer.
- No unobservable individual characteristics (talent, motivation, social skills, etc.) are correlated both with the probability to use computer and higher productivity.

Di Nardo and Pischke (1997)

- Di Nardo and Pischke (1997) question the analysis by Krueger (1993):
 - “The Returns to Computer Use Revisited: Have Pencils Changed the Wage Structure Too?”, *The Quarterly Journal of Economics*, Vol. 112, No. 1 (Feb., 1997), pp. 291-303.

Motivation

- They replicate the study of Krueger (1993), in this case for German workers, and introduce many more controls, including the ones capturing the effect of other specific tools used by workers at work such as writing materials like a pen or pencil, or sitting on the job.
- Include rich set of detailed occupation dummies

TABLE I
 PERCENT OF WORKERS IN VARIOUS CATEGORIES WHO USE DIFFERENT TOOLS
 ON THEIR JOB

Group	U. S. 1984	U. S. 1989	U. S. 1993	Germany 1979	Germany 1985–1986	Germany 1991–1992
Percentage that are computer users						
All workers	25.1	37.4	46.6	8.5	18.5	35.3
Men	21.6	32.2	41.1	7.9	18.5	36.4
Women	29.6	43.8	53.2	9.7	18.5	33.5
Less than high school	5.1	7.7	10.4	3.2	4.3	9.9
High school	19.2	28.4	34.6	8.5	18.3	32.7
Some college	30.6	45.0	53.1	8.5	24.8	48.4
College	42.4	58.8	70.2	13.4	30.5	61.6
Age 18–24	20.5	29.6	34.3	10.1	13.8	27.8
Age 25–39	29.6	41.4	49.8	9.6	21.6	39.9
Age 40–54	23.9	38.9	50.0	6.6	17.2	35.9
Age 55–64	17.7	27.0	37.3	5.9	13.5	23.7
Blue-collar	7.1	11.2	56.6	1.2	3.5	10.7
White-collar	39.7	56.6	67.6	12.8	28.9	50.2
Part-time	14.8	24.4	29.3	6.4	14.7	26.5
Full-time	29.3	42.3	51.0	8.7	19.1	37.0
Percentage of all workers who use a specific tool						
Computer	25.1	37.4	46.6	8.5	18.5	35.3
Calculator				19.6	35.7	44.2
Telephone				41.8	43.7	58.4
Pen/pencil				54.9	53.4	65.6
Work while sitting ^a				30.8	19.3	—
Hand tool (e.g., hammer)				29.4	32.9	30.5
Number of obs.	61,704	62,748	59,852	19,427	22,353	20,042

a. Variable definition differs in 1979 and 1985–1986. In 1979 it refers to "Never or rarely standing," and in 1985–1986 it refers to "Often or almost always sitting."

TABLE II
 OLS REGRESSIONS FOR THE EFFECT OF COMPUTER USE ON PAY
 DEPENDENT VARIABLE: LOG HOURLY WAGE
 (STANDARD ERRORS IN PARENTHESES)

Independent variable	U. S. 1984	U. S. 1989	U. S. 1993	Germany 1979	Germany 1985–1986	Germany 1991–1992
Computer	0.171 (0.008)	0.188 (0.008)	0.204 (0.008)	0.112 (0.010)	0.157 (0.007)	0.171 (0.006)
Years of schooling	0.068 (0.001)	0.075 (0.002)	0.081 (0.002)	0.073 (0.001)	0.063 (0.001)	0.072 (0.001)
Experience	0.028 (0.001)	0.028 (0.001)	0.026 (0.001)	0.030 (0.001)	0.035 (0.001)	0.030 (0.001)
Experience ^{2/} 100	-0.043 (0.002)	-0.043 (0.002)	-0.041 (0.003)	-0.052 (0.002)	-0.058 (0.002)	-0.046 (0.002)
R ²	0.444	0.448	0.424	0.267	0.280	0.336
Number of obs.	13,335	13,379	13,305	19,427	22,353	20,042

Columns 1 to 3 are from Table 4 in Autor, Katz, and Krueger [1996]. Data for columns 1 to 3 are from the October *Current Population Survey*; data for columns 4 to 6 are from the *Qualification and Career Survey*. All models also include an intercept, a dummy for part-time, large city/SMSA status, female, married, female*married. Regressions for the United States in columns 1 to 3 also include dummies for black, other race, veteran status, union membership, and three regions. Regressions for Germany in columns 4 to 6 also include a dummy for civil servants (*Beamter*).

- The estimated returns to computer use at work are similar to the ones of Autor, Katz, and Krueger (1996), when following the identification strategy of the latter study.

TABLE III
 OLS REGRESSION FOR THE EFFECT OF DIFFERENT TOOLS ON PAY
 DEPENDENT VARIABLE: LOG HOURLY WAGE
 (STANDARD ERRORS IN PARENTHESES)

Independent variable	Germany 1979	Germany 1985–86	Germany 1991–92	Germany 1979	Germany 1979	Germany 1985–1986	Germany 1991–1992
Occupation indicators	No	No	No	501	501	742	1071
Grades and father's	No	No	No	No	Yes	No	No
Occupation ^a							

	Tools entered together						
Computer	0.066 (0.010)	0.105 (0.008)	0.126 (0.007)	0.027 (0.011)	0.024 (0.011)	0.067 (0.008)	0.069 (0.007)
Calculator	0.017 (0.008)	0.053 (0.007)	0.044 (0.007)	0.015 (0.008)	0.014 (0.008)	0.032 (0.008)	0.022 (0.007)
Telephone	0.072 (0.007)	0.043 (0.008)	0.045 (0.008)	0.043 (0.008)	0.041 (0.008)	0.035 (0.008)	0.048 (0.008)
Pen/pencil	0.062 (0.007)	0.031 (0.008)	0.035 (0.008)	0.040 (0.008)	0.038 (0.008)	0.024 (0.008)	0.007 (0.008)
Work while sitting	0.058 (0.007)	0.050 (0.007)	—	0.036 (0.008)	0.035 (0.008)	0.032 (0.008)	— (0.008)

a. Two variables for self-reported grades in math and German and eleven dummy variables for father's education.

Data are from the *Qualification and Career Survey*. All regressions also include an intercept, years of schooling, experience and experience squared, dummies for part-time, city, female, married, married*female, and for civil servants (*Beamter*).

- The estimated returns to computer use at work decrease substantially when we include controls for the use of other tools at work and a more detailed set of occupation dummies.
- Moreover, pencil effect is as big, even though literacy is 99% in Germany. What does it proxy for? Selection of certain types of individuals into certain jobs.
- Overall, the findings of Di Nardo and Pischke (1997) suggest that Krueger's (1993) estimation of the returns to computer use is subject to an important omitted variable bias.

Bounds approach

- In the presence of selection, we might prefer to think in terms of lower bounds and upper bounds in order to interpret observational evidence.
- In this case, if we suspect that there is positive selection, the above results might be interpreted as an upper bound of the returns to computer use.
- It is also common to use the following heuristic for evaluating the robustness of results to omitted variable bias:
 - Check how the coefficient of interest varies when you include controls. (Intuition: if individuals in the treatment and the control group differ in observables, they are likely to differ in (relevant) unobservables.)
 - Building on this idea, some authors proposed a formal way to take into account the degree of selection on observed variables to assess the extent of selection on unobserved variables
 - Altonji, Elder and Taber 2005, Oster 2014)

Computer use vs. computer skills

- Beyond the potential existence of an omitted variable bias, what is the relevant ‘treatment’ we should be considering?
- To deal with this endogeneity problem, we could run the following RCT: randomly assign a computer to a group of previous nonusers and then compare their wages with those of an untreated comparison group.
- But this does not make too much sense: (i) computers are only productive in conjunction with a specific set of skills (e.g., programming); (ii) computers are of value only in certain jobs.
- The relevant treatment are computer skills!

EC902/EC907: Econometrics A

Lecture 8.3

Manuel Bagues

Warwick University

Roadmap

- ❶ Empirical strategies relying on observational data
- ❷ Identification based on observables
 - Conditional independence assumption
 - Ideal experiment
 - Threats to validity
 - Example 1: the effect of networks (slides 8.1)
 - Example 2: the returns to computer use (slides 8.2)
 - Standard errors ([slides 8.3](#))

Standard errors

- So far we have focused on how we can *identify* the effect of a treatment (**unbiasedness/consistency**)
- Let us now focus on the ‘accuracy’ of estimates (**inference**)
- Let us consider the following model:
 - $Y = XB + \epsilon$
- As you know:
 - $\hat{\beta}_{OLS} = (X'X)^{-1}(X'Y)$
- We want to find the variance-covariance matrix of $\hat{\beta}_{OLS}$ (for simplicity in what follows let us drop the OLS subindex)

$$Var(\hat{\beta}|X) = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'|X]$$

Reminder: Expected value of $\hat{\beta}_{OLS}$

This first step is just plugging in the population model: $y = X\beta + \epsilon$

$$E(\hat{\beta}) = E[(X'X)^{-1}X'y] = E[(X'X)^{-1}X'(X\beta + \epsilon)]$$

Expanding out:

$$= E[(X'X)^{-1}X'(X\beta) + (X'X)^{-1}X'\epsilon)]$$

Then most of the $X'X$ terms cancel out:

$$= E[\beta + (X'X)^{-1}X'\epsilon)]$$

The β is taken outside the operator since it is a constant:

$$= \beta + E[(X'X)^{-1}X'\epsilon)]$$

And therefore

$$E(\hat{\beta} - \beta) = E[(X'X)^{-1}X'\epsilon)]$$

The Variance-Covariance Matrix of the OLS Estimates

$$\begin{aligned}Var(\hat{\beta}|X) &= E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'|X] = E[(X'X)^{-1}X'\epsilon)(X'X)^{-1}X'\epsilon)'|X] \\&= E[(X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1}|X]\end{aligned}$$

where we took advantage of the fact that $(AB)' = B'A'$ and therefore we can rewrite $(X'X)^{-1}X'\epsilon)'$ as $\epsilon'X(X'X)^{-1}$.

Given that X is non-stochastic, we get:

$$E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'|X] = (X'X)^{-1}X'E[\epsilon\epsilon'|X]X(X'X)^{-1}$$

where

$$E(\epsilon\epsilon'|X) = \Omega = \begin{bmatrix} E[\epsilon_1^2|X] & E[\epsilon_1\epsilon_2|X] & \dots & E[\epsilon_1\epsilon_n|X] \\ E[\epsilon_2\epsilon_1|X] & E[\epsilon_2^2|X] & \dots & E[\epsilon_2\epsilon_n|X] \\ \vdots & \vdots & \vdots & \vdots \\ E[\epsilon_n\epsilon_1|X] & E[\epsilon_n\epsilon_2|X] & \dots & E[\epsilon_n^2|X] \end{bmatrix}$$

Variance of estimator β_{OLS}

We can consider several possible cases.

- ① Spherical errors (aka "Standard" standard errors)
- ② Robust standard errors (heteroskedasticity consistent standard errors)
- ③ Clustered standard errors

1. Spherical errors

No serial correlation and no heteroskedasticity

- Two assumptions:
 - No serial correlation
 - No heteroskedasticity

$$E(\epsilon\epsilon'|X) = \sigma^2 \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} = \sigma^2 I$$

Then

$$Var(\hat{\beta}|X) = (X'X)^{-1} X' (\sigma^2 I) X (X'X)^{-1} = \sigma^2 (X'X)^{-1}$$

- and we estimate σ^2 with $\hat{\sigma}^2$, where:

$$\hat{\sigma}^2 = \frac{e'e}{n - k}$$

- where $e = \hat{y} - y = X\hat{\beta} - y$ is the vector of residuals, n is the number of observations and k is the number of covariates.

$$\widehat{Var(\hat{\beta}|X)} = \frac{e'e}{n - k}(X'X)^{-1}$$

- the square root of which is the familiar standard error that we use to construct confidence intervals or perform significance tests

Practical notes

- This is the standard error that Stata calculates by default
 - `reg y x`
- However, because assuming spherical errors is often unrealistic (and because it is trivial to relax this assumption), these "standard" standard errors are rarely used in practice

2. Robust standard errors

- Also known as heteroskedasticity consistent standard errors

$$E(\epsilon\epsilon'|X) = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & & \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

- Common example of heteroskedasticity: the errors tend to be larger for larger values of y
- In stata `vce()` specifies how to estimate the variance-covariance matrix (VCE)
 - `reg y x, vce(robust)`

- Some remarks:
 - ❶ robust st. errors are (almost) always preferable to "standard" ones
 - ❷ the type of st. errors you calculate does not affect the point estimate

3. Clustered standard errors

- Very often we cannot assume that unobserved shocks are independent across individuals
- Within some groups or clusters, observations may be exposed to common shocks
- Example: in the STAR experiment on the impact of class size, students in the same classroom are likely to be exposed to common shocks such as being exposed to the same bad/good peer, teacher...
- We will assume independence across ‘clusters’ but ...
- ... allow for dependence within ‘clusters’ (or groups), and estimate variance and covariance of uncertainty within groups.

Example with m clusters, each one with n_j observations

$$E(\epsilon\epsilon') = \begin{bmatrix} \sigma_{(1,1)1}^2 & \cdots & \sigma_{(1,n_1)1}^2 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{(n_1,1)1}^2 & \cdots & \sigma_{(n_1,n_1)1}^2 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \sigma_{(1,1)2}^2 & \cdots & \sigma_{(1,n_2)2}^2 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \sigma_{(n_2,1)2}^2 & \cdots & \sigma_{(n_2,n_2)2}^2 \\ & & & & \ddots & \\ & & & & & \sigma_{(1,1)m}^2 & \cdots & \sigma_{(1,n_m)m}^2 \\ & & & & & \vdots & \ddots & \vdots \\ & & & & & \sigma_{(n_m,1)m}^2 & \cdots & \sigma_{(n_m,n_m)m}^2 \end{bmatrix}$$

Some remarks (i)

- ❶ typically clustered standard errors are larger than robust or "standard" errors
- ❷ the type of standard errors you calculate does not affect the point estimate
- ❸ how to choose the relevant level of clustering:
 - level at which common shocks are likely to operate
 - at least the level at which the treatment is defined (e.g. impact of a policy implemented at the class level → cluster at least the class level)
 - not always trivial to choose the correct level (e.g. STAR experiment - should we cluster at the class or at the school level?)
 - larger cluster level → more conservative standard errors
- ❹ In some sense, clustering implies acknowledging how many independent sources of information there are in the data (e.g. the number of clusters is essentially your N)

Some remarks (ii)

- How do you estimate clustered standard errors?
 - ① When the number of groups is large enough (rule of thumb: $N > 50$), use the ‘sandwich formula’
 - `reg y x, vce(cluster 'clustvar')`
 - ② When the number of groups is small, the corresponding asymptotic properties do not hold. There are some alternatives:
 - Randomization inference
 - Block-bootstrap

Example

- As an illustration let us use the PISA dataset and try different ways of calculating standard errors

```
clear
```

```
use "dataset pisa 2015.dta"
```

```
reg reading computer_at_home
```

Source	SS	df	MS	Number of obs	=	498,603
Model	402023729	1	402023729	F(1, 498601)	=	39153.16
Residual	5.1196e+09	498,601	10267.9766	Prob > F	=	0.0000
				R-squared	=	0.0728
				Adj R-squared	=	0.0728
Total	5.5216e+09	498,602	11074.2579	Root MSE	=	101.33

reading	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
:computer_at_home	91.67408	.4633009	197.87	0.000	90.76602 92.58213
_cons	386.4479	.4376915	882.92	0.000	385.5901 387.3058

- Now with robust standard errors:

```
. reg reading computer_at_home , vce(robust)
```

Linear regression

		Number of obs	=	498,603
		F(1, 498601)	=	46991.19
		Prob > F	=	0.0000
		R-squared	=	0.0728
		Root MSE	=	101.33

reading	Robust					
	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
computer_at_home	91.67408	.4229006	216.77	0.000	90.84521	92.50295
_cons	386.4479	.3940102	980.81	0.000	385.6757	387.2202

- Now with clustered standard errors:

```
. reg reading computer_at_home , vce(cluster school_id)

Linear regression                               Number of obs      =  498,603
                                                F(1, 17620)       =  9399.51
                                                Prob > F        =  0.0000
                                                R-squared        =  0.0728
                                                Root MSE         =  101.33

                                                (Std. Err. adjusted for 17,621 clusters in school_id)
```

reading	Robust					
	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
computer_at_home	91.67408	.9455707	96.95	0.000	89.82067	93.52749
_cons	386.4479	.9087378	425.26	0.000	384.6667	388.2292

EC902/EC907: Econometrics A

Lecture 9

Manuel Bagues

Warwick University

November 7, 2022
Lecture Slides

Logistics

- Midterm on Wed Nov 16 - 09:00-10:00 (in moodle)
 - Includes material covered until **today** (lecture 9)
 - Mock exam will be provided at the end of this week
- Q&A session before midterm:
 - Tuesday Nov 15: 12:00-13:00 (Room R0.21)

Roadmap

- Main threats to validity of identification based on observables:
 - ➊ Omitted variables
 - ➋ Bad controls
 - ➌ Measurement error in the independent variable
 - ➍ Measurement error in the dependent variable
 - ➎ Reverse causality
- Examples:
 - ➊ The Impact of Social Networks on Remuneration in Top Executive Jobs (slides 8.1)
 - ➋ The Returns to Computer Use (slides 8.2)
 - ➌ Gender Gaps in PISA Test Scores: Mother's Transmission of Role Attitudes (Problem set 4)
- Standard errors (slides 8.3)

Example 1

The Impact of Social Networks on the Remuneration in Top Executive Jobs
Lalanne and Seabright 2011

- This paper investigates the impact of social networks on earnings. The authors use a large data set of more than 80,000 individuals working in high positions in almost 4,000 US and UK firms over a 12 year period (from 1997 to 2009).
- The dataset contains information about individuals' demographic characteristics such as age, nationality and gender, about individuals' employment history such as earnings and position, about individuals' education characteristics such as degree obtained, field and university, and about firms' characteristics such as market capitalization, sector or number of employees.

- The authors use this original dataset to create a new variable, called **links**, which measures the number of individuals with whom an individual in the dataset has worked in the same firm at some point of time. The authors argue that this variable measures professional networks.
- They run the following regression:

$$\begin{aligned} \text{Log}(salary}_i) = & \beta_1 + \beta_2 * \text{links}_i + \beta_3 * \text{female}_i + \beta_4 * \text{age}_i + \\ & + \beta_5 * \text{education}_i + \epsilon_i \end{aligned}$$

- The OLS estimate of β_2 is positive and this coefficient is significantly different from zero at the standard levels. The authors conclude that links have a causal positive effect on the remuneration of top executives. The effect is larger for men (we will not address this part of the paper in our discussion)
- Note: This working paper was featured in **The Economist** in 2011. However, it has not been accepted for publication at the time of writing (November 2021).

Poll questions

- We will make some polls using [vevox.app](#)
 - ① Please, open [vevox.app](#) in your computer, in Teams or download the app in your mobile
 - ② Enter session ID: 144-921-069

Reply to the following questions:

- ① Discuss the possible existence of measurement error in the variable **links** and how this would affect results.
- ② Discuss the possible existence of measurement error in the variable **salary** and how this would affect results.
- ③ Let us assume for simplicity that there is no measurement error. Which additional assumptions do we need to make in order to give a **causal interpretation** to the statistical correlation observed by the authors?
- ④ Would it be a good idea to control for occupation? And for firm fixed effects?

Poll question 1

- Overall, how would you think about the results of the authors:
 - ① $\hat{\beta}_2$ provides the causal impact of the networks on earnings
 - ② $\hat{\beta}_2$ provides an upper bound for the causal impact of the networks on earnings (assuming positive selection)
 - ③ $\hat{\beta}_2$ provides an lower bound for the causal impact of the networks on earnings (assuming positive selection)

Poll question 1

- Overall, how would you think about the results of the authors:
 - ➊ $\hat{\beta}_2$ provides the causal impact of the networks on earnings
 - ➋ $\hat{\beta}_2$ provides an upper bound for the causal impact of the networks on earnings (assuming positive selection)
 - ➌ $\hat{\beta}_2$ provides an lower bound for the causal impact of the networks on earnings (assuming positive selection)

Another example

- "Does Peacekeeping Keep Peace? International Intervention and the Duration of Peace After Civil War" (**Fortna 2004**)
- Interesting question: does it help to send UN peacekeepers in the aftermath of civil war
- Abstract: *This article examines international interventions in the aftermath of civil wars to see whether peace lasts longer when peacekeepers are present than when they are absent. I attempt to control for factors that might affect both the likelihood of peacekeepers being sent and the ease or difficulty of maintaining peace so as to avoid spurious findings. I find that peacekeeping after civil wars does indeed make an important contribution to the stability of peace.*

First glance at the (post cold war) data: Another round of fighting between the same parties in about 37% when no peacekeepers were deployed, and in approximately 36% of those with peacekeeping.

TABLE 3. Victory, UN Peacekeeping, and the Resumption of War: Post-Cold War

	<i>No More War</i>	<i>More War</i>
No UN peacekeeping	Sri Lanka (JVP II) 1989 Romania 1989 V Iraq-Kurds 1991 V Eritrea 1991 V Iraq-Shiites 1994 V Ethiopia-ideology 1991 V India-Sikh 1994 V Bangladesh-CHT 1994 Mexico 1994 V Azerbaijan 1994 Yemen 1994 V Chad 1994 V Northern Ireland 1994 Moldova 1994 V Djibouti 1994 Mali 1995 Burma 1995 V Papua New Guinea 1997 Cambodia 1998	Sri Lanka (Tamil) 1987 Somalia 1991 V Afghanistan 1992 V Liberia 1993 Israel-Palestine 1993 Philippines-NPA 1993 Congo Brazzaville 1996 Russia-Chechnya 1996 Philip.-MNLF/MILF 1996 Sierra Leone 1996 Congo/Zaire 1997 V
	N = 19	N = 11

UN peacekeeping

Namibia 1989	
Nicaragua 1989	Angola 1991
Lebanon 1991	Cambodia 1991
Morocco/W. Sahara 1991	Yugoslavia-Croatia 1992
Mozambique 1992	Georgia-Abkhazia 1993
El Salvador 1992	Rwanda 1993
Georgia-Ossetia 1994	Yugoslavia-Croatia 1994
South Africa 1994	Rwanda 1994 V
Georgia-Abkhazia 1994	Angola 1994
Guatemala 1994	Sierra Leone 1999
Haiti 1994 V	
Tajikistan 1994	
Yugoslavia-Croatia 1995	
Yugoslavia-Bosnia 1995	
Liberia 1996	
Central Africa 1997	

N = 16

N = 9

V = War ends in victory for one side.

- Challenge: peacekeeping is not applied to cases of civil war at random
- Direction of the selection: possible hypotheses? How would you investigate in which direction it goes?

- Challenge: peacekeeping is not applied to cases of civil war at random
- Direction of the selection: possible hypotheses? How would you investigate in which direction it goes?
 - peacekeepers tend to deploy only to relatively easy cases
 - peacekeepers tend to be sent where they are most needed, when peace would otherwise be difficult to keep
- Using information on observables, the author finds that the latter is likely to be the case

Suggestive evidence of positive selection: Conflicts that ended up in victory less likely to get peacekeepers

TABLE 4. Where Do Peacekeepers Go?

Logistic Regressions

	<i>Post-WWII</i>		<i>Post-Cold War</i>	
	<i>All Peacekeeping</i>		<i>UN Peacekeeping</i>	<i>Non-UN Peacekeeping</i>
	<i>All Peacekeeping</i>	<i>UN Peacekeeping</i>		
Victory	– 3.53*** (1.01)	– 2.44** (1.14)	– 2.26* (1.33)	1.33 (1.66)
Treaty	– 1.04 (1.06)	– 1.44* (0.82)	1.15 (1.24)	0.26 (1.62)
Identity War	0.48 (0.42)	0.69 (0.86)	0.66 (0.82)	0.51 (0.69)
Cost of War	0.07 (0.16)	0.13 (0.19)	0.12 (0.19)	0.06 (0.19)
Duration of War	– 0.002 (0.003)	– 0.004 (0.004)	– 0.003 (0.005)	– 0.009** (0.004)
Many Factions	0.48 (0.55)	0.93 (0.82)	1.67* (0.99)	0.59 (0.79)
Primary Commodity Exports	1.45 (3.77)	1.27 (5.40)	– 9.35** (4.40)	2.30 (4.53)
Development	0.0006* (0.0003)	0.0004 (0.0003)	0.0002 (0.0006)	0.0005 (0.0004)
Prior Democracy	– 0.04 (0.06)	0.05 (0.09)	0.04 (0.13)	0.22** (0.10)
Government Army Size	– 0.003** (0.001)	– 0.003** (0.001)	– 0.005** (0.002)	– 0.003* (0.002)
Constant	0.73 (1.95)	0.37 (2.18)	– 0.88 (2.54)	– 2.89 (2.55)
N	110	52	52	52
Pseudo R ²	0.39	0.31	0.44	0.26
Log Likelihood	– 44.05	– 23.07	– 20.19	– 25.90

Coefficients are reported. Robust standard errors (cases clustered by country) are given in parentheses.

Impact of peacekeepers on probability of new round of fighting (hazard model)

TABLE 7. Effects on the Duration of Peace: Post–Cold War
Cox Proportional Hazards Model: Time-Varying Peacekeeping

	All Peacekeeping	UN Peacekeeping	Non-UN Peacekeeping
Peacekeeping	0.32** (0.18)	0.51* (0.19)	0.34 (0.23)
Victory	0.15 (0.20)	0.24 (0.29)	0.31 (0.35)
Treaty	0.54 (0.64)	0.87 (0.93)	0.78 (0.80)
Identity War	2.33 (1.90)	2.36 (1.80)	2.05 (1.54)
Cost of War	1.43* (0.29)	1.37* (0.23)	1.36* (0.24)
Duration of War	0.99* (0.005)	0.99* (0.005)	0.99 (0.01)
Many Factions	0.93 (0.60)	1.04 (0.60)	1.11 (0.66)
Primary Commodity Exports	9.07 (30.79)	5.52 (18.05)	7.68 (26.70)
Development	0.999* (0.0006)	0.999** (0.001)	0.999 (0.001)
Prior Democracy	1.02 (0.08)	1.01 (0.08)	1.07 (0.07)
Government Army Size	1.001 (0.001)	1.001 (0.001)	1.001 (0.002)
Number of Subjects	51	51	51

Poll question 2

- Let us focus on the coefficient in the first cell, 0.32, which captures the impact of peacekeeping on the duration of peace:
- If we assume that selection in unobservables is likely to mirror selection in observables, should we interpret this estimate as:
 - ① the causal impact of peacekeepers
 - ② as an upper bound for the causal impact of peacekeepers
 - ③ as a lower bound for the causal impact of peacekeepers

Poll question 2

- Let us focus on the coefficient in the first cell, $\downarrow 68\%$, which captures the impact of peacekeeping on the duration of peace:
- If we assume that selection in unobservables is likely to mirror selection in observables, should we interpret this estimate as:
 - ① the causal impact of peacekeepers
 - ② *as an upper bound* for the causal impact of peacekeepers (i.e. the true impact is larger in absolute terms)
 - ③ as a lower bound for the causal impact of peacekeepers (i.e. the true impact is smaller in absolute terms)

Slides 8.2

- Example: The Returns to Computer Use ([slides 8.2](#))
 - Krueger 1993: How computers have changed the wage structure: evidence from microdata, 1984-1989
 - DiNardo and Pischke (1997): “The Returns to Computer Use Revisited: Have Pencils Changed the Wage Structure Too?”

TABLE II
 OLS REGRESSIONS FOR THE EFFECT OF COMPUTER USE ON PAY
 DEPENDENT VARIABLE: LOG HOURLY WAGE
 (STANDARD ERRORS IN PARENTHESES)

Independent variable	U. S. 1984	U. S. 1989	U. S. 1993	Germany 1979	Germany 1985–1986	Germany 1991–1992
Computer	0.171 (0.008)	0.188 (0.008)	0.204 (0.008)	0.112 (0.010)	0.157 (0.007)	0.171 (0.006)
Years of schooling	0.068 (0.001)	0.075 (0.002)	0.081 (0.002)	0.073 (0.001)	0.063 (0.001)	0.072 (0.001)
Experience	0.028 (0.001)	0.028 (0.001)	0.026 (0.001)	0.030 (0.001)	0.035 (0.001)	0.030 (0.001)
Experience ^{2/}	-0.043 100	-0.043 (0.002)	-0.041 (0.002)	-0.052 (0.003)	-0.058 (0.002)	-0.046 (0.002)
R ²	0.444	0.448	0.424	0.267	0.280	0.336
Number of obs.	13,335	13,379	13,305	19,427	22,353	20,042

Columns 1 to 3 are from Table 4 in Autor, Katz, and Krueger [1996]. Data for columns 1 to 3 are from the October *Current Population Survey*; data for columns 4 to 6 are from the *Qualification and Career Survey*. All models also include an intercept, a dummy for part-time, large city/SMSA status, female, married, female*married. Regressions for the United States in columns 1 to 3 also include dummies for black, other race, veteran status, union membership, and three regions. Regressions for Germany in columns 4 to 6 also include a dummy for civil servants (*Beamter*).

TABLE III
 OLS REGRESSION FOR THE EFFECT OF DIFFERENT TOOLS ON PAY
 DEPENDENT VARIABLE: LOG HOURLY WAGE
 (STANDARD ERRORS IN PARENTHESES)

Independent variable	Germany 1979	Germany 1985–86	Germany 1991–92	Germany 1979	Germany 1979	Germany 1985–1986	Germany 1991–1992
Occupation indicators	No	No	No	501	501	742	1071
Grades and father's	No	No	No	No	Yes	No	No
Occupation ^a							

	Tools entered together						
Computer	0.066 (0.010)	0.105 (0.008)	0.126 (0.007)	0.027 (0.011)	0.024 (0.011)	0.067 (0.008)	0.069 (0.007)
Calculator	0.017 (0.008)	0.053 (0.007)	0.044 (0.007)	0.015 (0.008)	0.014 (0.008)	0.032 (0.008)	0.022 (0.007)
Telephone	0.072 (0.007)	0.043 (0.008)	0.045 (0.008)	0.043 (0.008)	0.041 (0.008)	0.035 (0.008)	0.048 (0.008)
Pen/pencil	0.062 (0.007)	0.031 (0.008)	0.035 (0.008)	0.040 (0.008)	0.038 (0.008)	0.024 (0.008)	0.007 (0.008)
Work while sitting	0.058 (0.007)	0.050 (0.007)	—	0.036 (0.008)	0.035 (0.008)	0.032 (0.008)	— (0.008)

a. Two variables for self-reported grades in math and German and eleven dummy variables for father's education.

Data are from the *Qualification and Career Survey*. All regressions also include an intercept, years of schooling, experience and experience squared, dummies for part-time, city, female, married, married*female, and for civil servants (*Beamter*).

- The estimated returns to computer use at work decrease substantially when we include controls for the use of other tools at work and a more detailed set of occupation dummies.
- Moreover, pencil effect is as big, even though literacy is 99% in Germany. What does it proxy for? Selection of certain types of individuals into certain jobs.
- Overall, the findings of Di Nardo and Pischke (1997) suggest that Krueger's (1993) estimation of the returns to computer use is subject to an important omitted variable bias.

Poll question 3

- Di Nardo and Pischke (1997) proposed controlling for occupation at a very detailed level
- Would you say that this might be:
 - ① potentially a good control, because otherwise there is too much unobserved heterogeneity
 - ② potentially a bad control, because it might be affected by the treatment
 - ③ all the above

Poll question 3

- Di Nardo and Pischke (1997) proposed controlling for occupation at a very detailed level
- Would you say that this might be:
 - ① potentially a good control, because otherwise there is too much unobserved heterogeneity
 - ② *potentially a bad control, because it might be affected by the treatment*
 - ③ *all the above*

Dilemma

- Very often we will face this dilemma: should we include a variable that is potentially a bad control?
- If we do not control for occupation, we are probably comparing apples and oranges. People using or not computers are likely to differ in many other dimensions, such as their field of study
- but if we control for occupation (i.e. working in IT), this might be actually capturing the impact of the treatment
- Solution: report results in with and without dubious controls

Bounds approach

- In the presence of selection, we might prefer to think in terms of lower bounds and upper bounds in order to interpret observational evidence.
- In this case, if we suspect that there is positive selection, the above results might be interpreted as an upper bound of the returns to computer use.
- It is also common to use the following heuristic for evaluating the robustness of results to omitted variable bias:
 - Check how the coefficient of interest varies when you include controls. (Intuition: if individuals in the treatment and the control group differ in observables, they are likely to differ in (relevant) unobservables.)
 - Building on this idea, some authors proposed a formal way to take into account the degree of selection on observed variables to assess the extent of selection on unobserved variables
 - Altonji, Elder and Taber 2005, Oster 2014

Standard errors

- So far we have focused on how we can *identify* the effect of a treatment (**unbiasedness/consistency**)
- Let us now focus on the ‘accuracy’ of estimates (**inference**)
- Let us consider the following model:
 - $Y = XB + \epsilon$
- As you know:
 - $\hat{\beta}_{OLS} = (X'X)^{-1}(X'Y)$
- We want to find the variance-covariance matrix of $\hat{\beta}_{OLS}$ (for simplicity in what follows let us drop the OLS subindex)

$$Var(\hat{\beta}|X) = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'|X]$$

The Variance-Covariance Matrix of the OLS Estimates

$$\begin{aligned}Var(\hat{\beta}|X) &= E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'|X] = E[(X'X)^{-1}X'\epsilon)(X'X)^{-1}X'\epsilon)'|X] \\&= E[(X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1}|X]\end{aligned}$$

where we took advantage of the fact that $(AB)' = B'A'$ and therefore we can rewrite $(X'X)^{-1}X'\epsilon)'$ as $\epsilon'X(X'X)^{-1}$.

Given that X is non-stochastic, we get:

$$E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'|X] = (X'X)^{-1}X'E[\epsilon\epsilon'|X]X(X'X)^{-1}$$

where

$$E(\epsilon\epsilon'|X) = \Omega = \begin{bmatrix} E[\epsilon_1^2|X] & E[\epsilon_1\epsilon_2|X] & \dots & E[\epsilon_1\epsilon_n|X] \\ E[\epsilon_2\epsilon_1|X] & E[\epsilon_2^2|X] & \dots & E[\epsilon_2\epsilon_n|X] \\ \vdots & \vdots & \vdots & \vdots \\ E[\epsilon_n\epsilon_1|X] & E[\epsilon_n\epsilon_2|X] & \dots & E[\epsilon_n^2|X] \end{bmatrix}$$

Variance of estimator β_{OLS}

We can consider several possible cases.

- ① Spherical errors (aka "Standard" standard errors)
- ② Robust standard errors (heteroskedasticity consistent standard errors)
- ③ Clustered standard errors

1. Spherical errors

Errors are uncorrelated and homoskedastic

- Two assumptions:
 - Errors are uncorrelated (aka no serial dependence)
 - Homoskedasticity (etymology: from Ancient Greek homo ‘same’ and skedasis ‘dispersion’)

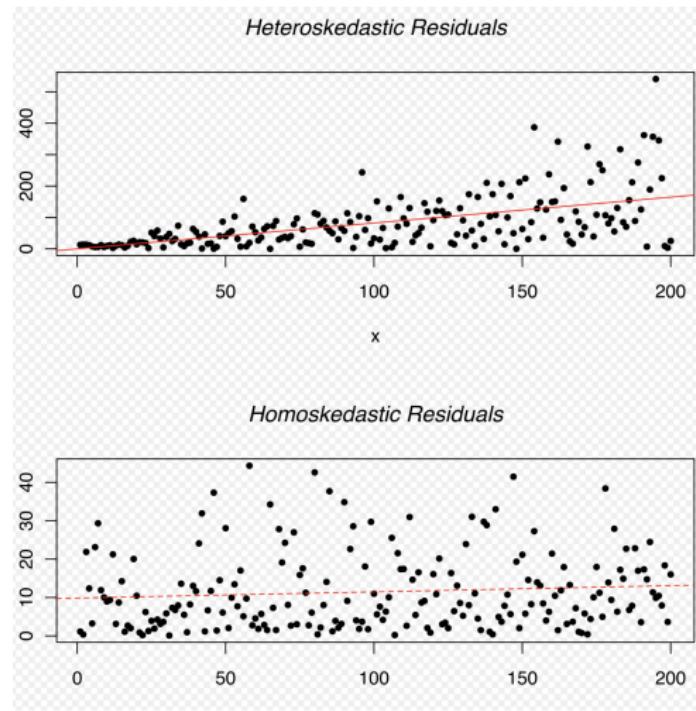
$$E(\epsilon\epsilon'|X) = \sigma^2 \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} = \sigma^2 I$$

Then

$$\text{Var}(\hat{\beta}|X) = (X'X)^{-1} X' (\sigma^2 I) X (X'X)^{-1} = \sigma^2 (X'X)^{-1}$$

2. Robust standard errors

Common example of heteroskedasticity: the errors tend to be larger for larger values of y



2. Robust standard errors

- Also known as heteroskedasticity consistent standard errors

$$E(\epsilon\epsilon'|X) = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & & \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

- In stata `vce()` specifies how to estimate the variance-covariance matrix (VCE)
 - `reg y x, vce(robust)`

3. Clustered standard errors

- Very often we cannot assume that unobserved shocks are independent across individuals
- Within some groups or clusters, observations may be exposed to common shocks
- Example: in the STAR experiment on the impact of class size, students in the same classroom are likely to be exposed to common shocks such as being exposed to the same bad/good peer, teacher...
- We will assume independence across ‘clusters’ but ...
- ... allow for dependence within ‘clusters’ (or groups), and estimate variance and covariance of uncertainty within groups.

Example with m clusters, each one with n_j observations

$$E(\epsilon\epsilon') = \begin{bmatrix} \sigma_{(1,1)1}^2 & \cdots & \sigma_{(1,n_1)1}^2 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{(n_1,1)1}^2 & \cdots & \sigma_{(n_1,n_1)1}^2 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \sigma_{(1,1)2}^2 & \cdots & \sigma_{(1,n_2)2}^2 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \sigma_{(n_2,1)2}^2 & \cdots & \sigma_{(n_2,n_2)2}^2 \\ & & & & \ddots & \\ & & & & & \sigma_{(1,1)m}^2 & \cdots & \sigma_{(1,n_m)m}^2 \\ & & & & & \vdots & \ddots & \vdots \\ & & & & & \sigma_{(n_m,1)m}^2 & \cdots & \sigma_{(n_m,n_m)m}^2 \end{bmatrix}$$

Some remarks (i)

- ❶ typically clustered standard errors are larger than robust or "standard" errors
- ❷ the type of standard errors you calculate does not affect the point estimate
- ❸ how to choose the relevant level of clustering:
 - level at which common shocks are likely to operate
 - at least the level at which the treatment is defined (e.g. impact of a policy implemented at the class level → cluster at least the class level)
 - not always trivial to choose the correct level (e.g. STAR experiment - should we cluster at the class or at the school level?)
 - larger cluster level → more conservative standard errors
- ❹ In some sense, clustering implies acknowledging how many independent sources of information there are in the data (e.g. the number of clusters is essentially your N)

Some remarks (ii)

- How do you estimate clustered standard errors?
 - ① When the number of groups is large enough (rule of thumb: $N > 50$), use the ‘sandwich formula’
 - `reg y x, vce(cluster 'clustvar')`
 - ② When the number of groups is small, the corresponding asymptotic properties do not hold. There are some alternatives:
 - Randomization inference
 - Block-bootstrap

Clustering standard errors

Thought experiment: common shocks

- Imagine that we want to estimate the **impact of taking a certain pill on individual happiness**. Individuals in the control group will receive a placebo
- Sample size: 1000 individuals from Leamington Spa and 1000 from Coventry
- Note 1: Let us also assume, for the sake of the argument, that some preliminary survey shows that the level of happiness in Leamington Spa and Coventry is statistically similar.
- Note 2: Let us assume, for the sake of the argument, that there is no concern about general equilibrium effects (individuals do not interact).

- How do we assign individuals to treatment and control? Two proposals
 - ① Flip a coin once: tail, individuals from Leamington Spa are treated, heads, individuals from Coventry are treated
 - ② Flip a coin 2000 times, once for each individual: tail, the individual is assigned to treatment; heads she is assigned to control
- Which of the two implementations would be more informative about the impact of the treatment? Why?
- However, the OLS standard errors are similar in both cases. What's wrong?

- How do we assign individuals to treatment and control? Two proposals
 - ① Flip a coin once: tail, individuals from Leamington Spa are treated, heads, individuals from Coventry are treated
 - ② Flip a coin 2000 times, once for each individual: tail, the individual is assigned to treatment; heads she is assigned to control
- Which of the two implementations would be more informative about the impact of the treatment? Why?
- However, the OLS standard errors are similar in both cases. What's wrong?
- The potential presence of a common random effect:
 - There might be some common shock affecting all individuals in the treatment group or in the control group (Moulton 1990).
 - OLS standard errors assume that all observations are independent realizations. Standard errors have to be corrected to account for the presence of a common random effect.

Poll question 4

- Impact of parental education on children's performance using PISA data
- Should we cluster at the:
 - ① individual level
 - ② class level
 - ③ school level

Poll question 4

- Impact of parental education on children's performance using PISA data
- Should we cluster at the:
 - ① individual level
 - ② class level
 - ③ *school level*

Poll question 5

- The impact of minimum wages on employment exploiting variation across states over time.
- Should we cluster at the:
 - ① individual level
 - ② individual*year level
 - ③ state*year level
 - ④ state level

Poll question 5

- The impact of minimum Wages on employment exploiting variation across states over time.
- Should we cluster at the:
 - ① individual level
 - ② individual*year level
 - ③ state*year level
 - ④ *state level*

EC902/EC907: Econometrics A

Lecture 10.1

Manuel Bagues

Warwick University

Last week:

- Examples of identification based on observables:
 - ① The Impact of Social Networks on Remuneration in Top Executive Jobs
 - ② The Returns to Computer Use
 - Krueger 1993: “How computers have changed the wage structure: evidence from microdata, 1984-1989”
 - DiNardo and Pischke (1997): “The Returns to Computer Use Revisited: Have Pencils Changed the Wage Structure Too?”
- OLS Standard errors
 - ① Spherical errors (aka "Standard" standard errors)
 - ② Robust standard errors (heteroskedasticity consistent standard errors)
 - ③ Clustered standard errors

This week: Instrumental variables

- Instrumental variables (IV)
 - Definition (slides 10.1)
 - Examples (slides 10.2)
 - OLS vs. IV (slides 10.3)
 - Wald estimator (slides 10.3)
 - Example: Angrist and Krueger 1991

Instrumental variables (IV)

- Sometimes (often) the regression we have is not the regression we want.
- That is, we do not have a rich enough data to eliminate the selection bias.
- Possible solution: look for an instrumental variable
- In other words, look for some source of exogenous variation in your treatment
 - As we will see, IV is essentially similar to an RCT without full compliance
- But good instruments are hard to find!

Example: Returns to schooling

- Our aim is to estimate the causal effect of schooling (S_i) on wages (Y_i)
- Let us assume that the following equation describes the process that determines wages:

$$Y_i = \alpha + \rho S_i + \gamma A_i + \nu_i$$

- where A_i is correlated with S_i
- Unfortunately, the econometrician (us!) cannot observe (perfectly) individual ability (A_i)

$$Y_i = \alpha + \rho S_i + \underbrace{\gamma A_i + \nu_i}_{\eta_i}$$

- If we use OLS to estimate the relationship between Y_i and S_i :
- $$\hat{\rho}_{OLS} = \frac{Cov(Y_i, S_i)}{V(S_i)}$$
- The OLS estimate would be subject to an omitted variable bias since

$$E[S_i \eta_i] \neq 0$$

- If we could observe the variables A_i we could simply include them to the regressions and estimate

$$Y_i = \alpha + \rho S_i + \gamma A_i + \nu_i$$

- How to estimate ρ without observing A_i ?
- Instrumental variable (IV) allows us to estimate ρ when A_i is unobserved

IV requirements

- We need a variable Z_i that is correlated with the treatment [$\text{Cov}(Z_i, S_i) \neq 0$] and which does not affect the treatment through any other channel [$\text{Cov}(Z_i, \eta_i) = 0$]
- Then we can write ρ in terms of the following population moments:

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

- We take the covariance with respect of Z:

$$\text{Cov}(Z_i, Y_i) = \rho \text{Cov}(Z_i, S_i) + \underbrace{\text{Cov}(Z_i, \eta_i)}_{=0}$$

- It follows that:

$$\rho = \frac{\text{Cov}(Z_i, Y_i)}{\text{Cov}(Z_i, S_i)}$$

- Note also that:

$$\rho = \frac{\text{Cov}(Z_i, Y_i)}{\text{Cov}(Z_i, S_i)} = \frac{\frac{\text{Cov}(Z_i, Y_i)}{V(Z_i)}}{\frac{\text{Cov}(Z_i, S_i)}{V(Z_i)}}$$

- The IV estimate, $\hat{\rho}_{IV}$, is the ratio between the β_{OLS} of regression of Y_i on Z_i (aka *reduced form* estimate) and β_{OLS} of a regression of S_i on Z_i (*first stage* estimate).
- In sum, the IV estimate captures the correlation between the outcome variable and the instrument, upscaled by the correlation between the treatment and the instrument

What is a valid instrumental variable?

- Two main assumptions:
 - ➊ **Relevance:** The instrument is correlated with the causal variable of interest, S_i ,
 $\text{Cov}(Z_i, S_i) \neq 0$
 - ➋ **Independence:** The instrument is uncorrelated with any other determinants of Y_i
 $\text{Cov}(Z_i, \eta_i) = 0$
- This requirement can be decomposed in two:
- 2.1 **Exogeneity:** The instrument is as good as random, none of the unobserved factors affects it [$\eta_i \not\rightarrow Z_i$]
 - 2.2 **Exclusion restriction:** Z_i only affects Y_i through its effect on S_i
[$Z_i \not\rightarrow \eta_i$]

Note: Some authors may refer to the independence assumption as the exogeneity condition or the exclusion restriction. However, it is useful to consider exogeneity and exclusion restriction as two distinct requirements.

Why IV works?

- Intuitive idea behind IV is as follows:
 - ➊ **Relevance:** You found a variable (the instrument) that affects who is assigned to the treatment
 - ➋ **Exogeneity:** The instrument is as good as randomly assigned
 - ➌ **Exclusion restriction:** And you know that your instrument only affect the outcome through its impact on the treatment (cannot be correlated with any other factor that affects the outcome, or affect directly the outcome).
- Note that an IV strategy is equivalent to an RCT where there is no full compliance

Can we test validity of IV?

- Can we test the assumptions required for the validity of a IV:
 - ① Is the instrument correlated with the treatment?
 $[\text{Cov}(Z_i, S_i) \neq 0]$
 - YES: Significance of first stage, F-statistics
 - ② Exogeneity $[\eta_i \not\rightarrow Z_i]$
 - SOMETIMES YES: Is the instrument as good as random?
 - ③ Exclusion restriction? $[\eta_i \not\rightarrow Z_i]$
 - NO!
 - The exclusion restriction relies on theory - is it plausible that the instrument only affects the outcome variable through its impact on the treatment?
 - There might be always some unobserved channel through which the instrument affects the outcome
 - Note: you can use always-takers and never-takers to check if the instrument affects the outcome variable (we will discuss this next week / do not worry about it now)

- **Exogeneity** is sufficient for a causal interpretation of the reduced form.
- The **exclusion restriction** is distinct from the claim that the instrument is (as good as) randomly assigned. Rather, it is a claim about a unique channel for causal effects of the instrument.
- Technical note: **tests of overidentification** just tell you whether instruments are consistent with each other, but instruments can still be invalid.

EC902/EC907: Econometrics A

Lecture 10.2

Manuel Bagues

Warwick University

This week: Instrumental variables

- Definition (slides 10.1)
- Examples (slides 10.2)
- OLS vs. IV (slides 10.3)
- Wald estimator (slides 10.3)
 - Example: Angrist and Krueger 1991

Good instruments are hard to find

- Are these good instruments for schooling?
 - What about the last digit of social security number?
 - What about IQ?
 - Family background?
 - Geographical proximity?
 - Altonji, Elder and Taber (JHR 2005)
 - Working status?

Good instruments are hard to find

Good instruments come from a combination of three ingredients:

- Good institutional knowledge
- Economic theory
- Last but not least: Originality

Some usual sources of instruments:

- Nature
 - Sometimes nature randomizes, providing exogenous variation for some variable of interest (e.g. gender of children, twins...)
- Assignment rules that rely on randomization
 - returns to medical school in Netherlands, time in prison, foster care, military service...
- ‘Natural’ experiments
 - Some exogenous variables may influence assignment to the treatment (the quarter of birth, electoral timing...)

Note that, in general, choice variables of the agent tend to be bad instruments

Sources of instruments: nature

- The effect of family size on children's education and female labor force participation
 - Twins, gender of the first born, gender of the two first born
(Black, Devereux and Salvanes, QJE 2005; Angrist, Lavy and Schlosser, JOLE 2009)
 - IVF treatments (*Lundborg, Plug and Wurtz Rasmussen 2014*)

Sources of instruments: assignment rules that rely on randomization

- Foster care
- Time in prison
- Children's custody
- Other ideas?

Sources of instruments: ‘natural’ experiments

- Immigration
 - Networks of immigrants (Card 1991)
- Does police decrease crime?
 - Electoral cycles (Levitt 1997)
- The impact of violent movies on crime
 - Blockbuster movies (Dahl and DellaVigna 2009)

- The effect of preschool television exposure on standardized test scores during adolescence:
 - Gentzkow and Shapiro 2008
- Influence of mass media on U.S. government response to natural disasters
 - Eisensee and Strömberg 2007

(Bad) instruments

- Parental socioeconomic characteristics as an instrument for children education
- ‘South of Italy’ as an instrument for CEO’s gender
- Students’ working status as an instrument for attendance
- Generally: Lagged variables as instruments

EC902/EC907: Econometrics A

Lecture 10.3

Manuel Bagues

Warwick University

This week: Instrumental variables

- Definition (slides 10.1)
- Examples (slides 10.2)
- OLS vs. IV (slides 10.3)
- Wald estimator (slides 10.3)
 - Example: Angrist and Krueger 1991

Identification based on observables vs. Instrumental variables

- Regression based on observables (e.g. OLS)
 - The consistency of the estimate relies on the “hope” that any unobserved factor that might affect the outcome variable is balanced across the treatment and the control group.
 - Therefore, any difference in outcomes between the control and the treatment group can be attributed to the treatment.
- Instrumental variables:
 - We identify some source of variation in the assignment to the treatment which, for some reason, we know that it is orthogonal to any relevant unobserved variable which might be affecting the outcome variables.
 - We compare individuals that, due to the instrument, are assigned to the control and the treatment group. Any difference in outcomes is attributed to the treatment.

More formally:

Let us consider the following model:

$$Y = X\beta + U$$

Let us examine the OLS and the IV estimators:

$$\begin{aligned} E[\beta_{OLS}] &= E[(X'X)^{-1}(X'Y)] = E[(X'X)^{-1}(X'(X\beta + U))] = \\ &\quad \beta + E[(X'X)^{-1}(X'U)] \end{aligned}$$

$$\begin{aligned} E[\beta_{IV}] &= E[(Z'X)^{-1}(Z'Y)] = E[(Z'X)^{-1}(Z'(X\beta + U))] = \\ &\quad \beta + E[(Z'X)^{-1}(Z'U)] \end{aligned}$$

- Very often β_{OLS} is likely to suffer from an omitted variables problem.
- At the same time, the β_{IV} estimator may be subject to a serious inconsistency problem whenever the first stage is weak and the orthogonality condition is not strictly satisfied (known as the *weak instruments problem*)
- Corollary: sometimes the cure may be worse than the disease

The Wald estimator

- Consider the case when we have:
 - Model with one endogenous regressor (x_i) and no covariates
 - Single binary instrument [$z_i \in \{0, 1\}$]
- Let us denote the *reduced form estimate* to:

$$E[y_i|z_i = 1] - E[y_i|z_i = 0]$$

- and the *first stage estimate* is equal to:

$$E[x_i|z_i = 1] - E[x_i|z_i = 0]$$

- The IV estimator is equal to:

$$\rho = \frac{\text{Cov}(y_i, z_i)}{\text{Cov}(x_i, z_i)} = \frac{E[y_i|z_i=1] - E[y_i|z_i=0]}{E[x_i|z_i=1] - E[x_i|z_i=0]} = \frac{\text{reduced form}}{\text{first stage}}$$

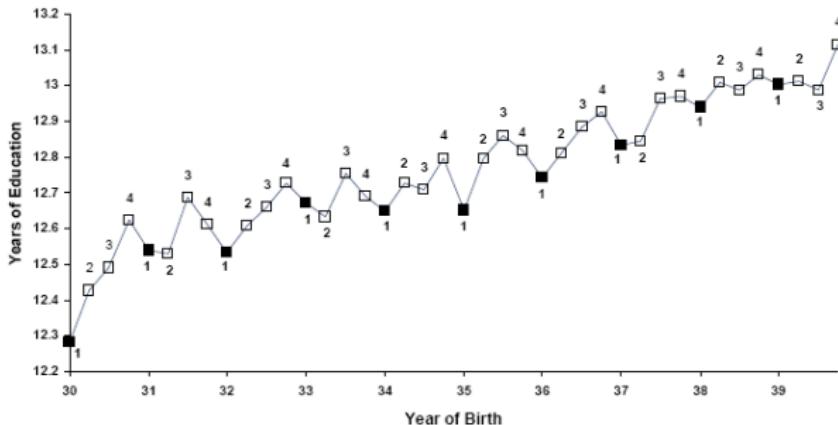
An interesting example: Angrist and Krueger, QJE 1991

“Does Compulsory School Attendance Affect Schooling and Earnings”, [\(Angrist and Krueger, QJE 1991\)](#)

- Quarter of birth as an instrument for schooling
- Students enter schooling in the calendar year in which they turn 6
- And compulsory school law requires them to remain in school until they become 16
- Hence people born late in the year are more likely to stay at school longer

Is the first stage right?

A. Average Education by Quarter of Birth (first stage)



The reduced form for earnings

B. Average Weekly Wage by Quarter of Birth (reduced form)

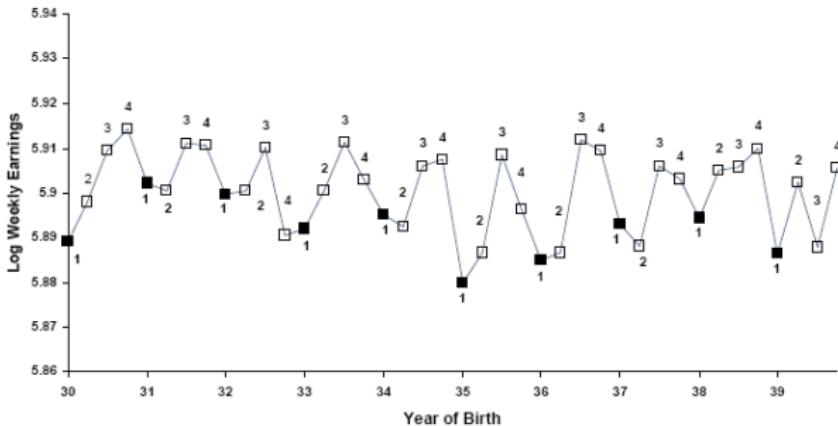


Table 4.1.2: Wald estimates of the returns to schooling using quarter of birth instruments

	(1) Born in the 1st or 2nd quarter of year	(2) Born in the 3rd or 4th quarter of year	(3) Difference (std. error) (1)-(2)
In (weekly wage)	5.8916	5.9051	-0.01349 (0.00337)
Years of education	12.6881	12.8394	-0.1514 (0.0162)
Wald estimate of return to education			0.0891 (0.0210)
OLS estimate of return to education			0.0703 (0.0005)

Notes: Adapted from a re-analysis of Angrist and Krueger (1991) by Angrist and Imbens (1995). The sample includes native-born men with positive earnings from the 1930-39 birth cohorts in the 1980 Census 5 percent file. The sample size is 329,509.

EC902/EC907: Econometrics A

Lecture 11

Manuel Bagues

Warwick University

November 14, 2022
Lecture Slides

This week

- Midterm on Wednesday, 9:00 a.m.
 - Includes material covered until lecture 9 (last Monday)
 - Last 2 years midterms available in moodle
- Today:
 - Clustering standard errors (included in the midterm!)
 - Instrumental variables (not included in midterm)
- Tomorrow (Nov 15):
 - Midterm Q&A session, 12:00-13:00, Room R0.21

Roadmap

- Identification based on observables
- Clustering standard errors (slides 8.3)
- Instrumental variables
 - Definition (slides 10.1)
 - Examples (slides 10.2)
 - OLS vs. IV (slides 10.3)
 - Wald estimator (slides 10.3)
 - Example: Angrist and Krueger 1991

Clustering standard errors

- Very often we cannot assume that unobserved shocks are independent across individuals
- Within some groups or clusters, observations may be exposed to common shocks
- Example: in the STAR experiment on the impact of class size, students in the same classroom are likely to be exposed to common shocks such as being exposed to the same bad/good peer, teacher...
- We will assume independence across ‘clusters’ but ...
- ... allow for dependence within ‘clusters’ (or groups), and estimate variance and covariance of uncertainty within groups.

Some remarks (i)

- ❶ typically clustered standard errors are larger than robust or "standard" errors
- ❷ the type of standard errors you calculate does not affect the point estimate
- ❸ how to choose the relevant level of clustering:
 - level at which common shocks are likely to operate
 - at least the level at which the treatment is defined (e.g. impact of a policy implemented at the class level → cluster at least the class level)
 - not always trivial to choose the correct level (e.g. STAR experiment - should we cluster at the class or at the school level?)
 - larger cluster level → more conservative standard errors (\uparrow probability of a false positive negative)
- ❹ In some sense, clustering implies acknowledging how many independent sources of information there are in the data (e.g. the number of clusters is essentially your N)

Some remarks (ii)

- Useful reference: Alberto Abadie, Susan Athey, Guido W Imbens, Jeffrey M Wooldridge, When Should You Adjust Standard Errors for Clustering?, *The Quarterly Journal of Economics*, 2022.
- How do you estimate clustered standard errors?
 - ➊ When the number of groups is large enough (rule of thumb: $N > 50$), use the ‘sandwich formula’
 - `reg y x, vce(cluster 'clustvar')`
 - ➋ When the number of groups is small, the corresponding asymptotic properties do not hold. There are some alternatives:
 - Randomization inference
 - Block-bootstrap

Clustering standard errors

1st thought experiment: common shocks

- How do we assign individuals to treatment and control? Two proposals
 - ① Flip a coin once: tail, individuals from Leamington Spa are treated, heads, individuals from Coventry are treated
 - ② Flip a coin 2000 times, once for each individual: tail, the individual is assigned to treatment; heads she is assigned to control
- Which of the two implementations would be more informative about the impact of the treatment? Why?
- However, the OLS standard errors are similar in both cases. What's wrong?

Clustering standard errors

1st thought experiment: common shocks

- How do we assign individuals to treatment and control? Two proposals
 - ① Flip a coin once: tail, individuals from Leamington Spa are treated, heads, individuals from Coventry are treated
 - ② Flip a coin 2000 times, once for each individual: tail, the individual is assigned to treatment; heads she is assigned to control
- Which of the two implementations would be more informative about the impact of the treatment? Why?
- However, the OLS standard errors are similar in both cases. What's wrong?
- The potential presence of a common random effect:
 - There might be some common shock affecting all individuals in the treatment group or in the control group (Moulton 1990).
 - OLS standard errors assume that all observations are independent realizations. Standard errors have to be corrected to account for

Clustering standard errors

2nd thought experiment: serial correlation

- RCT about the impact of providing feedback to students about their relative performance.
- Sample of 977 students, 354 in the treatment group and 623 in the control group.
- The treatment is implemented at the beginning of the 2nd year
- Our outcome of interest is students' average GPA during the 2nd year, as measured by their average GPA
 - We can compare the average yearly performance of the treatment and control group
 - Sample size: 977
 - Standard error=st.dev/ $\sqrt{977}$.

- Alternatively, you could consider observations at the term * course level
- You would end up with $2*977$ observations (there are two terms each year)
 - What would happen to standard errors?
 - By how much?
- The information I am exploiting it is essentially the same, but the OLS standard errors are now smaller. What is wrong?
 - Again, OLS standard errors assume that the observations are independent realizations
 - We need to account for serial correlation when we calculate the standard errors

Additional example: Does class size affect students' performance?

Tennessee STAR experiment

- How can we improve students' performance?
- Should we devote more resources to reduce **class size**?
 - Example: Should we split this course in two separate groups?
- A large number of observational studies tend to find that class size is not generally associated to better student performance
 - Hanushek (1997): “No strong or systematic relationship between school inputs and student achievement”

Percentage distribution of estimated effect of key resources on student performance, based on 376 studies

Resources	Number of estimates	Statistically significant		Statistically insignificant
		Positive	Negative	
Real classroom resources				
Teacher–pupil ratio	276	14%	14%	72%
Teacher education	170	9	5	86
Teacher experience	206	29	5	66
Financial aggregates				
Teacher salary	118	20	7	73%
Expenditure per pupil	163	27	7	66
Other				
Facilities	91	9	5	86
Administration	75	12	5	83

Example: Does class size affect students' performance?

Tennessee STAR experiment

Tennessee STAR experiment

- Cost: \$12 million
- A cohort of kindergartners in 1985/86: 11,600 children in 80 schools
- The study ran for four years
- Three treatments:
 - ① small classes with 13-17 children
 - ② regular classes with 22-25 children without a teacher's aide.
 - ③ regular classes with 22-25 children with a teacher's aide.
- Within each school, students are randomly assigned to one of these groups

Example: Does class size affect students' performance?

Tennessee STAR experiment

Krueger 1999 provides an econometric analysis of the short-run effects of the experiment

- Main findings:
 - ① performance on standardized tests increases by four percentile points the first year students attend small classes
 - ② teacher aides and measured teacher characteristics have little effect
- Note: Hanushek has written also a critical article about the pitfalls of the STAR experiment

TABLE I
COMPARISON OF MEAN CHARACTERISTICS OF TREATMENTS AND CONTROLS:
UNADJUSTED DATA

Variable	Small	Regular	Regular/Aide	Joint <i>P</i> -Value ^a
1. Free lunch ^c	.47	.48	.50	.09
2. White/Asian	.68	.67	.66	.26
3. Age in 1985	5.44	5.43	5.42	.32
4. Attrition rate ^d	.49	.52	.53	.02
5. Class size in kindergarten	15.1	22.4	22.8	.00
6. Percentile score in kindergarten	54.7	49.9	50.0	.00

Table 2.2.2: Experimental estimates of the effect of class-size assignment on test scores

Explanatory variable	(1)	(2)	(3)	(4)
Small class	4.82 (2.19)	5.37 (1.26)	5.36 (1.21)	5.37 (1.19)
Regular/aide class	.12 (2.23)	.29 (1.13)	.53 (1.09)	.31 (1.07)
White/Asian (1 = yes)	—	—	8.35 (1.35)	8.44 (1.36)
Girl (1 = yes)	—	—	4.48 (.63)	4.39 (.63)
Free lunch (1 = yes)	—	—	-13.15 (.77)	-13.07 (.77)
White teacher	—	—	—	-.57 (2.10)
Teacher experience	—	—	—	.26 (.10)
Master's degree	—	—	—	-0.51 (1.06)
School fixed effects	No	Yes	Yes	Yes
R ²	.01	.25	.31	.31

Example: Does class size affect students' performance?

Tennessee STAR experiment

- Several 'tricky' questions:
 - ➊ Would you prefer that your children are taught by a teacher with a Master's degree?
 - ➋ Does taking a Master's degree improve teachers' teaching skills?
- R-square - do we care for causal questions?

Poll questions

- We will make some polls using [vevox.app](#)
 - ➊ Please, open [vevox.app](#) in your computer, in Teams or in your mobile
 - ➋ Session ID: 144-921-069

Poll question 1

- Impact of class size on children's performance using STAR data
- Should we cluster at the:
 - ① individual level
 - ② class level
 - ③ city level

Poll question 1

- Impact of class size on children's performance using STAR data
- Should we cluster at the:
 - ① individual level
 - ② *class level*
 - ③ city level

Poll question 2

- Impact of parental education on children's performance using PISA data
- Should we cluster at the:
 - ➊ individual level
 - ➋ class level
 - ➌ school level

Poll question 2

- Impact of parental education on children's performance using PISA data
- Should we cluster at the:
 - ① individual level
 - ② class level
 - ③ *school level*

Poll question 3

- The impact of minimum wages on employment exploiting variation across states over time.
- Example: Card & Krueger AER 1994, ‘Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania’
- Should we cluster at the:
 - ① individual level
 - ② individual*year level
 - ③ state*year level
 - ④ state level

Poll question 3

- The impact of minimum Wages on employment exploiting variation across states over time.
- Should we cluster at the:
 - ① individual level
 - ② individual*year level
 - ③ state*year level
 - ④ *state level*

Instrumental variables (IV)

- Sometimes (often) the regression we have is not the regression we want.
- That is, we do not have a rich enough data to eliminate the selection bias.
- Possible solution: look for an instrumental variable
- In other words, look for some source of exogenous variation in your treatment
 - As we will see, IV is essentially similar to an RCT without full compliance
- But good instruments are hard to find!

Instrumental variables

- Let us consider the following model:

$$Y = X\beta + U$$

- where we are concerned about the possibility that $E[X \cdot U] \neq 0$
- Would it be a good idea to use OLS to estimate β ?
 - $\beta_{OLS} = [X'X]^{-1}X'Y$

Instrumental variables

- Let us consider the following model:

$$Y = X\beta + U$$

- where we are concerned about the possibility that $E[X \cdot U] \neq 0$
- Would it be a good idea to use OLS to estimate β ?
 - $\beta_{OLS} = [X'X]^{-1}X'Y$
- What about using an instrument Z ? Which features should it have?
 - $\beta_{IV} = [Z'X]^{-1}Z'Y$

Reminder: Instrumental variables

- Instrumental variable (Z) is a variable that:
 - ➊ **Relevance:** correlated with causal variable of interest, X_i ,
 $E[Z' \cdot X] \neq 0$
 - ➋ **Independence:** uncorrelated with any other determinants of Y_i
 $E[Z' \cdot U] = 0$This requirement can be decomposed in two:
 - 2.1 **Exogeneity:** The instrument is as good as random, none of the unobserved factors affects it $[U \not\rightarrow Z]$
 - 2.2 **Exclusion restriction:** Z_i only affects Y_i through its effect on S_i
 $[Z \not\rightarrow U]$

Note: Some authors may refer to the independence assumption as the exogeneity condition or the exclusion restriction. However, it is useful to consider exogeneity and exclusion restriction as two distinct requirements.

Some usual sources of instruments:

- Nature
 - Sometimes nature randomizes, providing exogenous variation for some variable of interest (e.g. gender of children, twins...)
- Assignment rules that rely on randomization
 - returns to medical school in Netherlands, time in prison, foster care, military service...
- ‘Natural’ experiments
 - Some exogenous variables may influence assignment to the treatment (the quarter of birth, electoral timing...)

Note that, in general, choice variables of the agent tend to be bad instruments

The Wald estimator

- Consider the case when we have:
 - Model with one endogenous regressor (x_i) and no covariates
 - Single binary instrument [$z_i \in \{0, 1\}$]
- Let us denote the *reduced form estimate* to:

$$E[y_i|z_i = 1] - E[y_i|z_i = 0]$$

- and the *first stage estimate* is equal to:

$$E[x_i|z_i = 1] - E[x_i|z_i = 0]$$

- The IV estimator is equal to:

$$\rho = \frac{Cov(y_i, z_i)}{Cov(x_i, z_i)} = \frac{E[y_i|z_i=1] - E[y_i|z_i=0]}{E[x_i|z_i=1] - E[x_i|z_i=0]} = \frac{\text{reduced form}}{\text{first stage}}$$

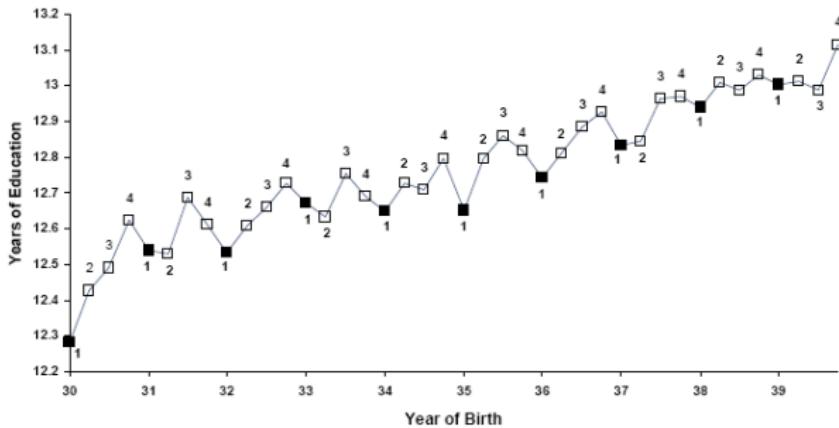
An interesting example: Angrist and Krueger, QJE 1991

“Does Compulsory School Attendance Affect Schooling and Earnings”, (Angrist and Krueger, QJE 1991)

- Quarter of birth as an instrument for schooling
- Students enter schooling in the calendar year in which they turn 6
- And compulsory school law requires them to remain in school until they become 16
- Hence people born late in the year are more likely to stay at school longer

Is the first stage right?

A. Average Education by Quarter of Birth (first stage)



The reduced form for earnings

B. Average Weekly Wage by Quarter of Birth (reduced form)

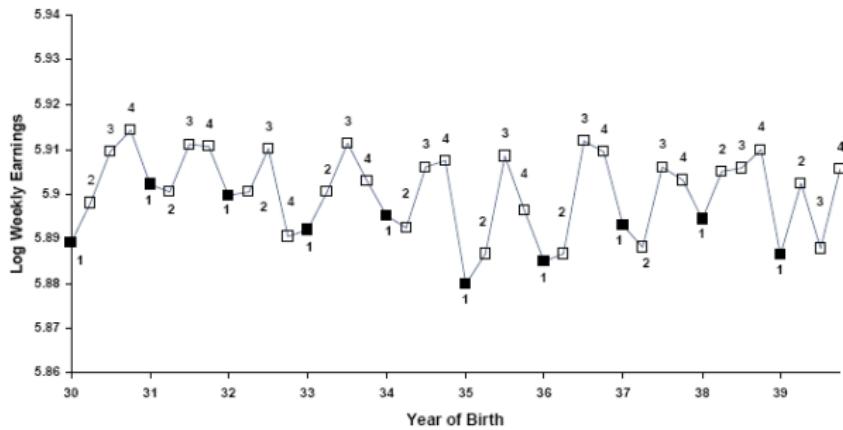


Table 4.1.2: Wald estimates of the returns to schooling using quarter of birth instruments

	(1) Born in the 1st or 2nd quarter of year	(2) Born in the 3rd or 4th quarter of year	(3) Difference (std. error) (1)-(2)
In (weekly wage)	5.8916	5.9051	-0.01349 (0.00337)
Years of education	12.6881	12.8394	-0.1514 (0.0162)
Wald estimate of return to education			0.0891 (0.0210)
OLS estimate of return to education			0.0703 (0.0005)

Notes: Adapted from a re-analysis of Angrist and Krueger (1991) by Angrist and Imbens (1995). The sample includes native-born men with positive earnings from the 1930-39 birth cohorts in the 1980 Census 5 percent file. The sample size is 329,509.

Poll question 4

- Is the instrument relevant?
 - ① Yes, because it is highly correlated with earnings
 - ② Yes, because it is highly correlated with educational attainment
 - ③ No, because there is no correlation with earnings

Poll question 4

- Is the instrument relevant?
 - ① Yes, because it is highly correlated with earnings
 - ② Yes, because it is highly correlated with educational attainment
 - ③ No, because there is no correlation with earnings

Poll question 5

- Would you expect the exogeneity condition to be (fully) satisfied?
 - ① No, because parental characteristics may affect the timing of births
 - ② Yes, because quarter of birth is predetermined
 - ③ No, because quarter or birth also affects relative age in the class, which may affect personality traits

Poll question 5

- Would you expect the exogeneity condition to be (fully) satisfied?
 - ① No, because parental characteristics may affect the timing of births
 - ② Yes, because quarter of birth is predetermined
 - ③ No, because quarter or birth also affects relative age in the class, which may affect personality traits

Poll question 6

- Would you expect the exclusion restriction to be satisfied?
 - ① Yes, because month of birth is as good as random.
 - ② No, because parental characteristics affect the timing of births
 - ③ No, because quarter or birth also affects relative age in the class, which may affect personality traits
 - ④ Yes, because quarter of birth does not affect educational attainment through any other channel

Poll question 6

- Would you expect the exclusion restriction to be satisfied?
 - ① Yes, because month of birth is as good as random
 - ② No, because parental characteristics affect the timing of births
 - ③ No, because quarter or birth also affects relative age in the class, which may affect personality traits
 - ④ Yes, because quarter of birth does not affect educational attainment through any other channel

Identification based on observables vs. Instrumental variables

- Regression based on observables (e.g. OLS)
 - The consistency of the estimate relies on the “hope” that any unobserved factor that might affect the outcome variable is balanced across the treatment and the control group.
 - Therefore, any difference in outcomes between the control and the treatment group can be attributed to the treatment.
- Instrumental variables:
 - We identify some source of variation in the assignment to the treatment which, for some reason, we know that it is orthogonal to any relevant unobserved variable which might be affecting the outcome variables.
 - We compare individuals that, due to the instrument, are assigned to the control and the treatment group. Any difference in outcomes is attributed to the treatment.

More formally:

Let us consider the following model:

$$Y = X\beta + U$$

Let us examine the OLS and the IV estimators:

$$\begin{aligned} E[\beta_{OLS}] &= E[(X'X)^{-1}(X'Y)] = E[(X'X)^{-1}(X'(X\beta + U))] = \\ &\quad \beta + E[(X'X)^{-1}(X'U)] \end{aligned}$$

$$\begin{aligned} E[\beta_{IV}] &= E[(Z'X)^{-1}(Z'Y)] = E[(Z'X)^{-1}(Z'(X\beta + U))] = \\ &\quad \beta + E[(Z'X)^{-1}(Z'U)] \end{aligned}$$

- Very often β_{OLS} is likely to suffer from an omitted variables problem.
- At the same time, the β_{IV} estimator may be subject to a serious inconsistency problem whenever the first stage is weak and the orthogonality condition is not strictly satisfied (known as the *weak instruments problem*)
- Corollary: sometimes the cure may be worse than the disease

EC902/EC907: Econometrics A

Lecture 12.1

Manuel Bagues

Warwick University

Last week: Instrumental variables (1/2)

- Instrumental variables (IV)
 - Definition
 - Examples
 - OLS vs. IV
 - Wald estimator
 - Example: Angrist and Krueger 1991

Today: Instrumental variables (2/2)

- Instrumental variables
 - Two-stage least squares ([slides 12.1](#))
 - Heterogeneity of the effect: local average treatment effect (LATE) ([slides 12.2](#))
 - Example: Doyle 2007 ([slides 12.3](#))
- Note: you may want to read Angrist and Pischke's and/or Cunningham's chapter on IV

Reminder: Instrumental variables

- Let us consider the following model:

$$Y = X\beta + U$$

- where we are concerned about the possibility that $E[X \cdot U] \neq 0$

Reminder: Instrumental variables

- Instrumental variable (Z) is a variable that:
 - ➊ **Relevance:** correlated with causal variable of interest, X_i ,
 $E[Z' \cdot X] \neq 0$
 - ➋ **Independence:** uncorrelated with any other determinants of Y_i
 $E[Z' \cdot U] = 0$
- This requirement can be decomposed in two:
- 2.1 **Exogeneity:** The instrument is as good as random, none of the unobserved factors affects it [$U \not\rightarrow Z$]
 - 2.2 **Exclusion restriction:** Z_i only affects Y_i through its effect on S_i
[$Z \not\rightarrow U$]

Note: Some authors may refer to the independence assumption as the exogeneity condition or the exclusion restriction. However, it is useful to consider exogeneity and exclusion restriction as two distinct requirements.

Three ways to calculate the IV estimator

- ① Standard formula: $\hat{\beta}_{IV} = [Z'X]^{-1}Z'Y$
- ② $\hat{\beta}_{IV} = \frac{\text{reduced form}}{\text{1st stage}}$
- ③ Two-stage least squares (2SLS)

Reduced form and 1st stage

- Consider the case the following model:

$$Y = X\beta + U$$

where Z is a potential instrument

- The *reduced form estimation* involves regressing the outcome variable on the instrument:

$$Y = Z\gamma + \eta$$

where $\hat{\gamma}_{OLS} = (Z'Z)^{-1}(Z'Y)$.

- And the *first stage estimation* involves regressing the treatment on the instrument:

$$X = Z\lambda + \mu$$

where $\hat{\lambda}_{OLS} = (Z'Z)^{-1}(Z'X)$

- The IV estimator is equal to:

$$\hat{\beta}_{IV} = \frac{\text{reduced form}}{\text{first stage}} = \frac{\hat{\gamma}}{\hat{\lambda}} = \frac{(Z'Z)^{-1}(Z'Y)}{(Z'Z)^{-1}(Z'X)}$$

Two-stage Least Squares (2SLS)

- In a model with a **single endogenous variable** s_i and a **single instrument** z_i , IV estimates are equivalent to a two stage procedure.
- Let us consider the following causal model with covariates X and treatment s :

$$Y_i = X'_i \alpha + \rho s_i + \eta_i \quad (1)$$

- and let z_i be a valid instrument for s_i
- The 2SLS procedure involves two stages.
- **First**, regress the endogenous variable s_i on the instrument, including also in the regression every covariate in equation (1):

$$s_i = \underbrace{X'_i \pi_1 + \pi_2 z_i}_{\hat{s}_i} + \epsilon_{1i}$$

Two-stage Least Squares (2SLS)

- **Second**, substitute the ‘predicted’ treatment (\hat{s}_i) in the main regression.

$$Y_i = X'_i \alpha + \rho \hat{s}_i + [\eta_i + \rho \epsilon_{1i}]$$

- and estimate by OLS!

Standard errors

Technical note

- With the manual two stage procedure, you do not get ‘automatically’ the correct standard errors
 - The residual that is used to calculate standard errors in second stage includes an extra error $Y_i - [X'_i \alpha - \rho \hat{s}_i] = [\rho \epsilon_{1i} + \eta_i]$
 - \hat{s} is a generated regressor and inflates the variance
 - Stata `ivreg` or `ivreg2` fixes it. It uses the original endogenous regressor to construct residuals:
$$Y_i - [X'_i \alpha - \rho s_i] = \eta_i$$

EC902/EC907: Econometrics A

Lecture 12.2

Manuel Bagues

Warwick University

Today: Instrumental variables (2/2)

- Instrumental variables
 - Two-stage least squares (slides 12.1)
 - Heterogeneity of the effect: local average treatment effect (LATE) ([slides 12.2](#))
 - Example: Doyle 2007 (slides 12.3)

Local average treatment effects

- The impact of the treatment might **heterogenous**
- Therefore, it is important to understand the type of treatment effect that we are identifying
- Let $D_{i1} \in \{0, 1\}$ indicate whether an individual **affected** by the instrument would receive or not the treatment. Similarly, let $D_{i0} \in \{0, 1\}$ indicate whether an individual **not affected** by the instrument would receive or not the treatment
- We can consider four instrument-dependent subgroups, defined by the manner in which members of the population react to the instrument:
 - Compliers: $D_{i1} = 1, D_{i0} = 0$
 - Always-takers: $D_{i1} = 1, D_{i0} = 1$
 - Never-takers: $D_{i1} = 0, D_{i0} = 0$
 - Defiers: $D_{i1} = 0, D_{i0} = 1$

Example: Angrist and Krueger 1991

- Verbalize who are the members of each group in the following setup:
 - Does compulsory school attendance affect schooling and earnings? (Angrist and Krueger 1991)
 - Outcome: earnings
 - Treatment: years of schooling
 - Instrument: quarter of birth
- Who are in this case the (i) compliers, (ii) always-takers, (iii) never-takers, and (iv) defiers?

- Groups
 - Compliers: kids who drop out or not depending on their month of birth (they only drop out if they were born at the beginning of the year). Eg: the instrument (*month of birth*) affects (in the expected way) whether they receive the treatment or not (*years of education*)
 - Always-takers: kids who would always finish the academic year, independently of their month of birth
 - Never-takers: kids who would never finish the academic year, independently of their month of birth
 - Defiers: kids who would drop out if born at the end of the year, but not if they were born at the beginning of the year

- Typically we assume that there are no defiers (*monotonicity assumption*)
- Under this assumption, with an instrumental variable strategy we learn about the impact of the treatment for the group of **compliers**.
- IV strategy identifies the **Local average treatment effect (LATE)**, in the sense that we learn about the impact of the treatment for a very particular group of individuals.
- **Discussion:** are compliers a relevant group?

Example: The Effect of Peer Salaries on Job Satisfaction

- Let us consider a simple set up where there is a binary instrument and a binary treatment:
 - Card et al. 2012 “Inequality at Work: The Effect of Peer Salaries on Job Satisfaction”
- Email to random sample of employees of University of California with link to a webpage with information on salaries.
- treatment: information; outcome: job satisfaction; instrument: email
- Impact of the instrument on the treatment:
 - People that receive the information email: 50% check the webpage
 - People that do not receive the information email: 20% check the webpage

Example: The Effect of Peer Salaries on Job Satisfaction

- Two questions
 - ➊ Explain verbally who are the compliers, always-takers, never-takers and defiers.
 - ➋ Let us assume that there are no defiers. What is the share of always-takers, never-takers and compliers?

EC902/EC907: Econometrics A

Lecture 12.3

Manuel Bagues

Warwick University

Today: Instrumental variables (2/2)

- Instrumental variables
 - Two-stage least squares (slides 12.1)
 - Heterogeneity of the effect: local average treatment effect (LATE) (slides 12.2)
 - Example: Doyle 2007 ([slides 12.3](#))

Effect of Foster Care on Criminal Behavior

Doyle, J. (2008), "Child Protection and Child Outcomes: Measuring the Effects of Foster Care", American Economic Review 97(5), pp. 1583-1610. ([link](#))

- Children who are abused or neglected are often put into foster care
 - Large numbers in the US: 500,000 kids!
- Three-quarters of these children live with substitute families (one third relatives of the children)
- These families are paid \$400 monthly
- Temporary arrangement: on average 2 years
- 60% return home, 15% are adopted, remainder age out

Motivation

- Children placed in foster care tend to have a higher propensity to commit crime, drop out of school, be on welfare... than the average kid.
- Obviously this does not tell us much about causal effect of foster care (does it help or harm the kids?)
- Let us consider the population of kids for whom foster care has been considered.

<i>Variable</i>		<i>Mean</i>
	Foster care placement	0.27
Initial reporter	Physician	0.12
	School	0.13
	Police	0.13
	Family	0.29
	Neighbor	0.06
	Other government	0.09
	Anonymous	0.15
	Other reporter	0.03
Age at report	Age	11.3
Sex	Boy	0.47

Race	White	0.11
	African American	0.76
	Hispanic	0.12
	Other race/ethnicity	0.01
Allegation	Physical abuse	0.17
	Substantial risk of harm	0.24
	Other abuse	0.02
	Lack of supervision	0.37
	Environmental neglect	0.15
	Other neglect	0.04
Location	Cook County	1.00
Outcome	Delinquency	0.17

IV-strategy

- Doyle exploits the fact there is a **rotation system** that assigns children to case managers, who decide who will be placed in foster care (removed from home).
- The decision to place a child in foster care is not trivial. As a result, some case managers have a higher tendency to place children in foster care.
- Children assigned to case managers with high tendency to place children to foster care have higher probability to be placed in foster care.

Instrument for Foster Care

The authors use as their instrument a measure of the *investigator placement propensity*:

$$Z_{c,k} = \frac{1}{\mu_{c,-k}} \sum \mu_{c,k} (\bar{R}^{c,k} - \bar{R}^k)$$

where c is the case manager, k is the case team (ZIP code x hispanic x year cells), $\mu_{c,k}$ is the number of children investigated by case manager c within case team k , and R is the fraction of children investigated who are eventually removed.

- Is there any variance in case managers' tendency to place children into foster care?
 - Mean(Z)=0 (by construction)
 - St. dev.(Z)= 9% (in the main subsample)

Assumptions for valid IV (i)

- Is there a first stage? (Do children who are assigned to a case manager with higher placement propensity have higher probability for foster care?)
- Let us examine the first stage, unconditional on controls:

$$D_i = \alpha_0 + \alpha_1 Z_i + \epsilon_i$$

- and also with controls:

$$D_i = \beta_0 + \beta_1 Z_i + \beta_2 X_i + \epsilon_i$$

Table 4

Dependent variable: Foster care placement

	Model:	Probit			Probit		
		Coefficient	S.E.	p-value	Coefficient	S.E.	p-value
Key explanatory variables	Case manager removal differential	0.30	0.07	0.00	0.27	0.05	0.00
Initial reporter (Other reporter excluded)	Physician				0.10	0.03	0.00
	School				-0.02	0.03	0.43
	Police				0.14	0.03	0.00
	Family				0.05	0.03	0.06
	Neighbor				0.02	0.03	0.53
	Other government				0.07	0.03	0.03
	Anonymous				-0.06	0.03	0.02
Age at report (Youngest age excluded)	Age 6				0.05	0.05	0.21
	Age 7				0.05	0.04	0.18
	Age 8				0.02	0.04	0.66
	Age 9				0.03	0.04	0.44
	Age 10				0.03	0.04	0.42

	Age 11	0.02	0.04	0.55
	Age 12	0.00	0.04	0.97
	Age 13	-0.02	0.04	0.63
	Age 14	-0.04	0.04	0.32
	Age 15	-0.07	0.03	0.08
Sex	Boy	-0.01	0.01	0.14
Race/ethnicity (Other race excluded)	White	0.00	0.05	0.95
	African American	0.11	0.04	0.02
	Hispanic	-0.03	0.05	0.50
Allegation (Other neglect excluded)	Physical abuse	-0.07	0.02	0.00
	Substantial risk of harm	0.00	0.02	0.88
	Other abuse	-0.09	0.02	0.00
	Lack of supervision	0.00	0.02	0.89
	Environmental neglect	-0.08	0.02	0.00
	Chi-squared (1) stat.	17.9		27.8
	Mean of dep. var.	0.27		
	Observations	15,039		

Note: Marginal effects and standard errors clustered at the case manager level are reported.

Assumptions for valid IV (ii)

- Does the exogeneity assumption hold?
- We can verify that case managers that are more strict are not being assigned to ‘special’ cases

TABLE 2—CHILD CHARACTERISTICS AND CASE MANAGER ASSIGNMENT: DELINQUENCY SAMPLE

<i>Dependent variable: Case manager removal differential</i>		Coefficient	<i>t</i>	<i>p</i> -value
Variable				
Initial reporter (Other reporter excluded)	Physician	−0.006	−0.81	0.416
	School	−0.005	−0.74	0.457
	Police	−0.008	−1.11	0.269
	Family	−0.003	−0.52	0.605
	Neighbor	−0.005	−0.73	0.464
	Other government	−0.007	−0.96	0.339
	Anonymous	−0.007	−1.07	0.287
Age at report (Youngest age excluded)	Age 6	0.005	0.41	0.679
	Age 7	0.012	1.07	0.284
	Age 8	0.009	0.90	0.367
	Age 9	0.015	1.42	0.156
	Age 10	0.008	0.72	0.470
	Age 11	0.009	0.94	0.346
	Age 12	0.010	0.99	0.324
	Age 13	0.013	1.26	0.207
	Age 14	0.009	0.91	0.366
	Age 15	0.009	0.89	0.373

		-	-	-
Sex	Boy	-0.002	-1.20	0.232
Race/ethnicity (Other race excluded)	White	-0.014	-1.32	0.186
	African American	-0.015	-1.22	0.224
	Hispanic	-0.012	-0.88	0.377
Allegation (Other neglect excluded)	Physical abuse	-0.002	-0.43	0.668
	Substantial risk of harm	-0.006	-0.94	0.348
	Other abuse	0.003	0.43	0.670
	Lack of supervision	-0.005	-0.98	0.325
	Environmental neglect	-0.007	-1.29	0.199
Mean of dependent variable		0.0001		
Standard deviation		0.0921		
<i>F</i> -statistic of joint significance		0.84		
<i>p</i> -value		0.75		
Number of case managers		409		
Observations		15,039		

Note: *t*-statistics and *F*-statistic are calculated using standard errors clustered by case manager.

Assumptions for valid IV (iii)

- Does the exclusion restriction hold?
- (Is the case manager placement propensity only affecting future outcomes of these children through the probability to be placed in foster care)
- What could go wrong? (some ideas?)
- Can you test this?
 - You can show that it does not hold, but you will never be sure of its validity.
- If the exclusion restriction may not hold, what can we make out of our results?
 - Interpret them as reduced from.

Main results

- Does foster care affect:
 - ① teenage motherhood
 - ② labor market outcomes
 - ③ juvenile delinquency
- We will focus on the latter one: juvenile delinquency
 - Wald estimator
 - Probit/OLS and IV estimates

The Wald estimator

- We could also consider a binary instrument: case manager with high or low previous placement propensity
- In this case we can write the IV estimator as a Wald estimator:

$$\begin{aligned}\beta_1 &= \frac{\text{Cov}(Y_i, Z_i)}{\text{Cov}(D_i, Z_i)} \\ &= \frac{E[Y_i | Z_i=1] - E[Y_i | Z_i=0]}{E[D_i | Z_i=1] - E[D_i | Z_i=0]} \\ &= \frac{\gamma_1}{\alpha_1} = \frac{\text{reduced form}}{\text{first stage}}\end{aligned}$$

Table 2

Table of means: instrumental variable estimation.

		Investigator placement propensity			
		High	Low	Difference	p-value
A. First stage	Foster care placement	0.316	0.224	0.092	<0.0001
B. Reduced form	Juvenile delinquency	0.171	0.158	0.013	0.043

Table 2

Table of means: instrumental variable estimation.

		Investigator placement propensity		Difference	p-value
		High	Low		
A. First stage	Foster care placement	0.316	0.224	0.092	<0.0001
B. Reduced form	Juvenile delinquency	0.171	0.158	0.013	0.043
C. IV estimate	Change in juvenile delinquency:	Difference in B ÷ difference in A:		0.142	
	Change in foster care placement	p-value:		0.035	
	Observations	7792	7889		

Juvenile Delinquency Sample: Children in Cook County who received an abuse/neglect report between July 1, 1990 and December 31, 2000 and were at least 15 in 2000.
 p-values calculated using standard errors clustered at the investigator level.

Juvenile delinquency

TABLE 6—FOSTER CARE PLACEMENT AND JUVENILE DELINQUENCY

Dependent variable: Juvenile delinquency

	Model:	Probit				IV Probit			
		Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.
	FC placement	0.01	0.01	0.00	0.01	0.26	0.14	0.35	0.14
Initial reporter (Other reporter excluded)	Physician			0.00	0.02			-0.02	0.02
	School			0.00	0.02			0.00	0.02
	Police			0.02	0.02			-0.01	0.03
	Family			0.00	0.02			-0.01	0.02
	Neighbor			0.01	0.03			0.00	0.03
	Other government			0.03	0.02			0.01	0.02
	Anonymous			0.01	0.02			0.03	0.02
Age at report (Youngest age excluded)	Age 5			—	—			—	—
	Age 6			0.06	0.05			0.04	0.05
	Age 7			0.10	0.05			0.08	0.05
	Age 8			0.13	0.05			0.12	0.05
	Age 9			0.13	0.05			0.12	0.05
	Age 10			0.17	0.06			0.15	0.05

	Age 11	0.19	0.06	0.18	0.05
	Age 12	0.22	0.06	0.21	0.05
	Age 13	0.23	0.06	0.23	0.06
	Age 14	0.23	0.06	0.23	0.06
	Age 15	0.12	0.05	0.14	0.05
Sex	Boy	0.19	0.01	0.19	0.01
Race/ethnicity (Other race excluded)	White	-0.07	0.03	-0.07	0.03
	African American	-0.02	0.04	-0.05	0.04
	Hispanic	-0.07	0.03	-0.07	0.03
Allegation (Other neglect excluded)	Physical abuse	-0.01	0.02	0.01	0.02
	Substantial risk of harm	-0.03	0.01	-0.03	0.02
	Other abuse	-0.02	0.02	0.01	0.03
	Lack of supervision	-0.02	0.02	-0.03	0.02
	Environmental neglect	-0.02	0.02	0.00	0.02
	Mean of dep. var.	0.17			
	Observations	15,039			

Note: Marginal effects and standard errors clustered at the case manager level are reported.

Conclusions from Doyle

- Foster care increases juvenile delinquency
- IV is even higher than OLS
- Doyle explains this by stating that for children on the margin to be placed in FC the impact is more harmful than for others (who benefit more)
- This argument rests on idea that the impact of foster care is heterogenous

EC902/EC907: Econometrics A

Lecture 13

Manuel Bagues

Warwick University

November 21, 2022
Lecture Slides

Logistics

- Start recording
- Midterm last week
 - Performance in line with expectations: mean=69 (last year mean=66)
 - Presumably many mistakes due to lack of time, misreading of questions...
 - Problems with multiple answers in questions 16, 17 and 18 - sorry about that!
 - Contact me if you have any doubts
- Project
 - EC902 (not for EC907)
 - Important deadlines
 - December 9 (by noon) - group composition (3 members)
 - March 20 (by noon) - project submission
 - We will discuss next week some examples of projects from previous years
- Any questions?

Roadmap

- Instrumental variables (1/2)
 - Definition (slides 10.1)
 - Examples (slides 10.2)
 - OLS vs. IV (slides 10.3)
 - Wald estimator (slides 10.3)
 - Example: Angrist and Krueger 1991
- Instrumental variables (2/2)
 - Two-stage least squares (slides 12.1)
 - Heterogeneity of the effect: local average treatment effect (LATE) (slides 12.2)
 - Example: Doyle 2007 (slides 12.3)

Instrumental variables (1/2)

- Sometimes (often) the regression we have is not the regression we want.
- That is, we do not have a rich enough data to eliminate the selection bias.
- Possible solution: look for an instrumental variable
- In other words, look for some source of exogenous variation in your treatment
 - As we will see, IV is essentially similar to an RCT without full compliance
- But good instruments are hard to find!

- Let us consider the following model:

$$Y = X\beta + U$$

- where we are concerned about the possibility that $E[X \cdot U] \neq 0$
- Would it be a good idea to use OLS to estimate β ?
 - $\beta_{OLS} = [X'X]^{-1}X'Y$

- Let us consider the following model:

$$Y = X\beta + U$$

- where we are concerned about the possibility that $E[X \cdot U] \neq 0$
- Would it be a good idea to use OLS to estimate β ?
 - $\beta_{OLS} = [X'X]^{-1}X'Y$
- What about using an instrument Z ? Which features should it have?
 - $\beta_{IV} = [Z'X]^{-1}Z'Y$

Reminder: Instrumental variables

- Instrumental variable (Z) is a variable that:
 - ❶ **Relevance:** correlated with causal variable of interest, X_i ,
 $E[Z' \cdot X] \neq 0$
 - ❷ **Independence:** uncorrelated with any other determinants of Y_i
 $E[Z' \cdot U] = 0$
- This requirement can be decomposed in two:
- 2.1 Exogeneity:** The instrument is as good as random, none of the unobserved factors affects it [$U \not\rightarrow Z$]
 - 2.2 Exclusion restriction:** Z_i only affects Y_i through its effect on S_i
[$Z \not\rightarrow U$]

Note: Some authors may refer to the independence assumption as the exogeneity condition or the exclusion restriction. However, it is useful to consider exogeneity and exclusion restriction as two distinct requirements.

Some usual sources of instruments:

- Nature
 - Sometimes nature randomizes, providing exogenous variation for some variable of interest (e.g. gender of children, twins...)
- Assignment rules that rely on randomization
 - returns to medical school in Netherlands, time in prison, foster care, military service...
- ‘Natural’ experiments
 - Some exogenous variables may influence assignment to the treatment (the quarter of birth, electoral timing...)

Note that, in general, choice variables of the agent tend to be bad instruments

The Wald estimator

- Consider the case when we have:
 - Model with one endogenous regressor (x_i) and no covariates
 - Single binary instrument [$z_i \in \{0, 1\}$]
- Let us denote the *reduced form estimate* to:

$$E[y_i|z_i = 1] - E[y_i|z_i = 0]$$

- and the *first stage estimate* is equal to:

$$E[x_i|z_i = 1] - E[x_i|z_i = 0]$$

- The IV estimator is equal to:

$$\rho = \frac{Cov(y_i, z_i)}{Cov(x_i, z_i)} = \frac{E[y_i|z_i=1] - E[y_i|z_i=0]}{E[x_i|z_i=1] - E[x_i|z_i=0]} = \frac{\text{reduced form}}{\text{first stage}}$$

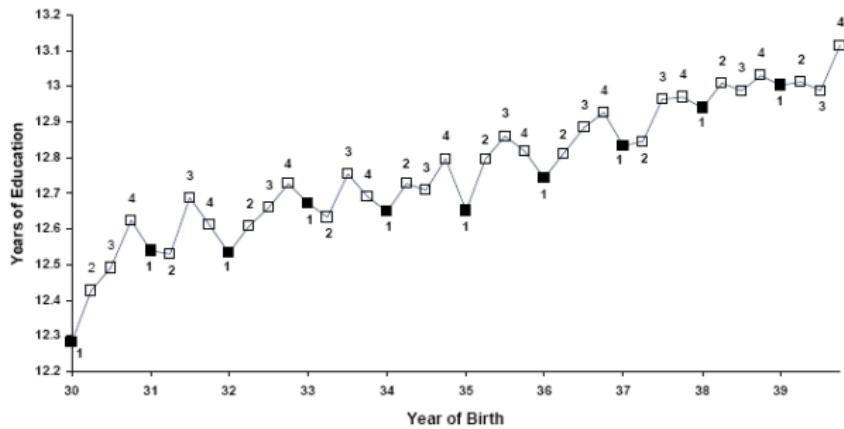
An interesting example: Angrist and Krueger, QJE 1991

“Does Compulsory School Attendance Affect Schooling and Earnings”, (Angrist and Krueger, QJE 1991)

- Quarter of birth as an instrument for schooling
- Students enter schooling in the calendar year in which they turn 6
- And compulsory school law requires them to remain in school until they become 16
- Hence people born late in the year are more likely to stay at school longer

Is the first stage right?

A. Average Education by Quarter of Birth (first stage)



The reduced form for earnings

B. Average Weekly Wage by Quarter of Birth (reduced form)

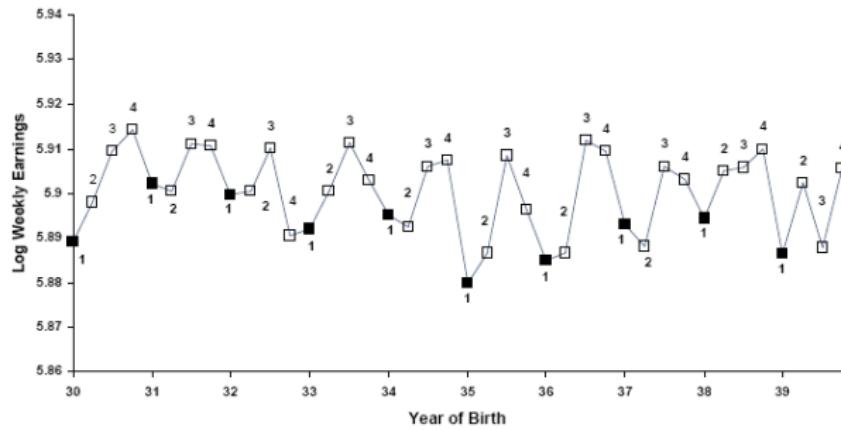


Table 4.1.2: Wald estimates of the returns to schooling using quarter of birth instruments

	(1) Born in the 1st or 2nd quarter of year	(2) Born in the 3rd or 4th quarter of year	(3) Difference (std. error) (1)-(2)
ln (weekly wage)	5.8916	5.9051	-0.01349 (0.00337)
Years of education	12.6881	12.8394	-0.1514 (0.0162)
Wald estimate of return to education			0.0891 (0.0210)
OLS estimate of return to education			0.0703 (0.0005)

Notes: Adapted from a re-analysis of Angrist and Krueger (1991) by Angrist and Imbens (1995). The sample includes native-born men with positive earnings from the 1930-39 birth cohorts in the 1980 Census 5 percent file. The sample size is 329,509.

Poll questions

- We will make some polls using [vevox.app](#)
 - ① Please, open [vevox.app](#) in your computer, in Teams or in your mobile
 - ② Session ID: 144-921-069

Poll question 1

- Is the instrument relevant?
 - ① Yes, because it is highly correlated with earnings
 - ② Yes, because it is highly correlated with educational attainment
 - ③ No, because there is no correlation with earnings

Poll question 1

- Is the instrument relevant?
 - ① Yes, because it is highly correlated with earnings
 - ② Yes, because it is highly correlated with educational attainment
 - ③ No, because there is no correlation with earnings

Poll question 2

- Would you expect the exogeneity condition to be (fully) satisfied?
 - ➊ No, because parental characteristics may affect the timing of births
 - ➋ Yes, because quarter of birth is predetermined
 - ➌ No, because quarter or birth also affects relative age in the class, which may affect personality traits

Poll question 2

- Would you expect the exogeneity condition to be (fully) satisfied?
 - ① No, because parental characteristics may affect the timing of births
 - ② Yes, because quarter of birth is predetermined
 - ③ No, because quarter or birth also affects relative age in the class, which may affect personality traits

Poll question 3

- Would you expect the exclusion restriction to be satisfied?
 - ① Yes, because month of birth is as good as random.
 - ② No, because parental characteristics affect the timing of births
 - ③ No, because quarter or birth also affects relative age in the class, which may affect personality traits
 - ④ Yes, because quarter of birth does not affect educational attainment through any other channel

Poll question 3

- Would you expect the exclusion restriction to be satisfied?
 - ① Yes, because month of birth is as good as random
 - ② No, because parental characteristics affect the timing of births
 - ③ No, because quarter or birth also affects relative age in the class, which may affect personality traits
 - ④ Yes, because quarter of birth does not affect educational attainment through any other channel

Identification based on observables vs. Instrumental variables

- Regression based on observables (e.g. OLS)
 - The consistency of the estimate relies on the “hope” that any unobserved factor that might affect the outcome variable is balanced across the treatment and the control group.
 - Therefore, any difference in outcomes between the control and the treatment group can be attributed to the treatment.
- Instrumental variables:
 - We identify some source of variation in the assignment to the treatment which, for some reason, we know that it is orthogonal to any relevant unobserved variable which might be affecting the outcome variables.
 - We compare individuals that, due to the instrument, are assigned to the control and the treatment group. Any difference in outcomes is attributed to the treatment.

More formally:

Let us consider the following model:

$$Y = X\beta + U$$

Let us examine the OLS and the IV estimators:

$$\begin{aligned} E[\beta_{OLS}] &= E[(X'X)^{-1}(X'Y)] = E[(X'X)^{-1}(X'(X\beta + U))] = \\ &\quad \beta + E[(X'X)^{-1}(X'U)] \end{aligned}$$

$$\begin{aligned} E[\beta_{IV}] &= E[(Z'X)^{-1}(Z'Y)] = E[(Z'X)^{-1}(Z'(X\beta + U))] = \\ &\quad \beta + E[(Z'X)^{-1}(Z'U)] \end{aligned}$$

- Very often β_{OLS} is likely to suffer from an omitted variables problem.
- At the same time, the β_{IV} estimator may be subject to a serious inconsistency problem whenever the first stage is weak and the orthogonality condition is not strictly satisfied (known as the *weak instruments problem*)
- Corollary: sometimes the cure may be worse than the disease

Instrumental variables (Part 2)

Three ways to calculate the IV estimator

① Standard formula: $\hat{\beta}_{IV} = [Z'X]^{-1}Z'Y$

② $\hat{\beta}_{IV} = \frac{\hat{\beta}_{\text{reduced form}}}{\hat{\beta}_{\text{1st stage}}}$

③ Two-stage least squares (2SLS)

- 1st stage: regress the treatment on the instrument (+ covariates)
- 2nd stage: regress the outcome variable on the predicted treatment (+ covariates)

Local average treatment effects

- The impact of the treatment might **heterogenous**
- Therefore, it is important to understand the type of treatment effect that we are identifying
- Let $D_{i1} \in \{0, 1\}$ indicate whether an individual **affected** by the instrument would receive or not the treatment. Similarly, let $D_{i0} \in \{0, 1\}$ indicate whether an individual **not affected** by the instrument would receive or not the treatment
- We can consider four instrument-dependent subgroups, defined by the manner in which members of the population react to the instrument:
 - Compliers: $D_{i1} = 1, D_{i0} = 0$
 - Always-takers: $D_{i1} = 1, D_{i0} = 1$
 - Never-takers: $D_{i1} = 0, D_{i0} = 0$
 - Defiers: $D_{i1} = 0, D_{i0} = 1$

Example: Angrist and Krueger 1991

- Verbalize who are the members of each group in the following setup:
 - Does compulsory school attendance affect schooling and earnings? (Angrist and Krueger 1991)
 - Outcome: earnings
 - Treatment: years of schooling
 - Instrument: quarter of birth
- Who are in this case the (i) compliers, (ii) always-takers, (iii) never-takers, and (iv) defiers?

- Groups
 - Compliers: kids who drop out or not depending on their month of birth (they only drop out if they were born at the beginning of the year). Eg: the instrument (*month of birth*) affects (in the expected way) whether they receive the treatment or not (*years of education*)
 - Always-takers: kids who would always finish the academic year, independently of their month of birth
 - Never-takers: kids who would never finish the academic year, independently of their month of birth
 - Defiers: kids who would drop out if born at the end of the year, but not if they were born at the beginning of the year

Example: the link between absenteeism and students' performance

- Andrietti, D'Addazio and Velasco (2008)
 - OLS estimates: students that attend class tend to obtain better grades
 - IV strategy:
 - distance to reach campus from the student's house
 - dummy variable that indicates if the student works
 - Both instruments correlate with absenteeism
 - But what about:
 - exogeneity?
 - exclusion restriction?

1st instrument: distance to reach campus

Sample exam question:

- ① Would you expect the exogeneity assumption to be satisfied?
How would you test this assumption?
- ② Propose some possible violation of the exclusion restriction
- ③ Discuss verbally who are the always-takers, the never-takers, the compliers and defiers.

Poll question 4

- Would you expect the exogeneity assumption to be satisfied?
 - ① No, because in this context distance is likely to correlate with (unobservable) socio-economic characteristics
 - ② No, because distance affects commuting time
 - ③ No, because distance does not correlate with students performance
 - ④ Yes, it is likely to be exogenous

Poll question 4

- Would you expect the exogeneity assumption to be satisfied?
 - ① No, because in this context distance is likely to correlate with (unobservable) socio-economic characteristics
 - ② No, because distance affects commuting time
 - ③ No, because distance does not correlate with students performance
 - ④ Yes, it is likely to be exogenous

Poll question 5

- Would you expect the exclusion restriction to be satisfied? (you can select more than one option)

Poll question 5

- Would you expect the exclusion restriction to be satisfied? (you can select more than one option)
 - ① No, because distance affects commuting time
 - ② No, because distance affects leisure opportunities
 - ③ No, because distance affects attendance
 - ④ Yes, it is likely to be exogenous

Poll question 5

- Would you expect the exclusion restriction to be satisfied? (you can select more than one option)
 - ① No, because distance affects commuting time
 - ② No, because distance affects leisure opportunities
 - ③ No, because distance affects attendance
 - ④ Yes, it is likely to be exogenous

Poll question 6

- Compliers?
 - ❶ Would perform well if they lived close, otherwise not
 - ❷ Would perform well if they attended lectures, otherwise not
 - ❸ Would attend lectures if they lived close, otherwise not

Poll question 6

- Compliers?
 - ❶ Would perform well if they lived close, otherwise not
 - ❷ Would perform well if they attended lectures, otherwise not
 - ❸ Would attend lectures if they lived close, otherwise not

2nd instrument: dummy variable that indicates if the student works

- ① Would you expect the exogeneity assumption to be satisfied?
How would you test this assumption?
- ② Propose some possible violation of the exclusion restriction
- ③ Discuss verbally who are the always-takers, the never-takers, the compliers and defiers.

Example: the link between absenteeism and students' performance

- Other possible instruments?
 - Arulampalam, Naylor and Smith (2012): “Am I missing something? The effects of absence from class on student performance”

- Arulampalam, Naylor and Smith (2012) study the causal effects of class absence on student performance in a paper titled ‘Am I missing something? The effects of absence from class on student performance’. They use data from 2nd year Economics undergraduate students and they focus on the absenteeism in tutorial classes. First, they show that being absent from 10% of classes is associated with around a 1.3 percentage point lower mark in the subject (st. error=0.4). Second, to deal with the potential endogeneity of absenteeism, they use an instrumental variables strategy which exploits (i) that students are randomly assigned to different tutorial groups and (ii) that attendance tends to be lower on certain days and on certain periods of the day (e.g. Monday morning). Their IV estimate suggests that being absent from 10% of classes has a causal negative impact on performance, which is reduced by 1.6 percentage points (st. error=0.7).

- ① Discuss possible explanations for why the OLS point estimate is smaller than the IV point estimate (1.3 vs. 1.6).
- ② Would you expect the exogeneity assumption to be satisfied? How would you test this assumption?
- ③ Propose some possible violation of the exclusion restriction
- ④ Discuss verbally who are the always-takers, the never-takers, the compliers and defiers.

Poll question 7

- Discuss possible explanations for why the OLS point estimate is smaller than the IV point estimate (1.3 vs. 1.6). (you can select more than one choice)

Poll question 7

- Discuss possible explanations for why the OLS point estimate is smaller than the IV point estimate (1.3 vs. 1.6). (you can select more than one choice)
 - ① The OLS estimate may suffer a downwards bias, if individuals living far away are negatively selected (in unobservables).
 - ② The OLS estimate may suffer a downwards bias, if individuals attending lectures are positively selected (in unobservables).
 - ③ Compliers benefit much more from attendance than the average individual ($LATE > ATE$)
 - ④ The IV estimate is biased, due to the violation of the exogeneity assumption.

Poll question 7

- Discuss possible explanations for why the OLS point estimate is smaller than the IV point estimate (1.3 vs. 1.6). (you can select more than one choice)
 - ① The OLS estimate may suffer a downwards bias, if individuals living far away are negatively selected (in unobservables).
 - ② The OLS estimate may suffer a downwards bias, if individuals attending lectures are positively selected (in unobservables).
 - ③ Compliers benefit much more from attendance than the average individual (LATE>ATE)
 - ④ The IV estimate is biased, due to the violation of the exogeneity assumption.

Poll question 8

- Would you expect the exogeneity assumption to be satisfied?
(you can select more than one choice)
 - ① Yes, because assignment is random
 - ② No, because attendance affects performance
 - ③ No, because time of the class affects students mood during lectures
 - ④ Yes, because the IV has been used in a published paper

Poll question 8

- Would you expect the exogeneity assumption to be satisfied?
(you can select more than one choice)
 - ① Yes, because assignment is random
 - ② No, because attendance affects performance
 - ③ No, because timing of the class affects students mood during lectures
 - ④ Yes, because the IV has been used in a published paper
- How would you test this assumption?

Poll question 9

- Propose some possible violation of the exclusion restriction (you can select more than one)

Poll question 9

- Propose some possible violation of the exclusion restriction (you can select more than one)
 - ① Timing of the class may affect students mood during lectures, which may influence performance
 - ② Timing of the class may affect teachers mood during lectures, which may influence performance
 - ③ Timing of the class may affect sleeping time, which may influence performance
 - ④ Timing of the class may affect instructors identity, which may influence performance, and the authors were unable to control for this variable

Poll question 9

- Propose some possible violation of the exclusion restriction (you can select more than one)
 - ① Timing of the class may affect students mood during lectures, which may influence performance
 - ② Timing of the class may affect teachers mood during lectures, which may influence performance
 - ③ Timing of the class may affect sleeping time, which may influence performance
 - ④ Timing of the class may affect instructors identity, which may influence performance, and the authors were unable to control for this variable

Poll question 10

- Who are the defiers?
 - ① They are more likely to attend classes when the timing is ‘good’ (e.g. Tuesday afternoon)
 - ② The timing of the class does not affect their attendance
 - ③ The timing of the class does not affect their performance
 - ④ They are more likely to attend classes when the timing is ‘bad’ (e.g. Monday 9:00 am)

Poll question 10

- Who are the defiers?
 - ① They are more likely to attend classes when the timing is ‘good’
(e.g. Tuesday afternoon)
 - ② The timing of the class does not affect their attendance
 - ③ The timing of the class does not affect their performance
 - ④ They are more likely to attend classes when the timing is ‘bad’
(e.g. Monday 9:00 am)

Main take away from IV

- IV estimates are a powerful tool to identify causal links
- But IV power relies on the quality of the instruments
- Always discuss instrument plausibility
- Three dimensions:
 - ① Power
 - Always report the first stage (F-test above 10)
 - Weak instruments have very unpleasant consequences
 - ② Exogeneity
 - Does it make sense to believe that the instrument is randomly assigned?
 - To be sure: check if the instrument is correlated with predetermined variables
 - ③ Exclusion restriction
 - Cannot be tested, but discuss the possible links between z and u
- Specify the group which is affected by the instrument (LATE)

EC902/907: Econometrics A

Lecture 14.1: Difference-in-differences (i)

Manuel Bagues

University of Warwick

This week

- Difference-in-differences (DiD)
 - Example: Death penalty ([slides 14.1](#))
 - Example: Card and Krueger (1994) ([slides 14.2](#))
 - Improving traditional DiD set up ([slides 14.3](#))

Roadmap so far:

- With every research question it is not possible to run a randomized controlled trial.
- Maybe we can look for an instrumental variable, but good instruments are difficult to find...
- We may also try to learn about the impact of a treatment using an empirical strategy based on observables:
 - We can compare individuals exposed to the treatment with other individuals that look alike in terms of observables.
 - Unfortunately, this evidence is likely to be subject to selection biases and often it is difficult to interpret.
- What else can we do → Difference-in-differences
 - We look for a control group such that its evolution provides a good counterfactual for how the treatment group would have evolved in the absence of the treatment
 - Note: the control is assumed to evolve similar, but does not need to be similar in levels to the treatment group

Example: the impact of death penalty

- Let us consider the case of a dychotomic treatment
- For instance, we can consider the following question:
 - Does death penalty reduce the homicide rate?
- How would address this question?
 - How would you exploit the fact that some US states have introduced/abolish death penalty?

Example: the impact of death penalty

Before-and-after approach

- Note: We will follow the review of the literature by **Donohue and Wolfers 2005** to analyze how different scholars have approached this question.
- Some authors have analyzed how the homicide rate evolves before and after the abolition (or the introduction) of death penalty.
- For instance **Dezhbakhsh and Shepherd (2004)** use data from US states which have either introduced or abolished the death penalty between 1960 and 2000. They show that:
 - when the death penalty is abolished, the homicide rate tends to increase
 - when the death penalty is reinstated, the homicide rate tends to decrease

Table 1: Estimating How Changes in Death Penalty Laws Effect Murder: Selected Before and After Comparisons: 1960-2000

Dependent Variable: % Change in State Murder Rates Around Regime Changes						
	Death Penalty Abolition			Death Penalty Reinstatement		
	1-Year Window	2-Year Window	3-Year Window	1-Year Window	2-Year Window	3-Year Window
	(1)	(2)	(3)	(4)	(5)	(6)
Panel B: Our Replication: Changes Around Death Penalty Shifts (Treatment)						
Mean Change	10.1% *** (2.9)	16.0% *** (2.3)	21.5% ** (2.6)	-6.3% * (3.4)	-7.0% ** (2.9)	-3.8% (2.9)
Median Change	8.5%	13.8%	18.5%	-9.3%	-8.5%	-7.4%
Number of States Where Homicide Increased	35/46	39/46	41/46	12/41	15/39	14/39

- As **Donohue and Wolfers (2005)** point out, there are two possible interpretations for this empirical evidence:
 - Causal effect: the introduction (abolition) of death penalty decreases (increases) homicides rate
 - Spurious correlation: there some confounding effects

Differences-in-differences

- How can we control for confounding effects?
- Differences-in-differences strategy: We can try to look for a control group which is similarly affected by these confounding effects.
- For instance, we can also examine the evolution of homicide rates during the same period in states that did not experience any policy change.
- Donohue and Wolfers 2005 show that this group exhibits very similar trends (Table 1, panel C).

Table 1: Estimating How Changes in Death Penalty Laws Effect Murder: Selected Before and After Comparisons: 1960-2000

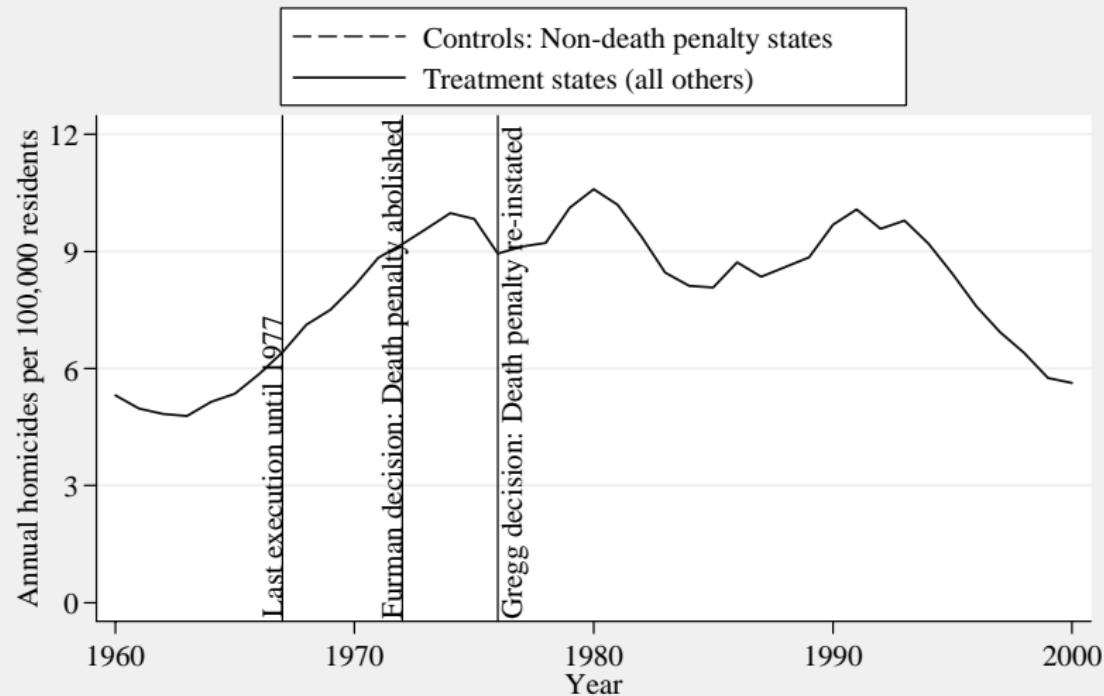
Dependent Variable: % Change in State Murder Rates Around Regime Changes						
	Death Penalty Abolition			Death Penalty Reinstatement		
	1-Year Window	2-Year Window	3-Year Window	1-Year Window	2-Year Window	3-Year Window
	(1)	(2)	(3)	(4)	(5)	(6)
Panel B: Our Replication: Changes Around Death Penalty Shifts (Treatment)						
Mean Change	10.1% *** (2.9)	16.0% *** (2.3)	21.5% *** (2.6)	-6.3% * (3.4)	-7.0% ** (2.9)	-3.8% (2.9)
Median Change	8.5%	13.8%	18.5%	-9.3%	-8.5%	-7.4%
Number of States Where Homicide Increased	35/46	39/46	41/46	12/41	15/39	14/39
Panel C: Our Innovation: Changes in Comparison States (Control)						
Mean Change	8.7% *** (0.5)	16.0% *** (0.8)	20.6% *** (1.1)	-7.5% *** (1.5)	-6.6% *** (1.5)	-3.7% *** (1.3)
Median Change	8.5%	16.1%	20.9%	-11.5%	-9.8%	-5.2%
Number of States Where Homicide Increased	44/46	44/46	44/46	7/41	8/39	8/39

- If we compare the two groups, states that introduced/abolished death penalty (Panel B) vs. states that did not make any changes (Panel C), there are not significant differences (panel D).

Table 1: Estimating How Changes in Death Penalty Laws Effect Murder: Selected Before and After Comparisons: 1960-2000

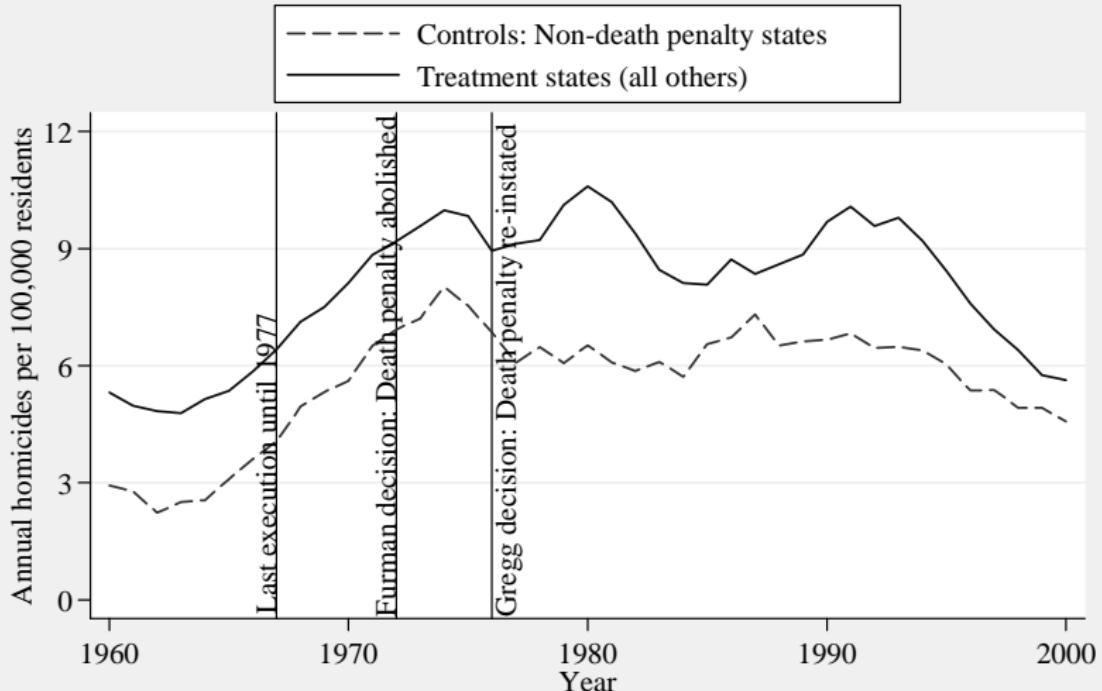
Dependent Variable: % Change in State Murder Rates Around Regime Changes						
	Death Penalty Abolition			Death Penalty Reinstatement		
	1-Year Window	2-Year Window	3-Year Window	1-Year Window	2-Year Window	3-Year Window
	(1)	(2)	(3)	(4)	(5)	(6)
Panel D: Difference-in-Difference Estimates (Treatment-Control)						
Mean Change	1.4% (2.9)	-0.1% (2.4)	0.9% (2.8)	1.2% (3.7)	-0.5% (3.2)	-0.1% (3.2)
Median Change	<0.001% (2.7)	-2.3% (2.5)	-2.4% (3.6)	2.2% (3.5)	1.3% (4.5)	-2.2% (2.0)

- We may be concerned that, at the state-level, any policy changes are part of a larger policy bundle (e.g. when democrat governors are elected, they tend to abolish death penalty and also implement many other policies)
- We can also focus on the death penalty moratorium between 1972 and 1978.
- First, let us see how the number of homicides varies when:
 - death penalty was abolished in 1972
 - death penalty was reinstated in 1978
- Let us consider separately:
 - states that had death penalty (and therefore were affected by the moratorium)
 - states that did not (and should not have been affected)



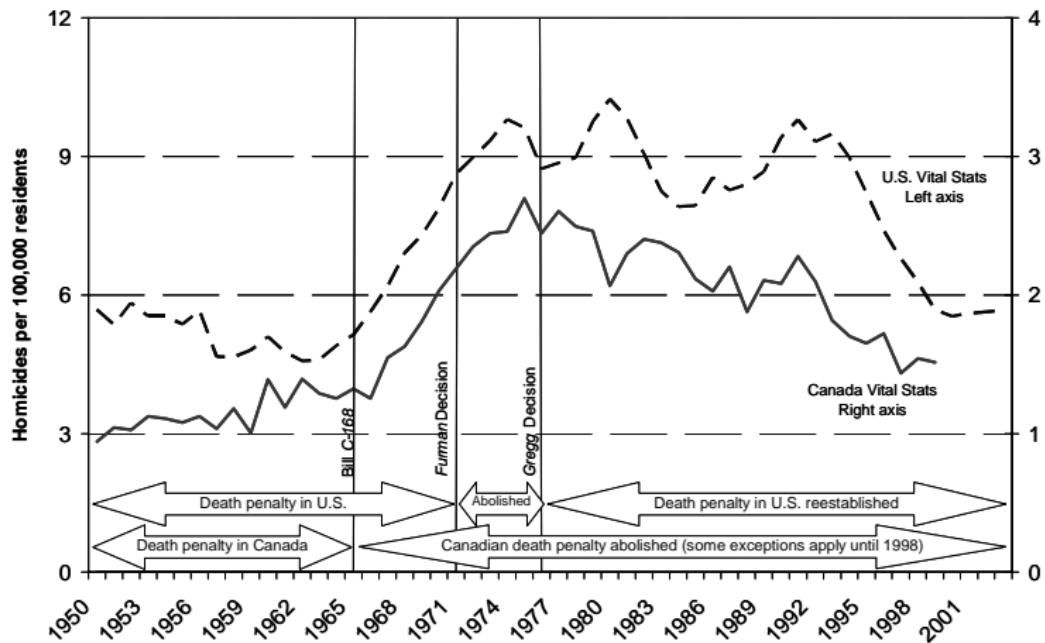
Non-death penalty states are those without a death penalty throughout 1960-2000: AK HI ME MI MN WI

Let us compare it to states that did not have death penalty



Non-death penalty states are those without a death penalty throughout 1960-2000: AK HI ME MI MN WI

Or if we compare the evolution of homicide rates in the US and Canada



Differences-in-differences (dif-in-dif)

- The above example captures the main intuition behind the **differences-in-differences** analysis.
- We use the evolution of the outcome variable in the control group to construct a counterfactual of what would have happened in the treatment group in the absence of the treatment.
- **Parallel trends assumption:** The fundamental identifying assumption is that, in the absence of the treatment, both groups would have followed **parallel trends**
- Note that this empirical strategy allows for the existence of time-invariant differences between the two groups, but it assumes that there are no time-variant relevant differences.

Main threats to validity of dif-in-dif estimates

- ① If the groups are different in levels, maybe they evolve differently?
- ② Why did the treatment group adopt the policy, and not the control group?
- ③ Policies are usually implemented in bundles (the timing of the treatment may not be by chance) → the outcome variable may be affected by these other policies
- ④ The treatment should not affect the control group
- ⑤ The composition of the treatment and control groups should not change as a result of treatment

Usual checks

- ① The two groups evolved similarly in the past (although note that this is neither a sufficient nor a necessary condition for the validity of the empirical strategy!)
- ② The timing of the adoption of the policy was as good as random
- ③ No other policies were adopted at the same time
- ④ Verify that there is no reason to believe that the control group might be affected

Fixed effects vs. Difference-in-differences

- The individual-fixed effect approach:
 - the variable of interest is often continuous and changes repeatedly over time (e.g.: years of experience)
- Difference-in-differences is a particular version of a fixed-effects models where:
 - The treatment is usually dichotomous
 - Units in the treatment group start being exposed to the treatment at time t (i.e.: a new law is implemented in a certain region, but not in the control regions)
- The Difference-in-differences framework helps us to think much more carefully about identification issues.

EC902/907: Econometrics A

Lecture 14.2: Difference-in-differences (ii)

Manuel Bagues

University of Warwick

This week

- Difference-in-differences (DiD)
 - Example: Death penalty example (slides 14.1)
 - Example: Card and Krueger (1994) ([slides 14.2](#))
 - Improving traditional DiD set up (slides 14.3)

Empirical Estimates of Minimum Wage Effects: Card and Krueger (1994)

- Until the 1990s, most empirical research suggested that minimum wages depressed employment.
- However, a highly influential paper by Card and Krueger (AER 1994) found positive employment effects using a difference-in-differences strategy.
- The employment impacts of minimum wages have been a controversial topic among labor economists ever since.
- Before covering this paper, let us discuss the theory of minimum wage effects on employment.

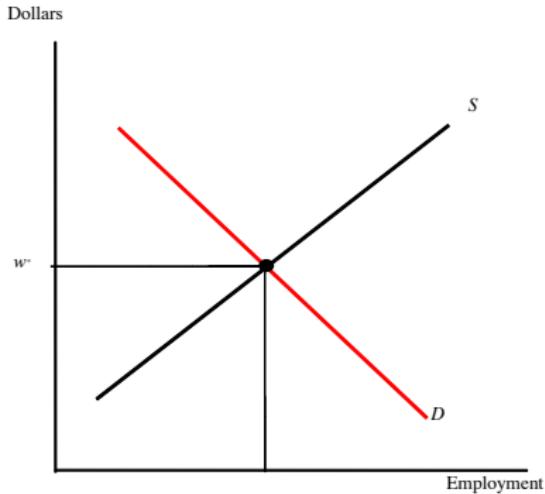
Effect of Minimum Wage on Employment: Theory

- Theory:

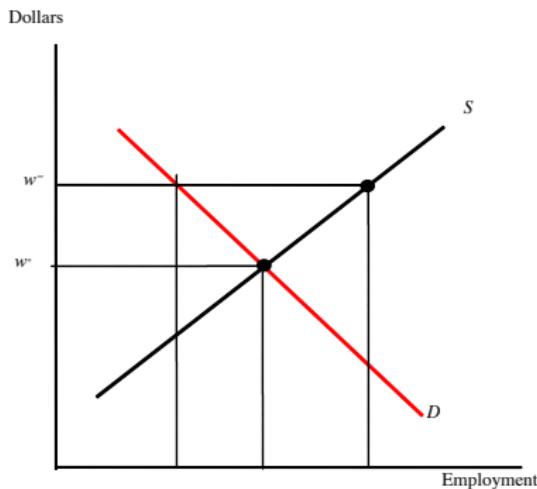
Effect of Minimum Wage on Employment: Theory

- Theory:
 - ➊ In a competitive model the result of increasing the minimum wage is to reduce employment.

The Impact of Minimum Wages: Perfect Competition



The Impact of Minimum Wages: Perfect Competition



A minimum wage set at w^- results in employers cutting employment from E^* to E^- . The higher wage also encourages $E_S - E^*$ workers to enter the market. Thus, under a minimum wage, $E_S - E^-$ workers are unemployed.

Effect of Minimum Wage on Employment: Theory

- Theory:
 - ① In a competitive model the result of increasing the minimum wage is to reduce employment.
 - ② However, in a monopsonistic model an increase in minimum wages can actually increase employment.
- Let us examine a simple ‘Mickey Mouse’ example to illustrate how this can happen

'Mickey Mouse' example

- Let us conduct a small experiment to understand how minimum wages affect employment under monopsony
- Two groups: employers and workers
- There are 10 workers, numbered from 1 to 10.
- Payoffs:
 - Employers: Productivity of workers - salaries
 - Workers: Salary - reservation wage
- Average Productivity = Marginal Productivity = 10 (same for all workers)
- Reservation wage of workers = i , where $i=1\dots 10$
- We will consider three scenarios
 - 1 Competitive labor market
 - 2 Monopsonistic labor market
 - 3 Monopsonistic labor market + minimum wage

Competitive labor market

- Sequential game:
 - ① Employers: select which salary you offer (you can hire more than one worker)
 - ② Workers: select which salary offer you accept
- Results:
 - Salary?
 - Employment?
 - Firm profits?

Competitive labor market

- Sequential game:
 - ① Employers: select which salary you offer (you can hire more than one worker)
 - ② Workers: select which salary offer you accept
- Results:
 - Salary=10
 - Employment=10
 - Firm profits=0

Monopsonistic labor market

- One single employer
- Sequential game:
 - ① Employers: select jointly which salary you offer
 - ② Workers: select whether you accept the offer or not
- Employers: make sure that you maximize your profit.
 - Example: Wage=8 → 8 workers take the offer,
 $\text{profit} = (\text{Prod}-\text{W}) * \text{N} = (10-8) * 8 = 16$
 - Example: Wage=2 → 2 workers take the offer,
 $\text{profit} = (\text{Prod}-\text{W}) * \text{N} = (10-2) * 2 = 16$
- Results:
 - Salary paid by the monopsonist?
 - Employment?
 - Firm profits?

Monopsonistic labor market

- One single employer
- Sequential game:
 - ① Employers: select jointly which salary you offer
 - ② Workers: select whether you accept the offer or not
- Employers: make sure that you maximize your profit.
 - Example: Wage=8 → 8 workers take the offer,
 $\text{profit}=(\text{Prod}-\text{W}) * \text{N} = (10-8) * 8 = 16$
 - Example: Wage=2 → 2 workers take the offer,
 $\text{profit}=(\text{Prod}-\text{W}) * \text{N} = (10-2) * 2 = 16$
- Results:
 - Salary paid by the monopsonist=5
 - Employment=5
 - Firm profits= $10 * 5 - 5 * 5 = 25$

Monopsonistic labor market + minimum wage

- One single employer
- Minimum wage fixed by the central planner. For instance, let us assume equal to 8.
- Sequential game:
 - ① Employers: select jointly which salary you offer (must be at least equal to the minimum wage)
 - ② Workers: select whether you accept the offer or not
- Results:
 - Minimum salary=8
 - Employment?
 - Firm profits?

Monopsonistic labor market + minimum wage

- One single employer
- Minimum wage fixed by the central planner. For instance, let us assume equal to 8.
- Sequential game:
 - ① Employers: select jointly which salary you offer (must be at least equal to the minimum wage)
 - ② Workers: select whether you accept the offer or not
- Results:
 - Minimum salary=8
 - Employment=8
 - Firm profits= $8 \cdot 10 - 8 \cdot 8 = 16$

Effect of Minimum wages on employment

- On April 1, 1992, New Jersey raised the state minimum from \$4.25 to \$5.05, whereas in the bordering state of Pennsylvania the minimum wage stayed at \$4.25 throughout this period.
- Card and Krueger (1994) evaluated the effect of this change on the employment of low wage workers.
- They conducted a survey to some 400 fast food restaurants from the two states just before the NJ reform, and a second survey to the same outlets 7-8 months after.

Treatment and Control Locations

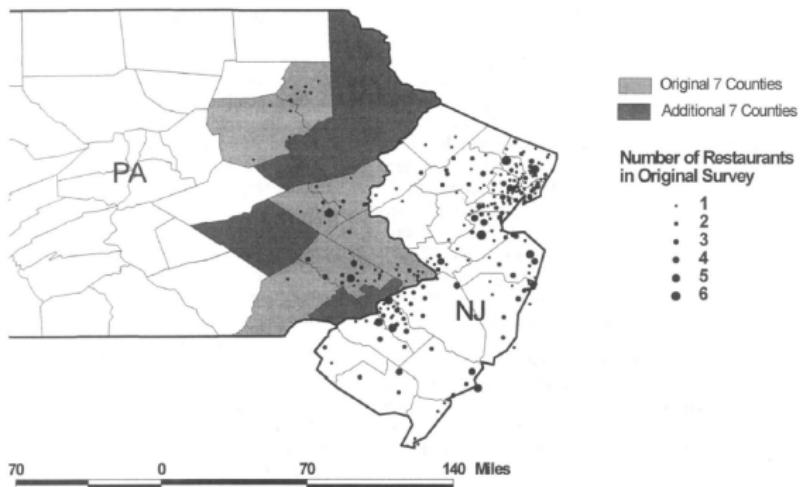
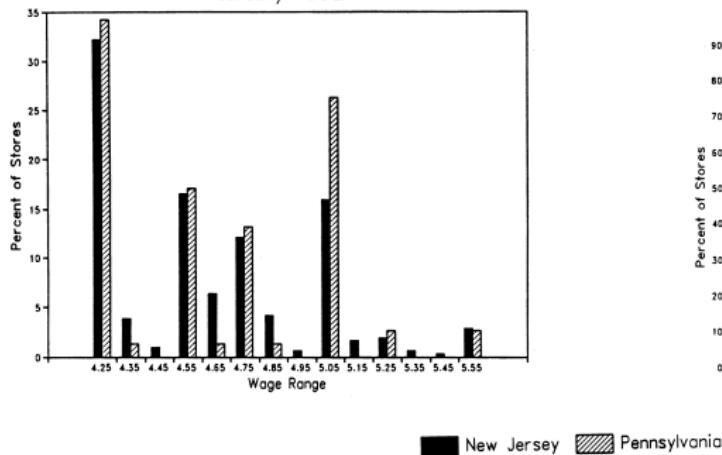


FIGURE 1. AREAS OF NEW JERSEY AND PENNSYLVANIA COVERED BY ORIGINAL SURVEY AND BLS DATA

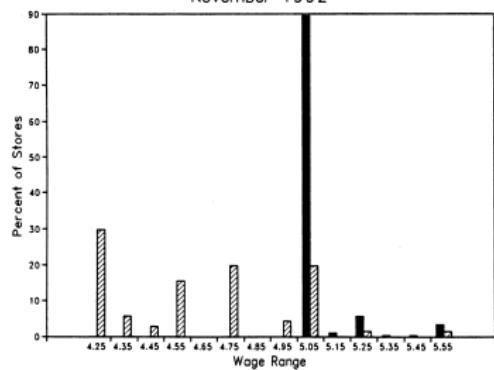
- Characteristics of fast food restaurants:
 - ❶ A large source of employment for low-wage workers.
 - ❷ They comply with minimum wage regulations (especially franchised restaurants).
 - ❸ Fairly homogeneous job, so good measures of employment and wages can be obtained.
 - ❹ Easy to get a sample frame of franchised restaurants (yellow pages) with high response rates.
 - ❺ Response rates 87% and 73% (less in Penn, because the interviewer was less persistent).

Distribution of wage rates, before and after

February 1992



November 1992



■ New Jersey ■ Pennsylvania

- The Dif-in-Dif coefficient is

$$\beta_3 = [E(Y_1|D = 1) - E(Y_0|D = 1)] - [E(Y_1|D = 0) - E(Y_0|D = 0)].$$

- where Y_0 and Y_1 denote employment before and after the reform, $D = 1$ denotes a store in NJ (treatment group) and $D = 0$ in Penn (control group).
- β_3 measures the difference between the average employment change in NJ and the average employment change in Penn.
- The **key identifying assumption** of the Difference-in-differences strategy is that, in the absence of treatment, NJ would have evolved in a similar way as PA

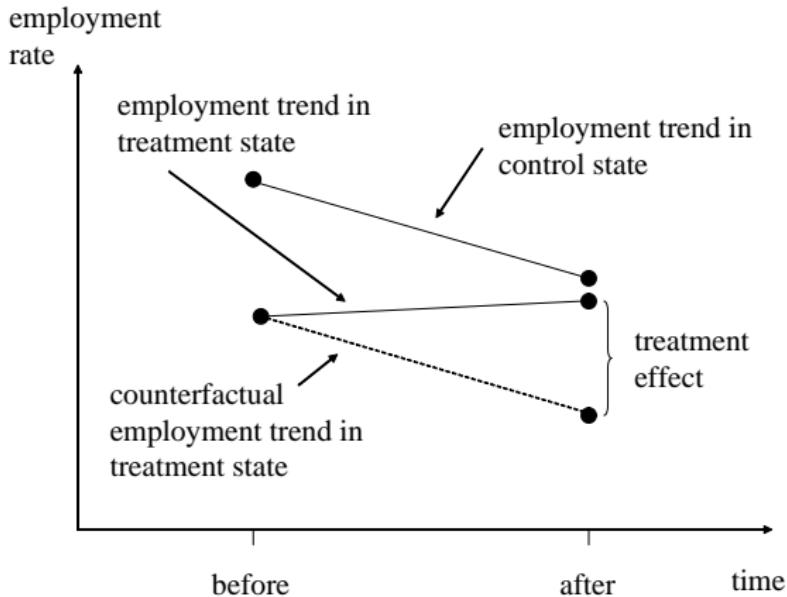


Figure 5.2.1: Causal effects in the differences-in-differences model

Table 5.2.1: Average employment per store before and after the New Jersey minimum wage increase

Variable	PA (i)	NJ (ii)	Difference, NJ-PA (iii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	-2.89 (1.44)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	-0.14 (1.07)
3. Change in mean FTE employment	-2.16 (1.25)	0.59 (0.54)	2.76 (1.36)

Notes: Adapted from Card and Krueger (1994), Table 3. The table reports average full-time equivalent (FTE) employment at restaurants in Pennsylvania and New Jersey before and after a minimum wage increase in New Jersey. The sample consists of all stores with data on employment. Employment at six closed stores is set to zero. Employment at four temporarily closed stores is treated as missing. Standard errors are reported in parentheses

Simple Regression Dif-in-Dif

We can also use the following regression to estimate the Difference-in-differences coefficient:

$$Y_{ist} = \beta_0 + \beta_1 TREAT_s + \beta_2 AFTER_t + \beta_3 (AFTER_t * TREAT_s) + \varepsilon_{ist}$$

- Y_{ist} is the number of full-time employees working in establishment i , located in state $s \in \{\text{NJ}, \text{PA}\}$, in period $t \in \{\text{Feb 1992}, \text{Nov 1992}\}$
- $TREAT_s$: dummy variable equal to 1 when $s=\{\text{NJ}\}$
- $AFTER_t$: dummy variable equal to 1 when $t=\{\text{Nov 1992}\}$
- $TREAT_s * AFTER_t$ interaction term that takes value one when $s=\{\text{NJ}\}$ & $t=\{\text{Nov 1992}\}$

Simple Regression DD: interpreting coefficients

$$Y_{ist} = \beta_0 + \beta_1 TREAT_s + \beta_2 AFTER_t + \beta_3 (AFTER_t * TREAT_s) + \varepsilon_{ist}$$

- $E(Y/TREAT=0, AFTER=0) = \beta_0$
- $E(Y/TREAT=1, AFTER=0) = \beta_0 + \beta_1$
- $E(Y/TREAT=0, AFTER=1) = \beta_0 + \beta_2$
- $E(Y/TREAT=1, AFTER=1) = \beta_0 + \beta_1 + \beta_2 + \beta_3$

Simple Regression DD: interpreting coefficients

$$Y_{ist} = \beta_0 + \beta_1 TREAT_s + \beta_2 AFTER_t + \beta_3 (AFTER_t * TREAT_s) + \varepsilon_{ist}$$

- β_0 :
- β_1 :
- β_2 :
- β_3 :

Simple Regression DD: interpreting coefficients

$$Y_{ist} = \beta_0 + \beta_1 TREAT_s + \beta_2 AFTER_t + \beta_3 (AFTER_t * TREAT_s) + \varepsilon_{ist}$$

- β_0 : average Y in the control group (PA) in the pre-treatment period
- β_1 : difference in Y between treatment group (NJ) and control group (PA) in the pre-treatment period
- β_2 : ΔY in the control group between the pre-treatment and the treatment period
- β_3 : ΔY in the treatment group between the pre-treatment and the treatment period, relative to the ΔY in the control group
→ captures effect of the policy!

```
. xi: reg EMPTOT i.NEWJERSEY*i.AFTER, cluster(ID)
i.NEWJERSEY      _INEWJERSEY_0-1      (naturally coded; _INEWJERSEY_0 omitted)
i.AFTER          _IAFTER_0-1        (naturally coded; _IAFTER_0 omitted)
i.NEW~Y*i.AFTER  _INEXWAFT_#_#      (coded as above)
```

Linear regression

Number of obs = 794
 F(3, 409) = 1.80
 Prob > F = 0.1462
 R-squared = 0.0074
 Root MSE = 9.4056

(Std. Err. adjusted for 410 clusters in ID)

EMPTOT	Robust					[95% Conf. Interval]
	Coef.	Std. Err.	t	P> t		
_INEWJERSEY_1	-2.891761	1.439546	-2.01	0.045	-5.721593	-.0619281
_IAFTER_1	-2.165584	1.218025	-1.78	0.076	-4.559954	.2287855
_INEXWAFT_1_1	2.753606	1.306607	2.11	0.036	.1851025	5.322109
_cons	23.33117	1.346536	17.33	0.000	20.68417	25.97816

Regression DD: Including controls

$$Y_{ist} =$$

$$\beta_0 + \beta_1 TREAT_s + \beta_2 AFTER_t + \beta_3 (AFTER_t * TREAT_s) + X'_{ist} \beta + \varepsilon_{ist}$$

- Sometimes you may want to control for certain time-varying covariates (X_{ist}).
- Including controls may help to obtain more precise estimates, but make sure not to include *bad controls*

Results of the CK Study

- Wages increased by 10% in NJ, remained constant PA
- ... but employment rose in NJ and decreased in PA
- The dif-in-dif estimate suggests that **the rise of minimum wage increased employment**
- Result robust to alternative specifications and to an alternative control group (workers with salaries above the minimum salary)

Reactions to the CK Study

- Angus Deaton: “*The reception accorded to Princeton faculty by their colleagues in other institutions is what might be expected by the friends and defenders of child-molesters*”
- James Buchanan in the Wall Street Journal:
“no self-respecting economist would claim that increases in the minimum wage increase employment. Such a claim, if seriously advanced, becomes equivalent to a denial that there is even minimum scientific content in economics, and that, in consequence, economists can do nothing but write as advocates for ideological interests.
Fortunately, only a handful of economists are willing to throw over the teaching of two centuries; we have not yet become a bevy of camp-following whores”

See Angus Deaton's “Letters from America” for more:

www.princeton.edu/~deaton/downloads/letterfromamerica_oct1996.html

Reactions to the CK Study

- Neumark and Wascher (2000, AER)
 - CK data has a lot of measurement error
 - data provided by Employment Policies Institute reveal that the minimum wage rise did decrease employment

Reactions to the CK Study

- Neumark and Wascher (2000, AER)
 - CK data has a lot of measurement error
 - data provided by Employment Policies Institute reveal that the minimum wage rise did decrease employment
- Card and Krueger (2000, AER)
 - administrative data from Bureau of Labor Statistics confirm the key findings of the 1994 paper
 - *“calls into question the representativeness of the sample assembled by Berman, Neumark and Wascher”*

See John Schimitt's "Cooked to Order" for more:

www.prospect.org/cs/articles?article=cooked_to_order

Card and Krueger (2000)

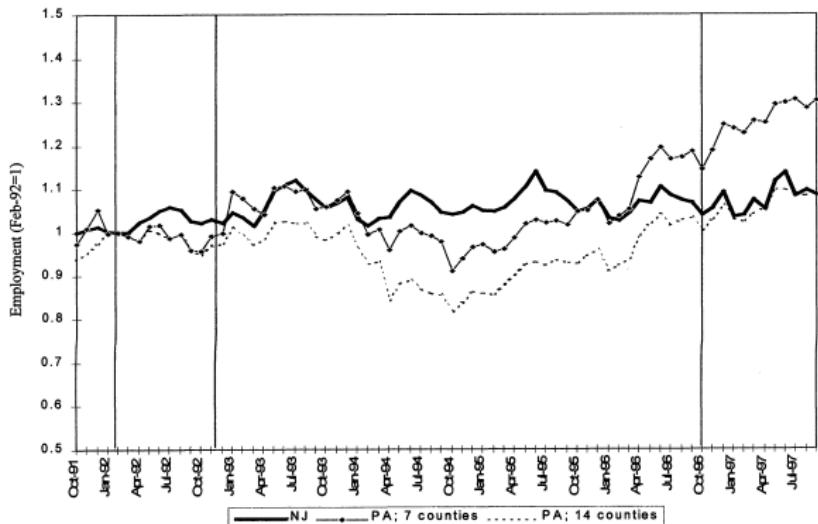


FIGURE 2. EMPLOYMENT IN NEW JERSEY AND PENNSYLVANIA FAST-FOOD RESTAURANTS, OCTOBER 1991 TO SEPTEMBER 1997

Note: Vertical lines indicate dates of original Card-Krueger survey and the October 1996 federal minimum-wage increase.
Source: Authors' calculations based on BLS ES-202 data.

- My own personal take away from this paper:
 - Historically it represented a methodological advance...
 - ...but the paper is weak according to current standards.
- Some potential methodological concerns:
 - ❶ The authors do not examine how the trends evolved in the past. Information from future trends suggests that they are not parallel.
 - ❷ At the end of the day, we only have two observations. Possible common shocks may affect the treatment or the control group.
 - ❸ Other policies?
 - ❹ Note also the tension between having observations that are geographically close and the potential existence of an impact of the treatment on the control group.

EC902/907: Econometrics A

Lecture 14.3: Difference-in-differences (iii)

Manuel Bagues

University of Warwick

This week

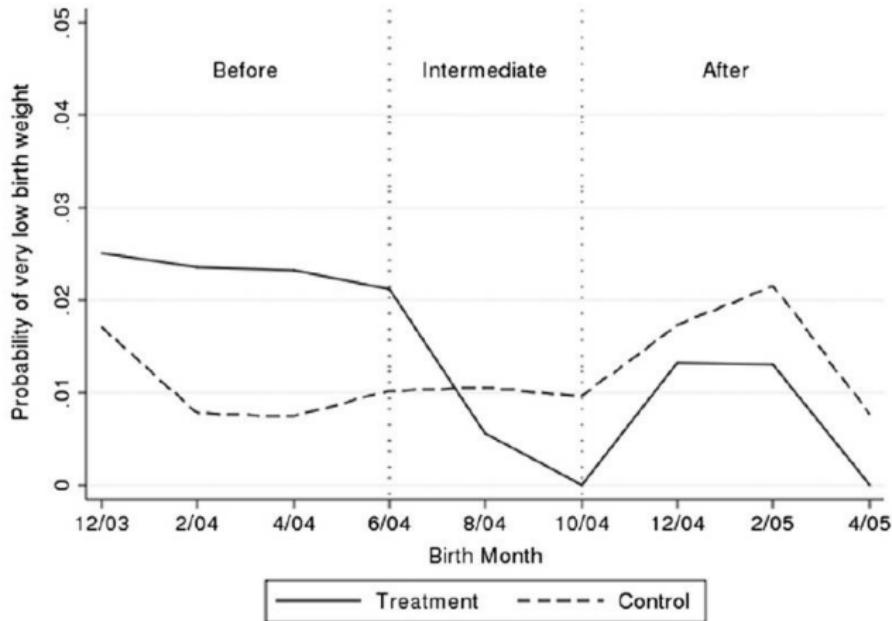
- Difference-in-differences (DiD)
 - Example: Death penalty example (slides 14.1)
 - Example: Card and Krueger (1994)(slides 14.2)
 - Improving traditional DiD set up ([slides 14.3](#))

Improving Traditional Dif-and-Dif Set Up

- The crucial assumption in DID set up is that the control group provides information about how would the treatment group evolved in the absence of treatment (parallel trends)
- With more than two periods this can be investigated in several ways...
 - ① Illustrate graphically that the evaluation of outcomes was similar in the years before the policy was implemented, e.g. plotting the means
 - ② In the same vein, estimate formally some placebo models: does the placebo policy introduced in t-1, t-2, etc. have any significant impact?
 - ③ Include group-specific trends

Graphical checks

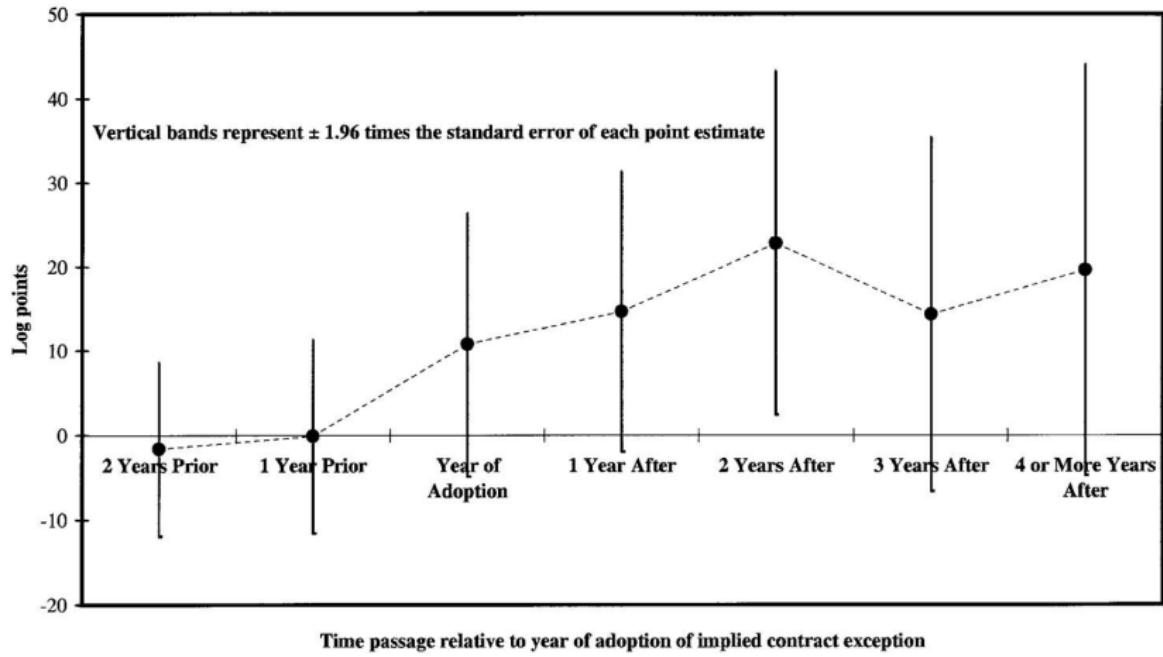
- A Difference-in-differences paper should always begin by showing graphically the evolution of the treatment and the control group before and after the introduction of the treatment
- Just plot the means for both groups before and after the treatment took place
- For instance, Bharadwaj, Johnsen and Loken (2014) study the impact of restaurant smoking bans on birth weights



Event study framework

$$Y_{ist} = \alpha_0(s) + \beta_t + \sum_{\tau=0}^m \beta_{-\tau} D_{st-\tau} + \sum_{\tau=1}^k \beta_\tau D_{st+\tau} + X'_{ist} \gamma + \varepsilon_{ist}$$

- If more than two time periods, one can check whether the difference between treatment (D) and control units discontinuously changes at the time of policy change
- Autor (2003) study the effect of employment protection rules on employer's use of temporary help



Group Trends

- When there are more than two time periods, we can add group specific trends to take into account differences in the trajectories of the treatment and the control group
- For instance, Stephens and Yang (2014) point out that estimates of the benefits of increased schooling using US state schooling laws as instruments typically rely on specifications which assume common trends across states in the factors affecting different birth cohorts.
- This assumption may fail to hold if there are differential changes across states during this period, such as relative school quality improvements.
- They show that, if we allow for state-specific trends, the positive effect observed in previous studies of education on wages becomes lower and statistically significant.

Stephens and Yang (2014)

Table 1 - The Effect of Schooling on log Weekly Wages

Sample	White males ages 40–49		White males ages 25–54	
	(1)	(2)	(3)	(4)
OLS	0.073 (0.0005)	0.073 (0.0005)	0.063 (0.0004)	0.063 (0.0004)
2SLS	0.095 [0.064, 0.126]	-0.020 [-0.163, 0.060]	0.097 [0.080, 0.117]	-0.014 [-0.066, 0.021]
F (first stage instruments)	42.8	8.2	81.4	23.6
Fixed effects:				
State of birth	Yes	Yes	Yes	Yes
Year of birth	Yes	Yes	Yes	Yes
Region × year of birth	No	Yes	No	Yes
Additional controls	None	None	Age quartic, census yr	Age quartic, census yr

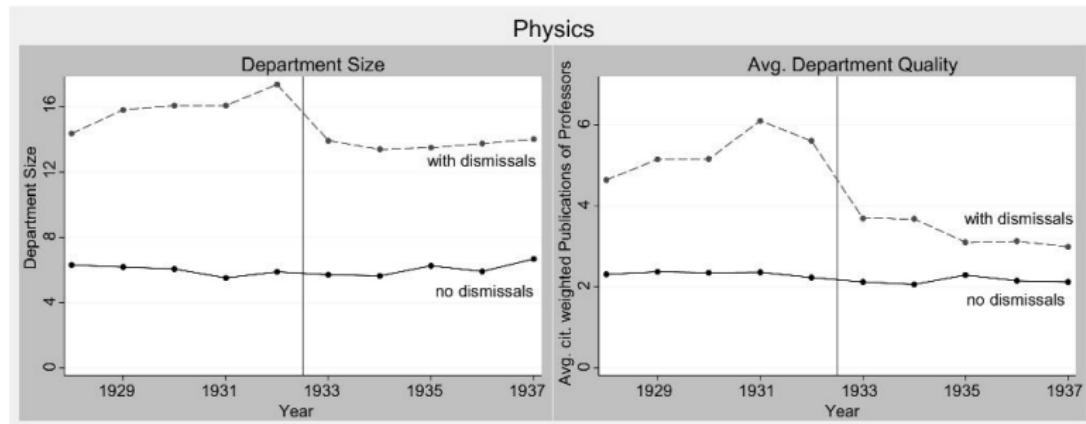
Other Examples

Peer effects in science

- Effect of top researchers on their colleagues and students
(Waldinger RES 2012, Waldinger JPE 2010)
- Only 66 days after Hitler's National Socialist party secured power, all Jewish and 'politically unreliable' scientists were dismissed: more than 1,000 researchers, about 15% of all researchers between 1933-1934 (among them, Albert Einstein, Georg von Hevesy, Johann von Neumann)
- Compare how the productivity of PhD graduates evolves in departments that are differentially affected by the dismissal of scientists by the Nazi government in 1933

Other Examples

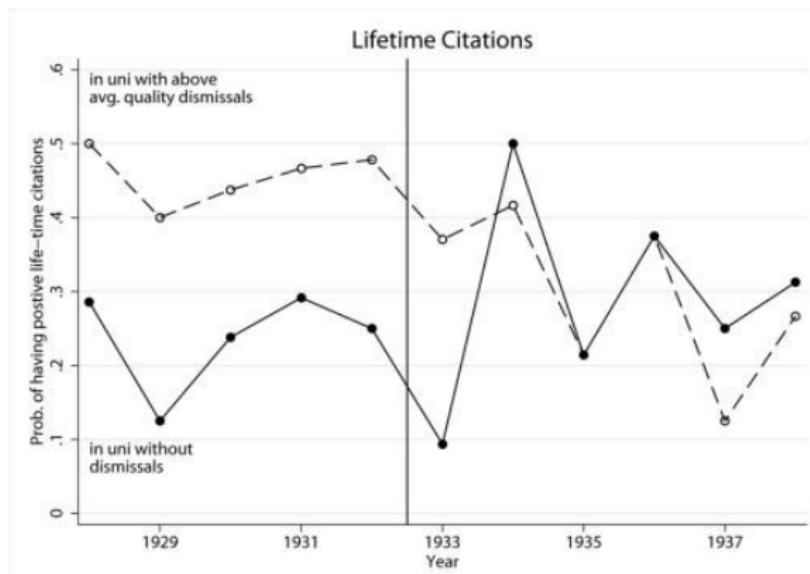
Peer effects in science



Other Examples

Peer effects in science

- Negative effect on PhD students:



Other Examples

- Impact of legalized abortion on crime ([Donohue and Levitt 2001](#)).
 - 5 states that allowed abortion in 1970 compared to the rest, which legalized in 1973.
- Economic diversity at school and social preferences and behavior ([Rao 2013](#)).
 - In 2007, a policy change in India that introduced quotas for poor children in new admissions.
- Redistribution policies and investment in education ([Abramitzky and Lavy 2014](#)).
 - Different Israeli kibbutzim shifted from equal sharing to productivity-based wages in different years

Final comments on diff-and-diff:

- Identification relies on a claim that is very often more of an act of faith than a assumption clearly grounded on theory
- Show that in the past trends were parallel (it does not guarantee that this will be satisfied in the future, but at least it is supportive)
- Discuss explicitly why it is a good assumption to believe that the timing of the treatment/policy was as good as random
- Discuss explicitly the existence of alternative policies that might contemporaneously affect the treatment or the control group
- Discuss the possibility that the control group is affected by the treatment.
- Note: next day we will discuss in detail how to calculate properly standard errors in a diff-and-diff setting.

EC902/EC907: Econometrics A

Lecture 15

Manuel Bagues

Warwick University

November 28, 2022
Lecture Slides

Logistics

- Project
 - EC902 (not for EC907)
 - Important deadlines
 - December 9 (by noon) - group composition (3 members)
 - February 6-10 - 15 min. presentation/discussion of the project with instructors
 - March 20 (by noon) - project submission
 - We will discuss today some examples of projects from previous years
- Any questions?

Roadmap so far:

- With every research question it is not possible to run a randomized controlled trial.
- Maybe we can look for an instrumental variable, but good instruments are difficult to find...
- We may also try to learn about the impact of a treatment using an empirical strategy based on observables:
 - We can compare individuals exposed to the treatment with other individuals that look alike in terms of observables.
 - Unfortunately, this evidence is likely to be subject to selection biases and often it is difficult to interpret.
- What else can we do → Difference-in-differences
 - We look for a control group such that its evolution provides a good counterfactual for how the treatment group would have evolved in the absence of the treatment
 - Note: the control is assumed to evolve similar, but does not need to be similar in levels to the treatment group

- Difference-in-differences (DiD)
 - Example: Death penalty example (slides 14.1)
 - Example: Card and Krueger (1994) (slides 14.2)
 - Improving traditional DiD set up (slides 14.3)
 - Extension: interpretation of **two-way fixed-effects in staggered DID** (**not for exam!**)
- Examples of projects

Differences-in-differences (dif-in-dif)

- Main intuition behind the **differences-in-differences** analysis: we use the evolution of the outcome variable in the control group to construct a counterfactual of what would have happened in the treatment group in the absence of the treatment.
- **Parallel trends assumption:** The fundamental identifying assumption is that, in the absence of the treatment, both groups would have followed **parallel trends**
- Note that this empirical strategy allows for the existence of time-invariant differences between the two groups, but it assumes that there are no time-variant relevant differences.

Use of differences-in-differences

Currie, Kleven and Zviers 2020

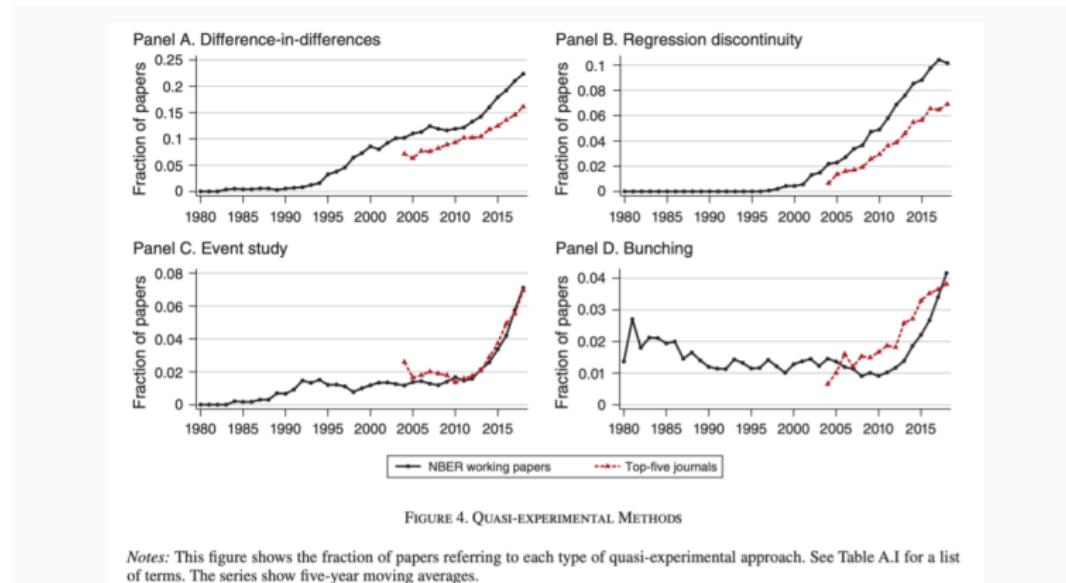


FIGURE 4. QUASI-EXPERIMENTAL METHODS

Notes: This figure shows the fraction of papers referring to each type of quasi-experimental approach. See Table A.I for a list of terms. The series show five-year moving averages.

Main threats to validity of dif-in-dif estimates

- ① If the groups are different in levels, maybe they evolve differently?
- ② Why did the treatment group adopt the policy, and not the control group?
- ③ Policies are usually implemented in bundles (the timing of the treatment may not be by chance) → the outcome variable may be affected by these other policies
- ④ The treatment should not affect the control group (SUTVA)
- ⑤ The composition of the treatment and control groups should not change as a result of treatment

Usual checks

- ① The two groups evolved similarly in the past (although note that this is neither a sufficient nor a necessary condition for the validity of the empirical strategy!)
 - Show it graphically
 - Report in a table the estimates from an event study estimation: interact ‘treatment dummy’ with lags and forwards of the ‘ time dummies’ (e.g. year dummies)
- ② Make sure that the timing of the adoption of the policy was as good as random (e.g. not driven by expectations of economic shocks)
- ③ No other policies were adopted at the same time
- ④ Verify that there is no reason to believe that the control group might be affected by the treatment

Effect of Minimum Wage on Employment: Theory

- Theory:
 - ① In a competitive model the result of increasing the minimum wage is to reduce employment.
 - ② However, in a monopsonistic model an increase in minimum wages can actually increase employment.
- Empirical evidence
 - Compare the evolution of employment in states that modified the minimum wage (treatment group) vs. states where it did not change (control group)
 - Seminal work by Card and Krueger (1994)

Effect of Minimum wages on employment

- On April 1, 1992, New Jersey raised the state minimum from \$4.25 to \$5.05, whereas in the bordering state of Pennsylvania the minimum wage stayed at \$4.25 throughout this period.
- Card and Krueger (1994) evaluated the effect of this change on the employment of low wage workers.
- They conducted a survey to some 400 fast food restaurants from the two states just before the NJ reform, and a second survey to the same outlets 7-8 months after.

Treatment and Control Locations

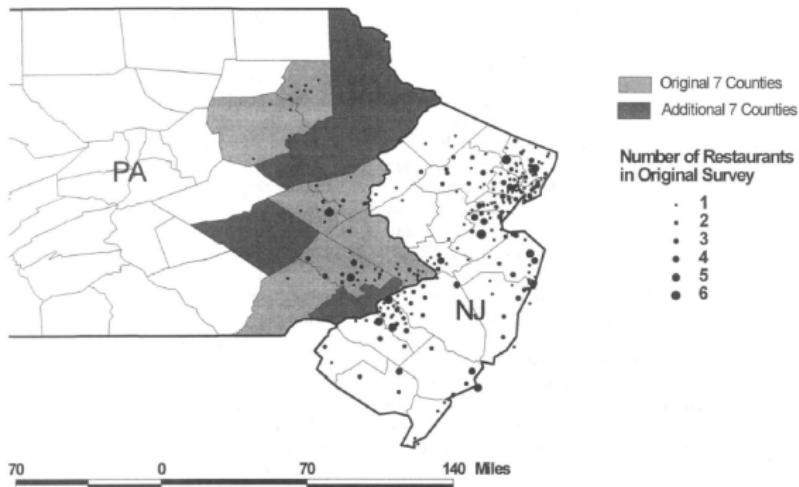
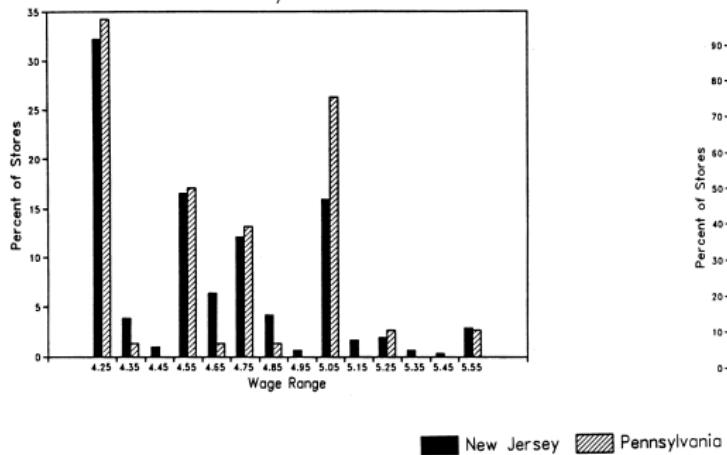


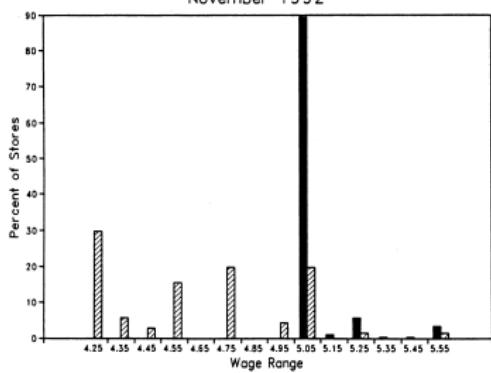
FIGURE 1. AREAS OF NEW JERSEY AND PENNSYLVANIA COVERED BY ORIGINAL SURVEY AND BLS DATA

Distribution of wage rates, before and after

February 1992



November 1992



■ New Jersey ■ Pennsylvania

- The Dif-in-Dif coefficient is

$$\beta_3 = [E(Y_1|D = 1) - E(Y_0|D = 1)] - [E(Y_1|D = 0) - E(Y_0|D = 0)].$$

- where Y_0 and Y_1 denote employment before and after the reform, $D = 1$ denotes a store in NJ (treatment group) and $D = 0$ in Penn (control group).
- β_3 measures the difference between the average employment change in NJ and the average employment change in Penn.
- The **key identifying assumption** of the Difference-in-differences strategy is that, in the absence of treatment, NJ would have evolved in a similar way as PA

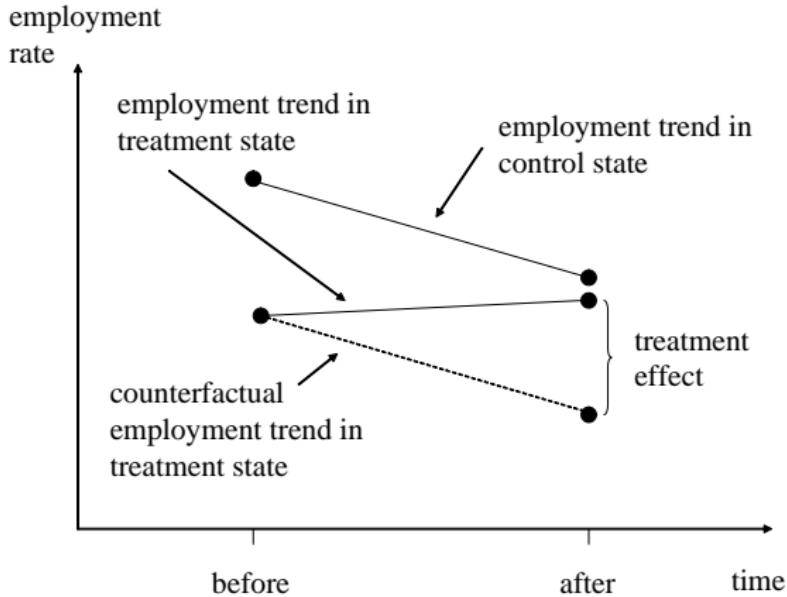


Figure 5.2.1: Causal effects in the differences-in-differences model

Table 5.2.1: Average employment per store before and after the New Jersey minimum wage increase

Variable	PA	NJ	Difference, NJ-PA
	(i)	(ii)	(iii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	-2.89 (1.44)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	-0.14 (1.07)
3. Change in mean FTE employment	-2.16 (1.25)	0.59 (0.54)	2.76 (1.36)

Notes: Adapted from Card and Krueger (1994), Table 3. The table reports average full-time equivalent (FTE) employment at restaurants in Pennsylvania and New Jersey before and after a minimum wage increase in New Jersey. The sample consists of all stores with data on employment. Employment at six closed stores is set to zero. Employment at four temporarily closed stores is treated as missing. Standard errors are reported in parentheses.

Simple Regression DD: interpreting coefficients

$$Y_{ist} =$$

$$\beta_0 + \beta_1 TREAT_s + \beta_2 AFTER_t + \beta_3 (AFTER_t * TREAT_s) + \varepsilon_{ist}$$

- β_0 : average Y in the control group (PA) in the pre-treatment period
- β_1 : difference in Y between treatment group (NJ) and control group (PA) in the pre-treatment period
- β_2 : ΔY in the control group between the pre-treatment and the treatment period
- β_3 : ΔY in the treatment group between the pre-treatment and the treatment period, relative to the ΔY in the control group
→ differential evolution of the treatment and control group
(which hopefully captures effect of the policy!)

```
. xi: reg EMPTOT i.NEWJERSEY*i.AFTER, cluster(ID)
i.NEWJERSEY      _INEWJERSEY_0-1      (naturally coded; _INEWJERSEY_0 omitted)
i.AFTER          _IAFTER_0-1        (naturally coded; _IAFTER_0 omitted)
i.NEW~Y*i.AFTER _INEXWAFT_#_#      (coded as above)
```

Linear regression

Number of obs = 794
 F(3, 409) = 1.80
 Prob > F = 0.1462
 R-squared = 0.0074
 Root MSE = 9.4056

(Std. Err. adjusted for 410 clusters in ID)

EMPTOT	Robust					[95% Conf. Interval]
	Coef.	Std. Err.	t	P> t		
_INEWJERSEY_1	-2.891761	1.439546	-2.01	0.045	-5.721593	-.0619281
_IAFTER_1	-2.165584	1.218025	-1.78	0.076	-4.559954	.2287855
_INEXWAFT_1_1	2.753606	1.306607	2.11	0.036	.1851025	5.322109
_cons	23.33117	1.346536	17.33	0.000	20.68417	25.97816

Results of the CK Study

- Minimum wage increased by 10% in NJ, remained constant PA
- ... but employment rose in NJ and decreased in PA
- The dif-in-dif estimate suggests that **the rise of minimum wage increased employment**
- Result robust to alternative specifications and to an alternative control group (workers with salaries above the minimum salary)

- My own personal take away from this paper:
 - Historically it represented a methodological advance...
 - ...but the paper is weak according to current standards.
- Some potential methodological concerns:
 - ① The authors do not examine how the trends evolved in the past. Information from future trends suggests that they are not parallel.
 - ② At the end of the day, we only have two observations. Possible common shocks may affect the treatment or the control group.
 - ③ Other policies?
 - ④ Note also the tension between having observations that are geographically close and the potential existence of an impact of the treatment on the control group.
- A recent article by [Cengiz et al \(QJE 2019\)](#) using 138 state-level minimum wage changes between 1979 and 2016 largely confirms Card and Krueger's findings

Other Examples

- Impact of legalized abortion on crime ([Donohue and Levitt 2001](#)).
 - 5 states that allowed abortion in 1970 compared to the rest, which legalized in 1973.
- Economic diversity at school and social preferences and behavior ([Rao 2013](#)).
 - In 2007, a policy change in India that introduced quotas for poor children in new admissions.
- Redistribution policies and investment in education ([Abramitzky and Lavy 2014](#)).
 - Different Israeli kibbutzim shifted from equal sharing to productivity-based wages in different years

Final comments on diff-and-diff:

- Identification relies on a claim that is very often more of an act of faith than a assumption clearly grounded on theory
- Show that in the past trends were parallel (it does not guarantee that this will be satisfied in the future, but at least it is supportive)
- Discuss explicitly why it is a good assumption to believe that the timing of the treatment/policy was as good as random
- Discuss explicitly the existence of alternative policies that might contemporaneously affect the treatment or the control group
- Discuss the possibility that the control group is affected by the treatment.
- Note: next day we will discuss in detail how to calculate properly standard errors in a diff-and-diff setting.

Staggered DID: Allowing for multiple periods and variation in treatment timing

- 2X2 comparisons (treatment, control, before, after) are well understood
- With staggered treatment implementation, two-way fixed effects estimators are commonly used.
- When treatments are heterogeneous across units or time → TWFE estimate does not have a meaningful interpretation
- Two problems:
 - ① Arbitrary weights (can be even negative)
 - ② ‘Forbidden’ comparisons (using already treated units)
- Recent papers propose solutions to diagnose and correct these issues.

Links to relevant material

- The following video presentations might be useful:
 - Goodman-Bacon
 - Callaway and Sant'Anna
- Recent review of the literature :
 - Jonathan Roth, Pedro H. C. Sant'Anna, Alyssa Bilinski, John Poe (2022), 'What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature'.

Differences-in-differences

Multiple time periods - TWFE

Staggered DID: more than two groups and two time periods

In the literature usually authors apply a two-way fixed-effects model:

$$y_{it} = \alpha_i + \alpha_t + \beta^{DD} D_{it} + \epsilon_{it} \quad (1)$$

Where α_i are unit-level fe, α_t time-trends, β^{DD} the coefficient of interest.

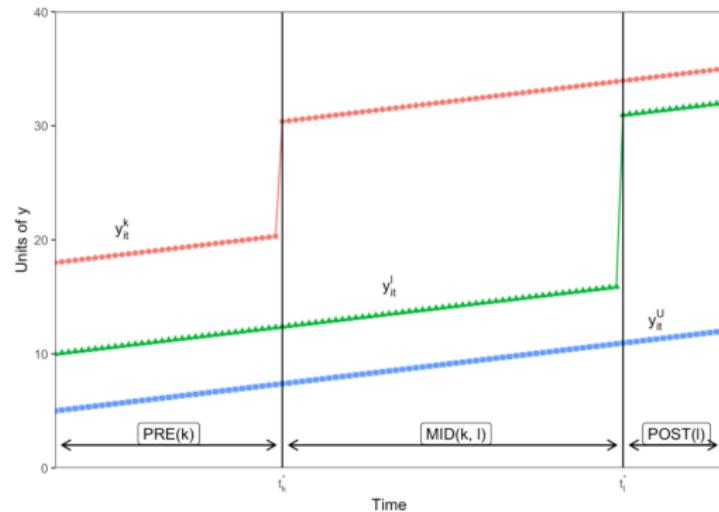
But the interpretation of β^{DD} is not straightforward. Why?

- The treatment effect is a weighted average of the comparison between different groups.
 - Not an issue of design per se, but with the use of a single coefficient to summarize treatment effects.
- The estimated effect is a weighted average of all possible comparisons across treatment cohorts and controls.

Example

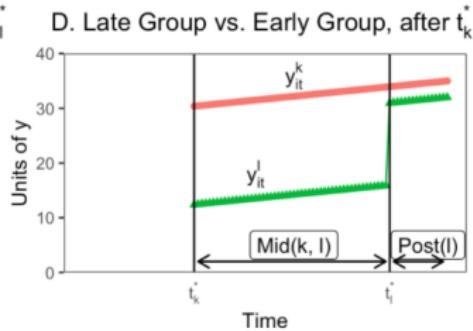
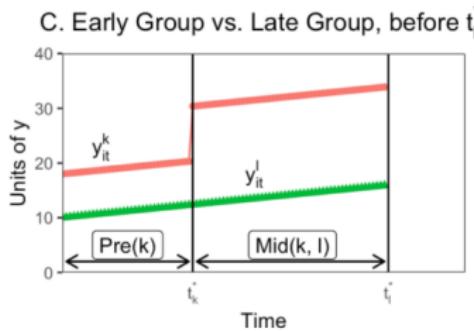
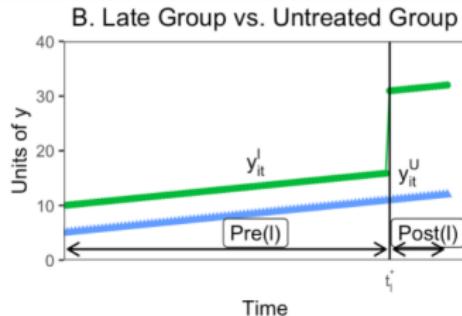
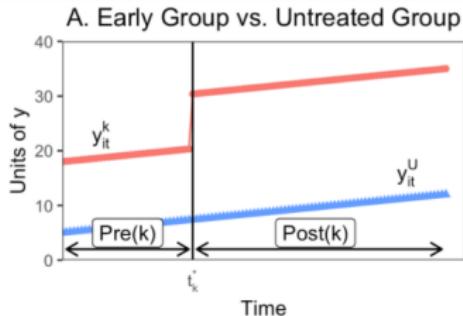
Goodman-Bacon 2019 - Multiple time periods

- 3 groups: Never treated U , early treatment k , late treated l .
- 3 windows: "PRE", "MID", "POST".



Decomposition of β^{DD}

Goodman-Bacon 2019



Decomposition of β^{DD}

Goodman-Bacon 2019

- Treated vs. Untreated comparisons:

$$\hat{\beta}_{jU}^{2x2} = \left(\bar{y}_j^{POST(j)} - \bar{y}_j^{PRE(j)} \right) - \left(\bar{y}_U^{POST(j)} - \bar{y}_U^{PRE(J)} \right), j = k, l \quad (2)$$

- Treated and not-yet treated comparison:

$$\hat{\beta}_{kl}^{2x2,k} = \left(\bar{y}_k^{MID(k,l)} - \bar{y}_k^{PRE(k)} \right) - \left(\bar{y}_l^{MID(k,l)} - \bar{y}_l^{PRE(k)} \right) \quad (3)$$

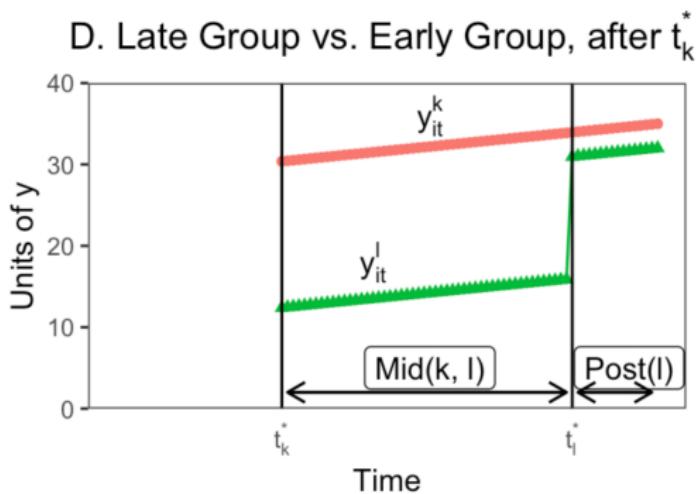
- Treated later and treated before comparison:

$$\hat{\beta}_{k,l}^{2x2,l} = \left(\bar{y}_l^{POST(l)} - \bar{y}_l^{MID(k,l)} \right) - \left(\bar{y}_k^{POST(l)} - \bar{y}_k^{MID(k,l)} \right) \quad (4)$$

Potentially problematic case

Using already treated as control, if there are dynamics

Already treated unit is treated as a control in the regression, their indicator does not switch from a period to the other.



Comparisons

- With K timing groups, you can form $K^2 - K$ timing only estimates comparing earlier and later treated groups.
- With an untreated group U you could form K treated/untreated 2×2 DiDs for a total of K^2 DiD estimates (4 in this case)
- If treatment effects are heterogeneous, using already treated as control can generate a bias. For instance, if the impact of units treated late is larger, and weights are negative, the TWFE could be negative even if all groups were being affected positively.

Decomposition Theorem

Comparisons and weights

The DiD decomposition theorem, 2 treatment groups and 2 time periods:

$$\hat{\beta}^{DD} = s_{kU}\hat{\beta}_{kU}^{2x2} + s_{\ell U}\hat{\beta}_{\ell U}^{2x2} + s_{k\ell}^k\hat{\beta}_{k\ell}^{2x2,k} + s_{k\ell}^\ell\hat{\beta}_{k\ell}^{2x2,\ell} \quad (5)$$

Note that this is a weighted average of all possible pairwise DD comparisons (slides before).

Weights

Estimated weights depend on sample shares and variance of treatment effects

Example: weight of the comparison between k and U will be:

$$s_{ku} = (n_k + n_u)^2 \frac{n_{ku} (1 - n_{ku}) \bar{D}_k (1 - \bar{D}_k)}{\widehat{\text{Var}}(\tilde{D}_{it})} \quad (6)$$

- n_{ku} refer to relative sample sizes (Concentration) ($n_{ku} = n_k / (n_k + n_u)$)
- \bar{D}_k = share of time group k spends treated.

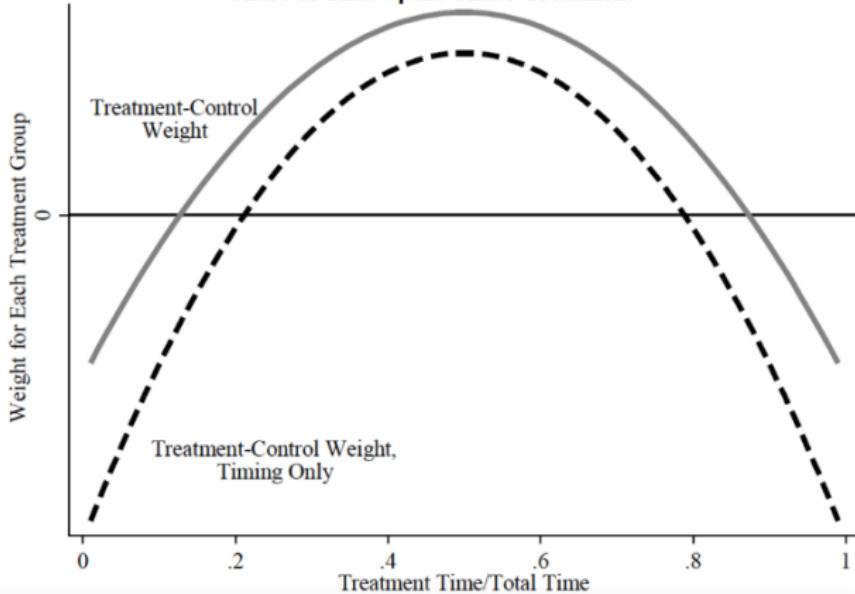
Notes:

- Each of the 2x2 DiDs identified by the treatment indicator variation in the sub-sample over which it is estimated.
- The share of the sample these observations represent also enter the weighting.
- At the middle of the panel (n_{ku} and $\bar{D}_k=1/2$) the weight is maximized.

Weights

Weights are maximized at 1/2 Treatment Time (Middle of panel)

Figure 4. Weighted Common Trends: The Treatment/Control Weights as a Function of the Share of Time Spent Under Treatment



Weights

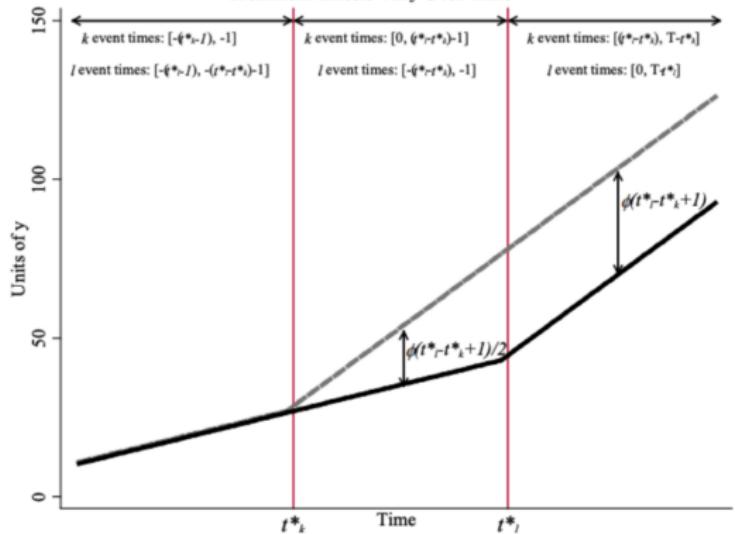
Take-aways

- **Weights are not equal to sample shares, in general.**
- Even if the treatment effects are constant, panel length by itself may affect the estimates.
- Estimates closer to the "middle" of the panel get more weight.
- Diagnostics proposed by Goodman-Bacon enable to assess which groups contribute the most to the observed treatment effect.
- So far we concentrated on weights, but we might also have dynamics (trends) in each unit's treatment effect.

Bias from dynamics

Goodman-Bacon 2019

Figure 3. Difference-in-Differences Estimates with Variation in Timing Are Biased When Treatment Effects Vary Over Time



Heterogeneity bias

The TWFE DiD estimator (when $N \rightarrow \infty$, T fixed):

$$\text{plim}_{N \rightarrow \infty} \hat{\beta}^{DD} = \text{VWATT} + \text{VWCT} - \Delta\text{ATT} \quad (7)$$

- VWATT is the ‘variance-weighted average treatment effect on the treated’
- VWCT is the ‘variance-weighted common trend’. Different groups might not have the same underlying trend in outcome dynamics.
- ΔATT weighted sum of the change in treatment effects within each unit’s post-period with respect to another unit’s treatment timing.
 - Last term enters because of the comparison between later-earlier treated, only when the treatment effect is not stable across the same unit over time, otherwise it is 0.
 - Note: heterogeneity across cohorts is not the problem here, rather dynamics for a given treatment unit over time.

New estimators for staggered DID

- Diagnostic approaches: assess how relevant is the problem!
 - Chaisemartin and D'Haultfoeuille (2020)
 - heterogeneity in treatment effects that would reverse the sign of the estimate
 - Goodman-Bacon (2021)
 - report weights for each group of comparisons: how much weight on ‘forbidden’ comparisons?
 - Jakielka (2021)
 - negative weights?
 - test the constant treatment assumption
- Several new estimators for staggered DID, with some common features:
 - Use only ‘clean’ comparisons between treated and non-treated groups
 - Aggregate them using some type of user-specified weights
- Let us see one of them in practice: Callaway and Sant'Anna (2021)

E.g. Callaway and Sant'Anna (2021)

- Let us assume parallel trends, no anticipation, and SUTVA
- Two possible control groups:
 - never-treated units
 - all not-yet treated units
- Several options for weights available
- When the number of periods and groups is small you may report all relevant $\text{ATT}(g,t)$
- Example from problem set 3:
 - Impact of minimum wage on teen employment
 - Sample of US counties, years 2003-2007, $N=2,500$
 - Some states treated in 2004, 2006 and 2007

```
. csdid lemp lpop, i(countyreal) t(year) g(first_treat)
.....
difference-in-difference with Multiple Time Periods
```

Number of obs = 2,500

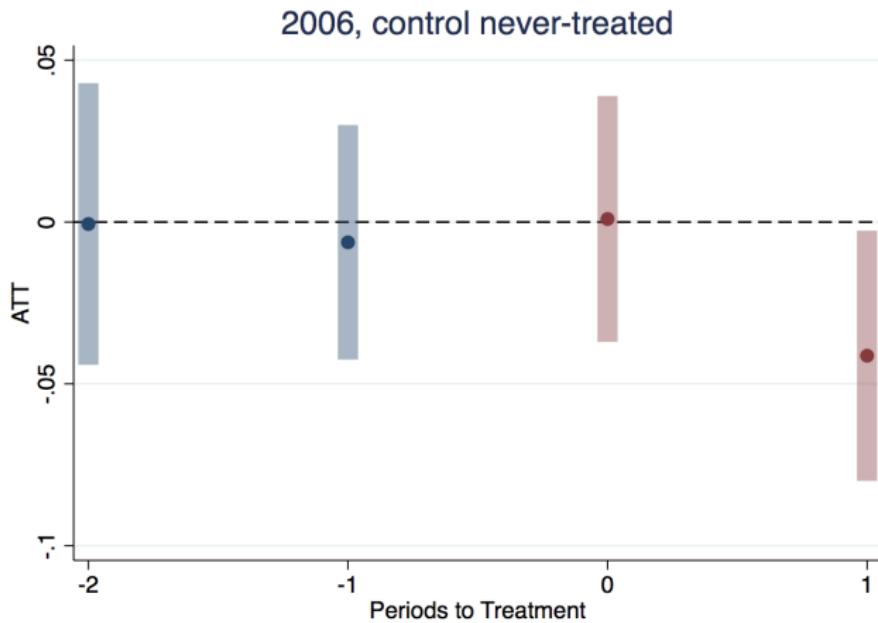
Outcome model : weighted least squares
Treatment model: inverse probability tilting

	Coefficient	Std. err.	z	P> z	[95% conf. interval]
j2004					
t_2003_2004	-.0145329	.0221264	-0.66	0.511	-.0579 .0288341
t_2003_2005	-.0764267	.0286661	-2.67	0.008	-.1326113 -.0202422
t_2003_2006	-.1404536	.035373	-3.97	0.000	-.2097834 -.0711239
t_2003_2007	-.1069093	.0328863	-3.25	0.001	-.1713652 -.0424533
j2006					
t_2003_2004	-.0006112	.022198	-0.03	0.978	-.0441185 .0428961
t_2004_2005	-.006267	.018481	-0.34	0.735	-.042489 .0299551
t_2005_2006	.0009473	.0193812	0.05	0.961	-.0370391 .0389337
t_2005_2007	-.0413123	.0197171	-2.10	0.036	-.0799571 -.0026674
j2007					
t_2003_2004	.0266993	.0140628	1.90	0.058	-.0008633 .0542619
t_2004_2005	-.0045906	.0157101	-0.29	0.770	-.0353818 .0262007
t_2005_2006	-.0284515	.0181775	-1.57	0.118	-.0640787 .0071758
t_2006_2007	-.0287821	.0162333	-1.77	0.076	-.0605988 .0030347

Control: Never Treated

Plot results for the 2006 group:

csdid_plot, group(2006)



- Aggregate result for ATE

```
. // Aggregate result for ATE
```

```
. estat simple
```

```
Average Treatment Effect on Treated
```

	Coefficient	Std. err.	z	P> z	[95% conf. interval]
ATT	-.0417577	.0115008	-3.63	0.000	-.0642989 -.0192165

- Test for pretrend

```
. // Test for pretrend
```

```
. estat pretrend
```

```
Pretrend Test. H0 All Pre-treatment are equal to 0
```

```
chi2(5) =        6.8436
```

```
p-value =        0.2325
```

Project: some examples from 2020

- Effects of cannabis legalisation on SAT scores
- The Effect of Divorce Legalization on Female Labour Force Participation Rate in Chile
- Does Extending Parental Leave Influence Children's Educational Outcomes?
- Can Cigarette Taxes Affect Low Birth Weight Infants in the United States?
- The Consequences of Re-criminalisation of Indoor Prostitution: An Empirical Study on Rhode Island
- Do Congestion Charges Affect Driving Decisions? Evidence from Norway

EC902/907: Econometrics A

Lecture 16: Regression discontinuity design

Manuel Bagues

University of Warwick

Last week

- Difference-in-differences (DiD)
 - Example: Death penalty
 - Example: Card and Krueger (1994)
 - Improving traditional DiD set up

This week

- Introduction to Regression Discontinuity Design
- Example: Bagus and Campa (2020)

Regression discontinuity design

Rules create experiments

- Often people use rules to assign individuals to “treatments” which can be exploited for estimating causal effects
- The most notorious case are **threshold rules** that are based on some ex-ante variable, typically, correlated with the expected effectiveness of the treatment
- Example: Need-based Grant Programs for low income students (**Fack and Grenet 2015**)
 - there are discontinuities in the grant eligibility formula
- This ex-ante variable is called the **running variable** (or also, forcing or assignment variable).
- This threshold divides individuals into “treated” and “not treated”. The idea in RDD design is to exploit the **randomness of this threshold**: we compare individuals just below (non-treated) and just above (treated).

Main idea

- Two conditions for **consistent** and **precise** RDD estimates:
 - ① Variation in the treatment status near the threshold is **as good as random**
 - If agents can anticipate the threshold they may try to manipulate the running variable. Manipulation may be a threat to the validity of RDD if it is precise enough
 - Moreover, there should not be any other treatments at the same threshold
 - ② Sufficient **number of observations** around the threshold

Regression discontinuity design

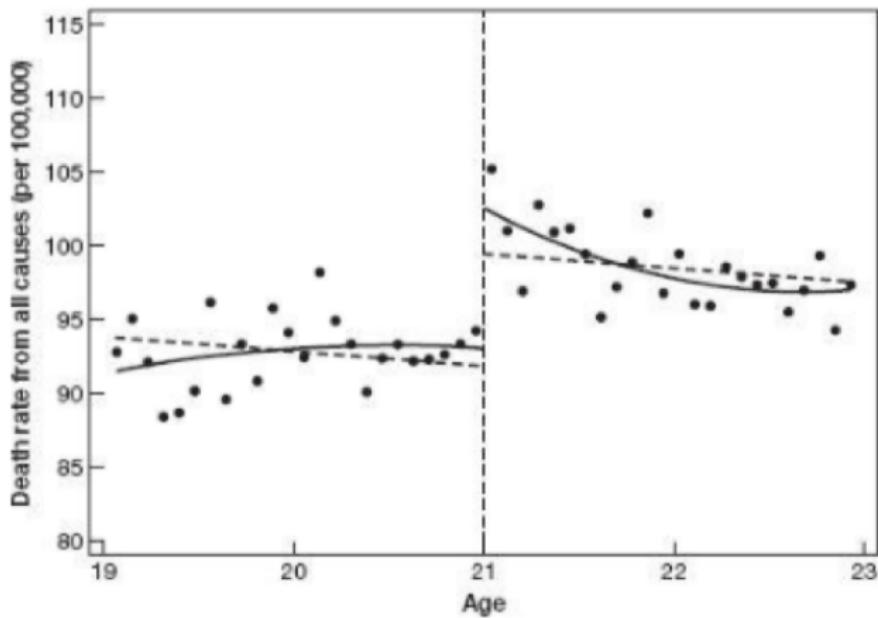
- Examples:
 - Effect of being able to drink alcohol legally on death rates

Effect of the Minimum Legal Drinking Age (MLDA) on death rates

Carpenter and Dobkin (2009)

- ① outcome variable y_i : death rate
- ② treatment D_i : legal drinking status
- ③ running variable x_i : age
- ④ cutoff: MLDA transforms 21-year-olds from underage minors to legal alcohol consumers.

RDD estimation: Visual example



Note: we need an out-of-range prediction to get a counterfactual!

- Other examples: (Treatment / Outcome variable / Threshold)
 - Attending a certain university degree / Future labor outcomes / Score in entry exams
 - Length of maternity leave / Children cognitive ability / date of policy change
 - R&D public subsidies / firms R&D investments / threshold in project quality score
 - Survivor Benefit Program / Widow(er)s labor supply / date of policy change
 - Tuberculosis vaccine / Immunity against covid / date of policy change

Main choices

- An RD paper typically starts showing the RD plot (as the ones you saw earlier)
 - Stata command: *rdplot*
- The RD plot provides merely suggestive evidence
- The main estimates are then reported in a table.
- Several choices need to be made:
 - ① How to control for the running variable
 - ② Bandwidth
 - ③ Kernel

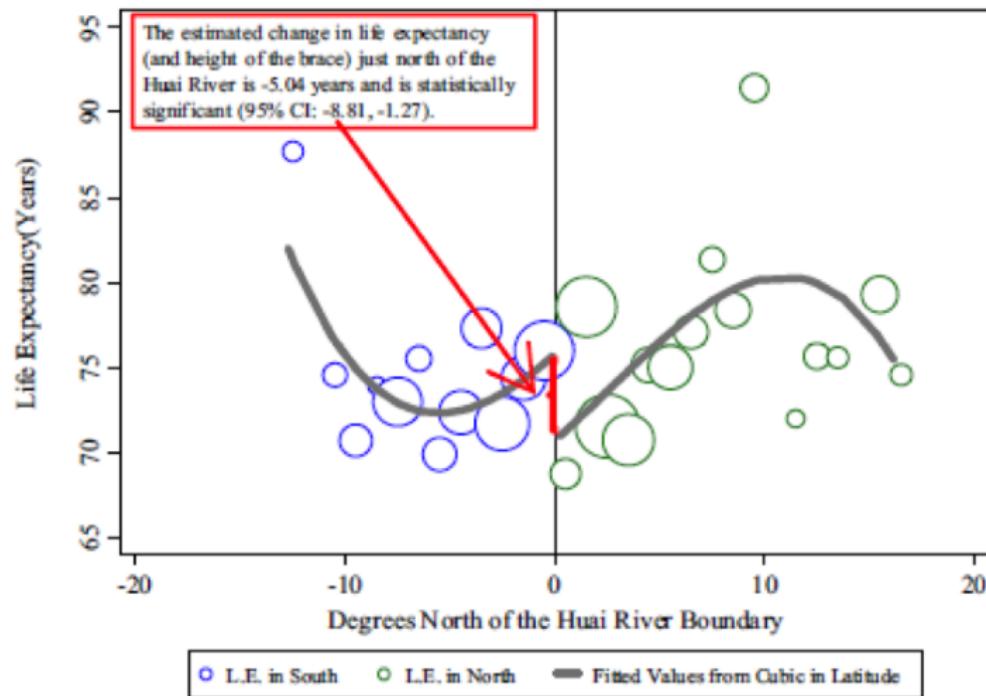
Control for the running variable

- Traditionally:
 - Global polynomials of higher order
- Gelman and Imbens (2014) critique:
 - ‘Why high order polynomials should not be used in regression discontinuity designs’
 - 1st or 2nd order
- State of the art: Local linear regression within a given window (bandwidth) of width h around the cutoff point

Control for the running variable

- Example :
 - 'Yuyu Chen, Avraham Ebenstein, Michael Greenstone and Hongbin Li (PNAS 2013) 'Evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai river policy'

Regression Discontinuity Design



Regression Discontinuity Design

Table S9

Robustness checks of choice of functional form for latitude

	Linear & Controls	Quadratic & Controls	Cubic & Controls	Quartic & Controls	Quintic & Controls
	(1)	(2)	(3)	(4)	(5)
Panel 1: Impact of "North" on the Listed Variable, Ordinary Least Squares					
Life Expectancy (years)	-1.62 (1.66) [0.101] {757.1}	-1.29 (1.68) [0.6] {758}	-5.52** (2.39) [0.001] {746.8}	-5.67** (2.36) [0.737] {748.2}	-5.43* (2.94) [0.984] {750.2}

Bandwidth

- How to choose the bandwidth?
- Trade-off: the closer you get the better it is for identification, but the less data you have...
- Two steps to follow:
 - ① Use existent statistical algorithms for selecting the “optimal bandwidth” (e.g.: Imbens-Kalyanaraman 2011, Calonico, Cattaneo and Titiunik (2014)).
 - ② Explore the robustness of results to different bandwidths
- In Stata, you may install `rdrobust` package to estimate RDD.
 - *net install rdrobust,
from(<https://sites.google.com/site/rdpackages/rdrobust/stata>)
replace*

Choice of kernel

- Choice of kernel
 - rectangular vs. triangular

Testing the validity of RDD set up

- RD Design can be invalid if individuals can precisely manipulate the assignment variable x_i in order to get (or to avoid) the treatment.
- Testing for the validity:
 - ➊ Predetermined characteristics should have the same distribution just above and just below the cut off
 - ➋ Density of the running variable should be continuous
 - McCrary test (`DCdensity`)
 - Cattaneo, Jansson and Ma (2017b): `rddensity`

Sharp vs. Fuzzy Regression Discontinuity Design

- So far we have assumed that treatment status (D_i) is deterministic and discontinuous function of the running (assignment, forcing) variable (x_i):
 - $D_i = 1$ if $x_i > c$
 - $D_i = 0$ if $x_i < c$
- In this case, we have a **sharp RDD**.
- When there is not full compliance we have a **fuzzy RDD**.
 - the increase in the probability of receiving the treatment increases by less than one when you cross the cutoff
 - the estimation is similar to the Wald estimator

$$\beta_{fuzzyRDD} = \frac{\beta_{RDD}}{\beta_{firststage}} \quad (1)$$

Summary

- Useful method to analyze the impact of treatment when the assignment varies discontinuously due to some rules! (test score, electoral results, income threshold, etc.)
- Only feasible when the number of observations around the threshold is sufficiently large
- Condition for consistency: no precise **manipulation** at the threshold
 - This may not hold if agents anticipate the threshold
 - Make sure there are no other treatments at the same threshold
- Usually graphical analysis is already very convincing
- Note that RDD provides information about the impact of the treatment only for individuals around the threshold.
 - Sometimes this might be different from the effect for other individuals (e.g.: entry to university)

EC902/EC907: Econometrics A

Lecture 17

Manuel Bagues

Warwick University

December 5, 2022
Lecture Slides

Road map

- Problem sets
 - This week's tutorial: Problem set 8 - based on lecture 16 (RDD)
- Extra problem set on staggered DID (not for exam!)
- This week we finish the term
 - EC907: You are done!
 - EC902: Next term with Subham Kailthya
- Student feedback: any comments and suggestions on how to improve are very welcome!
- Any questions?

Roadmap so far:

- RCT
- Identification based on observables
- IV
- DID
- This week: RDD

Previous lecture: Regression Discontinuity design

- Main requirements
 - No (perfect) manipulation
 - Enough observations around the threshold
- Local treatment effects: impact of the treatment around the threshold
- Some technical issues:
 - Bandwidth
 - Controlling for the running variable
 - Kernel

Previous lecture: Regression Discontinuity design

- Standard tests of the validity:

Predetermined covariates do not exhibit a sharp discontinuity
Density function is smooth around the threshold (e.g. McCrary test)

- Standard robustness checks:

Different bandwidths
Different functional forms

Today

- Presentation of a research paper using RDD:

Bagues and Campa (2021)

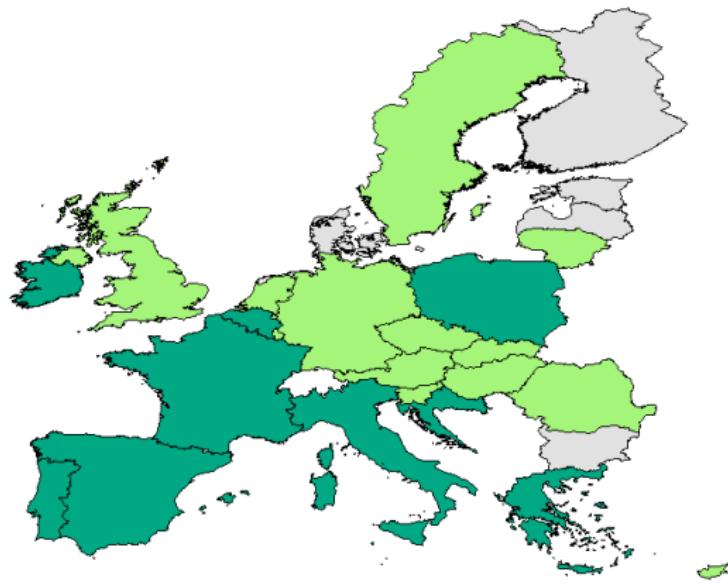
Do Gender Quotas in Candidate Lists Empower Women?

Introduction

- Few women in politics
 - 19% of Lower House and 22% of Senate members in US
 - 32% of UK House of Commons (historical record), 22% of Cabinet
 - 27% of MPs in the European Union
 - 25% of mayors in Spain (focus of this study)

Introduction

- Many countries worldwide have adopted *quotas* to close the gender gap in political institutions:
 - a. mandated representation
 - b. candidate gender quotas
 - open lists (e.g. double preference voting conditioned on gender)
 - closed lists



■ Legislated quotas ■ Voluntary quotas ■ No quota

Source: www.quotaproject.org (IDEA, Inter-Parliamentary Union and Stockholm University)

In the UK

- Voluntary party quotas
- Lib. Democrats:
 - 40 percent target of women candidates in 2001
 - Prior to the 2005 elections, women in 40 percent of the “winnable seats”
 - ‘Zipping’ system on their candidate lists for the European election in 1999
- Labor Party
 - “All Women Shortlists” select candidates in half of all winnable seats

In this paper:

- Evaluation of **legislated candidate gender quotas** in a closed list system
- Evidence from the introduction of quotas in **local elections in Spain**:

At least **40%** of candidates of each gender

It applies both to the whole **party list** and to **every five posts**

Since 2007 in municipalities above 5,000 inhabitants.

Extended in 2011 to municipalities above 3,000 inhabitants

Main features

- **Comprehensive evaluation**
 - Presence of women in the ballot
 - Voting behavior
 - Women elected and access to leadership positions
 - Characteristics of politicians
 - Local public finance
- **Short- and medium-term impact:** three electoral cycles (2007-2015).
- We rely on a **RDD**:
 - it requires weaker assumptions than other methods (e.g. DID)...
 - ... but it comes at the cost of **lower accuracy** and **local treatment effects**
- **Small municipalities:**
 - Typically excluded from quotas (e.g. France, Italy, Spain)
 - ↓ education, ↑ age, ↓ FLFP
 - ↓ concern gender discrimination

1 Institutional Background

- Electoral system and gender quotas
- Local Government

2 Data

- Sample
- Electoral data
- Local Public Finance Data
- Voters' preferences

3 Empirical Analysis

- Threats to validity
- Results
 - Candidates
 - Women's leadership
 - Voting behavior
 - Women elected
 - Characteristics of politicians

4 Conclusion

Electoral System

- Local elections in Spain:
Proportional representation with closed lists

The Ballot

ELECCIONES LOCALES 2011

MUNICIPIO DE SAHAGÚN



UNIÓN DEL PUEBLO LEONÉS
(UPL)

CANDIDATURA AL AYTO. DE SAHAGÚN

- 1.- Virgilio Buiza Diez
- 2.- Rosa María Quintanilla González
- 3.- Domingo Vallejo González
- 4.- José Mauricio Mencía Blanco
- 5.- María Yolanda Mayo Collantes
- 6.- Paulino Estébanez Espeso
- 7.- Soraya Ruiz García
- 8.- Luis Alberto Contreras Yanes
- 9.- Amparo Contreras González
- 10.- María Victoria Martínez Lazo
- 11.- Esteban Martínez Bartolomé

The Ballot

ELECCIONES LOCALES 2011

MUNICIPIO DE SAHAGÚN



UNIÓN DEL PUEBLO LEONÉS
(UPL)

CANDIDATURA AL AYTO. DE SAHAGÚN

- M** 1.- Virgilio Buiza Diez
- F** 2.- Rosa María Quintanilla González
- M** 3.- Domingo Vallejo González
- M** 4.- José Mauricio Mencía Blanco
- F** 5.- María Yolanda Mayo Collantes
- M** 6.- Paulino Estébanez Espeso
- F** 7.- Soraya Ruiz García
- M** 8.- Luis Alberto Contreras Yanes
- F** 9.- Amparo Contreras González
- F** 10.- María Victoria Martínez Lazo
- M** 11.- Esteban Martínez Bartolomé

Local government

- Around 15% of public expenditure

Data: sample

- Overall in Spain there are 8000+ municipalities
- We focus on municipalities with more than 250 inhabitants and less than 10,000
 - 5,000 municipalities
 - 20% of the Spanish population

Electoral data

- Elections: 2003, 2007, 2011, 2015
 - Candidate lists: gender, experience
 - Votes
 - Council composition: gender, experience, education, occupation

Table: Electoral data

Election year:	(1) 2003	(2) 2007	(3) 2011	(4) 2015
A. Candidate lists				
Number of parties	3.1	3.2	3.1	3.1
Lists with at least 40% of candidates of either gender	26%	43%	57%	62%
<i>Share of women:</i>				
all candidates	29%	35%	38%	40%
upper positions candidates	28%	33%	35%	38%
bottom positions candidates	32%	38%	42%	44%
party leaders	17%	19%	22%	25%
<i>Experience:</i>				
all candidates	32%	41%	40%	40%
female candidates	24%	32%	34%	35%
male candidates	35%	46%	44%	43%
B. Electoral results				
Turnout	78%	76%	78%	75%
<i>Vote share:</i>				

	Institutional Background Data	Sample Electoral data	Local Public Finance Data	Voters' preferences
	Empirical Analysis	Conclusion		
		(1)	(2)	(3)
C. Local council				(4)
Parties in the council	2.6	2.6	2.6	2.6
<i>Share of women:</i>				
among councilors	25%	29%	32%	35%
among mayors	13%	15%	17%	20%
<i>Experience:</i>				
all councilors	36%	48%	46%	46%
male councilors	39%	52%	50%	50%
female councilors	27%	38%	39%	39%
<i>Years of education:</i>				
all councilors	11	11.3	11.7	12
male councilors	10.7	11.1	11.4	11.7
female councilors	11.9	12.1	12.5	12.8
<i>Age:</i>				
all councilors	43	45	46	47
male councilors	44	46	47	48
female councilors	39	41	43	44
Sample size				
Number of party lists	14,930	15,230	14,773	14,161

Local Public Finance Data

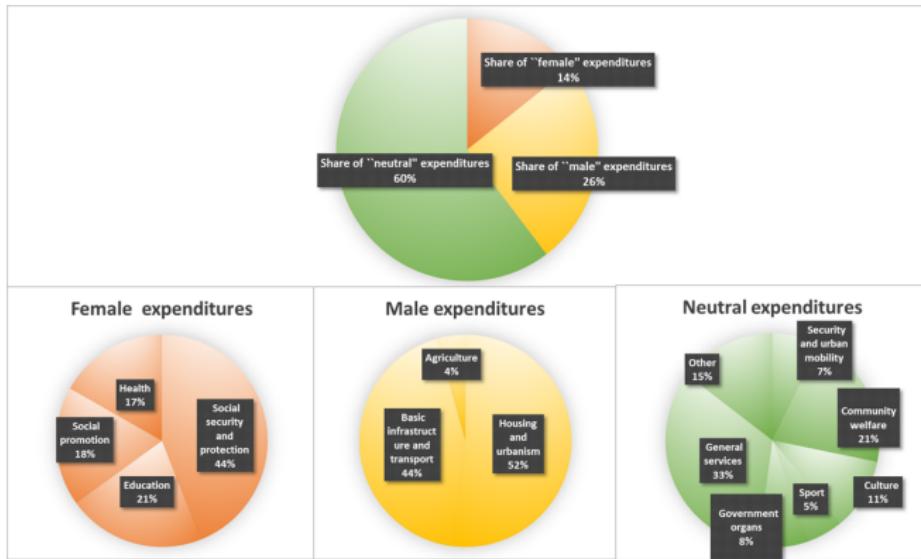
- Local Public Finance Data: 2004 - 2014
 - Total expenditures and revenue
 - Subgroups of expenditures by function

Data: Voters' preferences

- Data on male and female voters' preferences: 2000-2006
 - 57,000 individuals
 - three problems that affect you the most.*

	(1) Women	(2) Men	(3) Difference	(4) Women	(5) Men	(6) Difference
	<i>Full sample</i>			<i>Less than 10,000 inhabitants</i>		
Unemployment	0.30	0.28	0.02***	0.28	0.25	0.03***
Pensions	0.08	0.06	0.02***	0.10	0.07	0.02***
Education	0.06	0.05	0.02***	0.05	0.03	0.02***
Health system	0.07	0.05	0.01***	0.07	0.06	0.01**
Drugs	0.04	0.03	0.01***	0.04	0.03	0.01***
Youth problems	0.02	0.01	0.01***	0.02	0.01	0.01***
Violence against women	0.01	0.01	0.01***	0.01	0.00	0.01***
Women's issues	0.01	0.00	0.01***	0.01	0.00	0.01***
Social problems	0.03	0.02	0.01***	0.02	0.02	0.01**
War	0.01	0.00	0.00***	0.01	0.00	0.00**
Crisis of values	0.02	0.01	0.00***	0.01	0.01	0.00*
Terrorism	0.12	0.12	-0.00	0.11	0.10	0.01
Public services	0.01	0.01	0.00	0.01	0.01	0.00
Racism	0.00	0.00	0.00	0.00	0.00	0.00
Crime	0.12	0.12	-0.00	0.09	0.09	-0.00
Agriculture, hunting, and fishing	0.01	0.01	-0.00***	0.02	0.03	-0.01***
Judiciary system	0.01	0.01	-0.00***	0.01	0.01	-0.00**
Environmental degradation	0.01	0.02	-0.00***	0.01	0.02	-0.01**
Economic problems	0.16	0.17	-0.01***	0.17	0.18	-0.01
Infrastructure	0.02	0.03	-0.01***	0.02	0.02	-0.00*
Corruption	0.01	0.01	-0.01***	0.01	0.02	-0.01***
Politics	0.02	0.03	-0.01***	0.01	0.03	-0.01***
Work conditions	0.05	0.06	-0.01***	0.03	0.05	-0.01***
Immigration	0.06	0.08	-0.01***	0.05	0.07	-0.02***
Housing	0.12	0.14	-0.02***	0.09	0.10	-0.01***

Figure: Municipal expenditure, Years 2004 - 2009



- “Female expenditures”:
 - Social-security and protection
 - Education
 - Social promotion
 - Health
 - Employment services
 - Pensions
- In cross-sectional regressions, share of female councilors positively correlated with municipal share of female expenditures, and negatively correlated with share of male expenditures

Regression Discontinuity Design

$$Y_{i,2007+k} = \beta_0 + \beta_1 I[\text{population}_{i,2006} > 5000] + \beta_2 f(\text{population}_{i,2006}) + \epsilon_i \quad (1a)$$

$$Y_{i,2011+k} = \gamma_0 + \gamma_1 I[\text{population}_{i,2010} > 3000] + \gamma_2 g(\text{population}_{i,2010}) + \epsilon_i \quad (1b)$$

- i = list or municipality
- $I[\cdot]$ takes value one if population is above corresponding threshold
- $k = 0$: impact of quotas the first election after their introduction.
- Long-term impacts of quota:
 - equation (1a), 2011 election ($k = 4$, second versus first round of implementation of the quota)
 - equation (1a), 2015 election ($k = 8$, third versus second round of implementation of the quota)

Threats to validity

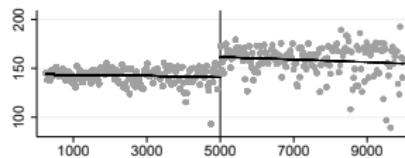
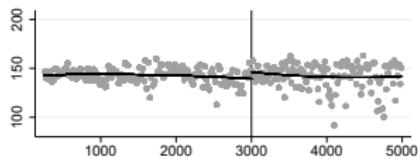
① Other policies at the same threshold:

3,000 inhabitants threshold: Only relevant for quotas during this period

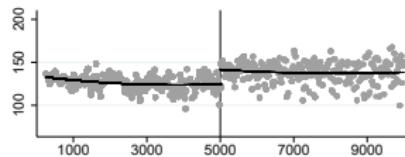
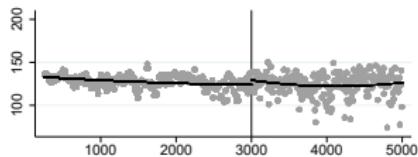
5,000 inhabitants threshold: relevant for (i) funding, (ii) competences and (iii) council size

To alleviate concerns we consider the **outcome variable in differences**

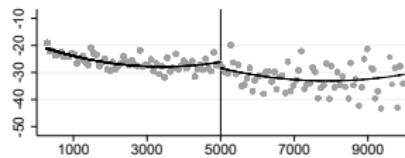
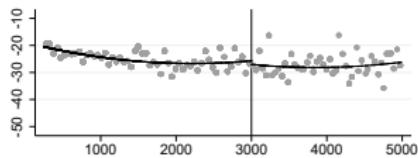
Panel A: Before-Quota



Panel B: After-Quota



Panel C: After-Quota minus Before-Quota



Threats to validity

① Other policies at the same threshold:

3,000 inhabitants threshold: Only relevant for quotas during this period

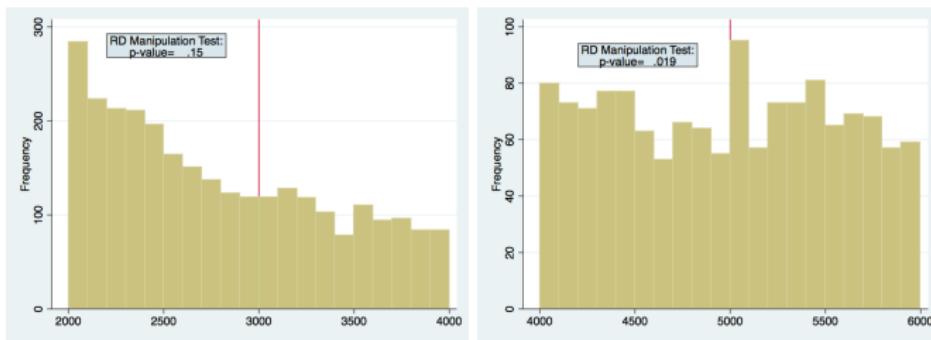
5,000 inhabitants threshold: relevant for (i) funding, (ii) competences and (iii) council size

② Manipulation:

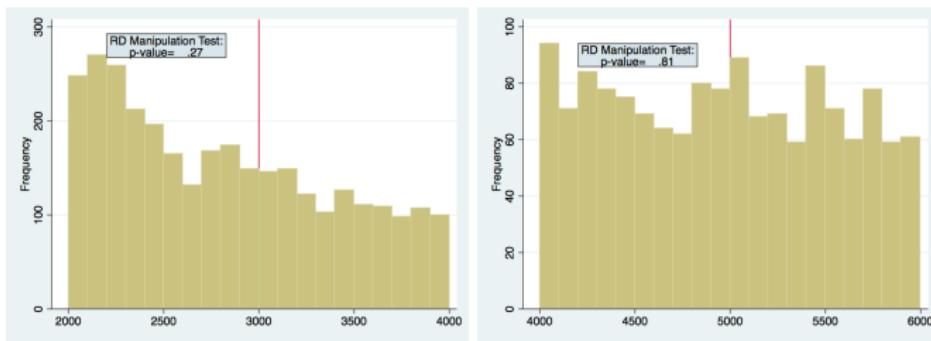
5,000 inhabitants threshold: density function of the forcing variable discontinuous before 2007, no manipulation afterwards

3,000 inhabitants threshold: density function continuous

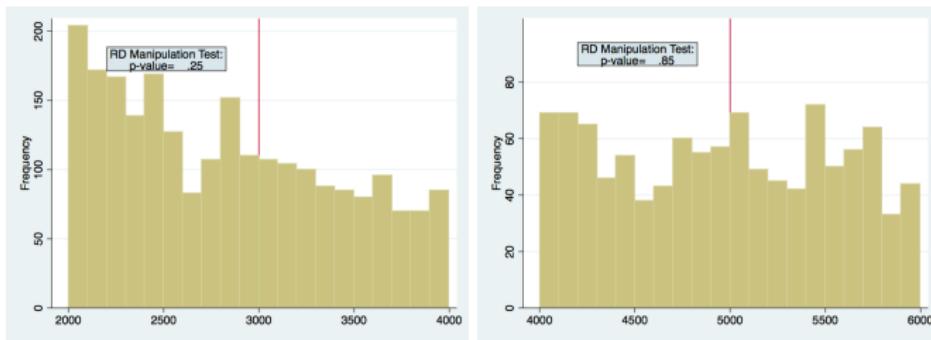
Histograms of population, years 2003-2006



Histograms of population, years 2007-2010



Histograms of population, years 2011-2013



Our take away from this:

- Causal effect of gender quotas identified at 3,000 cutoff
- 5,000 cutoff might be more problematic
 - To alleviate concerns we consider the outcome variable in differences
- Results are generally similar at both thresholds
- When we analyze short-term outcomes we also pool the two thresholds together

Results

- We analyze the impact of candidate quotas on:
 - ❶ Characteristics of candidates
 - ❷ Voting behavior
 - ❸ Presence of women in the local council
 - ❹ Public policies

① Female candidates

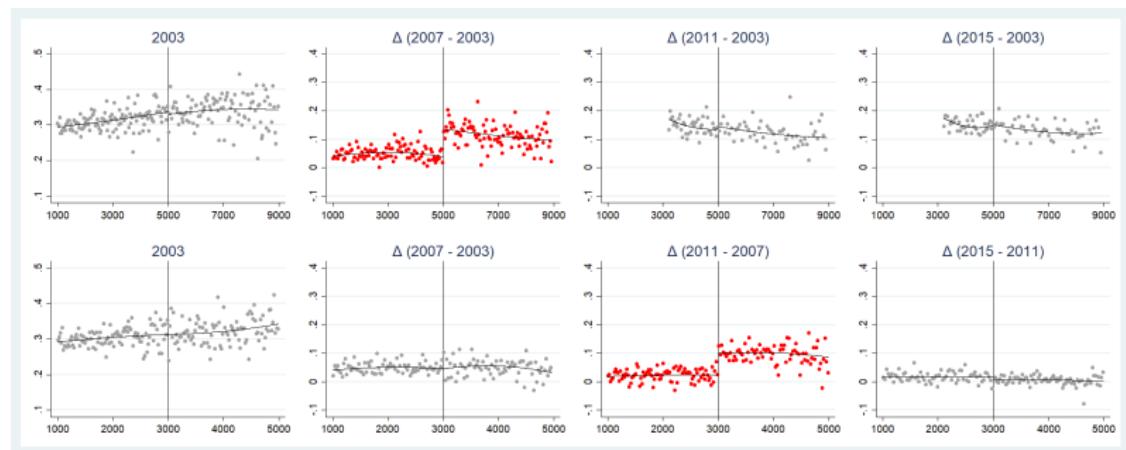
- Theory:

- Requirements of the quota

- Cost of not complying

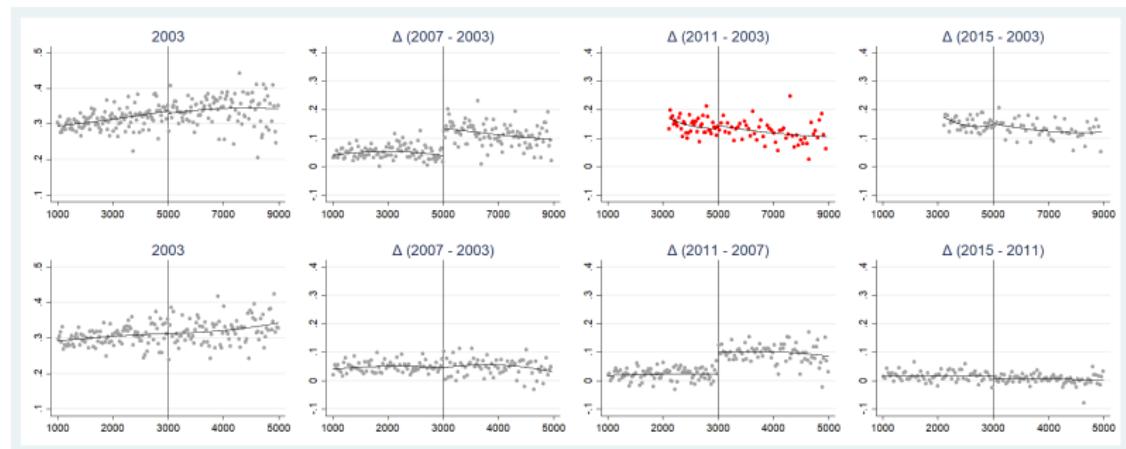
Share of female candidates

Short term impact: 0.08 (s.e. 0.01)



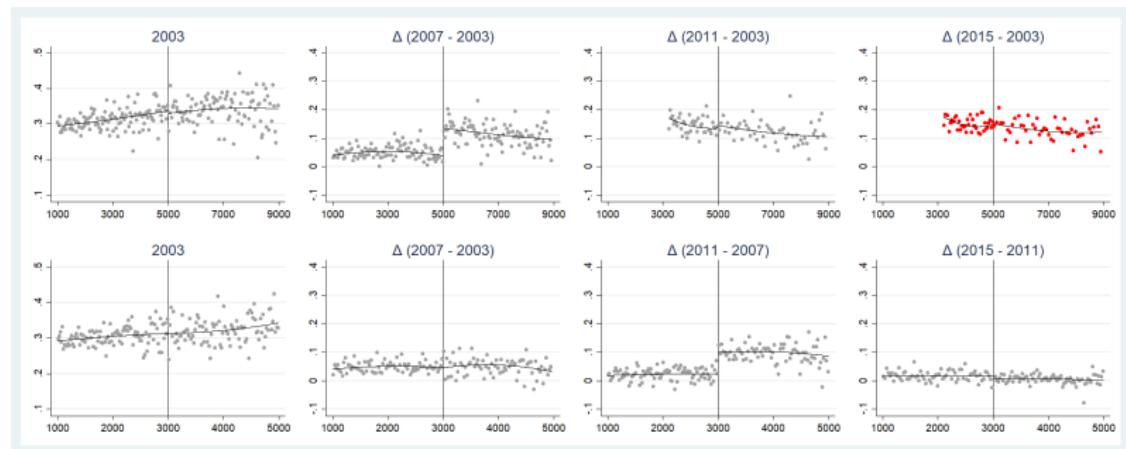
Share of female candidates

2nd term vs. 1st term: 0.00 (s.e. 0.01)



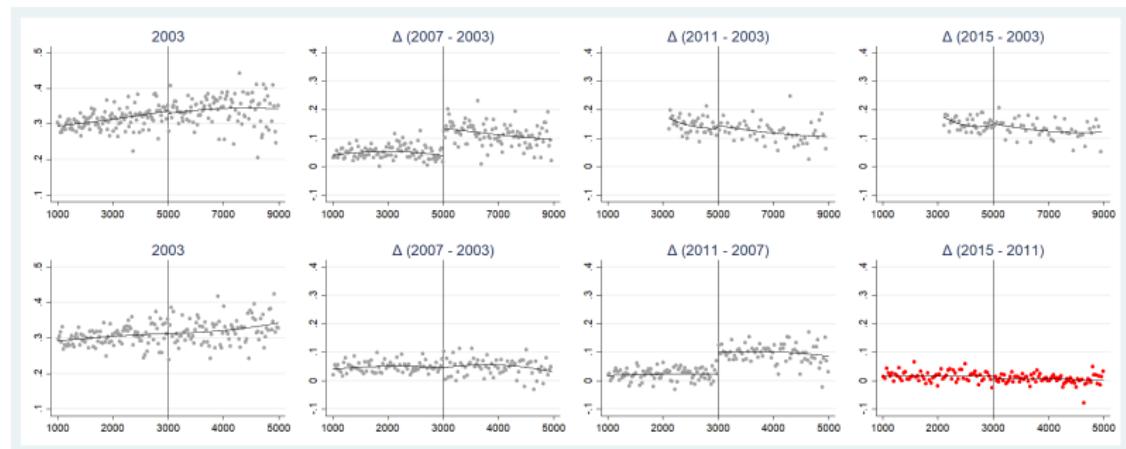
Share of female candidates

3rd term vs. 2nd term: 0.00 (s.e. 0.01)



Share of female candidates

2nd term vs. no quota: 0.01 (s.e. 0.02)



① Women's access to leadership positions

- Theory:

- Quotas may break down negative stereotypes regarding female politicians (Beaman et al. 2009)

- Creation of female-friendly political networks

- On the flip side: backlash, creation/strengthening of negative stereotypes if pool of female candidates is limited

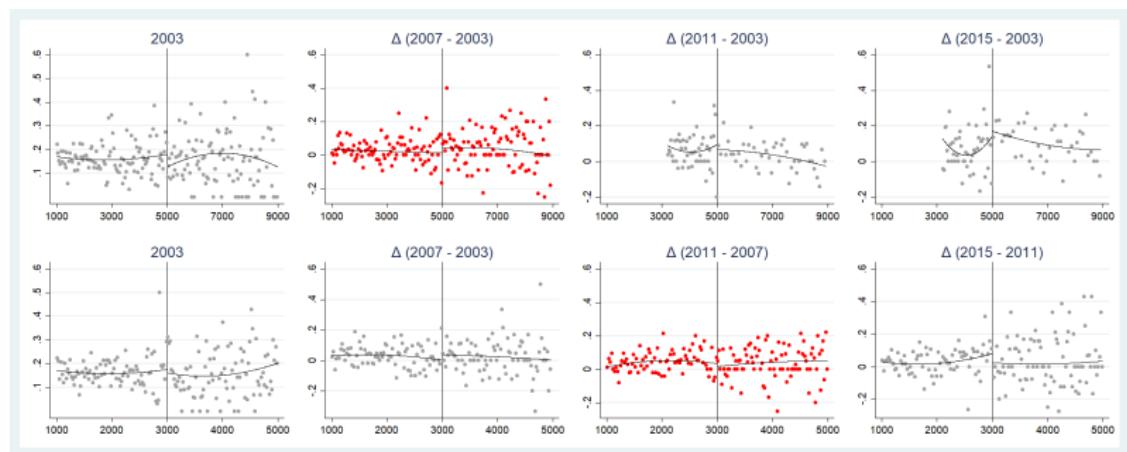
- Previous evidence:

- Sweden 1991-1994 (O'Brien and Rickne 2016):

- $\uparrow 10$ p.p. women elected $\rightarrow \uparrow 5$ p.p. female leaders (baseline $\approx 20\%$)

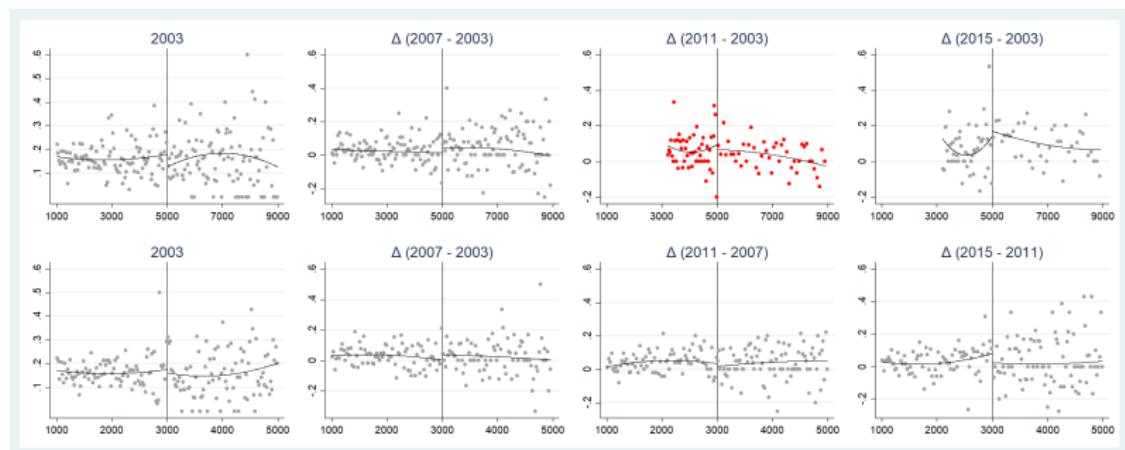
Women at the top of the ticket

Short term impact: 0.02 (0.03)



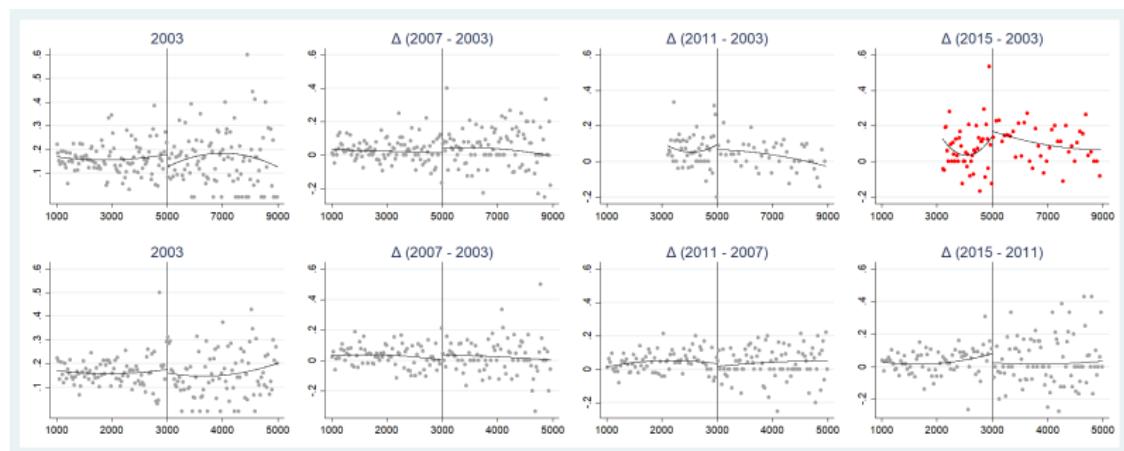
Women at the top of the ticket

2nd term vs. 1st term: 0.01 (0.08)



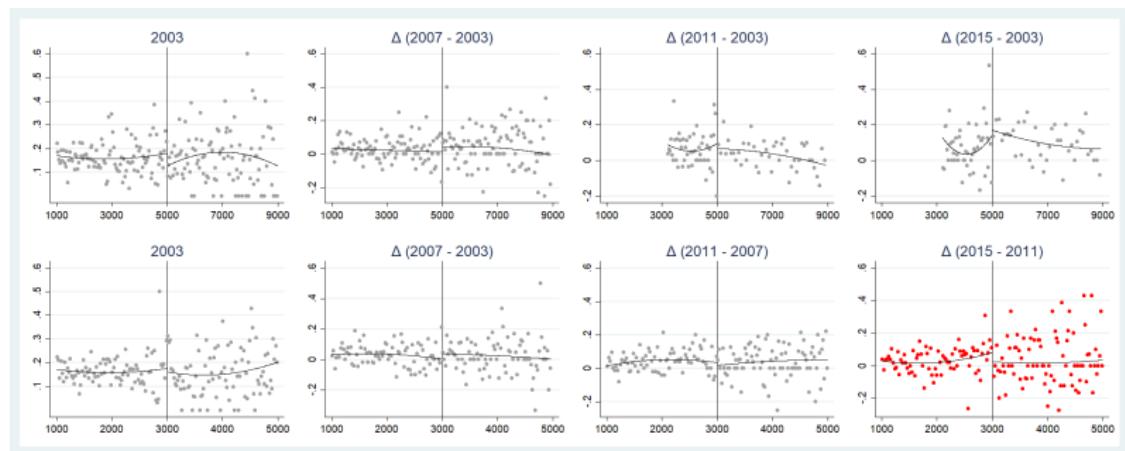
Women at the top of the ticket

3rd term vs. 2nd term: 0.04 (0.08)



Women at the top of the ticket

2nd term versus no quota: -0.06 (0.06)



Summary: presence of women in candidates' lists

- Share of women in lists

Immediate impact on the share of women in the list ($\uparrow 8$ p.p.)

No additional impact in the two following elections (precise 0)

- Women in 'winning' positions (not shown)

Strategic positioning: increase more modest than in the share of female candidates

Top positions: $\uparrow 2$ p.p. (st. error=1)

Bottom: $\uparrow 12$ p.p. (st. error=2)

- Women on top of the list:

Short term: $\uparrow 2.4$ p.p. (st. error=2.6) ; baseline $\approx 21\%$

Mid term: -6 p.p. (st. error=6)

- Candidates' experience (not shown)

Negative short-term effect (candidates 4 p.p. less likely to have run in previous elections)

① Voting behavior

- Theory:

If the lack of women is demand driven → parties that are most affected by the quota may lose votes

If the lack of women is supply driven → parties that are most affected by the quota may lose votes

If political parties discriminate against women → quotas may improve the selection of candidates

- Previous evidence:

2007 Spanish local elections (Casas-Arce and Saiz JPE 2015)

Lists forced to include women attract more votes

Δ 40 p.p. female candidates ⇒ ↑ 6.6 p.p. votes (54% of a st. dev.)

Impact on voting behavior

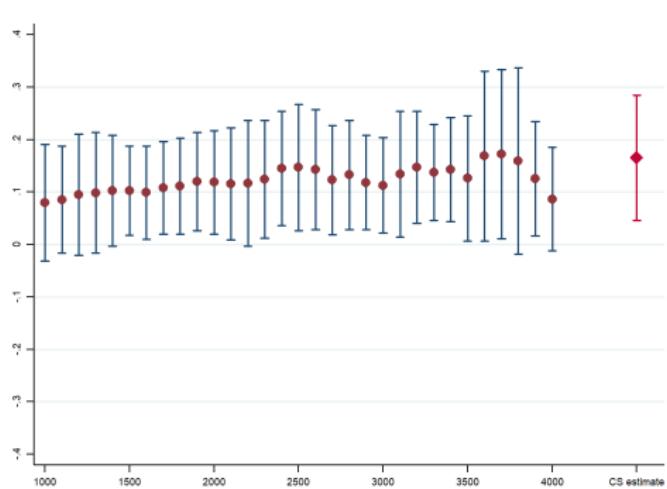
- ① Impact of quotas on voters turnout X

Impact on voting behavior

- ① Impact of quotas on voters turnout X
- ② Do voters favor parties that are forced by the quota to increase their share of female candidates?

Testing identifying assumption in Casas-Arce and Saiz (2015)

Placebos: sample of municipalities with less than 5,000 (from Bagues and Campa, 2017b)



Impact on voting behavior

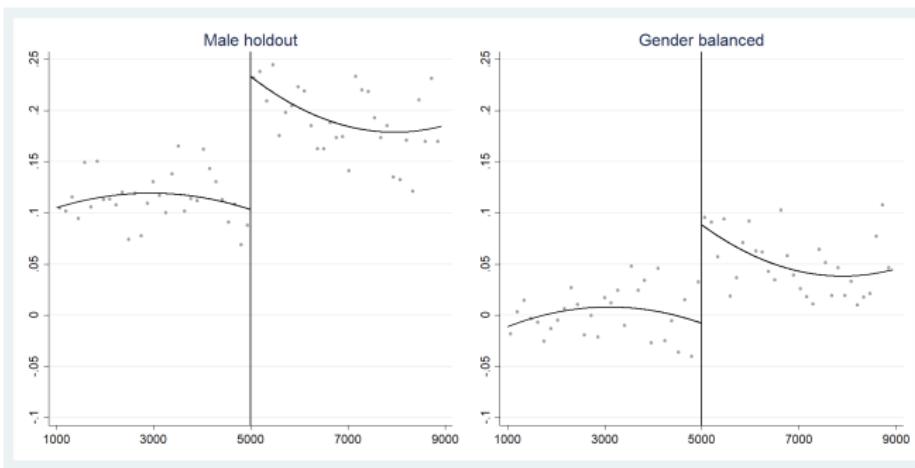
RDD analysis

- We identify two lists with the largest share of votes in their municipality
- We drop municipalities where both lists have the same share of female candidates
- Divide lists in two groups:

male holdouts: 18% of candidates are women

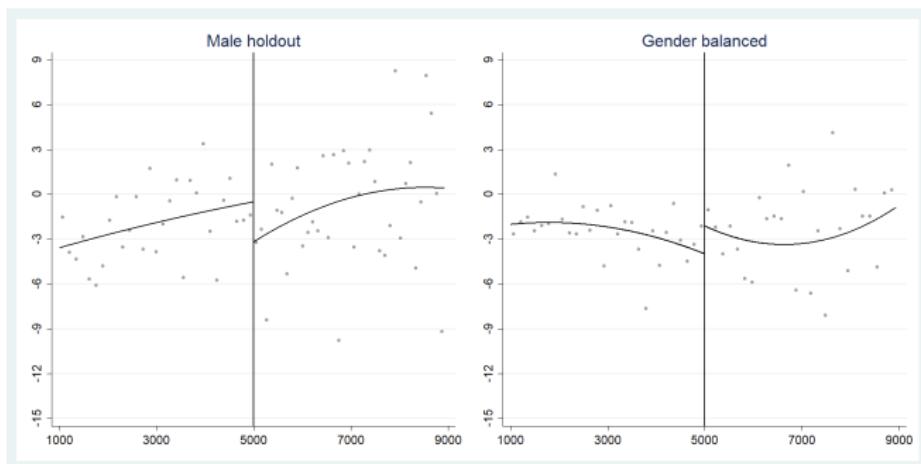
rival lists: 38% of candidates are women

△ Share of female candidates - Year 2007 - 5,000 threshold

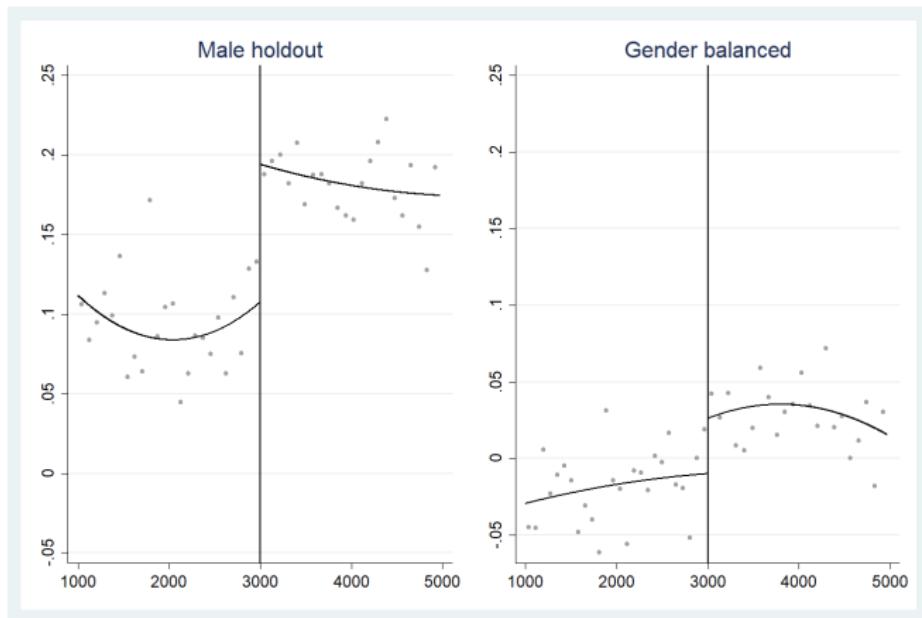


Δ Votes (%) - Year 2007 - 5,000 threshold

Male hold-outs in large municipalities: 0.00 (2.11) Δ Vote share



△ Share of female candidates - Year 2011 - 3,000 threshold



Δ Votes (%) - Year 2011 - 3,000 threshold

Male hold-outs in large municipalities: -5.18 (2.79) Δ Vote share



- RD analysis pooling the two thresholds:

Due to the quota, male holdout lists increase their share of female candidates by 4 p.p. more than their rival list

No evidence that the quota increased electoral support for male-holdouts versus their rival list (-4 p.p., 95% C.I. btw -9.6 p.p. and 1.2 p.p.)

So far:

① Candidate lists:

- female candidates: short term effect ✓
- female candidates: medium term effect X
- women at the top of the ticket X
- experience ✓

② Voters' behavior X

③ Local Council ?

- gender and educational attainment of councilors
- gender of mayors

① Number of women elected

- Theory:

- Position of women in lists

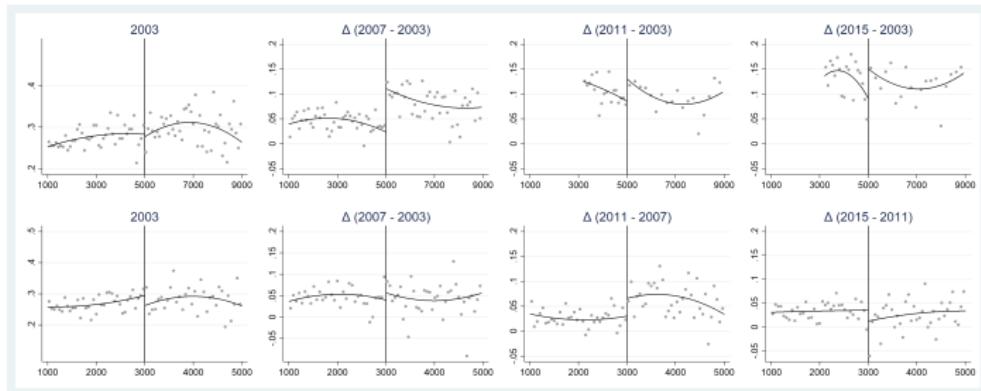
- Impact of quotas on voting behavior

- Previous evidence:

- ↑ women elected, but the magnitude depends on how easy it is to game the regulation (e.g.: Baltrunaite et al. 2016, Dahlerup and Freidenvall 2013, Esteve-Volart and Bagues 2012, Jones 2008, Matland 2006)

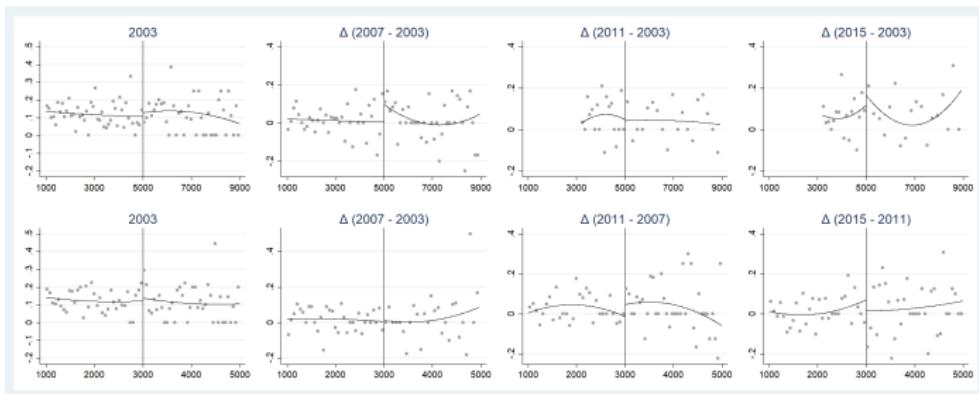
Share of women elected

Short term: 4 p.p. (s.e. 1); No additional impact in the following elections



Female mayors

Short term: 0.10 (0.03), not robust; No additional impact in the following elections



① Characteristics of politicians

• Theory:

If the lack of women is supply driven → quotas may also lead to a decrease in the quality of candidates

If political parties discriminate against women → quotas may improve the selection of candidates

• Previous evidence:

New council members tend to have a better educational and professional background

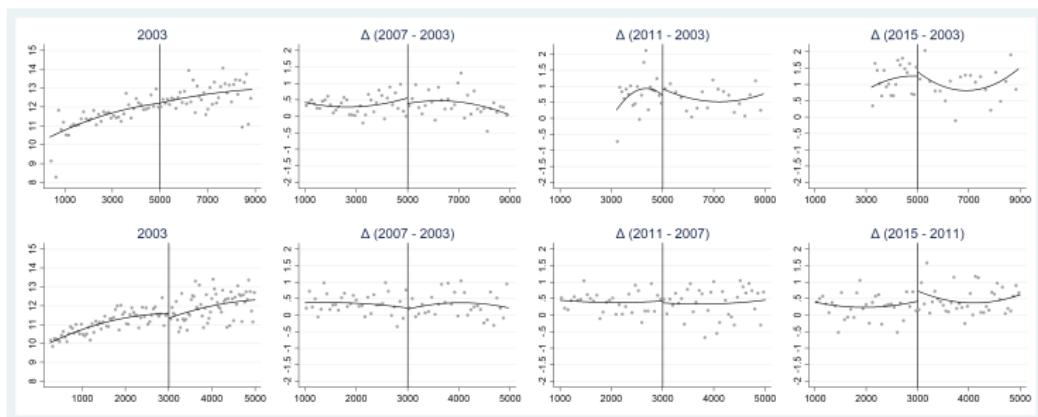
Sweden (Besley et al. 2017): ↑10 p.p. women elected → ↑ 3 p.p. competent politicians (baseline = 50%)

Italy (Baltrunaite et al. 2014): ↑4.7 p.p. women elected ↑ 0.12-0.24 years of education

Councilors' education

Short term impact: -0.05 (0.22)

Medium term impact: 0.46 (0.49) at 5,000; 0.59 (0.45) at 3,000



① Local public finance

- Theory:

Standard citizen candidate models (Osborne and Slivinski 1996,
Besley and Coate 1997)

Median voter (Downs 1957)

- Previous evidence:

Evidence from mandated representation (Chattopadhyay and
Duflo 2004)

Candidate gender quotas in Italy: ↑ capital account expenditures
in education and environment (Baltrunaite et al., 2016)

Impact on local public finance

- Outcome variables:
Size of government **X**

Impact on local public finance

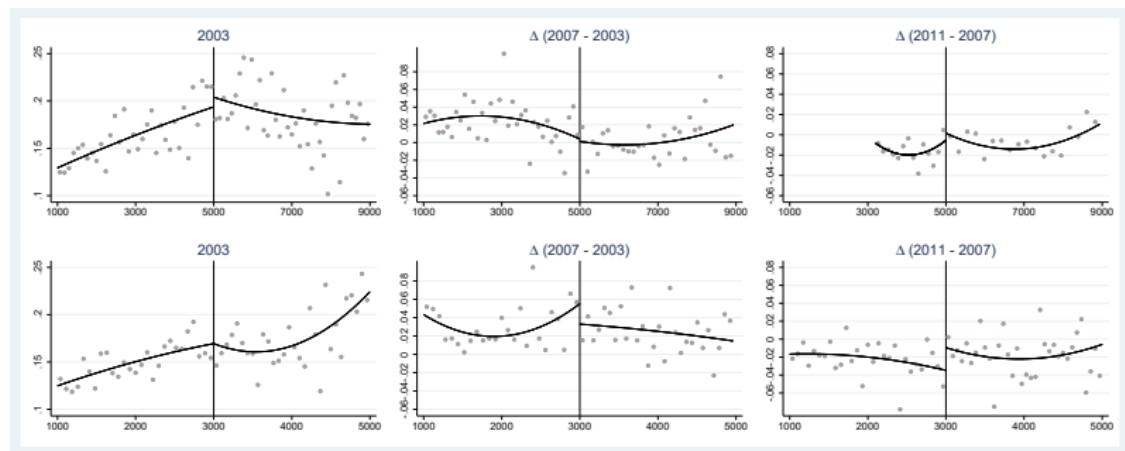
- Outcome variables:

Size of government **X**

Type of expenditures

Local budget

Composition of the budget: female expenditures up by 1.3 p.p. (s.e. 1)



Summary of results

① ↑ Number of women elected

Mechanic increase in the share of female candidates by 8 p.p.

Women tend to be placed on the bottom of the ballot

Modest increase in female councilors: 4 p.p.

Mid term (up to three elections): No additional impact

② ✗ Electoral support for ‘male hold-outs’ (unlike in Casas-Arce and Saiz 2015)

③ ✗ Women’s access to leadership positions

④ ✗ Expenditure preferred by women

No statistically significant effect.

Upper bound: up to a 3 p.p. increase in ‘female’ expenditures

⑤ ✗ Observable quality of elected candidates

Conclusion

- In Spanish local elections, candidate quotas might fail to remove the barriers that prevent women from playing an influential role in politics.
- Limitations of our study:
 - ① Accuracy
 - ② Small impact on the composition of the local council (although comparable to previous studies)
 - ③ Short/mid term vs. long term
 - ④ Context
- More research is needed:
 - Information from a larger number of contexts
 - Better understanding of why quotas work

Thank you!

EC902: Term 2 Module Outline

Introduction to Time Series Econometrics

Subham Kailthya

University of Warwick

Spring 2023

Organisation – Term 2

- Lecturer: Subham Kailthya, [webpage]
- Email: Subham.Kailthya@warwick.ac.uk
- A&F Hours:
 - Monday, 3pm–4pm
 - Friday, 2pm–3pm
 - book online on my webpage
- Every week you will have:
 - Asynchronous material
 - Synchronous material
 - Seminar classes

Module Outline

1. **Univariate time series models:** Stationary ARMA processes, Box and Jenkins methodology, forecasting with ARMA models.
2. **Dynamic regression models with stationary variables:** Autoregressive distributed lag (ADL) models, VAR models
3. **Nonstationary time series:** Deterministic and stochastic trends; testing for nonstationarity.
4. Spurious regressions, **cointegration** and **error correction models.**
5. Analysis of **panel data**

Textbook (Alternatives)

- Wooldridge, J. M. (2009) Introductory Econometrics: A modern approach, 4th Edition.
- Gujarati D.N. and D.C. Porter, (2009). Basic Econometrics, 5th Edition
- **Verbeek, Marno. (2012) A Guide to Modern Econometrics, 4th Edition**
- Stock J.H. and M.W. Watson (2012) Introduction to Econometrics, 3rd Edition, Part Four: chapters 14-16.

References (1/3)

1. Univariate time series models: stationary series

- Gujarati and Porter, Ch. 21-22
- Wooldridge, Ch. 10 - 12.
- Verbeek, Ch.8
- Stock and Watson Ch. 14.1-14.3, 14.5

2. Dynamic regression models with stationary variables: ADL models and VAR models

- Gujarati and Porter, Ch. 17.1-17.3, 17.14 and Ch. 22.9 (VAR)
- Wooldridge, Ch. 10
- Verbeek, Ch.9
- Stock and Watson Ch. 14.4, Ch. 15, Ch. 16.1 (VAR)

References (2/3)

3. Non stationary time series: trends and testing for nonstationarity

- Gujarati and Porter, Ch. 21
- Wooldridge, Section 18.2
- Verbeek, Ch.8-9
- Brooks, Ch. 7, sections 7.1-7.2
- Stock and Watson Ch. 14.6

4. Spurious regressions, cointegration and error correction models

- Gujarati and Porter, Ch. 21, section 21.11
- Wooldridge, Sections 18.3-18.4
- Verbeek, Ch. 9
- Stock and Watson Ch. 14.6, 16.4
- Brooks, Ch. 7, sections 7.3-7.7.

5. Analysis of panel data

- Gujarati and Porter, Ch. 16
- Wooldridge, Ch. 13 and Ch. 14
- Verbeek Ch. 10
- Stock and Watson, Ch. 10
- Baltagi, B. H. (2013) Econometric Analysis of Panel Data

Assessment

EC902

Coursework (45%) + Examination (55%)

- Coursework details
 - Test 1 (4%)
 - Test 2 (6%)
 - Midterm (10%)
 - 3000 word project (25%)
- Examination (55%)
 - Summer

Introduction to Time Series

EC902: Econometrics A

Subham Kailthya

University of Warwick

Organisation

Topic 1 consists of the following video lectures:

1. **Introduction to time series data and univariate time series models**
2. Theoretical properties of ARMA processes
3. Empirical modelling of univariate time series: Box-Jenkins approach
4. Forecasting with ARMA models

Introduction to Time Series

Charateristics of Time Series

Time series data are characterized by:

- Temporal dependence
- Trends
- Seasonality
- Volatility clustering (typical of financial time series)

Time Series Concepts

Time series process

A sequence of random variables $\{Y_t\}_{t=1}^T$ indexed by time t :

$$\{Y_1, Y_2, \dots, Y_T\} = \{Y_t\}$$

This is also known as a **stochastic process** or a **random process**.

A time series process can be described by a T -dimensional probability distribution $F(Y_1, Y_2, \dots, Y_T)$ and characterized by:

- time ordering
- systematic correlation between elements in the sequence

Sample Path

If $\{Y_t\}$ is a stochastic process its **realization** or **sample path** is an assignment to each t of a possible value of Y_t :

$$\{Y_1 = y_1, Y_2 = y_2, \dots, Y_T = y_t\}$$

The observed time series $\{y_t\}$ is a particular realization of the stochastic process, the data we observe, the **sample**

Stochastic process – statistical model, data generating process, population

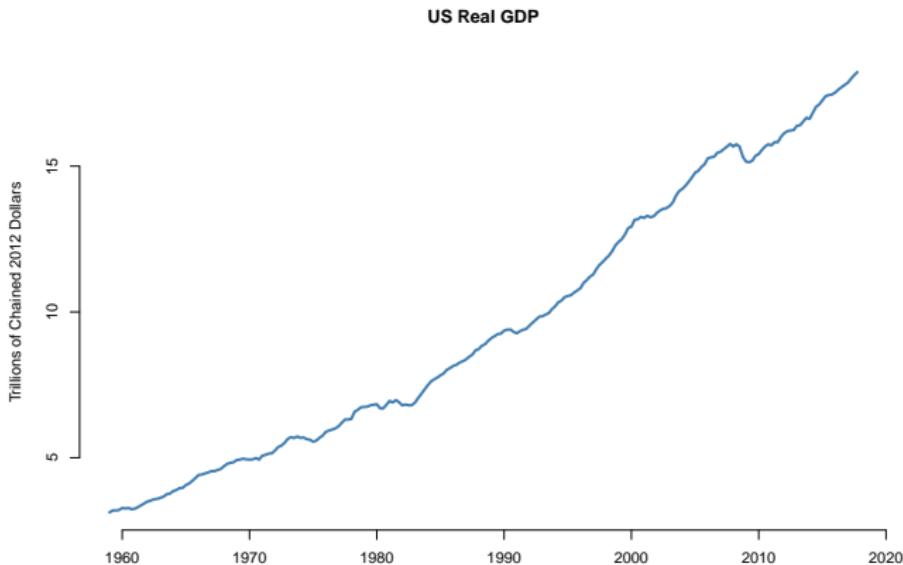
Sample size

Sample size for a time series: the number of time periods over which the variable is observed, T

Example: Suppose aggregate expenditures on consumer goods are observed for the years 1960 ($t = 1$) through to 2000 ($t = T$), the sample size is 41.

Time series data can be observed at different **frequencies** – yearly, quarterly, monthly, or at higher frequency e.g. weekly, daily, hourly (financial time series data)

Examples



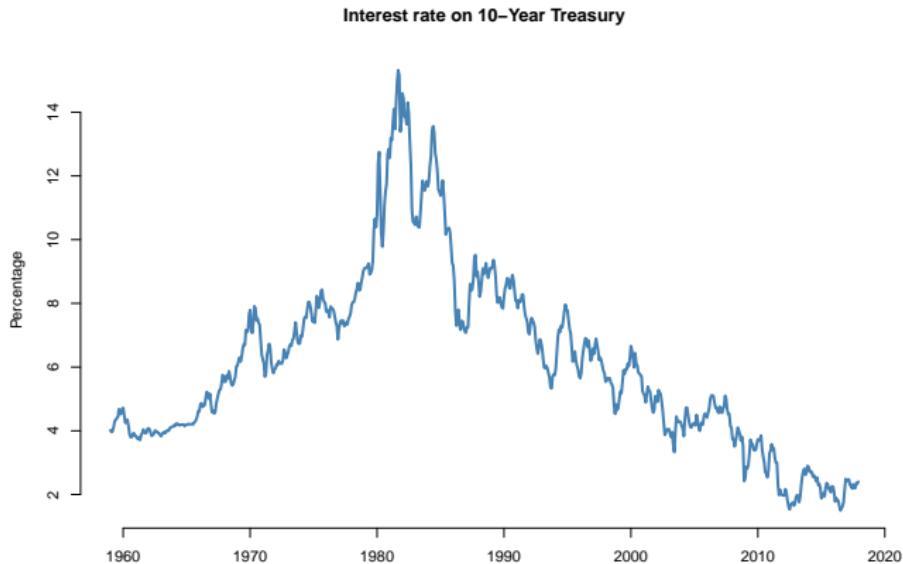
Plot of quarterly US GDP series against time. Displays a strong upward trend.

Examples

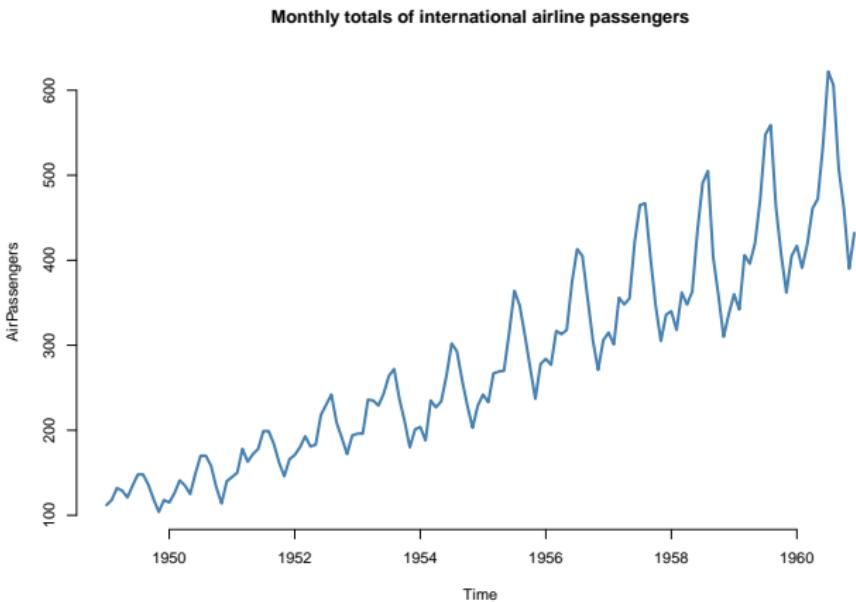


Series exhibits persistence. Deviations take some time to revert back to the mean.

Examples



Examples



Air passenger traffic is seasonal.

DGP Assumptions: Cross section v Time series

Cross section econometrics relies on **random sampling** assumption.
Data is obtained by *random sampling* from a fixed population which ensures that observations are statistically independent.

In time series, the sequence of data is important. Random sampling assumption no longer applies.

The **fundamental problem** in time series analysis – we can observe the realization of the process only once. E.g. the US GDP in the figure is only one realization of several possible histories we do not observe.

Stationarity and Weak Dependence

Main assumptions in time series

Impose conditions under which we can treat the stochastic process as a random sample as the sample size becomes very large.

Under such conditions, the **ensemble mean** at time $t = t_0$

$$\frac{1}{N} \sum_{k=1}^N Y_{t_0}^{(k)}$$

will converge to the sample **time average**:

$$\frac{1}{T} \sum_{t=1}^T Y_t$$

as N and T becomes very large. If this property holds, the stochastic process is said to be *ergodic*.

The stochastic process that has generated the data is **stationary** and **weakly dependent**

Important concepts

- Stationarity (strict and weak stationarity)
- weak dependence
- autocorrelation

Examples of stationary time series processes:

- White noise
- AR
- MA
- ARMA

Strict stationarity

A stochastic process $\{Y_t\}$ is **strictly stationary** if, for a given finite integer r and for any set of subscripts t_1, t_2, \dots, t_r the joint distribution of:

$$(Y_t, Y_{t_1}, Y_{t_2}, \dots, Y_{t_r})$$

depends only on $t_1 - t, t_2 - t, t_r - t$ but not on t .

The distribution of Y_t does not depend on the absolute position t , of Y_t and what matters for the distribution is the relative position in the sequence.

Strict stationarity – remarks

- E.g. the distribution of (Y_1, Y_4) is the same as the distribution of (Y_{15}, Y_{18}) .
- For a strictly stationary process, Y_t has the same mean, variance, and other higher moments, if they exist remain the same across t .
- Any transformation $g(\cdot)$ of a strictly stationary process. $\{g(Y_t)\}$ is also strictly stationary.

Example:

- i.i.d sequence
- constant series, $Y_t = y_1$

Covariance (weak) stationarity

A stochastic process $\{Y_t\}$ is **weakly** (or **covariance**) stationary if:

- $E[Y_t] = \mu < \infty$ does not depend on t
- $V[Y_t] = E[(Y_t - \mu)^2] = \gamma_0 < \infty$
- $Cov(Y_t, Y_{t-k}) = E[(Y_t - \mu)(Y_{t-k} - \mu)] = \gamma_k$ exists, is finite, and depends only on k but not on t .

The mean, variance and autocovariances are independent of time for a weakly stationary process.

If a sequence is strictly stationary and if the variance and autocovariances are finite, then the sequence is weakly stationary.

Stationarity – remarks

- strict stationarity (with finite second moments) implies weak stationarity, but the converse does not hold.
- If, however, joint normality could be assumed so that the distribution was entirely characterized by the first two moments, weak stationarity would imply strict stationarity.

Autocovariances

Autocovariance of order k , i.e. the autocovariance between Y_t and Y_{t-k} :

$$\text{cov}(Y_t, Y_{t-k}) = E[(Y_t - \mu)(Y_{t-k} - \mu)] = \gamma_k$$

depends on k , the time distance and not on t . It measures the direction of linear dependence between Y_t and Y_{t-k} .

Remarks:

- joint distribution is time invariant

$$\text{cov}(Y_t, Y_{t-k}) = \text{cov}(Y_t, Y_{t+k}) = \text{cov}(Y_{t-j}, Y_{t-j-k})$$

- Autocovariance of order zero ($k = 0$) is the **variance** of the process:

$$\text{cov}(Y_t, Y_t) = E[(Y_t - \mu)(Y_t - \mu)] = \gamma_0 = V[Y_t]$$

Autocorrelation

The k^{th} order **autocorrelation** is defined as:

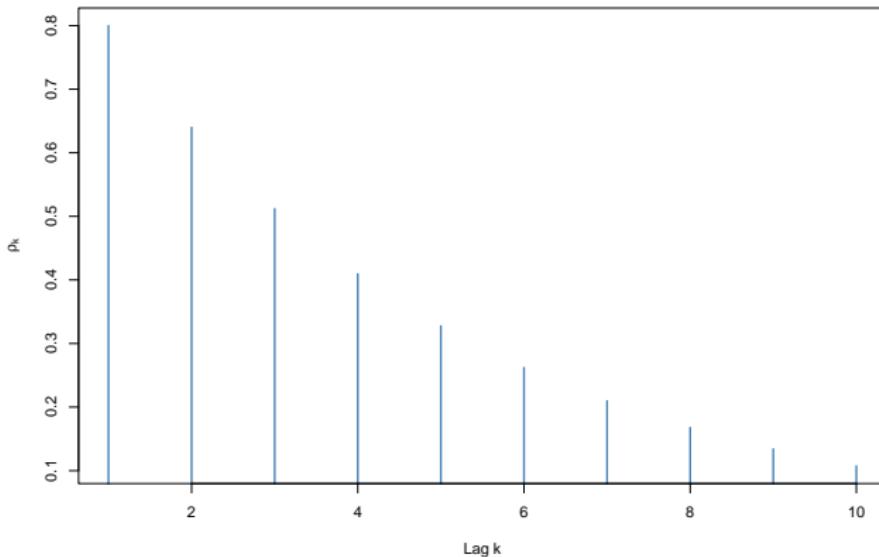
$$\rho_k = \frac{\text{cov}(Y_t, Y_{t-k})}{\sqrt{\text{var}(Y_t)\text{var}(Y_{t-k})}} = \frac{\gamma_k}{\gamma_0}$$

ρ_k measures both the direction and the strength of linear dependence between Y_t and Y_{t-k}

ACF plots

Autocorrelation function is a plot of ρ_k against k .

Figure shows an ACF for a hypothetical covariance stationary process with $\rho_k = (0.8)^k$ for $j = 1, 2, \dots, 10$.



Weak dependence

This concept is about the strength of the dependence between two observations of the time series as the time distance between them becomes larger.

A stationary time series process is said to be weakly dependent if the autocovariances decline sufficiently rapidly as the time separation increases:

$$\text{Cov}(Y_t, Y_{t-k}) \rightarrow 0 \quad \text{as} \quad k \rightarrow \infty$$

Y_t and Y_{t-k} are asymptotically independent i.e. variables positioned far apart in the sequence are almost independently distributed.

Time series assumptions

The assumptions of stationarity and weak dependence of a time series process replace the assumption of random sampling in cross section analysis, so that

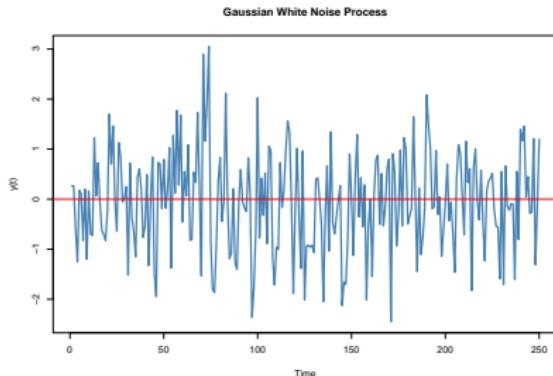
- the law of large numbers (LLN) and the central limit theorem (CLT) can be applied, and
- standard statistical results (consistency and asymptotic normality of estimators) can be established.

Stationary time series processes

Gaussian white noise $GN(0, \sigma^2)$

$$Y_t = \varepsilon_t, \quad \varepsilon_t \sim iid \ N(0, \sigma^2)$$

Feature: Lack of any predictable pattern over time in the realized value of the process.



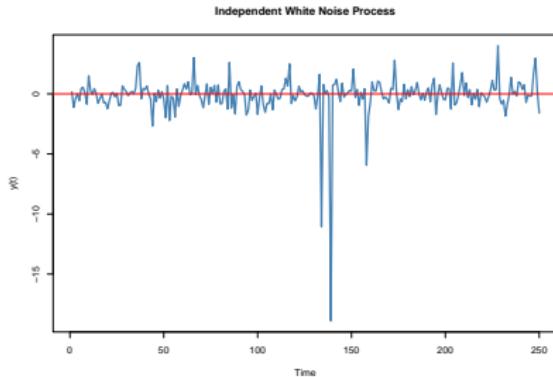
Stationary time series processes

Independent white noise / $WN(0, \sigma^2)$

$$Y_t = \varepsilon_t, \quad \varepsilon_t \sim iid(0, \sigma^2)$$

$$E[Y_t] = 0, \quad V[Y_t] = \sigma^2, \quad \gamma_j = 0, \quad j \neq 0$$

Suppose that $Y_t = \frac{1}{\sqrt{3}}t_3$ where t_3 denotes a Students t distribution with 3 d.f. $E[Y_t] = 0$, $V[Y_t] = \frac{\nu}{\nu-2} = 1$



Stationary time series processes

White noise $WN(0, \sigma^2)$

$$Y_t = \varepsilon_t$$
$$E[Y_t] = 0, \quad V[Y_t] = \sigma^2, \quad \gamma_j = 0, \quad j \neq 0$$

Basic ARMA Models

Wold's Decomposition Theorem

Any covariance stationary, purely non-deterministic, time series $y_t - \mu$ can be written as a linear combination of a sequence of uncorrelated random variables (or **linear filter**).

The *linear filter* representation is given by:

$$\begin{aligned} Y_t - \mu &= \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} + \dots \\ &= \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} \\ \psi_0 &= 1, \quad \sum_{j=0}^{\infty} \psi_j^2 < \infty, \quad \varepsilon \sim WN(0, \sigma^2) \end{aligned}$$

Using the lag operator notation, this can be rewritten as:

$$y_t - \mu = \psi(L) \varepsilon_t$$

where $\psi(L) = \sum_{j=0}^{\infty} \psi_j L^j$.

Properties

- Mean

$$E[y_t] = \mu$$

- Variance

$$\begin{aligned}\gamma_0 &= \text{var}[y_t] = E(y_t - \mu)^2 \\&= E(\varepsilon_t + \psi_1\varepsilon_{t-1} + \psi_2\varepsilon_{t-2} + \dots)^2 \\&= E(\varepsilon_t)^2 + \psi_1^2 E(\varepsilon_{t-1})^2 + \psi_2^2 E(\varepsilon_{t-2})^2 + \dots \\&= \sigma^2 + \psi_1^2 \sigma^2 + \psi_2^2 \sigma^2 + \dots \\&= \sigma^2 \sum_{j=0}^{\infty} \psi_j^2\end{aligned}$$

by using the white noise result that $E(\varepsilon_{t-i}\varepsilon_{t-j}) = 0$ for $i \neq j$

Properties

- Autocovariance

$$\begin{aligned}\gamma_k &= E(y_t - \mu)(y_{t-k} - \mu) \\&= E(\varepsilon_t + \psi_1\varepsilon_{t-1} + \dots + \psi_k\varepsilon_{t-k} + \dots)(\varepsilon_{t-k} + \psi_1\varepsilon_{t-k-1} + \dots) \\&= \sigma^2(1 \cdot \psi_k + \psi_1\psi_{k+1} + \psi_2\psi_{k+2} + \dots) \\&= \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+k}\end{aligned}$$

- Autocorrelation

$$\rho_k = \frac{\sum_{j=0}^{\infty} \psi_j \psi_{j+k}}{\sum_{j=0}^{\infty} \psi_j^2}$$

where $\sum_{j=0}^{\infty} |\psi_j| < \infty$

ARMA Models

Most economic variables evolve with some degree of dependence.
ARMA models can be used to capture serial dependence.

Autoregressive process of order p , AR(p):

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t; \quad \varepsilon_t \sim iid(0, \sigma^2)$$

y_t is regressed on past values of itself. $\phi_1, \phi_2, \dots, \phi_p$ are unknown parameters to be estimated.

MA(q)

Moving average process of order q , MA(q):

$$y_t = \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q} + \varepsilon_t; \quad \varepsilon_t \sim iid(0, \sigma^2)$$

y_t is a weighted average of past innovations $\varepsilon_t, \varepsilon_{t-1}, \dots, \varepsilon_{t-q}$ and the weights θ_i s are to be estimated.

ARMA(p, q)

Autoregressive Moving Average process of order p, q ,
ARMA(p, q):

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} \\ + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

where, $\varepsilon_t \sim iid(0, \sigma^2)$.

y_t combines an AR(p) with an MA(q) model.

Goal of Univariate Time Series

Main objectives of univariate time series models are to **analyse** and **forecast** a series y_t given:

- its own past pattern
- a stochastic error term (current and past)
- with no exogenous variables included in the model

Exercise

Write down the following processes:

- AR(1), AR(2), AR(4)
- MA(1), MA(2), MA(3)
- ARMA(2, 1)
- ARMA(1, 2)

Next

Theoretical properties of ARMA processes.

Theoretical Properties of ARMA processes

EC902: Econometrics A

Subham Kailthya

University of Warwick

Univariate Time Series

Topic 1 consists of the following video lectures:

1. Introduction to time series data and univariate time series models
2. **Theoretical properties of ARMA processes** (stationary time series)
3. Empirical modelling of univariate time series: Box-Jenkins approach
4. Forecasting with ARMA models

Overview of ARMA Models

Theoretical Properties of ARMA Processes

ARMA processes are characterized by their:

- mean
- variance
- autocorrelations

Autocorrelation: important role in describing the characteristics of the time series process.

Summary of Properties

Property	AR(1)	MA(1)
Sequence Assumption	$y_t = \phi y_{t-1} + \varepsilon_t$ $\varepsilon_t \sim iid(0, \sigma_\varepsilon^2)$	$y_t = \theta \varepsilon_{t-1} + \varepsilon_t$ $\varepsilon_t \sim iid(0, \sigma_\varepsilon^2)$
$E(y_t)$	0	0
$\text{var}(y_t) = \sigma_y^2$	$\frac{\sigma_\varepsilon^2}{1-\phi^2}$	$(1+\theta^2)\sigma_\varepsilon^2$
$\rho_k =$	ϕ^k	$\frac{\theta}{1+\theta^2}$ if $k=1$, and 0 for $k=2, 3, \dots$
$\text{corr}(y_t, y_{t-k})$		

Summary of Autocorrelation Patterns

Process	ACF	PACF
AR(p)	Infinite: damps out	Finite: cuts-off after lag p
MA(q)	Finite: cuts-off after lag q	Infinite: damps out
ARMA(p,q)	Infinite: damps out	Infinite: damps out

The lag operator

Often it is convenient to use the **lag operator**, denoted by L . It can be written as:

$$Ly_t = y_{t-1}$$

We can repeatedly apply L to represent higher order lags. For example:

$$L^2 y_t = L(Ly_t) = Ly_{t-1} = y_{t-2}$$

More generally,

$$L^p y_t = y_{t-p}; \quad L^0 \equiv 1$$

First-order Auroregressive Process

Important Properties of AR(1)

For a zero-mean AR(1) process:

$$y_t = \phi y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim iid(0, \sigma_\varepsilon^2)$$

- $E(y_t) = 0$
- $\text{var}(y_t) = \sigma_y^2 = \frac{\sigma_\varepsilon^2}{1-\phi^2}$
- $\rho_k = \frac{\gamma_k}{\sigma_y^2} = \phi^k$ where $\gamma_k = \text{cov}(y_t, y_{t-k}) = \phi^k \sigma_y^2$

Remarks:

- AR(1) process is stationary if $|\phi| < 1$
- AR(1) can be expressed as an infinite moving average process, MA(∞).

MA(∞) representation of AR(1) (1/2)

Consider a zero-mean AR(1) process:

$$y_t = \phi y_{t-1} + \varepsilon_t$$

With time shifts this can be rewritten as:

$$y_{t-1} = \phi y_{t-2} + \varepsilon_{t-1}$$

$$y_{t-2} = \phi y_{t-3} + \varepsilon_{t-2}$$

⋮

MA(∞) representation of AR(1) (2/2)

Transform into an MA(∞) process by repeated substitution:

$$\begin{aligned}y_t &= \phi y_{t-1} + \varepsilon_t \\&= \phi(\phi y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\&= \phi^2 y_{t-2} + \phi \varepsilon_{t-1} + \varepsilon_t \\&= \phi^2(\phi y_{t-3} + \varepsilon_{t-2}) + \phi \varepsilon_{t-1} + \varepsilon_t \\&\vdots \\&= \varepsilon_t + \phi \varepsilon_{t-1} + \phi^2 \varepsilon_{t-2} + \dots + \phi^k y_{t-k} \\&= \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j} \equiv \text{MA}(\infty)\end{aligned}$$

Note: Here $\phi^k y_{t-k} \rightarrow 0$ as $k \rightarrow \infty$ since $|\phi| < 1$

Mean of AR(1)

Consider the MA(∞) representation of an AR(1) process:

$$y_t = \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \phi^3\varepsilon_{t-3} + \dots, \quad \varepsilon_t \sim iid(0, \sigma_\varepsilon^2)$$

Mean of an AR(1)

$$\begin{aligned} E(y_t) &= E(\varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \phi^3\varepsilon_{t-3} + \dots) \\ &= E(\varepsilon_t) + \phi E(\varepsilon_{t-1}) + \phi^2 E(\varepsilon_{t-2}) + \phi^3 E(\varepsilon_{t-3}) + \dots \\ &= 0. \end{aligned}$$

Since $E(\varepsilon_t) = 0$ by assumption

The mean of an AR(1) process is **finite** and **time independent**.

Variance of AR(1) (1/2)

Consider an AR(1) process:

$$y_t = \phi y_{t-1} + \varepsilon_t$$

$$\begin{aligned}\text{var}(y_t) &= \text{var}(\phi y_{t-1}) + \text{var}(\varepsilon_t) + 2\text{cov}(\phi y_{t-1}, \varepsilon_t) \\ &= \phi^2 \text{var}(y_{t-1}) + \text{var}(\varepsilon_t) + 2\phi\text{cov}(y_{t-1}, \varepsilon_t) \\ &= \phi^2 \text{var}(y_{t-1}) + \text{var}(\varepsilon_t)\end{aligned}$$

Note: $\text{cov}(y_{t-1}, \varepsilon_t) = 0$ as $y_{t-1} = \phi y_{t-2} + \varepsilon_{t-1}$

Variance of AR(1) (2/2)

As y_t is covariance stationary,

$$\text{var}(y_t) = \text{var}(y_{t-1}) = \sigma_y^2 \text{ and } \text{var}(\varepsilon_t) = \sigma_\varepsilon^2$$

Then, we can rewrite the last equation as:

$$\begin{aligned}\sigma_y^2 &= \phi^2 \sigma_y^2 + \sigma_\varepsilon^2 \\ (1 - \phi^2)\sigma_y^2 &= \sigma_\varepsilon^2 \\ \sigma_y^2 &= \frac{\sigma_\varepsilon^2}{(1 - \phi^2)} \equiv \gamma_0\end{aligned}$$

Remarks:

Variance of an AR(1) process is **finite** and **time independent**. But, only if the **stationarity condition** is met i.e. $|\phi| < 1$. Otherwise, variance may not be finite.

ACF and PACF of AR(1)

The autocorrelations ρ_k as a function of k are referred to as the *autocorrelation function (ACF)* or *correlogram* of the series.

Autocovariance of order 1

$$\gamma_1 = \text{cov}(y_t, y_{t-1})$$

$$AR(1) : y_t = \phi y_{t-1} + \varepsilon_t$$

$$\begin{aligned}\gamma_1 &= \text{cov}(y_t, y_{t-1}) = \text{cov}((\phi y_{t-1} + \varepsilon_t)y_{t-1}) \\ &= \phi \text{cov}(y_{t-1}, y_{t-1}) + \text{cov}(\varepsilon_t y_{t-1}) \\ &= \phi \text{var}(y_{t-1}) + 0 \quad [\text{since } y_{t-1} \text{ and } \varepsilon_t \text{ are uncorrelated}]\end{aligned}$$

$$\gamma_1 = \phi \sigma_y^2$$

Autocorrelation coefficients of AR(1)

Autocorrelation at order 1 = $\frac{\text{Autocovariance at order 1}}{\text{Variance}}$

$$\begin{aligned}\rho_1 &= \frac{\text{cov}(y_t y_{t-1})}{\text{var}(y_t)} = \frac{\gamma_1}{\gamma_0} \\ &= \frac{\phi \sigma_y^2}{\sigma_y^2} = \phi\end{aligned}$$

Higher order AC

In the same way, we can show that:

$$\rho_2 = \frac{\gamma_2}{\gamma_0} = \phi^2$$

⋮

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \phi^k, \text{ for } k = 0, 1, 2, \dots$$

where $\rho_0 = \phi^0 = 1$

Remarks – AR(1)

- ACF for AR(1) decays to zero. If $|\phi| < 1$,

$$\rho_k = \phi^k \rightarrow 0 \quad \text{as} \quad k \rightarrow \infty$$

i.e if the series is weakly dependent.

- Thus the influence of any shock goes to zero at a rate which depends on the value of ϕ .
- Mean, variance, and autocovariances are all **finite** and **time independent** – the process is stationary and weakly dependent

Representing AR(1) using lag operator

For an AR(1) process:

$$y_t = \phi y_{t-1} + \varepsilon_t \quad \varepsilon_t \sim iid(0, \sigma_\varepsilon^2)$$

$$y_t = \phi L y_t + \varepsilon_t$$

$$(1 - \phi L) y_t = \varepsilon_t$$

$$y_t = (1 - \phi L)^{-1} \varepsilon_t$$

$$= (1 + \phi L + \phi^2 L^2 + \dots) \varepsilon_t$$

$$= \varepsilon_t + \phi \varepsilon_{t-1} + \phi^2 \varepsilon_{t-2} + \dots$$

$$= \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}$$

This linear filter will converge if $|\phi| < 1$ – **stationarity condition**

First-order Moving Average Process

Properties of MA processes

An MA(1) process takes the form:

$$y_t = \varepsilon_t + \theta \varepsilon_{t-1}, \quad \varepsilon_t \sim iid(0, \sigma_\varepsilon^2)$$

Important Properties

- $E(y_t) = 0$
- $\text{var}(y_t) = (1 + \theta^2)\sigma_\varepsilon^2 = \gamma_0$
- Autocovariances:

$$\begin{aligned}\gamma_1 &= \text{cov}(y_t, y_{t-1}) \\ &= \text{cov}[(\varepsilon_t + \theta \varepsilon_{t-1})(\varepsilon_{t-1} + \theta \varepsilon_{t-2})] \\ &= \theta \text{var}(\varepsilon_{t-1}) = \theta \sigma_\varepsilon^2\end{aligned}$$

- $\rho_1 = \frac{\gamma_1}{\gamma_0} = \frac{\theta \sigma_\varepsilon^2}{(1+\theta^2)\sigma_\varepsilon^2} = \frac{\theta}{1+\theta^2}$

Remarks:

- $\rho_k = 0$ for lag $k > 1$
- The memory of an MA(1) process lasts only **one** period.

Invertibility of MA(1)

- All MA models are stationary.
- To obtain a converging autoregressive representation, it has to meet the restriction $\theta < 1$. This is known as the **invertibility condition** and implies that the process can be written in terms of an infinite AR representation.

Invertibility of MA(1)

Consider an MA(1) process

$$y_t = \varepsilon_t - \theta \varepsilon_{t-1}$$

Using lag operator, we can write this as:

$$\begin{aligned} y_t &= (1 - \theta L) \varepsilon_t \\ \varepsilon_t &= \left[\frac{1}{1 - \theta L} \right] y_t, \quad \text{provided } |\theta| < 1 \\ &= (1 - \theta L + \theta^2 L^2 - \theta^3 L^3 + \dots) y_t \\ &= y_t - \theta y_{t-1} + \theta^2 y_{t-2} - \theta^3 y_{t-3} + \dots \end{aligned}$$

Rewriting this as an AR(∞) process:

$$\begin{aligned} y_t &= \theta y_{t-1} - \theta^2 y_{t-2} + \theta^3 y_{t-3} + \dots + \varepsilon_t \\ &= \sum_{j=1}^{\infty} (-1)^{j+1} \theta^j y_{t-j} + \varepsilon_t \end{aligned}$$

Autoregressive Moving Average Models

ARMA

Combinations of autoregressive and moving average models.

Example:

First-order autoregressive moving average, ARMA(1,1)

$$y_t = \phi y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}, \quad \varepsilon_t \sim WN(0, \sigma^2)$$

Using lag operator, we can rewrite this as:

$$(1 - \phi L)y_t = (1 + \theta L)\varepsilon_t$$

where, $|\phi| < 1$ necessary for stationarity and $|\theta| < 1$ is required for invertibility.

ARMA(p, q)

A general ARMA(p, q) can be written as:

$$y_t = \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q}$$

where $\varepsilon_t \sim WN(0, \sigma^2)$.

Using lag operators, this can be rewritten as:

$$\Phi(L)y_t = \Theta(L)\varepsilon_t$$

Remarks:

- If all roots of $\Phi(L)$ are outside the unit circle, then the process is **stationary** and has a convergent MA representation:

$$y_t = \left(\frac{\Phi(L)}{\Theta(L)}\right) \varepsilon_t.$$

- If all roots of $\Theta(L)$ are outside the unit circle, then the process is **invertible** and has a convergent infinite AR representation:

$$\left(\frac{\Phi(L)}{\Theta(L)}\right) y_t = \varepsilon_t$$

Partial Autocorrelations

Partial autocorrelation coefficients (PAC)

- PACs are the last coefficients in a sequence of successively longer autoregressions:

$$y_t = \phi_{k1}y_{t-1} + \phi_{k2}y_{t-2} + \dots + \phi_{kk}y_{t-k} + \varepsilon_t$$

obtained by running a series of regressions:

$$y_t = \phi_{11}y_{t-1} + \varepsilon_t$$

$$y_t = \phi_{21}y_{t-1} + \phi_{22}y_{t-2} + \varepsilon_t$$

$$y_t = \phi_{31}y_{t-1} + \phi_{32}y_{t-2} + \phi_{33}y_{t-3} + \varepsilon_t$$

⋮

$$y_t = \phi_{k1}y_{t-1} + \phi_{k2}y_{t-2} + \dots + \phi_{kk}y_{t-k} + \varepsilon_t$$

- PACFs are used to help discriminating between AR processes of different orders.

$AR(1) : \phi_{11} \neq 0, \phi_{22} = \phi_{33} = \dots = \phi_{kk} = 0,$ for $k > 1$

$AR(2) : \phi_{11} \neq 0, \phi_{22} \neq 0, \phi_{33} = \phi_{44} = \dots = \phi_{kk} = 0,$ for $k > 2$

⋮

$AR(p) : \phi_{11} \neq 0, \phi_{22} \neq 0, \dots, \phi_{pp} \neq 0, \phi_{kk} = 0,$ for $k > p$

- The estimate ϕ_{11} of the first equation is the lag-1 sample PACF of y_t , the estimate ϕ_{22} of the second equation is the lag-2 sample PACF of y_t , and so forth.
- PACs are **zero** for **lags larger than the order** of the process.
- For **moving average** processes: PACs patterns are very similar to the ACFs of AR processes: exponentially decaying, damping out with increasing k .

Next

Empirical modelling of univariate time series: Box-Jenkins approach

Empirical Modelling with Stationary Time Series

EC902: Econometrics A

Subham Kailthya

University of Warwick

Univariate Time Series

Topic 1 consists of the following video lectures:

1. Introduction to time series data and univariate time series models
2. Theoretical properties of ARMA processes (stationary time series)
3. **Empirical modelling of univariate time series: Box-Jenkins approach**
4. Forecasting with ARMA models

ARMA Model Building and Estimation

Question

If we observe a time series process

This could be, for example, inflation, interest rate, exchange rate, GDP growth etc.

How do we know which statistical model it follows?

- AR process? What is the value of p ?
- MA process? What is the value of q ?
- ARMA process? What are the values of p and q ?
- Later on: ARIMA (p, d, q). What are the values of p , d and q ? [here 'l' stands for **integrated** and d is the order of integration of the time series]

The **Box-Jenkins methodology** enables us to answer these questions.

Specification strategy for ARMA models

1. **Identification:** inspection of estimated ACF and PACF to identify reasonably simple ARMA structures (i.e. appropriate values of p and q)
2. **Estimation** of the parameters of the terms included in the various models
3. **Diagnostic checking** of the model residuals to check if they are approximately white noise. Is the model a good fit for the data?

If inspection of the estimated residuals suggest possible modifications, the specification process (identification, estimation, diagnostic checking) starts again.

4. The selected model is then used for **forecasting**.

Preliminary analysis

The first step involves preliminary **statistical** and **graphical** analysis.
These include:

- A **time plot** of the series
- Plot **histogram** – sample estimate of the probability density of a random variable
- Examine **summary statistics** – sample **mean**, **variance**, **autocorrelations**
 - these will be estimate of the mean, variance, autocorrelations of the stochastic process, μ , σ^2 and ρ_k , respectively

Under assumptions of **stationarity** and **weak dependence**, these unknown parameters can be estimated by their sample counterparts.

Sample statistics

- Mean:

$$\bar{y} = \frac{1}{T} \sum_{t=1}^T y_t$$

- Variance:

$$s^2 = \frac{1}{T} \sum_{t=1}^T (y_t - \bar{y})^2 = \hat{\gamma}_0$$

- autocorrelation coefficients:

$$r_k = \frac{\sum_{t=k+1}^T (y_t - \bar{y})(y_{t-k} - \bar{y})}{Ts^2} = \frac{\hat{\gamma}_k}{\hat{\gamma}_0}, \quad k = 1, 2, \dots$$

Use Stata command `summarize y` for descriptive statistics.

Identification

Identification stage

Match the behavior of the sample ACF and the sample PACF of a time series with that of various **theoretical** ACFs and PACFs.

Assess **individual** sample autocorrelations and partial autocorrelations for significance by comparing them to their respective standard errors.

$$H_0 : \rho_k = 0$$

Conduct a 'portmanteau' statistic to test joint significance of r_k s.

$$H_0 : \rho_1 = \rho_2 = \dots = \rho_m = 0$$

Test for statistical significance of individual AC

For an iid series the variance of r_k is given by $1/T$. If T is large, $\sqrt{T}r_k \sim^a N(0, 1)$ and $r_k \sim^a N(0, T^{-1})$.

Thus an absolute value of r_k in excess of $2/\sqrt{T}$ may be regarded as significantly different from zero at the 95% confidence level.

The approximate **95% confidence interval** for ρ_k is given by:

$$r_k \pm \frac{2}{\sqrt{T}}$$

Test for statistical significance of individual AC

The test is conducted by checking whether the estimated r_k falls within the two SE bounds $\pm \frac{2}{\sqrt{T}}$.

The null hypothesis $H_0 : \rho_k = 0$ is rejected if r_k is outside the two SE bounds:

$$|r_k| > \frac{2}{\sqrt{T}}$$

Example:

$$T = 400, \sqrt{400} = 20.$$

$$\text{Two SE bound: } \frac{2}{\sqrt{400}} = 0.1.$$

Any $|r_k| > 0.1$ would be statistically significantly different from zero.

Bartlett standard error

If $\rho_k = 0$ for $k > q$, the variance of r_k for $k > q$ is:

$$V(r_k) = T^{-1}(1 + 2\rho_1^2 + \dots + 2\rho_q^2)$$

The variance of the sequence r_1, r_2, \dots, r_k is estimated as $T^{-1}, T^{-1}(1 + 2r_1^2), \dots, T^{-1}(1 + 2r_1^2 + \dots + 2r_{k-1}^2)$. Taking the square root yields the standard error.

PAC

The sample ACF and sample PACF is usually calculated by fitting autoregressive models of increasing order. The estimate of the last coefficient in the each model is the sample PACF, $\hat{\phi}_{kk}$.

Recall: ϕ_{kk} measures the **additional** correlation between y_t and y_{t-k} , holding constant the effect of intermediate lags $(y_{t-1}, y_{t-2}, \dots, y_{t-k+1})$

If data follows an AR(p) process then for lags greater than p the variance of $\hat{\phi}_{kk}$ is approximately T^{-1} and $\hat{\phi}_{kk} \sim^a N(0, T^{-1})$.

Thus, $H_0 : \phi_{kk} = 0$ is rejected if $|\hat{\phi}_{kk}| > \frac{2}{\sqrt{T}}$

Comparing SACF and PACF with ACF and PACF

The behavior of the sample autocorrelations and partial autocorrelations is then compared with that of theoretical ACFs and PACFs, to identify a possible ARMA structure.

Examples:

1. If ACF dies off smoothly, at a geometric rate, and PACs are zero after one lag, this pattern suggests an AR(1) model.
2. If PACF dies off smoothly, at a geometric rate, and ACs are zero after one lag, then this pattern suggests an MA(1) model

Ljung-Box Portmanteau Q-statistic

The two-standard-error bands ($\pm 2/\sqrt{T}$) provide bounds for the **individual** sample autocorrelations.

The **Ljung and Box (Portmanteau Q) statistic** tests the joint significance of *all autocorrelations up to lag m*:

$$H_0 : \rho_1 = \rho_2 = \dots = \rho_m = 0$$

the series is white noise and all autocorrelations are jointly zero.

Test statistic:

$$Q_{LB}(m) = T(T+2) \sum_{k=1}^m \left(\frac{r_k^2}{(T-k)} \right) \sim^a \chi_m^2$$

i.e. $Q_{LB}(m)$ is approximately distributed as χ_m^2 random variable with m degrees of freedom under H_0 of white noise.

Stata commands – Portmanteau test

```
wntestq y, lags(m)
```

m usually selected to be around \sqrt{T} . For example, for $T = 400$, choose $m = 20$.

Stata commands – SACF and SPACF plots

Stata commands used in Box-Jenkins identification:

```
ac y
```

```
pac y
```

```
corrgram y
```

ac and pac graph autocorrelations and partial autocorrelations with confidence intervals.

corrgram tabulates and graphs autocorrelations and partial autocorrelations (SACF and SPACF) and Portmanteau (Q) statistics

Estimation

Model estimation

After selecting the best match (or set of matches) the next stage is to **estimate** the unknown parameters of the model $(\phi_i, \theta_i, \mu, \sigma^2)$.

AR(p):

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t$$

MA(q):

$$y_t = \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

ARMA(p,q):

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

where $\varepsilon_t \sim iid(0, \sigma_\varepsilon^2)$.

Estimating model parameters

AR models can be estimated with Ordinary Least Squares (`reg` command in Stata)

MA models require Nonlinear estimation methods (`arima` command)

This is done by Stata using these commands

```
reg y L.y          # AR(1) model  
arima y L. y, ma(1) # MA(1) model
```

For mathematical details, see Verbeek, section 8.6, if interested.

Stata commands – ARMA models

AR(1) model: $y_t = \phi_1 y_{t-1} + \varepsilon_t$

```
reg y L.y  
arima y, ar(1)  
arima y, arima(1 0 0)
```

AR(2) model: $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t$

```
reg y L(1/2).y  
arima y, ar(1 2)  
arima y, arima(2 0 0)
```

AR(p) model: $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t$

```
reg y L(1/p).y  
arima y, ar(1/p)  
arima y, arima(p 0 0)
```

Stata commands – ARMA models

MA(1) model: $y_t = \theta_1 \varepsilon_{t-1} + \varepsilon_t$

```
arima y, ma(1)  
arima y, arima(0 0 1)
```

ARMA(2,1) model: $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \theta_1 \varepsilon_{t-1} + \varepsilon_t$

```
arima y, ar(1/2) ma(1)  
arima y, arima(2 0 1)
```

If y is non-stationary, then we model the first difference of y
i.e. $\Delta y_t = y_t - y_{t-1}$. For an ARMA(1,1) in first difference use:

```
arima D.y, arima(1 0 1)  
arima y, arima(1 1 1)
```

Diagnostic Checking

Diagnostic checks

Is the chosen model a good fit for the time series?

Check for **model adequacy** i.e. if the model has been correctly specified.

Two possible ways:

- residual analysis
- overfitting

Diagnostic – residual analysis

Residuals of an adequate model should be **approximately white noise**.

If inspection of the estimated residuals suggests possible modifications, the specification process (identification, estimation, diagnostics) starts again.

Informal methods: plot residuals (to look for some indication of autocorrelation or outliers).

Formal diagnostics: tests for residual autocorrelation in the residuals.

Diagnostic checks – residual analysis

If the model residuals are not white noise, some dynamics in y_t has not been taken into account by the model.

For a white noise process, the autocorrelations are zero.

Compare significance of individual ACs \hat{r}_k s with the two SE bound $\pm 2/\sqrt{T}$.

Check for overall acceptability of the residual ACs, conduct the Ljung-Box portmanteau test.

$$Q_{LB}(m) = T(T+2) \sum_{k=1}^m \left(\frac{\hat{r}_k^2}{T-k} \right)$$

But now the test statistic $Q_{LB}(m) \sim \chi^2_{m-p-q}$ i.e. lower degrees of freedom.

Stata command – residual analysis

Suppose we select an ARMA(1,1) model as a potential candidate.

```
arima y, arima(1 0 1)      # Estimate model parameters  
predict res, residuals      # save ARMA residuals  
wntestq res, lags(4)        # Q test on ARMA residuals
```

Diagnostic checks – overfitting

If the model estimated is ARMA(p, q),

- **overfit** by estimating an ARMA($p + 1, q$) or an ARMA($p, q + 1$). and
- check if the additional fitted parameter is statistically significant.

If any deficiencies are encountered, the model must be refined until a well-specified model with no obvious deficiencies is obtained.

Model Selection

Criteria for Model Selection

Economic theory does not provide any guidance on the appropriate choice of model. How do we select from alternative models that are acceptable from a statistical point of view?

All model selection criteria, provide a **trade-off** between **goodness-of-fit** and the **number of parameters** used to obtain the fit. Note that a more general model will always provide better in-sample fit than a restricted model.

A variety of selection criteria may be used to choose an appropriate model:

- Akaike's Information Criterion (AIC)
- Schwarz Information Criterion (SIC) (also known as Bayesian information criterion BIC)

Akaike's Information Criterion (AIC)

$$AIC = \underbrace{\frac{-2}{T} \ln (\text{likelihood})}_{\text{goodness-of-fit}} + \underbrace{\frac{2}{T} r}_{\text{penalty function}}$$

where the likelihood function is evaluated at the maximum likelihood estimates, T is the sample size, and r the number of independent parameters that are fitted for the model that is assessed. For an ARMA(p, q), $r = p + q + 1$

AIC selects the model with the **best fit** as measured by the likelihood function ($\log \hat{\sigma}^2$)

Schwarz Information Criterion (AIC)

$$SIC = \frac{-2}{T} \ln (\text{likelihood}) + \frac{\ln(T)}{T} r$$

Compared with AIC, SIC tends to select a lower AR model when the sample size is moderate or large.

Selection rule

In practice, to use AIC in selecting an AR model, one computes $AIC(l)$ for $l = 0, \dots, P$ where P is a pre-specified positive integer and select the order k that has the **minimum** AIC value.

The same rule applies to BIC.

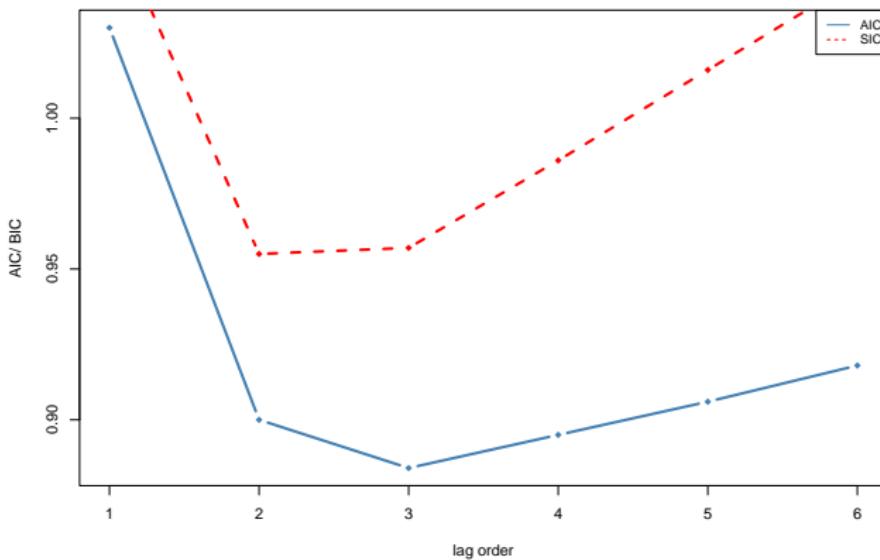
Example

A researcher estimates various AR models for the series y_t with different lag orders $p = 1, 2, \dots, 6$, and for each model computes SIC and AIC criteria in order to select the model that fits the data best.

The results are reported in the table below:

lags	SIC	AIC
1	1.067	1.030
2	0.955	0.900
3	0.957	0.884
4	0.986	0.895
5	1.016	0.906
6	1.046	0.918

Plot of AIC and BIC



SIC selects 2 AR terms while AIC selects 3 AR terms.

Next

Forecasting with ARMA Models

Forecasting with ARMA Models

EC902: Econometrics A

Subham Kailthya

University of Warwick

Univariate Time Series

Topic 1 consists of the following video lectures:

1. Introduction to time series data and univariate time series models
2. Theoretical properties of ARMA processes (stationary time series)
3. Empirical modelling of univariate time series: Box-Jenkins approach
4. **Forecasting with ARMA models**

Introduction to Forecasting

Forecasting is an important application of time series analysis.

Having built a time series model, an important goal is predicting the future path of economic variables.

*"ARMA models usually perform quite well in this respect
and often outperform more complicated structural models"*
– Verbeek

Definitions, Notations and Terminology

Suppose one or more models have been selected for forecasting.

Assume that the model parameters are estimated using data up to time T .

T is the **forecast origin**.

Let y_{T+h} be the period $T + h$ value of the process $\{y_t\}$. At time T , the value of y_{T+h} is unknown.

Denote by \hat{y}_{T+h} the h -step ahead forecast of the future value of y_{T+h} . \hat{y}_{T+h} is the **point forecast at horizon h** .

Examples:

one-step ahead forecast \hat{y}_{T+1}

two-step ahead forecast \hat{y}_{T+2}

... and so on

Definitions, Notations and Terminology

The **forecast error** at horizon h is $e_{T+h} = y_{T+h} - \hat{y}_{T+h}$.

Since there is no model uncertainty, the **conditional expectation** of a future value given available information is the optimal forecast:

$$\hat{y}_{T+h} = E_T[y_{T+h}] = E[y_{T+h} | \mathcal{I}_T]$$

Remark:

The **unconditional expectation** of y_t is the expected value of y_t
i.e. $E[y_t] = \frac{1}{T} \sum_{t=1}^T y_t = \mu$, with no reference to time.

The Optimal Forecast

Optimal Forecast

Suppose we are at time T and we would like to predict Y_{T+h} , h periods ahead.

A forecast for y_{T+h} will be based on all available information up to and including time T – the **information set** – denoted by \mathcal{I}_T .

In a univariate setting, \mathcal{I}_T includes the values of y_t and all its lags.

$$\mathcal{I}_T = \{y_{-\infty}, \dots, y_{T-1}, y_T\}$$

In general, \hat{y}_{T+h} is a function of the information set, \mathcal{I}_T .

Optimal Forecast as MMSE

Objective is to obtain a forecast that is close to the true future value.

We choose the forecast that **minimizes** forecast error:

$$E \left[(y_{T+h} - \hat{y}_{T+h|T})^2 | \mathcal{I}_T \right] = E[e_{T+h}^2 | \mathcal{I}_T]$$

The conditional expectation $E_T[y_{T+h}] = E[y_{T+h} | \mathcal{I}_T]$ is called the **minimum mean squared error (MMSE)** forecast.

Remarks:

- **predicted values** and **Residuals** are *in-sample* concepts
- **forecasts** and **forecast errors** are *out-of-sample* concepts

Forecasting an AR(1) Process

For an AR(1) model:

$$y_T = \phi y_{t-1} + \varepsilon_t, \quad t = 1, 2, \dots, T$$

and assume that the **parameters are known** (no model uncertainty).

Given the sample information at time T ,

- first obtain the one-step ahead forecast
- then the two-step ahead forecast ... and so on.

Forecasting an AR(1) Process

At time T , terms in the present or in the past are known, thus:

- $E_T[y_t] = y_t, \quad t \leq T$
- $E_T[\varepsilon_t] = \varepsilon_t, \quad t \leq T$

while terms with $t > T$ are unknown.

Since ε_t are iid with zero-mean, the best forecast of their future value is their mean of zero:

$$E_T[\varepsilon_t] = E[\varepsilon_T] = 0 \quad t > T$$

AR(1) – One-step Ahead Forecast

The value of y_{T+1} is:

$$y_{T+1} = \phi y_T + \varepsilon_{T+1}$$

The optimal (MMSE) forecast is its **conditional expectation**:

$$\begin{aligned}\hat{y}_{T+1} &= E_T[y_{T+1}] \\ &= E_T[\phi y_T] + E_T[\varepsilon_{T+1}] \\ &= \phi y_T\end{aligned}$$

Since $\varepsilon_t \sim iid(0, \sigma^2)$, the best forecast of their future value is their mean of zero – white noise process is unpredictable.

AR(1) – Two-step Ahead Forecast

The value of y_{T+2} is:

$$\begin{aligned}y_{T+2} &= \phi y_{T+1} + \varepsilon_{T+2} \\&= \phi(\phi y_T + \varepsilon_{t+1}) + \varepsilon_{T+2} \\&= \phi^2 y_T + \phi \varepsilon_{t+1} + \varepsilon_{T+2}\end{aligned}$$

The conditional expectation (optimal forecast) is:

$$\hat{y}_{T+2} = E_T[y_{T+2}] = \phi^2 y_T$$

Forecast Accuracy

AR(1) – one-step Ahead Forecast Error

The one-step ahead optimal forecast for an AR(1) process is:

$$\hat{y}_{T+1} = \phi y_T$$

How accurate is the forecast?

The one-step ahead **forecast error** is:

$$\begin{aligned} e_{T+1} &= y_{T+1} - \hat{y}_{T+1} \\ &= (\phi y_T + \varepsilon_{T+1}) - \phi y_T \\ &= \varepsilon_{T+1} \end{aligned}$$

AR(1) – one-step Ahead Forecast Error Variance

The **variance** of the forecast error is:

$$\text{var}(e_{T+1}) = \sigma_\varepsilon^2$$

A 95% confidence interval around the one-step ahead point forecast is given by (assuming ε_t is normally distributed):

$$\hat{y}_{T+1} \pm 1.96 \sqrt{\text{var}(e_{T+1})}$$

that is:

$$\hat{y}_{T+1} \pm 1.96 \sqrt{\sigma_\varepsilon^2}$$

AR(1) – two-step Ahead Forecast Error

The two-step forecast error for an AR(1) process is:

$$\begin{aligned} e_{T+2} &= y_{T+2} - \hat{y}_{T+2} \\ &= (\phi^2 y_T + \phi \varepsilon_{t+1} + \varepsilon_{T+2}) - \phi^2 y_T \\ &= \phi \varepsilon_{t+1} + \varepsilon_{T+2} \end{aligned}$$

Variance of the two-step forecast error is:

$$\text{var}(e_{T+2}) = (1 + \phi^2) \sigma_\varepsilon^2$$

95% confidence interval is given by:

$$\hat{y}_{T+2} \pm 1.96 \sqrt{(1 + \phi^2) \sigma_\varepsilon^2}$$

Forecast Performance as h Increases

For an AR(1) process, the three-step ahead forecast is:

$$\hat{y}_{T+3} = E_T[y_{T+3}] = \phi^3 y_T$$

The forecast error is:

$$e_{T+3} = \varepsilon_{T+3} + \phi \varepsilon_{T+2} + \phi^2 \varepsilon_{T+1}$$

The forecast error variance is:

$$V(e_{T+3}) = (1 + \phi^2 + \phi^4) \sigma_\varepsilon^2$$

95% confidence interval is: $\hat{y}_{T+3} \pm 1.96 \sqrt{(1 + \phi^2 + \phi^4) \sigma_\varepsilon^2}$

The width of the interval reflects forecast uncertainty.

Forecast Error Variance as h Increases

For an AR(1) process, the 95% confidence interval at different h :

$$h = 1, \quad \hat{y}_{T+1} \pm 1.96\sqrt{\sigma_{\varepsilon}^2}$$

$$h = 2, \quad \hat{y}_{T+2} \pm 1.96\sqrt{(1 + \phi^2)\sigma_{\varepsilon}^2}$$

$$h = 3, \quad \hat{y}_{T+3} \pm 1.96\sqrt{(1 + \phi^2 + \phi^4)\sigma_{\varepsilon}^2}$$

As $h \rightarrow \infty$,

- \hat{y}_{T+h} converges to the **unconditional mean** of the AR(1) process
- $\text{var}(e_{T+h}) \rightarrow \frac{\sigma_{\varepsilon}^2}{1 - \phi^2}$ converges to the **unconditional variance** of the AR(1) process.

Consequently, the informational value contained in an AR(1) process slowly decays over time. Thus, it makes sense to focus on smaller values of h .

Remarks

For any stationary ARMA model:

- the conditional forecast of y_{T+h} converges to the unconditional mean as $h \rightarrow \infty$.
- the forecast error variance converges to the unconditional variance

MA processes are less useful for forecasting

- The MA(q) process has memory of only q periods.
- MA(1) process produces forecast up to 1 step ahead.
- The best forecast two or more periods ahead is the mean of the process.

Evaluating Forecasts

Forecast Performance

We examined cases where there is no parameter uncertainty. But genuine out-of-sample forecasting is more challenging (and less optimistic).

Why might models fall short in their forecast performance?

- AIC or BIC model selection criterion selects models with best in-sample fit, but not necessarily best out-of-sample forecasting performance
- Parameter uncertainty. Becomes more severe with over-parameterized models
- Model uncertainty. E.g. model misspecification
- True process that generates the data may vary over time. E.g. structural breaks

Criteria for Evaluating Forecasts

It is common practice to estimate the models on the first n observations, and keep the last observations, say m , to evaluate the selected models.

This evaluation can be based on a number of different criteria:

$$\text{Mean Squared Error (MSE)} : \frac{1}{m} \left[\sum_{h=1}^m (y_{T+h} - \hat{y}_{T+h})^2 \right]$$

$$\text{Root Mean Squared Error (RMSE)} : \sqrt{\frac{1}{m} \left[\sum_{h=1}^m (y_{T+h} - \hat{y}_{T+h})^2 \right]}$$

Criteria for Evaluating Forecasts

$$\text{Mean Absolute Error (MAE)} : \frac{1}{m} \left[\sum_{h=1}^m |y_{T+h} - \hat{y}_{T+h}| \right]$$

$$\text{Mean Absolute Percentage Error (MAPE)} : \frac{1}{m} \left[\sum_{h=1}^m \left| \frac{y_{T+h} - \hat{y}_{T+h}}{y_{T+h}} \right| \right]$$

Forecasts for the same series across different models are compared:
the model with the lowest measure is selected.

Comparing Forecasts

Suppose we get our forecasts from two models – model 1 and model 2 – and we focus on their respective *MSEs*.

If $\widehat{MSE}_1 < \widehat{MSE}_2$ in a particular sample realization does not imply that model 1 provides better forecasts than model 2 in population.

Define a **loss function** $L(e_t) \geq 0$ which captures the cost of making a wrong forecast, $t > T$.

Assume loss is quadratic: $L(e_T) = e_T^2$.

Diebold-Mariano Test

To compare two or more competing forecasting models, Diebold and Mariano (1995) propose to compare the difference in the average loss functions of the two forecasts.

The time t **loss differential** between forecasts from model 1 and 2 is $d_{12t} = L(e_{1t}) - L(e_{2t})$. Under the null hypothesis of equal forecast accuracy, \bar{d}_{12} is zero i.e. $E[d_{12t}] = 0$ and a set of assumptions, the test statistic:

$$DM_{12} = \frac{\bar{d}_{12}}{\hat{\sigma}_{\bar{d}_{12}}} \rightarrow N(0, 1)$$

where,

- $L(e_{1t})$ is the loss function from model 1 at time t .
- $\bar{d}_{12} = \frac{1}{T} \sum_{t=1}^T d_{12t}$ is the sample mean loss differential
- $\hat{\sigma}_{\bar{d}_{12}}$ is a consistent estimate of the standard deviation of \bar{d}_{12} . Computing this is complicated due to serial correlation.

DM Assumptions

DM assumes that:

1. $E[d_{12t}] = \mu$ for all t
2. $\text{cov}(d_{12t}, d_{12(t-r)}) = \gamma(r)$ for all t
3. $0 < \text{var}(d_{12t}) = \sigma^2 < \infty$

Next

Topic 2: Dynamic regression models with stationary variables

Dynamic Regression Models with Stationary Variables

EC902: Econometrics A

Subham Kailthya

University of Warwick

References

- Gujarati and Porter, Ch. 17.1-17.3, 17.14
- Wooldridge, Ch. 10
- Verbeek, Ch.9
- Stock and Watson Ch. 14.4, Ch. 15, Ch. 16.1 (VAR)

Summary – Static and Dynamic Models

Static Model

$$y_t = \alpha + \beta_0 x_t + u_t, \quad t = 1, 2, \dots, T$$

Distributed Lag Model

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \dots + \beta_k x_{t-k} + u_t$$

Autoregressive Distributed Lag Model

$$\begin{aligned} y_t = & \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \dots + \beta_p y_{t-p} \\ & + \gamma_0 x_t + \gamma_1 x_{t-1} + \gamma_2 x_{t-2} + \dots + \gamma_q x_{t-q} + u_t \end{aligned}$$

Static Model

Static Models

Consider a static model with a single explanatory variable:

$$y_t = \alpha + \beta_0 x_t + u_t, \quad t = 1, 2, \dots, T$$

The model is **static** because x_t and y_t are dated **contemporaneously**.

y reacts immediately to one unit change in x : $\frac{\partial y_t}{\partial x_t} = \beta_0$

We can learn the trade-off between y and x .

Example: Static Model

The static Phillips curve is given by:

$$inf_t = \alpha + \beta_0 unem_t + u_t, \quad t = 1, 2, \dots, T \quad (1)$$

can be used to study the contemporaneous trade-off between inflation (y_t) and unemployment (x_t).

Assumption:

- constant *natural rate of unemployment*
- constant inflationary expectations

But, what if the effect shows up with a lag or if the impact persists beyond one period?

OLS Assumptions

- **contemporaneous exogeneity** of x_t : $E(u_t|x_t) = 0$
- errors at time t are **uncorrelated** with the explanatory variable dated at time t :

$$\text{cov}(u_t, x_t) = 0, \quad t = 1, 2, \dots, T$$

- rules out omitted variables that are in u_t and are correlated with x_t .

Dynamic Models

In time-series models, we usually consider not only how much effect x has on y , but also the **dynamic effects** of (temporary and permanent) changes in the explanatory variable.

What is the effect of a change in x in period t on the value of y in periods $t, t+1, \dots$?

- Is the effect immediate?
- Does it emerge slowly?
- Is there an initial effect that goes away after a few periods?

DL, ADL models are fit for this.

Distributed Lag (DL) Models

DL(1)

Consider a **distributed lag** model of **order 1**, **DL(1)**:

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim WN(0, \sigma^2)$$

Assume that:

- x_t is **covariance stationary** and **weakly dependent**
- x_t is **strictly exogenous**, $E[\varepsilon_t | x_s] = 0$, for all t, s .

Example: Dynamic demand function

Consider a DL(1) dynamic demand function:

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$$

y_t = **log quantity** and x_t = **log price**.

Dynamic multipliers (or impulse responses)

- **contemporaneous** price elasticity: $\frac{\partial y_t}{\partial x_t} = \beta_0$
- lag 1 price elasticity: $\frac{\partial y_t}{\partial x_{t-1}} = \frac{\partial y_{t+1}}{\partial x_t} = \beta_1$
- lag j price elasticity ($j > 1$): $\frac{\partial y_t}{\partial x_{t-j}} = \frac{\partial y_{t+j}}{\partial x_t} = 0$
- **Cumulative** or long-run price elasticity:

$$\sum_{j=0}^{\infty} \frac{\partial y_{t+j}}{\partial x_t} = \beta_0 + \beta_1$$

Meaning of long-run price elasticity

Assume x_t is at **equilibrium**. Its steady-state value is its **unconditional mean**:

$$x_t = x_{t-1} = \mu_x$$

and $\varepsilon_t = 0$.

The **unconditional mean** of y_t is:

$$E[y_t] = \alpha + \beta_0 E[x_t] + \beta_1 E[x_{t-1}] + E[\varepsilon_t]$$

$$\mu_y = \alpha + \beta_0 \mu_x + \beta_1 \mu_x$$

$$= \alpha + (\beta_0 + \beta_1) \mu_x$$

Thus,

$$\frac{\partial \mu_y}{\partial \mu_x} = \beta_0 + \beta_1$$

General DL(p) process

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \dots + \beta_p x_{t-p} + \varepsilon_t, \quad \varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$$

$$E[\varepsilon_t | x_s] = 0 \text{ for all } t, s.$$

Dynamic multipliers

$$\frac{\partial y_t}{\partial x_{t-j}} = \frac{\partial y_{t+j}}{\partial x_t} = \beta_j \quad \text{for } j < p$$

$$\frac{\partial \mu_y}{\partial \mu_x} = \beta_0 + \beta_1 + \dots + \beta_p$$

Autoregressive Distributed Lag Models

Autoregressive Distributed Lag Model

The **autoregressive distributed lag (ADL)** model with p lags of y_t and q lags of x_t denoted $\text{ADL}(p,q)$ is:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \dots + \beta_p y_{t-p} + \gamma_0 x_t + \gamma_1 x_{t-1} + \dots + \gamma_q x_{t-q} + u_t$$

$\text{ADL}(1,1)$:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \gamma_0 x_t + \gamma_1 x_{t-1} + u_t$$

Remarks:

For $\text{ADL}(p,q)$, the assumption

$$E(u_t | y_{t-1}, y_{t-2}, \dots, x_t, x_{t-1}, x_{t-2}, \dots) = 0$$

implies that the lag lengths p and q are the **true** lag lengths and the coefficients on additional lags are zero.

Selecting the lag orders p and q

General-to-specific approach:

Start with a general model (long lags), and choose p and q by considering whether:

- the estimated coefficients are significant, and
- the residuals are uncorrelated.

Can also be determined by **Information Criteria**: AIC, SIC

Impact and Dynamic Effects of ADL

Consider an ADL(1,1) model:

$$y_t = \alpha_0 + \beta_0 x_t + \beta_1 x_{t-1} + \gamma y_{t-1} + u_t$$

Short-run impact: $\frac{\partial y_t}{\partial x_t} = \beta_0$

But $\frac{\partial y_t}{\partial x_{t-1}} \neq \beta_1$

Remarks:

β_1 **cannot** be interpreted as the effect on y_t of a unitary change in x_t at time $t - 1$ holding all other variables constant.

Why? Because of the presence of a **lagged dependent variable** y_{t-1} .

Impact and dynamic effects for ADL(1,1) model

After one period: $\frac{\partial y_t}{\partial x_{t-1}} = \gamma\beta_0 + \beta_1$

After two periods: $\frac{\partial y_t}{\partial x_{t-2}} = \gamma(\gamma\beta_0 + \beta_1)$

After three periods: $\frac{\partial y_t}{\partial x_{t-3}} = \gamma^2(\gamma\beta_0 + \beta_1)$

⋮

After k periods: $\frac{\partial y_t}{\partial x_{t-k}} = \gamma^{k-1}(\gamma\beta_0 + \beta_1)$

Due to stationarity, $|\gamma| < 1$, shocks have transitory effects:

$$\frac{\partial y_t}{\partial x_{t-k}} \rightarrow 0 \quad \text{as } k \rightarrow \infty$$

Long run multiplier, ADL(1,1)

long-run multiplier (or equilibrium multiplier): The long-run effect of a unit change in x_t is given by:

$$\begin{aligned}\frac{\partial y_t}{\partial x_t} &+ \frac{\partial y_{t+1}}{\partial x_t} + \frac{\partial y_{t+2}}{\partial x_t} + \dots \\&= \beta_0 + (\gamma\beta_0 + \beta_1) + \gamma(\gamma\beta_0 + \beta_1) \dots \\&= \beta_0 + (1 + \gamma + \gamma^2 + \dots)(\gamma\beta_0 + \beta_1) \\&= \beta_0 + \frac{\gamma\beta_0 + \beta_1}{1 - \gamma} \\&= \frac{\beta_0 + \beta_1}{1 - \gamma}\end{aligned}$$

If the increase in x_t is **permanent**, the long run multiplier $\frac{\beta_0 + \beta_1}{1 - \gamma}$ can be interpreted as the expected long-run permanent increase in y_t .

Long run multiplier, ADL(1,1)

At the long-run equilibrium, $E(x_t) = E(x_{t-1})$ and $E(y_t) = E(y_{t-1})$ which yields:

$$E(y_t) = \alpha + \beta_0 E(x_t) + \beta_1 E(x_t) + \gamma E(y_t)$$

$$E(y_t) = \frac{\alpha}{1 - \gamma} + \frac{\beta_0 + \beta_1}{1 - \gamma} E(x_t)$$

$\frac{\beta_0 + \beta_1}{1 - \gamma}$ is the long-run multiplier, or the long-run elasticity of y with respect to x (if variables are in logs).

ADL model for prediction

An ADL model **without** the contemporaneous regressor(s) is useful for prediction:

$$y_t = \alpha + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + u_t$$

A question that frequently arises in time series analysis is whether or not one variable can help forecast another variable.

One way to address this question was proposed by Granger (Econometrica, 1969). *Investigating causal relations by econometric models and cross spectral methods.*

Granger Causality Tests

Granger Causality

Concept of Granger causality (predictability):

- whether or not one variable can help forecast another variable.
- Granger causality simply refers to predictive content (cannot imply exogeneity).

How to implement the test:

- Context of two stationary time series y and x
- Can be generalised to more than 2 variables

Granger Causality

A test of temporal precedence (or incremental predictability) rather than a test of causality

× **Granger causes** y if:

$$E(y_t | \mathcal{I}_{t-1}) \neq E(y_t | \mathcal{J}_{t-1})$$

where \mathcal{I}_{t-1} contains past information on y and x and \mathcal{J}_{t-1} contains only past information on y .

If past values of a variable x contain information that helps predict y **in addition to** the information contained in past values of y alone, then x is said to **Granger-cause** y .

Example

Consider an ADL(2, 2) model:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \gamma_1 x_{t-1} + \gamma_2 x_{t-2} + u_t$$

where $E(u_t | y_{t-1}, y_{t-2}, x_{t-1}, x_{t-2}) = 0$

We say that x **does not Granger-cause** y if:

$$E(y_t | y_{t-1}, y_{t-2}, x_{t-1}, x_{t-2}) = E(y_t | y_{t-1}, y_{t-2})$$

Testing for Granger causality

Null hypothesis: Absence of Granger causality

$$H_0 : \gamma_1 = \gamma_2 = 0$$

$$H_1 : \text{at least one } \gamma_i \neq 0$$

F test for the joint linear hypothesis.

Test statistic:

$$F_{m, T-k} = \frac{(RSS_R - RSS_{UR})/m}{RSS_{UR}/(T - k)} \sim F_{m, T-k}$$

where m is the number of lagged x , T the number of observations, and k the number of estimated parameters. RSS is the residual sum of squares.

Rejection of H_0 implies that x Granger-causes y

Example contd.

In a similar way, we can test if y Granger-causes x :

$$x_t = \delta_0 + \delta_1 y_{t-1} + \delta_2 y_{t-2} + \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + u_t$$

Relevant hypotheses:

$$H_0 : \delta_1 = \delta_2 = 0$$

$$H_1 : \text{at least one } \delta_i \neq 0$$

F test for the joint linear hypothesis

Rejection of H_0 implies that y Granger-causes x

Interpretation of Results

Four possible results

1. Both sets of lags are significant, there is **bi-directional Granger-causality**. (the computed F value exceeds the critical F value at the chosen level of significance), that is H_0 is rejected in both equations.
2. H_0 rejected only in eq.(12), there is **uni-directional Granger-causality** from x to y .
3. H_0 rejected only in eq.(16), there is **uni-directional Granger-causality** from y to x .
4. H_0 cannot be rejected in either equation: there is **no Granger-causality**.

Next

Topic 3: Nonstationarity and Unit Root Tests

Nonstationary Time Series and Unit Root Tests

EC902: Econometrics A

Subham Kailthya

University of Warwick

Topic 3 Nonstationarity and Cointegration

Topic 3 will include the following:

1. Nonstationary time series
2. Testing for nonstationarity
3. Cointegration and error correction models

References

- Gujarati and Porter, Ch. 21
- Wooldridge, Section 18.2
- Verbeek, Ch.8-9
- Stock and Watson Ch. 14.6.

Nonstationary Time Series

Economic Time Series

Economic and financial time series are often not well described by the assumption of stationarity i.e. they are **non-stationary**

Recall, a time series process is stationary if its mean, variance and autocovariances do not depend on the particular time period.

Economic time series usually have:

- time varying mean
- time varying variance
- persistence over time etc.

Overview

- We will see examples of time series with different types of **trends**
 - deterministic
 - stochastic
- How non stationary time series can be made stationary
 - trend stationary series
 - difference stationary series
- Random walk processes and unit roots

Example 1

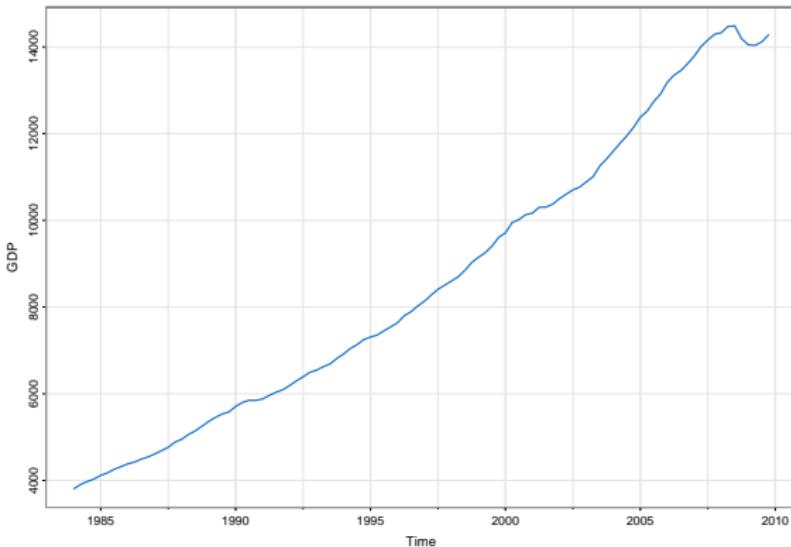


Figure 1: Quarterly US GDP

Example 2

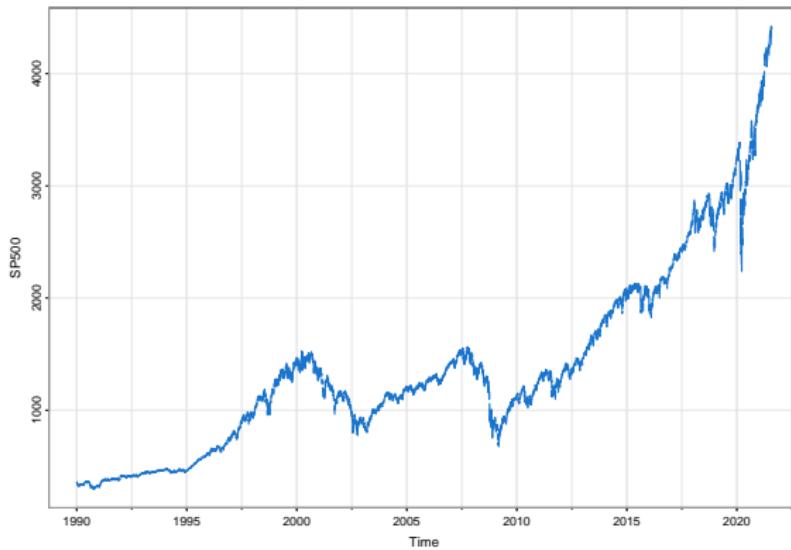


Figure 2: SP 500 index

Figures 1 and 2 have a strong trend component.

Example 3

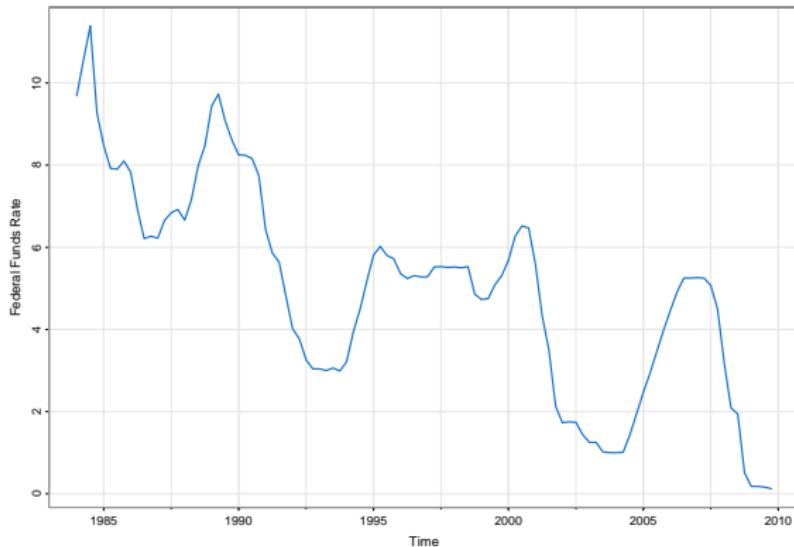


Figure 3: US Fed Funds Rate

Example 4

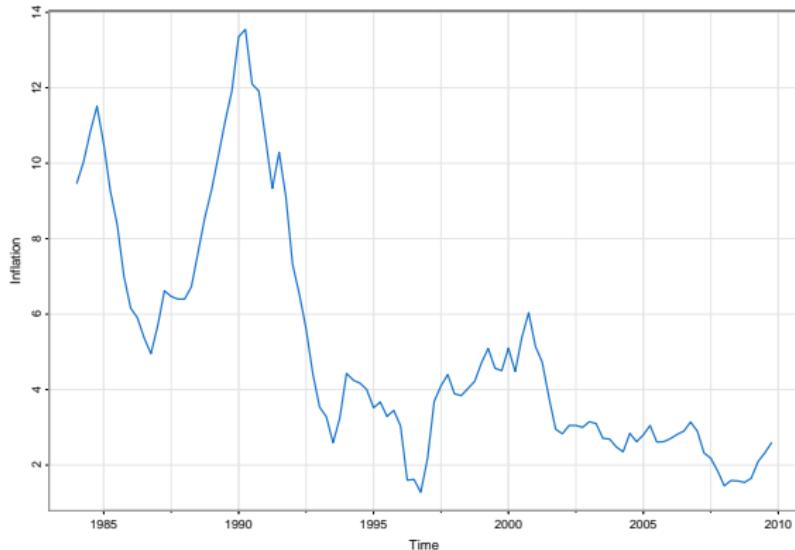


Figure 4: US Inflation

Figures 3 and 4 exhibit persistence.

Example 5

Compare Figure 5 with Figures 1-4. What do you observe?

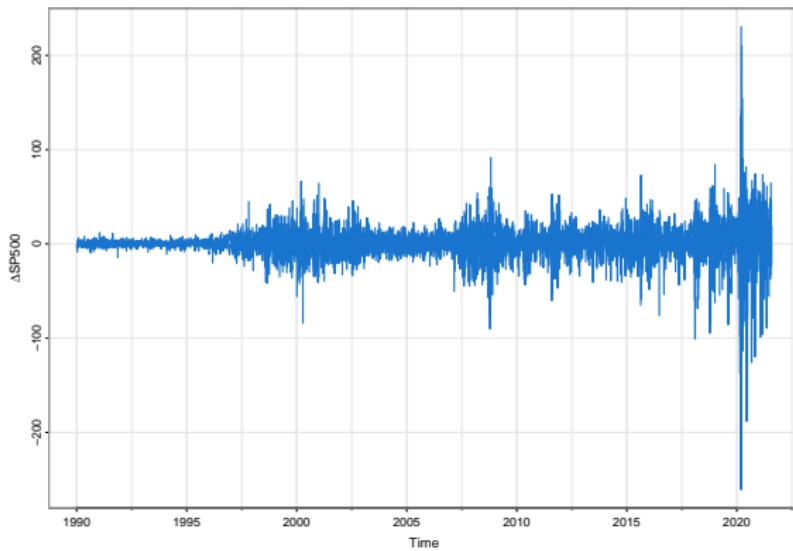


Figure 5: SP500 returns

Importance of Stationarity

- Stationary processes are better understood than non-stationary processes.
- The test statistic of certain non-stationary processes do not follow usual distributions.
- Knowing **how** a process is non-stationary will enable us to make the necessary corrections.
- Estimating regression models with nonstationary series can lead to **spurious regression** (next topic 3.3)

Trend

What is a **trend**?

A persistent long-term movement of a variable over time.

Trends can be:

- deterministic
- stochastic

Trend Stationary Series

Trend Stationary Series

Trend stationarity

A trend stationary series fluctuates around a deterministic trend (the mean of the series) with no tendency for the amplitude of the fluctuations to increase or decrease.

Example

$$y_t = \alpha + \beta t + u_t, \quad t = 1, \dots, T$$

where u_t can be:

- white noise error, or
- any stationary ARMA process

Example: $y_t = \text{deterministic trend} + AR(p)$

Trend Stationary Series

Trend stationary processes are:

- **nonstationary in the mean** but variance is time invariant.
- called **trend stationary** since the short-run fluctuations (u_t) are stationary around the deterministic trend.

How can we remove trends?

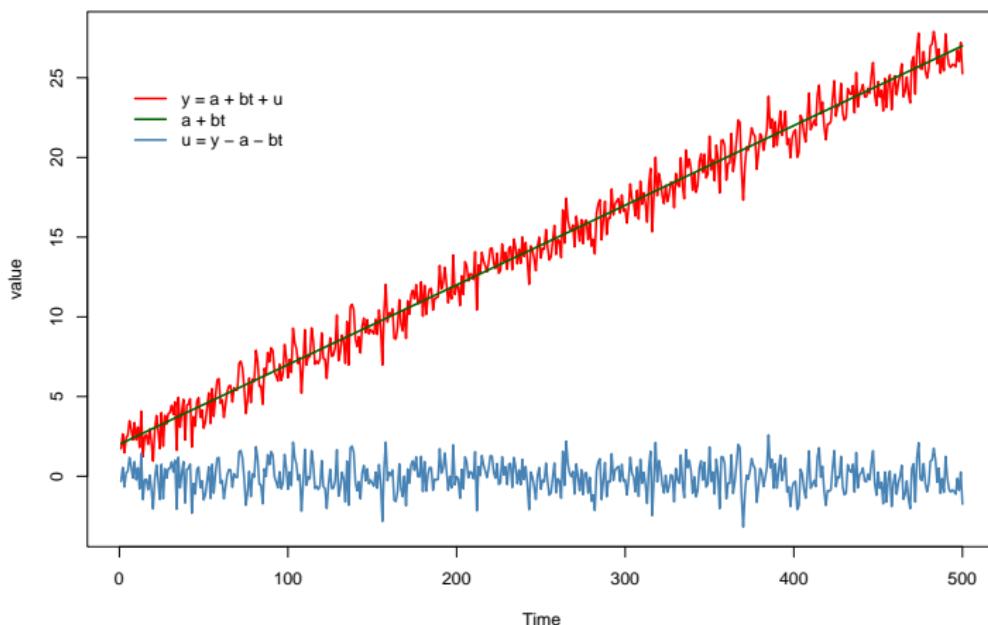
By fitting a deterministic time-trend to the series:

$$y_t = \hat{\alpha} + \hat{\beta}t + u_t$$

The residuals from the fitted trend will give the **de-trended** series:

$$\hat{u} = y_t - \hat{\alpha} - \hat{\beta}t$$

Detrending a Trend Stationary Process



Difference Stationary Series

Difference Operator

Δ is the **difference operator**.

$$\begin{aligned}\Delta y_t &= y_t - y_{t-1} \\ \Delta^2 y_t &= \Delta(\Delta y_t) = \Delta(y_t - y_{t-1}) \\ &= (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) \\ &= y_t - 2y_{t-1} + y_{t-2}\end{aligned}$$

Difference Stationary Process

Difference stationary processes are also known as **integrated** series or **stochastic trends**.

If a non-stationary series can be made stationary by differencing d times we say that the series is **integrated of order d** and denoted by $I(d)$.

y_t is $I(1)$ if Δy_t is stationary

y_t is $I(2)$ if $\Delta^2 y_t$ is stationary

y_t is $I(d)$ if $\Delta^d y_t$ is stationary

A **stationary** process is integrated of order zero, $I(0)$.

So, a white noise process, a stationary AR or ARMA prepossess are all $I(0)$.

Simple Random Walk

y_t is a **simple random walk** if:

$$y_t = y_{t-1} + u_t, \quad u_t \sim iid(0, \sigma_u^2)$$

This is an AR(1):

$$y_t = \phi y_{t-1} + u_t \quad \text{with } \phi = 1$$

This series is said to have a **unit root**, or to be integrated of order 1 i.e. an $I(1)$ process.

The first difference of the random walk process is stationary $I(0)$:

$$\Delta y_t = u_t$$

Simple Random Walk

Let the process start at $t = 0$ with a value y_0 assumed to be fixed.

By recursive substitution we get:

$$y_1 = y_0 + u_1$$

$$y_2 = y_1 + u_2 = y_0 + u_1 + u_2$$

⋮

$$y_t = y_0 + u_1 + u_2 + \dots + u_t = \underbrace{y_0}_{\text{Initial value}} + \underbrace{\sum_{i=1}^t u_i}_{\text{Stochastic trend}}$$

- y_t is expressed as a function of its initial value y_0 and a partial sum series $\sum_{i=1}^t u_i$ called the **stochastic trend**.
- If y_t is stationary, the effect of a shock at time t , u_t , decays as time passes.
- For a **unit-root** process, shocks can change the level of the series permanently.

Example: RW Process

Draw 500 observations from:

$$u_t \sim N(0, 0.04); \quad y_t = \sum_{t=1}^{500} u_t$$

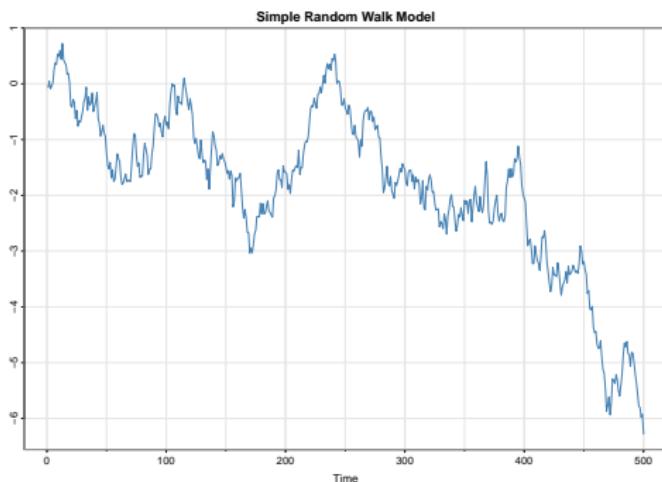


Figure 6: Simple Random Walk Model

Properties of Simple RW

- $E(y_t) = y_0$ constant, time independent,
- Simple RW: **non-stationary in variance.**

$$V(y_t) = V\left(\sum_{i=1}^t u_i\right) = t\sigma_u^2$$

- \implies the simple random walk process is **non-stationary**.
- Covariances/ correlations are **time dependent**

$$\rho(1) = \sqrt{\frac{t-1}{t}}; \quad \rho(s) = \sqrt{\frac{t-s}{t}}$$

- But, $\Delta y_t = u_t$ is stationary.

This process is also called **random walk without a drift**.

RW with a drift

Add a constant term to simple RW:

$$y_t = \mu + y_{t-1} + u_t$$

RW with drift is stationary after first differencing:

$$\Delta y_t = \mu + u_t$$

We can show that:

$$y_t = y_0 + t\mu + \sum_{i=1}^t u_i$$

which consists of:

- a **deterministic trend**, $y_0 + t\mu$, and
- a **stochastic trend**, $\sum_{i=1}^t u_i$

RW with drift

$$y_t = y_0 + t\mu + \sum_{i=1}^t u_i$$

where, $y_0 = 2$, $\mu = 0.05$, $u_t \sim (0, 0.04)$.

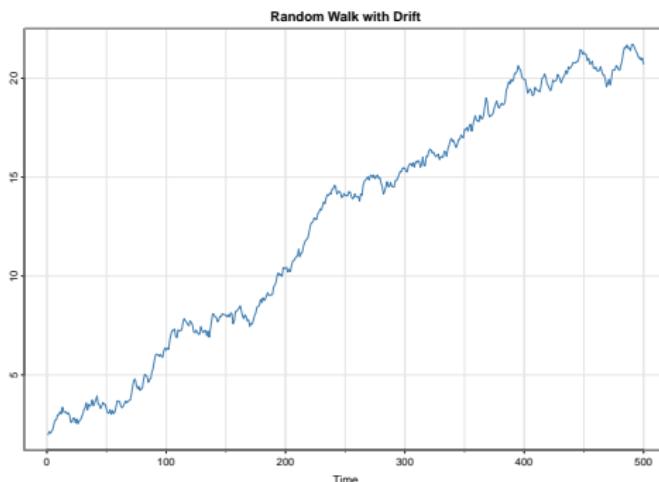


Figure 7: Random Walk with Drift

Properties of Random Walk with Drift

- Mean increases linearly with t (**a trend in the mean**):

$$E(y_t) = y_0 + t\mu$$

- Variance increases linearly with t (**a trend in the variance**)

$$V(y_t) = V\left(\sum_{i=1}^t u_i\right) = t\sigma_u^2$$

- As in the simple RW, the covariances/correlations are time dependent.

A **RW with drift** has a trend in **both** its mean and variance while a **simple RW** has a trend in **only** its variance but not its mean.

Remarks

- Both RW and RW with drift are not useful for forecasting.
- For a simple RW:

$$\hat{y}_{T+h} = E_T(y_{T+h}) = y_T, \quad \text{flat forecast function}$$

- For a RW with drift:

$$\hat{y}_{T+h} = E_T(y_{T+h}) = y_T + \mu h, \quad \text{trend}$$

- Forecast error variance increases linearly with the forecast horizon:

$$V(e_{T+1}) = \sigma_\varepsilon^2; \quad V(e_{T+2}) = 2\sigma_\varepsilon^2; \quad \dots; \quad V(e_{T+h}) = h\sigma_\varepsilon^2$$

- Unbounded forecast confidence intervals!

Remarks

- Random walks are special cases of $I(1)$ series in which the first differences are white noise and thus serially uncorrelated:
 - $\Delta y_t = \varepsilon_t$, if y_t is a simple RW.
 - $\Delta y_t = \mu + \varepsilon_t$, if y_t is a RW with drift.
- More generally, the differenced stationary process can be an ARMA process:
 - $\Delta y_t \sim ARMA(p, q)$, in which case,
 - $y_t \sim ARIMA(p, 1, q)$, i.e. $y_t \sim I(1)$ an autoregressive **integrated** moving average (ARIMA) process.

Problems Caused by Integrated Series

- Standard distribution theory not valid.
- Distribution of OLS and t statistics not normal even in large samples.
- So cannot use standard normal critical values and confidence intervals.
- Can lead to the problem of **spurious regression** (more on this later).

Summary: Difference Between I(0) and I(1) Series

I(0) stationary	I(1) integrated
Effects of a shock $\rightarrow 0$ as time passes	Effects of shock lasts forever
Observations will fluctuate around a mean and they will cross this value frequently	Observations will wander widely and will only rarely return to an earlier value
$\rho_k \rightarrow 0$ more or less rapidly	ρ_k will stay around 1 for even very large k
Standard distribution theory can be used	Standard distribution theory not valid

Differencing and Detrending Appropriately

Suppose we **de-trend** a RW with drift by fitting a linear trend:

$$y_t = y_0 + t\mu + \sum_{t=1}^t \varepsilon_t$$

$$y_t - t\mu = y_0 + \sum_{t=1}^t \varepsilon_t$$

This does **not** eliminate the stochastic trend.

Overdifferencing

Similarly, transforming a trend stationary series by first differencing produces a **unit moving average root**:

$$y_t = \alpha + \beta t + \varepsilon_t$$

$$\Delta y_t = \beta + \epsilon_t, \quad \epsilon_t = \varepsilon_t - \varepsilon_{t-1}$$

An non-invertible MA(1) process with $\theta = -1$

Differencing a Trend Stationary Series

Suppose y_t is a trend stationary process of the form:

$$y_t = \alpha + \beta t + \varepsilon_t$$

At time $t - 1$, this can be written as:

$$y_t = \alpha + \beta(t - 1) + \varepsilon_{t-1}$$

Thus, first difference of y_t yields:

$$\begin{aligned}\Delta y_t &= y_t - y_{t-1} \\ &= (\alpha + \beta t + \varepsilon_t) - (\alpha + \beta(t - 1) + \varepsilon_{t-1}) \\ &= \beta + \varepsilon_t - \varepsilon_{t-1} \\ &= \beta + \epsilon_t\end{aligned}$$

where $\epsilon_t = \varepsilon_t - \varepsilon_{t-1}$ is an MA(1) process with $\theta = 1$ (it is not invertible and can cause problems in estimation)

Tests for Nonstationarity

Detecting nonstationarity

The presence of unit roots means that standard distribution theory is not valid.

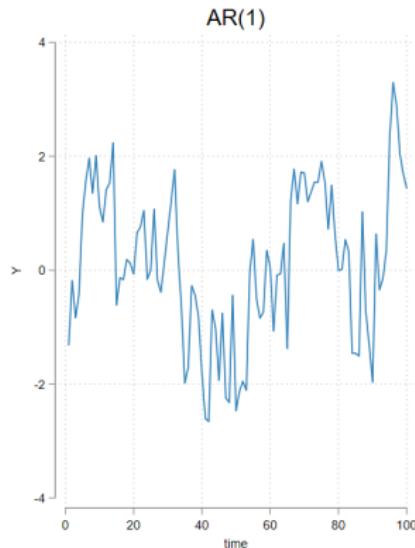
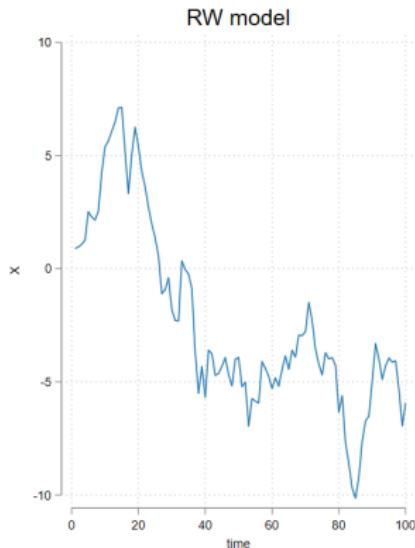
So it is important to test for stationarity of the series prior to any estimation in order to use the appropriate procedure for de-trending.

An **informal** method for detecting nonstationarity is based on inspection of the autocorrelation function.

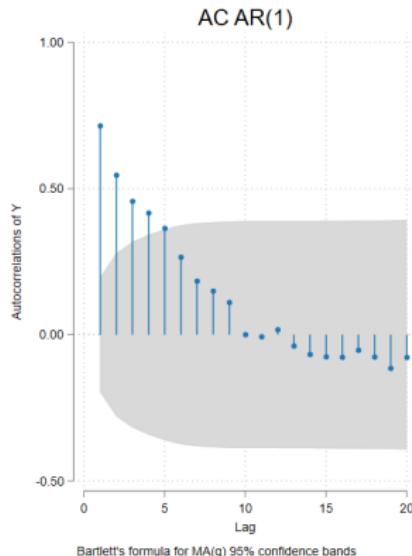
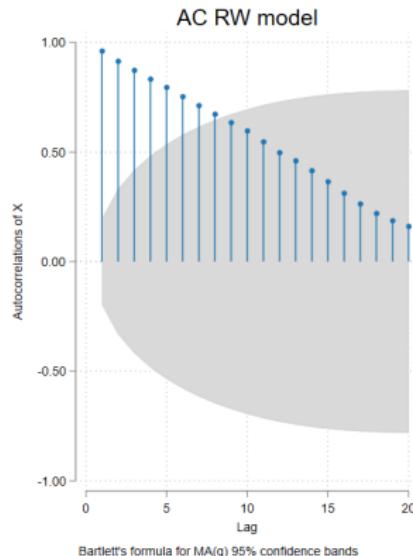
RW and stationary AR(1)

Consider two series X and Y .

X is RW. Y is AR(1).



ACF of RW and stationary AR(1)



ACF of RW declines very slowly whereas ACF of AR(1) declines rapidly.

Knowing the Source of Non-stationarity is Critical

A time series process can be non-stationary without being a unit-root. For example, there could be seasonality, deterministic trend, time-varying variance etc.

Whether a series has a **deterministic trend** or a **stochastic trend** has important implications.

Example

Suppose the stock price of a company follows a stochastic growth process. A temporary shock (e.g. energy crisis) will affect the stock price indefinitely into the future.

What happens if the stock price follows a deterministic trend model? Stock prices recover from the shock and rebound.

Dickey and Fuller (DF) Test

Unit Root Tests for Stationarity

Consider an AR(1) process:

$$y_t = \rho y_{t-1} + u_t, \quad u_t \sim iid(0, \sigma_u^2)$$

If $\rho = 1$, the equation defines a simple random walk and y_t is nonstationary.

Dickey and Fuller (DF) test in the AR(1) model

The hypothesis that y_t has a stochastic trend corresponds to:

$$H_0 : \rho = 1 \text{ (nonstationarity)}$$

$$H_1 : \rho < 1 \text{ (stationarity)}$$

The test of this hypothesis is called a **unit root test**.

Dickey-Fuller Unit Root Test

A simple way to test the null hypothesis of a unit root is to re-specify the AR(1) equation as:

$$\Delta y_t = \gamma y_{t-1} + u_t, \text{ where } \gamma = (\rho - 1)$$

The relevant hypotheses are

$$H_0 : \gamma = 0 \text{ (nonstationarity)}$$

$$H_1 : \gamma < 0 \text{ (stationarity)}$$

The test is a t statistic on $H_0 : \gamma = 0$

$$t_\gamma = \frac{\hat{\gamma} - 0}{se(\hat{\gamma})}$$

With the rejection region on the left of the t-distribution ($H_1: \gamma < 0$).

Cases where $\rho > 1$ are not considered since economic and financial series are not expected to be explosive.

DF Test

DF test can include:

1. **no constant or trend**

$$\Delta y_t = \gamma y_{t-1} + u_t \quad (1)$$

2. a **constant** term

$$\Delta y_t = \alpha + \gamma y_{t-1} + u_t \quad (2)$$

3. **constant** and a **trend** variable

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + u_t \quad (3)$$

Important:

DF tests have a non-standard t distribution under $H_0 : \rho = 1$ (non stationary).

Performing Dickey-Fuller test

Estimate one of the three regressions with OLS.

Compare the t-statistic on the coefficient of γ :

$$t_\gamma = \frac{\hat{\gamma} - 0}{se(\hat{\gamma})}$$

with **appropriate critical values** (see Table below).

Stata command: DF Test

```
// DF test with no constant or trend  
dfuller y, reg noconstant
```

```
// DF test with a constant term  
dfuller y, reg
```

```
// DF test with a constant and trend term  
dfuller y, reg trend
```

Remarks

- Regressions (2) and (3) used more frequently.
- Regression (2) if the series does not have an obvious trend:
 - inflation, interest rate, etc.
 - if the series looks like a simple random walk model
 - so under H_0 the series will be stationary AR.
- Regression (3) for series with an obvious trend
 - GDP, CPI, stock price index (levels), etc.
 - series looks like a random walk with drift.
 - so under H_0 , series will be trend stationary
- Regression (1) for zero-mean series.

Remarks

- DF tests have a non-standard t-distribution under $H_0 : \rho = 1$ (non-stationary)
- The distribution of the test is shifted to the left, relative to that of a Student's t distribution.
- Larger negative values are needed to reject the null hypothesis.
- From table below:
 - To reject H_0 at the 5% level, we need a value of the test ≤ -1.94
 - The 5% one-sided critical value for a $N(0, 1)$ is -1.645
- Inappropriate use of standard normal values will lead to **over-rejection** of the null.

Limiting distribution of DF τ and DF τ_μ

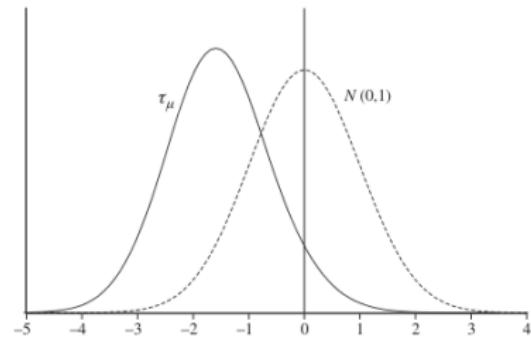
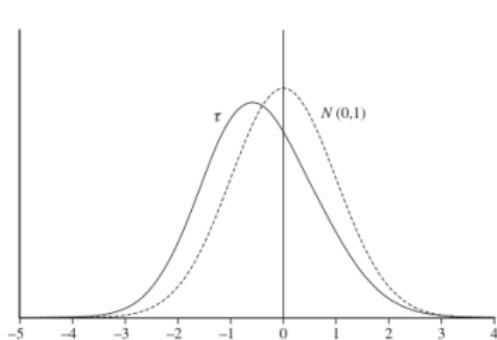


Figure source: Mills (2019)

Critical Values for Dickey-Fuller Test

Critical Values for Dickey-Fuller Test

Test St.	Model	1%	5%	10%
τ_{nc}	$\Delta y_t = \gamma y_{t-1} + v_t$	-2.56	-1.94	-1.62
τ_c	$\Delta y_t = \alpha + \gamma y_{t-1} + v_t$	-3.43	-2.86	-2.57
τ_{ct}	$\Delta y_t =$ $\alpha + \lambda t + \gamma y_{t-1} + v_t$	-3.96	-3.41	-3.13
$N(0,1)$	-	-2.33	-1.65	-1.28

Source: Davidson and MacKinnon (1993)

Augmented Dickey and Fuller (ADF) Test

Augmented DF Test

ADF test is used if the residuals in equations (1)-(3) are serially correlated i.e. if y_t is a higher order autoregressive process. Expands the DF regressions with Δy_t s on the right hand side

ADF test regression:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{j=1}^{p-1} \delta_j \Delta y_{t-j} + u_t$$

Include sufficient ΔY_{t-j} $j = 1, 2, \dots$ to secure an approximate WN error, u_t .

Test $H_0 : \gamma = 0$; Reject $H_0 \implies y_t \sim I(0)$.

What if we are unable to reject H_0 ? $\implies y_t \sim$ at least $I(1)$; repeat ADF test on Δy_t to determine order of integration d .

ADF Test Regressions

$$\Delta y_t = \gamma y_{t-1} + \sum_{j=1}^{p-1} \delta_j \Delta y_{t-j} + u_t \quad (4)$$

$$\Delta y_t = \alpha + \gamma y_{t-1} + \sum_{j=1}^{p-1} \delta_j \Delta y_{t-j} + u_t \quad (5)$$

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{j=1}^{p-1} \delta_j \Delta y_{t-j} + u_t \quad (6)$$

Include sufficient lagged first differences to secure an appropriate white noise error term in the ADF regression.

Test if the coefficient on y_{t-1} is zero, $H_0 : \gamma = 0$.

Interpretation of the ADF Test

- If the null of unit-root is rejected, we conclude that y_t is $I(0)$ (Stationary)
- But if H_0 is not rejected:
 - The series y_t is non-stationary
 - What is the order of integration of y_t ?
 - Is it $I(1)$? $I(2)$?

Remarks:

$y_t \sim I(1)$ if $\Delta y_t \sim I(0)$.

Repeat the DF test on the **first-difference** of y_t .

Interpretation of the ADF Test

- Null hypothesis of DF test is non-stationarity

$$H_0 : \Delta y_t \sim I(1), \quad H_1 : \Delta y_t \sim I(0)$$

- If H_0 is rejected, we conclude that $y_t \sim I(1)$.
- If H_0 cannot be rejected, conclude that y_t is at least $I(2)$.

Stata command: ADF Test

Suppose including two lags of the dependent variable is appropriate, then:

```
// ADF test with no constant or trend  
dfuller y, lags(2) reg noconstant
```

```
// ADF test with a constant term  
dfuller y, lags(2) reg
```

```
// ADF test with a constant and trend term  
dfuller y, lags(2) reg trend
```

Limitations of DF/ ADF Tests

- Low powered test
 - in small samples
 - for ρ close to 1 (near unit root)
 - in the presence of structural breaks
- Uncertainty about lag length p in the DF test regression
 - Extra regressors of lagged first-differences does not affect the **size of the test** (the level of significance of the test or Type I error) but decreases **power** of the test.
 - Too few lags may affect the size of the test (Perron 1989)
- Uncertainty about the test version to use i.e. whether to include intercept, trend term, etc.

Remarks

- Power is the probability of rejecting the null hypothesis when it is false.
- Failure to reject H_0 provides only weak evidence in favour of the random walk hypothesis.
- Power = $(1 - \text{Type II error})$ where Type II error is the probability of accepting a false H_0 .

Next

Cointegration and Error Correction Models

Cointegration and Error Correction Models

EC902: Econometrics A

Subham Kailthya

University of Warwick

Spurious Regression

Spurious Regression

If X_t and Y_t are two integrated series but they are otherwise unrelated, a regression of Y_t on X_t can find a statistically significant relationship between these variables, even though none exists.

That is, the result can be **spurious**. This result is due to Granger and Newbold (1974).

Relationship between two non stationary series

- **Stationarity** is a key assumption in time series regression analysis.
- Regression analysis with non-stationary time series may lead to the phenomenon of **spurious** or nonsense regressions.
- High R-squared value and statistically significant coefficients even if the series are completely unrelated.
- Usual t and F tests are **not reliable** in regressions with nonstationary series.

Spurious Regressions

Consider two random walk series:

$$Y_t = Y_{t-1} + v_t, \quad v_t \sim iid(0, \sigma_v^2)$$

$$X_t = X_{t-1} + u_t, \quad u_t \sim iid(0, \sigma_u^2)$$

where v_t and u_t independent.

If we regress Y_t on X_t :

$$Y_t = \alpha + \beta X_t + \epsilon_t$$

what would you expect?

Spurious Regressions

We would expect:

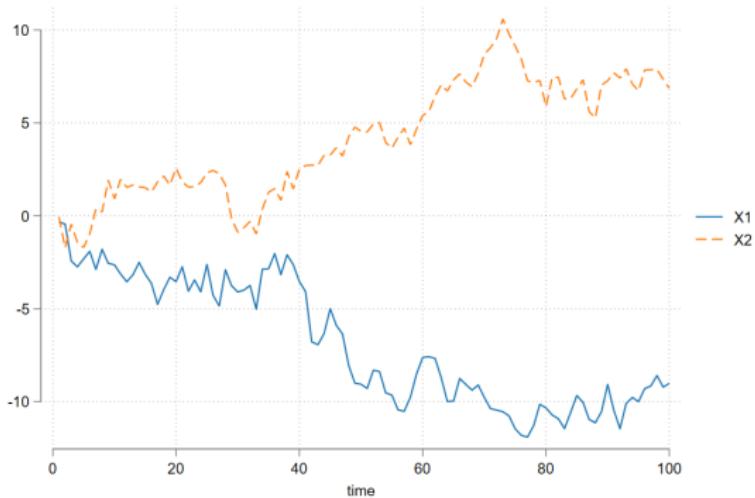
- No **true** relationship between Y_t and X_t
 - β not statistically significant
- low R-squared

Instead, despite lack of true relationship, we are likely to find:

- highly significant coefficients
 - a significant t ratio for $H_0 : \beta = 0$.
- high R-squared and F statistic
- Resembles a very reasonable model but is completely **spurious**.

Example: Spurious Regression

Consider two RWs X_1 and X_2 . Time series plot of X_1 and X_2 .



Example: Spurious regression

Regress X1 on X2:

```
reg X1 X2  
predict X1hat //fitted values
```

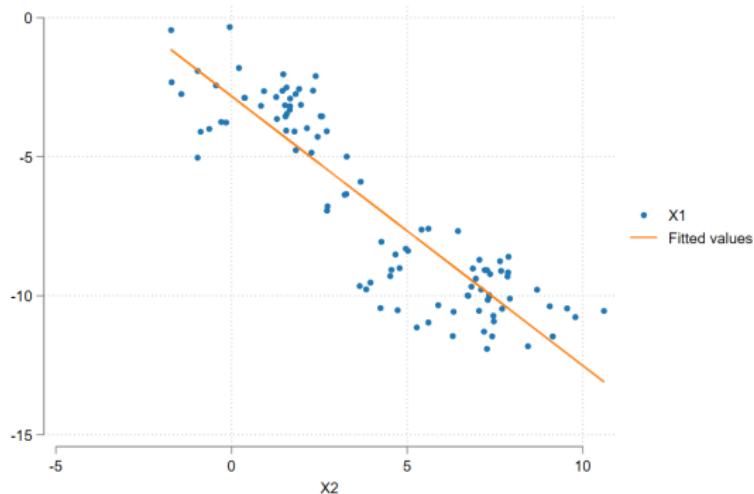
Source	SS	df	MS	Number of obs	=	100
Model	907.860794	1	907.860794	F(1, 98)	=	382.50
Residual	232.600982	98	2.37347941	Prob > F	=	0.0000
Total	1140.46178	99	11.5198159	R-squared	=	0.7960
				Adj R-squared	=	0.7940
				Root MSE	=	1.5406

	X1	Coefficient	Std. err.	t	P> t	[95% conf. interval]
	X2	-.9693983	.0495662	-19.56	0.000	-1.067761 - .8710358
	_cons	-2.822829	.2568452	-10.99	0.000	-3.33253 -2.313128

The relationship is highly significant. R-squared is high.

Example: Spurious regression

Scatter Plot of X_1 and X_2 overlayed with fitted regression line.



This result corresponds to a single sample. What if we replicated this exercise many times?

Simulation

Repeat the exercise 200 times and observe the results. In each run:

- Draw two RWs
- Regress X_1 on X_2
- Save R^2 and p-value
- Durbin-Watson autocorrelation test.

Summarizing the result yields:

Variable	Obs	Mean	Std. dev.	Min	Max	
DWstat	200	.3387572	.2059569	.0482281	1.065985	. count if Pval<0.05 132
R2	200	.2598471	.2422696	5.70e-06	.8869797	. di r(N)/_N
Pval	200	.139014	.2515517	2.29e-24	.9868725	.66

Remarks:

- average R^2 of 0.26 is high given that there is **no true** relationship between X_1 and X_2 .
- 66% of the 200 p-values are less than 0.05!

Remarks

- **standard asymptotic distribution theory does not apply** with nonstationary variables
- t test of $H_0 : \beta = 0$ is **not normally** distributed even asymptotically.
- standard results for OLS do not hold.

Most economic variables are non-stationary $I(1)$ variables:

- In general, regression models for non-stationary variables give spurious results.
- Only exception is if the variables are **cointegrated**.

Cointegration

Overview

- Genuine relationships arise only when the time series are cointegrated.
- **Cointegrated** time series are series which are:
 - non-stationary
 - but move together over time (common stochastic trend)
 - in this case we talk about **cointegrating regressions**.
- How can we detect if regression is spurious?
- How can we test if variables are cointegrated?

Example:

Drunk walking her dog on a lead (**Murray 1994**).

Linear Combination of Non Stationary Series

- In most cases, a linear combination of two $I(1)$ series will also be $I(1)$.
- In general, if $X_{i,t} \sim I(d_i)$ for $i = 1, 2, \dots, k$ so that there are k variables each integrated of order d_i , and:

$$z_t = \sum_{i=1}^k \alpha_i X_{i,t} \quad \text{then} \quad z_t \sim I(\max d_i)$$

- Rearranging, we get:

$$X_{1,t} = \sum_{i=2}^k \beta_i X_{i,t} + z'_t$$

where $\beta_i = \frac{-\alpha_i}{\alpha_1}$ and $z'_t = \frac{z_t}{\alpha_1}$, $i = 1, 2, \dots, k$

- The disturbance z'_t will have **undesirable properties**: nonstationary and autocorrelated if all of the X_i are $I(1)$.

Cointegration

In economic terms:

- two variables are cointegrated if they are bound by some **long-run or equilibrium** relationship
- In the short-run there may be **disequilibrium** but any deviation from their equilibrium relationship **must be temporary**.

Cointegration: Formal definition

Let w_t be a $k \times 1$ vector of variables, then the components of w_t are integrated of order (d, b) if:

- All components of w_t are $I(d)$
- There is at least one vector of coefficients α such that:

$$\alpha' w_t \sim I(d - b)$$

If $d = b = 1$, then a set of variables is cointegrated if a linear combination of them result in a stationary series.

Cointegration

Consider a regression model:

$$y_t = \beta x_t + u_t$$

where y_t , x_t are both $I(1)$.

We would typically expect $u_t = y_t - \beta x_t \sim I(1)$ regardless of the value of β .

However, there **may be** a β such that u_t (i.e. the deviation from the equilibrium) is stationary:

$$u_t = y_t - \beta x_t \sim I(0)$$

i.e. variables share a common trend. This is a **linear combination** which connects the variables in the long run.

We say y_t and x_t are **cointegrated** if $u_t \sim I(0)$ and call β (or any multiple of it) a cointegrating vector.

Examples of possibly cointegrating relationships

- Interest rates of bonds of different maturities share common trend
- Personal consumption expenditure and disposable personal income
- Corporate profits and dividends
- Money demand, income and interest rates
- Co-movements in asset prices
- Spot or future prices for a given commodity or asset
- Ratio of relative prices and exchange rates

Testing for Cointegration

Error Correction Mechanism

If two variables Y and X are cointegrated, then there is a steady relationship between them in levels.

What is the statistical mechanism that keeps these variables moving together?

- Suppose X_t is a RW:

$$X_t = X_{t-1} + u_t$$

and Y_t and X_t are cointegrated in their levels:

$$Y_t = \beta X_t + \epsilon_t$$

- According to **Granger's representation theorem**, two variables X and Y are cointegrated if and only if they have an **Error Correction Mechanism (ECM)** representation.

Error Correction Mechanism

- Simplest ECM representation:

$$\Delta Y_t = \gamma \Delta X_t - \alpha \underbrace{(Y_{t-1} - \beta X_{t-1})}_{\text{error correction term}} + e_t$$

- Provided that Y_t and X_t are cointegrated with cointegrating coefficient β , then $Y_{t-1} - \beta X_{t-1}$ will be $I(0)$ even though Y_t and X_t are both $I(1)$.
- Thus, it is valid to use OLS and standard procedures for statistical inference on ECM.
- ECM is also referred to as **equilibrium correction model**.

Engle-Granger 2-step Method

EG 2-Step Method

Two step method proposed by Engle and Granger (1987).

First Step

- Ensure that all the individual variables are $I(1)$.
- Estimate the cointegrating regression using OLS:

$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

- Estimate parameter values, cannot perform any inferences on the coefficient estimates.
- Test to see if the residuals are stationary:

$$\hat{u}_t = Y_t - \hat{\beta}_1 - \hat{\beta}_2 X_t$$

- OLS residuals \hat{u}_t are a **measure of disequilibrium**.

EG - First Step

- The DF or ADF test regression would be:

$$\Delta \hat{u}_t = \gamma \hat{u}_{t-1} + \varepsilon_t, \quad \text{DF}$$

$$\Delta \hat{u}_t = \gamma \hat{u}_{t-1} + \delta \Delta \hat{u}_{t-1} + \varepsilon_t, \quad \text{ADF}$$

- Null hypotheses: $H_0 : \gamma = 0$, \hat{u}_t is non-stationary and X_t and Y_t are **not** cointegrated.
- Alternative hypothesis: $H_0 : \gamma < 0$, \hat{u}_t is stationary and X_t and Y_t are cointegrated.
- Test statistic:

$$t_{DF-EG} = \frac{\hat{\gamma}}{SE(\hat{\gamma})}$$

EG - First Step

- If H_0 is rejected, X_t and Y_t are cointegrated \rightarrow Step 2.
- If there is insufficient evidence to reject H_0 , estimate a model containing only first differences.

Remarks:

Under H_0 , X_t and Y_t are **not** cointegrated i.e. under H_0 the relationship is **spurious** \rightarrow use the appropriate critical values for the cointegration test.

Critical values for EG-ADF statistic

When there are **two** cointegrated variables i.e. a **single** regressor:

Critical Values for Engle-Granger ADF statistic:

Model	1%	5%	10%
No constant or trend	-3.39	-2.76	-2.45
With constant, no trend	-3.96	-3.37	-3.07
Both constant and trend	-3.98	-3.42	-3.13

Source: Hamilton 1994, Table B.9

EG 2-Step Method

Second Step

In its simplest form,

$$\Delta Y_t = \phi_0 + \theta_0 \Delta X_t + \delta \hat{u}_{t-1} + u_t$$

(can add lag values of Δy_t and Δx_t to rhs)

The term $\hat{u}_{t-1} = Y_{t-1} - \hat{\alpha} - \hat{\beta} X_{t-1}$ is the **error** or **disequilibrium** in the previous period.

- Deviations from equilibrium are corrected each period
- Only $I(0)$ variables, so standard statistical inference valid.

Remarks

- Coefficient δ measures the **speed of adjustment** of Δy_t to the error (deviation from long run equilibrium) in the previous period.
- δ must be < 0 and significant in the presence of cointegration.
- if **error in the previous period is positive**, $\hat{u}_{t-1} > 0$, so that, $Y_{t-1} > \hat{\alpha} + \hat{\beta}X_{t-1}$, then Y_t should fall and ΔY_t should be negative.
- if **error in the previous period is negative**, $\hat{u}_{t-1} < 0$, so that, $Y_{t-1} < \hat{\alpha} + \hat{\beta}X_{t-1}$, then Y_t should rise and ΔY_t should be positive.
- $\delta = 0$ would mean no cointegration

Remarks

- A model with only differenced variables (without the error correction term) would throw useful long-run information away as it would be solely concerned with short-run movements.
- However, if the variables are $I(1)$ but not cointegrated then a model in the first differences of the variables is the way to avoid the spurious regression problem.

Limitations of Engle-Granger 2-step Method

- Usual finite sample problem of **lack of power** in unit root and cointegration tests.
- There could be **simultaneous equation bias** if the causality between Y and X runs in both directions.
- It is not possible to perform any **hypothesis tests** about the actual cointegrating relationship in step 1.
- There may be more than one cointegrating relationships.

Next

VAR Modelling

References

- Engle, Robert F, and Clive WJ Granger. 1987. "Co-Integration and Error Correction: Representation, Estimation, and Testing." *Econometrica: Journal of the Econometric Society*, 251–76.
- Granger, Clive WJ, and Paul Newbold. 1974. "Spurious Regressions in Econometrics." *Journal of Econometrics* 2 (2): 111–20.
- Murray, Michael P. 1994. "A Drunk and Her Dog: An Illustration of Cointegration and Error Correction." *The American Statistician* 48 (1): 37–39.

Vector Auto-Regression (VAR)

EC902: Econometrics A

Subham Kailthya

University of Warwick

Introduction to VAR

VAR as a System of Equations

- We have looked at **univariate models** i.e. models for single time series. Only **one** endogenous variable.
- We may be interested in the interaction of **several endogenous** time series.
 - e.g. inflation, output, employment influence and interact - affect future paths of each other.
- VAR is a **system** regression model i.e. there is more than one dependent variable.
- VAR extends univariate AR to a multivariate setting
- Advocated as an alternative to large-scale simultaneous equations structural models.

Bivariate VAR

Consider two endogenous variables y_{1t} and y_{2t} – current value depends on different combinations of p lags of both variables plus an error terms:

$$\begin{aligned}y_{1t} = & \beta_{10} + \beta_{11}y_{1t-1} + \dots + \beta_{1p}y_{1t-p} + \\& \alpha_{11}y_{2t-1} + \dots + \alpha_{1p}y_{2t-p} + u_{1t} \\y_{2t} = & \beta_{20} + \beta_{21}y_{2t-1} + \dots + \beta_{2p}y_{2t-p} + \\& \alpha_{21}y_{1t-1} + \dots + \alpha_{2p}y_{1t-p} + u_{2t}\end{aligned}$$

where $E(u_{it}) = 0$ is a white noise (serially uncorrelated or independent) with time invariant covariance matrix Σ .

For a bivariate VAR, $\text{cov}(u_{1t}, u_{2t}) = \sigma_{12}$ for $t = s$ and 0 otherwise.

VAR Benefits

- Flexible and easy to generalize
- Extend to include moving average errors – multivariate ARMA (VARMA).
- Expand to include n endogenous variables.

VAR History

- Concept of a **general equilibrium**
 - everything in the economy is related to everything else.
 - impossible to say which variable is exogenous.
- Cowles Commission approach
 - set up large systems of equations to model the economy
 - too many equations, unidentified parameters
 - Solve by:
 - arbitrarily setting some parameters to zero
 - estimate the component equations one-by-one (biased, Lucas (1976) critique)
- Sims (1980) introduced **vector autoregression** to model the interaction among several variables without imposing arbitrary assumptions on the data.
- Sims and Sargent were awarded the Nobel Prize in 2011 for their “empirical research on cause and effect in the macroeconomy” [\[link\]](#)

Understanding VAR Models

Vector Autoregressive Models

- Extension of the univariate autoregressive models.
- Each variable depends on its own lags and the lags of every other variable in a VAR.

$$\text{AR}(1) : \quad y_t = \phi_1 y_{t-1} + \varepsilon_t$$

$$\text{VAR}(1) : \quad \mathbf{y}_t = \boldsymbol{\Phi}_1 \mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t$$

where \mathbf{y}_t is an $n \times 1$ vector of variables, and $\boldsymbol{\Phi}_1$ is a matrix of coefficients of dimension $n \times n$.

VAR(1) with Two Variables

VAR(1) with $n = 2$ variables, y_{1t} and y_{2t} , $t = 1, 2, \dots, T$.

$$y_{1t} = \phi_{11}y_{1t-1} + \phi_{12}y_{2t-1} + \epsilon_{1t}$$

$$y_{2t} = \phi_{21}y_{1t-1} + \phi_{22}y_{2t-1} + \epsilon_{2t}$$

In matrix form:

$$\mathbf{y}_t = \Phi_1 \mathbf{y}_{t-1} + \boldsymbol{\epsilon}_t$$

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix} \begin{pmatrix} y_{1t-1} \\ y_{2t-1} \end{pmatrix} + \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{pmatrix}$$

VAR(2) with Two Variables

VAR(2) with $n = 2$ variables:

$$\mathbf{y}_t = \Phi_1 \mathbf{y}_{t-1} + \Phi_2 \mathbf{y}_{t-2} + \epsilon_t$$

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{bmatrix} \phi_{1,11} & \phi_{1,12} \\ \phi_{1,21} & \phi_{1,22} \end{bmatrix} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \end{pmatrix} + \begin{bmatrix} \phi_{2,11} & \phi_{2,12} \\ \phi_{2,21} & \phi_{2,22} \end{bmatrix} \begin{pmatrix} y_{1,t-2} \\ y_{2,t-2} \end{pmatrix} + \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{pmatrix}$$

Extended to p lags:

$$\mathbf{y}_t = \Phi_1 \mathbf{y}_{t-1} + \Phi_2 \mathbf{y}_{t-2} + \dots + \Phi_p \mathbf{y}_{t-p} + \epsilon_t$$

This is a VAR(p) model, because there are p lags of each variable on the right hand side.

VAR(p) with n Variables

$$\mathbf{Y}_t = \Phi_1 \mathbf{Y}_{t-1} + \dots + \Phi_p \mathbf{Y}_{t-p} + \boldsymbol{\epsilon}_t$$

where,

$$\mathbf{Y}_t = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \Phi_i = \begin{bmatrix} \phi_{i,11} & \phi_{i,12} & \dots & \phi_{i,1n} \\ \phi_{i,21} & \phi_{i,22} & \dots & \phi_{i,2n} \\ \vdots & \vdots & & \vdots \\ \phi_{i,n1} & \phi_{i,n2} & \dots & \phi_{i,nn} \end{bmatrix}, \quad \boldsymbol{\epsilon}_t = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

where,

$$E(\boldsymbol{\epsilon}_t) = 0, \text{ and } E(\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_s') = \begin{cases} \Sigma & , t = s \\ 0 & , t \neq s \end{cases}$$

VAR(p) with n Variables

$$\mathbf{y}_t = \Phi_1 \mathbf{y}_{t-1} + \Phi_2 \mathbf{y}_{t-2} + \dots + \Phi_p \mathbf{y}_{t-p} + \epsilon_t$$

Represent it compactly as:

$$\Phi(L) \mathbf{Y}_t = \epsilon_t$$

where Φ is a matrix polynomial in the lag operator i.e.

$$\Phi(L) \equiv I_n - \Phi_1(L) - \dots - \Phi_p(L)$$

For a VAR to be stable:

$$\det(I_n - \Phi_1 z - \dots - \Phi_p z^p) \neq 0$$

for $|z| \leq 1$.

Important Features

- No current values of the variables are included on right hand side of any of the VAR equations.
- Coefficients can be estimated by applying OLS to each equation (as all right hand side variables are predetermined)
- Only use **small scale VARs**, as the number of coefficients increase with n (number of variables) and p (lag order).

Number of Coefficients to Estimate

Examples

- $n = 2, p = 2$ lags will have:
 - in each equation: $1(\text{constant}) + 2 \times 2 = 5$ coefficients
 - in total for 2 equations: $5 \times 2 = 10$ coefficients
- $n = 3, p = 2$ lags will have:
 - in each equation: $1(\text{constant}) + 3 \times 2 = 7$ coefficients
 - in total for 3 equations: $7 \times 3 = 21$ coefficients
- $n = 3, p = 3$ lags will have:
 - in each equation: $1(\text{constant}) + 3 \times 3 = 10$ coefficients
 - in total for 3 equations: $10 \times 3 = 30$ coefficients

Number of coefficients in VAR: $n + pn^2$

Issues in VAR Modelling

- Which variables to include and how many. Variables are selected according to the relevant economic model.
- Lag lengths? There are tests (likelihood ratio test) and various selection criteria (AIC, SC) to select the appropriate lag length
- Levels of differences of the variables:
 - **levels**, if all variables are **stationary** i.e. $I(0)$.
 - **first differences** if the variables have a unit root i.e. they are $I(1)$ and they are **not** cointegrated.

VAR in First Differences

If y_{1t} and y_{2t} are both $I(1)$ but **not** cointegrated, and suppose that including 2 lags is appropriate, a VAR in first differences:

$$\Delta y_{1t} = \phi_{11}\Delta y_{1t-1} + \phi_{12}\Delta y_{2t-1} + \epsilon_{1t}$$

$$\Delta y_{2t} = \phi_{21}\Delta y_{1t-1} + \phi_{22}\Delta y_{2t-1} + \epsilon_{2t}$$

Remarks

- Even a small VAR contains many parameters
- Difficult to interpret the individual estimated parameters due to multicollinearity
- Ways to summarize the information contained in the estimated parameters in a VAR:
 - **Granger causality** tests
 - **Impulse Response Functions (IRF)**

Vector Error Correction Models (VECM)

Vector Error Correction Model (VECM)

- If y_{1t} and y_{2t} are both $I(1)$ and **cointegrated**, estimate VECM
- Multivariate counterpart of the Error Correction Model

$$\begin{aligned}\Delta y_{1t} = & \phi_{11} \Delta y_{1t-1} + \phi_{12} \Delta y_{2t-1} + \\ & \alpha_{11}(y_{1t-1} - \hat{\beta}_0 - \hat{\beta}_1 y_{2t-1}) + \epsilon_{1t}\end{aligned}$$

$$\begin{aligned}\Delta y_{2t} = & \phi_{21} \Delta y_{1t-1} + \phi_{22} \Delta y_{2t-1} + \\ & \alpha_{21}(y_{1t-1} - \hat{\beta}_0 - \hat{\beta}_1 y_{2t-1}) + \epsilon_{2t}\end{aligned}$$

VECM is a VAR in differences with the lag of the error correction term included as an additional regressor.

VECM

- The equilibrium relationship between the variables y_{1t} and y_{2t} :

$$y_{1t} = \beta_0 + \beta_1 y_{2t} + u_t$$

- Error from the equilibrium is:

$$\hat{u}_t = y_{1t} - \hat{\beta}_0 - \hat{\beta}_1 y_{2t}$$

Remarks

- With more than two variables there can be multiple cointegrating relationships.
- In general, in an n -variable VAR there can be up to $(n - 1)$ cointegrating relationships.
 - In a three-variable VAR there can be up to 2 cointegrating relationships
- Stata command to determine the number of cointegrating relationships: `vecrank y1 y2 y3`
- Tests are performed using system methods (Johansen methods)

Granger Causality and IRFs

VAR Models

- Granger causality tests
 - tell us whether the variables are significantly related to each other through time
- Impulse response analysis
 - show how the variables react dynamically to shocks
- Forecasting
 - often better than forecasts from complex “traditional structural” (large scale) econometric models.

Granger Causality

- With several lags of the same variables, each estimated coefficient may not be statistically significant, possibly because of multicollinearity.
- But jointly they may be significant.
- Granger causality tests ask: Is the variable y_2 helpful in predicting y_1 after controlling for the presence of past observations on y_1 ?
 - If the answer is yes, then the variable y_2 **Granger causes the variable** y_1 .
- Granger causality has little to do with causality in a standard sense.
- These tests characterise temporal relationships in terms of **predictability**

Granger Causality

- Consider a VAR in matrix form (see slide 12):

$$\Phi(L) \mathbf{Y}_t = \boldsymbol{\epsilon}_t$$

- Partition the vector of endogenous variables into two parts: \mathbf{Y}_1 and \mathbf{Y}_2 and rewrite the system as:

$$\begin{bmatrix} \Phi_{11}(L) & \Phi_{12}(L) \\ \Phi_{21}(L) & \Phi_{22}(L) \end{bmatrix} \begin{bmatrix} \mathbf{Y}_{1t} \\ \mathbf{Y}_{2t} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\epsilon}_{1t} \\ \boldsymbol{\epsilon}_{2t} \end{bmatrix}$$

- If \mathbf{Y}_2 does not Granger cause \mathbf{Y}_1 implies that $\Phi_{12} = 0$. Thus, the system becomes:

$$\begin{bmatrix} \Phi_{11}(L) & 0 \\ \Phi_{21}(L) & \Phi_{22}(L) \end{bmatrix} \begin{bmatrix} \mathbf{Y}_{1t} \\ \mathbf{Y}_{2t} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\epsilon}_{1t} \\ \boldsymbol{\epsilon}_{2t} \end{bmatrix}$$

Testing a VAR for Stationarity

- Recall: a univariate autoregressive process is stationary (or it is a stable process) if all the roots of $\phi(z) = 0$ lie outside the unit circle.
- A VAR is stationary if all the roots of:

$$\Phi(z) = 0$$

lie outside the (complex) unit circle or, equivalently if the eigenvalues of the **companion matrix** (next slide) have modulus less than one.

Any VAR(p) can be written as VAR(1)

Consider a VAR(p):

$$\mathbf{Y}_t = \Phi_1 \mathbf{Y}_{t-1} + \dots + \Phi_p \mathbf{Y}_{t-p} + \epsilon_t$$

Define:

$$\tilde{\mathbf{Y}}_t = \begin{bmatrix} Y_t \\ Y_{t-1} \\ \vdots \\ Y_{t-p+1} \end{bmatrix}, \tilde{\Phi} = \begin{bmatrix} \Phi_1 & \Phi_2 & \dots & \Phi_{p-1} & \Phi_p \\ I_n & 0 & \dots & 0 & 0 \\ 0 & I_n & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & I_n & 0 \end{bmatrix}, \tilde{\epsilon}_t = \begin{bmatrix} \epsilon_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

where $\tilde{\mathbf{Y}}_t$ is an np element vector, $\tilde{\Phi}$ is an $np \times np$ **companion matrix**, and $\tilde{\epsilon}_t$ stacks the n -element vector of random disturbances on top of $n(p - 1)$ zeroes.

VAR(1):

$$\tilde{\mathbf{Y}}_t = \tilde{\Phi} \tilde{\mathbf{Y}}_{t-1} + \tilde{\epsilon}_t$$

Impulse Response Functions

Trace out the expected responses of current and future values of each of the variables to a shock in one of the innovations (or surprise movements) of the VAR equations.

Interpretation of impulse responses in a two-variable VAR

- The response of y_1 (at time $t + 1, t + 2, \dots$)
 - to shocks in y_1
 - to shocks in y_2
- The response of y_2 (at time $t + 1, t + 2, \dots$)
 - to shocks in y_1
 - to shocks in y_2

IRFs obtained by rewriting the VAR in moving average form (see slide 30).

(Similar to obtaining the MA representation of a univariate AR process)

Impulse response functions (IRFs)

- Learn about dynamic properties of VAR
- How does a unit innovation to a series affect it over time?

Example

Let $Y_1 = Y_2 = \dots = Y_{T-1} = 0$. $\epsilon_1 = \epsilon_2 = \dots = \epsilon_{T-1} = 0$. At time $= T$, realise a unit shock: $\epsilon_T = \sigma$, and then $\epsilon_{T+1} = \dots = 0$

- For a white noise process, $Y_t = \epsilon_t \Rightarrow Y_T = \sigma$, $Y_{T+1} = \epsilon_{T+1} = 0$ and $Y_{T+h} = \epsilon_{T+h} = 0 \Rightarrow$ no dynamics
- For an AR(1) process, $Y_t = \phi Y_{t-1} + \epsilon_t$, $|\phi| < 1 \Rightarrow$ $Y_T = \phi \cdot 0 + \sigma = \sigma$. $Y_{T+1} = \phi\sigma + 0 = \phi\sigma$, $Y_{T+h} = \phi^h\sigma \Rightarrow$ impulse response dampens out.

Impulse response functions (IRFs)

Rewrite **stable** VAR(p):

$$\mathbf{Y}_t = \boldsymbol{\alpha} + \Phi_1 \mathbf{Y}_{t-1} + \dots + \Phi_p \mathbf{Y}_{t-p} + \boldsymbol{\epsilon}_t$$

in moving average form:

$$\mathbf{Y}_t = \mathbf{v} + \sum_{i=0}^{\infty} \Psi_i \boldsymbol{\epsilon}_{t-i}$$

where $\mathbf{v} = (I_n - \Phi_1 - \dots - \Phi_p)^{-1} \boldsymbol{\alpha}$ and $\Psi_0 = I_n$

The Ψ_i are the simple IRFs \rightarrow trace out the impact of single random shock (holding all else constant).

Identifying Shocks in VAR

Identification of Shocks

Major technical difficulty in the interpretation of impulse response functions has been in the identification of the shocks.

Can the VAR residuals (known as innovations) be unambiguously identified as shocks to y_1 or as shocks to y_2 ?

In general:

- errors can be correlated with each other
- this complicates the identification of the nature of shocks
- and hence the interpretation of the impulse response functions

Need Identification Scheme

A number of methods exist, including:

- Variables ordering (orthogonalised IRFs, Cholesky decomposition)
- Structural VARs (used for causal inference)
 - ‘structural’ because they are used to model the underlying structure of the economy
 - need knowledge of what is exogenous and what is not.
 - use economic theory to impose ‘identifying’ restrictions (ways of recovering structural shocks from reduced form residuals)
 - better studied in the context of simultaneous equations (beyond the scope of this module!)

We will only briefly illustrate the variables ordering. See the Stata Time Series Manual for a more detailed discussion of the topic of identification and SVAR models and how to implement it in Stata.

Variables ordering and orthogonalised IRFs

- A widely used solution is to transform the innovations to produce a new set of orthogonal innovations (uncorrelated with each other)
- This is obtained by imposing a particular ordering to the variables in the VAR (Cholesky decomposition, Cholesky ordering).
- The ordering is usually suggested by economic theory.

For example, if we use the ordering: y_1, y_2, y_3 , we are assuming that:

- y_1 does not immediately react to shocks to y_2 and y_3
- y_2 only reacts immediately to shock to y_1 , but not to shocks to y_3 , and
- y_3 reacts immediately to shocks to y_1 and y_2 .

Remarks

- When the residuals are almost uncorrelated, the ordering of the variables will make little difference.
- Usual practice is to observe the sensitivity of the results to changing in the ordering of the variables.

Example: Suppose we would like to estimate a VAR with two policy variables (government spending) and other endogenous variables (output).

If the policy variable (government spending) is in first position followed by output:

```
var govspending output
```

The assumption is that:

- government spending does not respond contemporaneously (at time t) to shocks in the other variables.
- but output (and the other variables) respond contemporaneously to shocks in government spending

Remarks

If policy variable in last position:

```
var output govspending
```

Then, the assumption is that:

- output does not respond contemporaneously to shocks in the policy variable (sluggish responses of output to shocks in the policy variable)
- but policy variable responds contemporaneously to shocks in output.

Stata for VAR Modelling

- Obtain VAR estimates by using the Stata command: `var y1 y2 y3 y4, lags(1/2)`
- Choose the lag length by using `varsoc`
 - Too few lagged terms will lead to specification errors
 - Too many lagged terms will consume degrees of freedom and will create problems of collinearity
- Alternative criteria for finding **best** model include:
 - AIC: Akaike information criterion
 - SIC: Schwarz information criterion
 - LR: Likelihood ratio test The “best” fitting model is the one that **maximizes** the LR, or **minimizes** AIC, SIC
- Additional requirement is that the VAR **residuals are not autocorrelated**. After the VAR estimation compute the LM test for autocorrelation by using the command: `varlmar, mlag(4)`

Stata for VAR Modelling

- For **Granger causality tests** after the estimation of the VAR use:

```
vargranger
```

This will test restrictions for each equation in the VAR that coefficients for selected variable(s) all equal zero.

- For **impulse-response functions** use:

```
varbasic infln pcwage // estimate VAR model  
irf graph irf          // plot IRF  
// specify impulse and response variable  
irf graph irf, impulse(pcwage) response(infln)
```

- To compute **orthogonalised impulse-response functions** use:

```
irf graph oirf, impulse(pcwage) response(infln)
```

Forecasting

To obtain dynamic forecasts after VAR estimation use:

```
fcast compute dyn_ , steps(10)      // Obtain forecasts  
fcast graph dyn_                      // plot forecasts
```

Consider the VAR(p) model:

$$\mathbf{y}_t = \Phi_1 \mathbf{y}_{t-1} + \Phi_2 \mathbf{y}_{t-2} + \dots + \Phi_p \mathbf{y}_{t-p} + \epsilon_t$$

At the forecast origin T , the relevant information set now includes the vectors $\mathbf{y}_T, \mathbf{y}_{T-1}, \dots$

Forecasting

The optimal one-step-ahead forecast is given by:

$$\hat{\mathbf{y}}_{T+1} = E(\mathbf{y}_{T+1} | \mathbf{y}_T, \mathbf{y}_{T-1}, \dots, \mathbf{y}_1) = \Phi_1 \mathbf{y}_T + \Phi_2 \mathbf{y}_{T-1} + \dots + \Phi_p \mathbf{y}_{T-p+1}$$

The 2-step-ahead forecast can be obtained recursively:

$$\hat{\mathbf{y}}_{T+2} = \Phi_1 \hat{\mathbf{y}}_{T+1} + \Phi_2 \mathbf{y}_T + \dots + \Phi_p \mathbf{y}_{T-p+2}$$

Next

Panel Data

Cholesky decomposition

Not in the exam

- Decompose the correlation matrix Σ into the product of a lower triangular matrix L and its transpose.

$$\Sigma = LL'$$

Thus,

$$L^{-1}\Sigma L'^{-1} = I_n$$

- Use L^{-1} to convert ϵ_t to a vector of uncorrelated random shocks:

$$E[L^{-1}\epsilon_t(L^{-1}\epsilon_t)'] = L^{-1}E(\epsilon_t\epsilon_t')L'^{-1} = L^{-1}\Sigma L'^{-1} = I_n$$

Rewriting as MA representation:

$$Y_t = v + \sum_{i=0}^{\infty} \Psi_i LL^{-1}\epsilon_{t-i} = v + \sum_{i=0}^{\infty} \Xi v_{t-i}$$

where $\Xi = \Psi_i L$ and $v_t = L^{-1}\epsilon_t$

OIRFs

We obtained that:

$$v_t = L^{-1} \epsilon_t$$

Pre-multiply both sides by L to obtain:

$$L v_t = \epsilon_t$$

This can be written as:

$$\begin{bmatrix} \ell_{11} & 0 & \dots & 0 \\ \ell_{21} & \ell_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \ell_{n1} & \ell_{n2} & \dots & \ell_{nn} \end{bmatrix} \begin{bmatrix} v_{1t} \\ v_{2t} \\ \vdots \\ v_{nt} \end{bmatrix} = \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \\ \vdots \\ \epsilon_{nt} \end{bmatrix}$$

Consider an **orthogonalized one-unit shock** to y_{2t} i.e. $v_{2t} = 1$ and all the other elements of $v_t = 0$.

OIRFs with $v_{2t} = 1$

The first row in \mathbf{L} contains all zeroes except ℓ_{11} . Thus, v_{2t} has no impact on ϵ_{1t} .

In the second equation, the impact on y_{2t} from $v_{2t} = 1$ is:

$$\epsilon_{2t} = \ell_{21}v_{1t} + \ell_{22}v_{2t} = \ell_{21} \times 0 + \ell_{22} \times 1 = \ell_{22}$$

For the third and later equations we observe that the impact is:
 $\epsilon_{3t} = \ell_{32}$, $\epsilon_{4t} = \ell_{42}$ and so on.

Thus the **initial impact** of a **one unit shock to** v_{2t} affects
 $\epsilon_{2t}, \epsilon_{3t}, \dots$ and y_{2t}, y_{3t}, \dots but not ϵ_{1t} or ϵ_{2t} .

Dynamic impact: Impact on variable i of an orthogonalized shock
to equation j that occurred k periods in the past is given by the (i,j)
th element of $\boldsymbol{\Xi}_k = \Psi_k \mathbf{L}$.

References

- Lucas, Robert E. 1976. "Econometric Policy Evaluation: A Critique." In *Carnegie-Rochester Conference Series on Public Policy*, 1:19–46. North-Holland.
- Sims, Christopher A. 1980. "Macroeconomics and Reality." *Econometrica: Journal of the Econometric Society*, 1–48.

Panel Data Analysis

EC902: Econometrics A

Subham Kailthya

University of Warwick

References

- Gujarati and Porter, Ch. 16
- Wooldridge, Ch. 13 and Ch. 14
- Verbeek Ch. 10
- Stock and Watson, Ch. 10
- Baltagi, B. H. (2013) Econometric Analysis of Panel Data

Introduction to PDMs

Overview

- Contains repeated observations over the same units (individuals, households, firms, etc.) collected over multiple time periods.
- Allows researchers to specify and estimate more complicated and realistic models than a single cross section or a single time series.

Estimation:

- Repeatedly observe the same units → can no longer assume that observations are independent → adjust standard errors.
- Panel data sets can suffer from **missing observations** (e.g. due to panel attrition) → adjust standard analysis

Benefits of PDM

Why use panel data?

- To control for **individual heterogeneity** in a regression model.

Benefits of PDM

Why use panel data?

- To control for **individual heterogeneity** in a regression model.
- More informative data, more variability, less collinearity among the variables, more degrees of freedom and more efficiency.

Benefits of PDM

Why use panel data?

- To control for **individual heterogeneity** in a regression model.
- More informative data, more variability, less collinearity among the variables, more degrees of freedom and more efficiency.
- Better suited to study **dynamics of adjustment**.

Why use panel data?

- To control for **individual heterogeneity** in a regression model.
- More informative data, more variability, less collinearity among the variables, more degrees of freedom and more efficiency.
- Better suited to study **dynamics of adjustment**.
- Better able to identify and measure effects that are not detectable in pure cross-section or pure time-series etc.

Limitations of PDM

- Design and data collection problems.

Limitations of PDM

- Design and data collection problems.
- Distortion of measurement errors.

Limitations of PDM

- Design and data collection problems.
- Distortion of measurement errors.
- Selectivity problems:

Limitations of PDM

- Design and data collection problems.
- Distortion of measurement errors.
- Selectivity problems:
 - self-selectivity

Limitations of PDM

- Design and data collection problems.
- Distortion of measurement errors.
- Selectivity problems:
 - self-selectivity
 - non-response

Limitations of PDM

- Design and data collection problems.
- Distortion of measurement errors.
- Selectivity problems:
 - self-selectivity
 - non-response
 - attrition etc.

Basics

- Panel data is usually observed at **regular time intervals**.

Basics

- Panel data is usually observed at **regular time intervals**.
- Panel data can be:
 - **Balanced** - all individual units are observed in all time periods ($T_i = T$ for all i).
 - **Unbalanced** - $T_i \neq T$ for all i
 - estimator consistency requires that missingness is for random reasons - rules out sample selection bias.

- Panel data is usually observed at **regular time intervals**.
- Panel data can be:
 - **Balanced** - all individual units are observed in all time periods ($T_i = T$ for all i).
 - **Unbalanced** - $T_i \neq T$ for all i
 - estimator consistency requires that missingness is for random reasons - rules out sample selection bias.
- Datasets may be:
 - **Short panel** (large N , small T) – **our focus**
 - **Long panel** (small N , large T)
 - **Both** (large N , large T)

- Panel data is usually observed at **regular time intervals**.
- Panel data can be:
 - **Balanced** - all individual units are observed in all time periods ($T_i = T$ for all i).
 - **Unbalanced** - $T_i \neq T$ for all i
 - estimator consistency requires that missingness is for random reasons - rules out sample selection bias.
- Datasets may be:
 - **Short panel** (large N , small T) – **our focus**
 - **Long panel** (small N , large T)
 - **Both** (large N , large T)
- Model errors are likely to be correlated → microeconomics – cluster individuals over time.

Panel Data Models

Consider the model:

$$y_{it} = \alpha_{it} + \mathbf{x}'_{it}\beta_{it} + u_{it}, \quad i = 1, \dots, N, \quad t = 1, \dots, T$$

where,

- y_{it} is the dependent variable.
- α_{it} is the intercept.
- β_{it} is $K \times 1$ vector of coefficients.
- \mathbf{x}_{it} is $K \times 1$ vector of explanatory variables.
- u_{it} is the disturbance term.
- i indexes individuals, firms, etc. t indexes time.

Model is **not** estimable \rightarrow more parameters to estimate than observations.

Need restrictions on the extent to which α_{it} and β_{it} vary with i and t and on the behaviour of the error u_{it} .

Pooled Model

Pooled Model

- Pooled model is **most restrictive** and specifies **constant coefficients**.

$$y_{it} = \alpha + \mathbf{x}'_{it}\beta + u_{it}$$

- α, β are the same for all individuals ($i = 1, \dots, N$) and time periods ($t = 1, \dots, T$).
- u_{it} varies over individuals and time and captures all unobservable factors that affect y_{it} .
- **Assumption:** If $E(x_{it}, u_{it}) = 0$ (i.e. regressors x_{it} are uncorrelated with the error term u_{it}) and either $N \rightarrow \infty$ or $T \rightarrow \infty$ is sufficient for **consistency** of α and β .

Pooled Model

- But u_{it} is likely to be correlated over time for the i th individual
→ usual SE will be downward biased and **panel-robust standard errors** are to be used.
- NT correlated observations have less information than NT independent observations.
- Pooled OLS is inconsistent if the **fixed effects** model is appropriate.

$$y_{it} = \alpha + \mathbf{x}'_{it}\beta + (\alpha_i - \alpha + u_{it})$$

If α_i is correlated with \mathbf{x}_{it} , then $(\alpha_i - \alpha + u_{it})$ is correlated with \mathbf{x}_{it} .

Individual and Time Dummies

- Variant of pooled model that permits intercepts to vary across individuals and over time while slope parameters do not.

$$y_{it} = \alpha_i + \gamma_t + \mathbf{x}'_{it}\beta + u_{it}$$

$$y_{it} = \sum_{j=1}^N \alpha_j d_{j,it} + \sum_{s=2}^T \gamma_s d_{s,it} + \mathbf{x}'_{it}\beta + u_{it}$$

- N individual dummies** $d_{j,it}$ equal one if $i = j$ and equal zero otherwise and **$(T - 1)$ time dummies** $d_{s,it}$ equal one if $t = s$ and equal zero otherwise.
- \mathbf{x}_{it} does not include an intercept.

Fixed and Random Effects Models

Fixed effects Models

Consider the **individual-specific effects model**:

$$y_{it} = \alpha_i + \mathbf{x}'_{it}\beta + u_{it}$$

where u_{it} is iid over i and t .

- α_i are random variables that capture **unobserved heterogeneity**.
- time dummies may be included in \mathbf{x}_{it} .
- Assume **strong exogeneity** or **strict exogeneity**

$$E(u_{it}|\alpha_i, \mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}) = 0, \quad t = 1, \dots, T.$$

- Strong exogeneity rules out models with lagged dependent variables or with endogenous variables as regressors.

Fixed effects Models

For the model:

$$y_{it} = \alpha_i + \mathbf{x}'_{it}\beta + u_{it}, \quad u_{it} \sim iid(0, \sigma_u^2)$$

- FE treats α_i as an **unobserved random variable that is potentially correlated with x_{it}** .
- FE is a linear regression model where the intercept terms vary over the individual units.
- In short panels, FE model permits only **identification** of the marginal effects i.e. of β for only **time-varying** regressors but the individual effects and the conditional mean are not identified.

Random Effects Model

RE assumes that α_i are random variables that are distributed independently of the regressors (**random intercept model**).

$$\alpha_i \sim [\alpha, \sigma_\alpha^2], \quad u_{it} \sim [0, \sigma_u^2]$$

and both α_i and u_{it} are iid.

RE can be viewed as:

$$y_{it} = \mathbf{x}'_{it}\beta + \underbrace{(\alpha_i + u_{it})}_{\text{composite error}}$$

Random Effects Model

Assumptions on α_i and u_{it} imply:

$$\text{Cov}[(\alpha_i + u_{it}), (\alpha_i + u_{is})] = \begin{cases} \sigma_\alpha^2 & \text{if } t \neq s \\ \sigma_\alpha^2 + \sigma_u^2 & \text{if } t = s \end{cases}$$

RE model imposes that u_{it} is **equicorrelated** since

$$\text{Cor}[u_{it}, u_{is}] = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_u^2}, \quad \text{for } t \neq s$$

does not vary with the time difference $|t - s|$.

RE Model – Remarks

$$\rho_u = \text{Cor}[u_{it}, u_{is}] = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_u^2}, \quad \text{for } t \neq s$$

This is the intra-class correlation of the error.

Thus, RE model permits serial correlation in the model error.

$\rho_u \rightarrow 1$ if the random effect (σ_α^2) is large relative to the idiosyncratic effect (σ_u^2).

Panel Data Estimators

Between Estimators

- P-OLS: exploits variation over both time and cross-sectional units.
- Between estimators in short panels uses only the cross-sectional variation.
- Consider the individual specific model:

$$y_{it} = \alpha_i + \mathbf{x}'_{it}\beta + u_{it}$$

- Averaging over all years yields:

$$\bar{y}_i = \alpha + \bar{\mathbf{x}}'_i\beta + (\alpha_i - \alpha + \bar{u}_i), \quad i = 1, \dots, N.$$

- where, $\bar{y} = T^{-1} \sum_t y_{it}$, $\bar{u} = T^{-1} \sum_t u_{it}$, and $\bar{\mathbf{x}} = T^{-1} \sum_t \mathbf{x}_{it}$
- Between estimator is consistent if $\bar{\mathbf{x}}_i$ are independent of $(\alpha_i - \alpha + \bar{u}_i)$.

Within or Fixed Effects Estimator

Consider the individual specific model:

$$y_{it} = \alpha_i + \mathbf{x}'_{it}\beta + u_{it}$$

Subtracting the time averaged variables yield:

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \mathbf{x}_i)' \beta + (u_{it} - \bar{u}_i)$$

as α_i cancels out.

Stata fits:

$$(y_{it} - \bar{y}_i + \bar{\bar{y}}) = \alpha + (\mathbf{x}_{it} - \bar{\mathbf{x}}_i + \bar{\bar{\mathbf{x}}})' \beta + (u_{it} - \bar{u}_i + \bar{\bar{u}})$$

where $\bar{\bar{y}} = (1/N)\bar{y}_i$ is the **grand mean** of $y_{it} \rightarrow$ provides an intercept estimate α (the average of individual effects α_i).

Remarks - WE

- WE yields consistent estimates of β in the FE model.
- FE can be considered as **nuisance parameters** that can be ignored.
- If N is not too large, estimate WE by LSDV.
- But coefficients of time-invariant regressors are not identified in the within model.

Consider the model:

$$y_{it} = \alpha_i + \mathbf{x}'_{it}\beta + u_{it}, \quad u_{it} \sim iid(0, \sigma_u^2)$$

Rewrite this as:

$$y_{it} = \sum_{j=1}^N \alpha_j d_{ij} + \mathbf{x}'_{it}\beta + u_{it}$$

where,

$$d_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

Remarks - LSDV

- includes N dummy variables or $(N - 1)$ dummies if the intercept term is included.
- estimate $\alpha_1, \dots, \alpha_N$ and β by OLS.
- implied estimator for β is called **least squares dummy variable (LSDV)** estimator.
- numerically unattractive if N is large.

First Difference (FD) Estimator

In short periods, FD estimator measures the association between individual-specific one-period changes in regressors and individual specific one-period changes in the dependent variable.

Consider the model:

$$y_{it} = \alpha_i + \mathbf{x}'_{it}\beta + u_{it}, \quad u_{it} \sim iid(0, \sigma_u^2)$$

Lagging it one period yields:

$$y_{i,t-1} = \alpha_i + \mathbf{x}'_{i,t-1}\beta + u_{i,t-1}$$

Subtracting this from the above equation yields:

$$y_{it} - y_{i,t-1} = (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\beta + (u_{it} - u_{i,t-1})$$

where $i = 1, \dots, N$ and $t = 2, \dots, T$ as the α_i cancels out.

Remarks - FD

Rewrite FD as:

$$\Delta y_{it} = \Delta \mathbf{x}'_{it} \beta + \Delta u_{it}$$

- FD estimator is the OLS estimator in the last equation.
- Yields consistent estimates of β in the FE model.
- Coefficients of time-invariant regressors are not identified.
- FD is less efficient than WE for $T > 2$ if u_{it} is iid.

RE Estimator

Consider a RE model:

$$y_{it} = \mathbf{x}'_{it}\beta + (\alpha_i + u_{it})$$

where α_i and u_{it} are both iid.

Treat α_i as random variables independently and identically distributed over units. See appendix.

The error term is composed of:

- α_i is an individual specific component.
- u_{it} is an idiosyncratic component

RE Model is consistent and fully efficient if the RE model is appropriate but inconsistent if FE model is appropriate.

RE is consistent but inefficient if errors exhibit within-panel correlation → use cluster-robust SE.

Correlated RE Model

- Relax the assumption that the random effect (α_i) is purely random and uncorrelated with exogenous variables \mathbf{x}_i
- Assume that random effect is a linear function of \mathbf{x} and an error term.

Mundlak correction

Mundlak (1978) assumes

$$E(\alpha_i | x_{i1}, \dots, x_{iT}) = \bar{\mathbf{x}}'_{1i} \gamma$$

where $\bar{\mathbf{x}}_{1i}$ are the time averages of the subset of regressors that have within variation.

Individual specific effect:

$$\alpha_i = \bar{\mathbf{x}}'_{1i} \gamma + \eta_i$$

where η_i is an independent error. The RE model becomes:

$$y_{it} = \mathbf{x}'_{it} + \bar{\mathbf{x}}'_{it} \gamma + (\eta_i + u_{it})$$

Fixed versus Random Effects

FE versus RE

If the FE model is correct, pooled OLS and RE estimators are inconsistent while WE is consistent but less efficient (uses only within variation).

Hausman Test

$$H_0 : \text{RE Model} \quad [\alpha_i \sim iid(0, \sigma_\alpha^2), E(u_{it}, \alpha_i) = 0]$$

$$H_1 : \text{FE Model}$$

Let $\hat{\beta}_{FE}$ and $\hat{\beta}_{RE}$ be the fixed and random effects estimates of β .

General form of the test is based on the **Wald statistic**:

$$WD = (\hat{\beta}_{FE} - \hat{\beta}_{RE})'[\hat{V}(\hat{\beta}_{FE}) - \hat{V}(\hat{\beta}_{RE})]^{-1}[(\hat{\beta}_{FE} - \hat{\beta}_{RE})] \sim \chi_R^2$$

where R is the number of explanatory variables in the model excluding the constant, and \hat{V} is the estimate of the true covariance matrix.

Hausman Test

Interpretation

- Under H_0 , the RE and FE estimators are both consistent and should not be statistically different from each other \implies prefer the more efficient RE estimator (GLS)
- Under H_1 , the RE estimator is not consistent \implies prefer FE estimator.

Limitation

Hausman test is not compatible with clustered standard errors. How can we choose between FE and RE when standard errors are clustered?

Cluster-Robust Hausman Test

- Hausman test requires that RE estimator is efficient $\implies \alpha_i$ and u_{it} be iid.
- If there is significant difference between clustered standard errors and the default standard errors, this assumption is not valid.
- Approach:
 - bootstrap Hausman test
 - Test $H_0 : \gamma = 0$ in the auxiliary OLS regression:

$$y_{it} = \mathbf{x}'_{it}\beta + \bar{\mathbf{x}}'_{1i}\gamma + \nu_{it}$$

where \mathbf{x}_1 denotes only time varying regressors.

Common PDM Stata commands

Stata command	Description
Data summary	<code>xtset; xtdescribe; xtsum; xtdata</code>
Tabulate and Plot	<code>xttab; xtline</code>
Pooled OLS	<code>regress, vce(cluster id)</code>
Random effects	<code>xtreg, re vce(cluster id)</code>
Fixed effects	<code>xtreg, fe; xtreg, fe vce(cluster id)</code>
Between	<code>xtreg, be vce(bootstrap)</code>
First-differences	<code>regress (with differenced data)</code>
Static IV	<code>xtivreg; xtaylor</code>
Dynamic IV	<code>xtabond; xtdpdsys; xtdpd</code>
Unit root tests	<code>xtunitroot</code>
Cointegration tests	<code>xtcointtest</code>

Here, 'id' is the cluster variable.

Appendix I - RE Estimator (1/4)

Content in the appendix will not be assessed

- The random effects model can be written as:

$$y_{it} = \alpha + \mathbf{x}'_{it}\beta + w_{it}$$

$$w_{it} = v_i + u_{it}$$

where $u_{it} \sim (0, \sigma_u^2)$ and $v_i \sim iid(0, \sigma_v^2)$, $E(v_i u_{it}) = 0$ and $E(v_i x_{it}) = 0$ for all i and t .

- The covariance structure of w_{it} is:

$$E(w_{it}) = E(u_{it}) + E(v_i) = 0$$

$$\text{var}(w_{it}) = E(u_{it}^2) + E(v_i^2) + 2E(u_{it}v_i) = \sigma_u^2 + \sigma_v^2$$

$$\text{cov}(w_{it} w_{is}) = E[(v_i + u_{it})(v_i + u_{is})]$$

$$= E(u_{it}u_{is}) + E(u_{it}v_i) + E(u_{is}v_i) + E(v_i^2) = \sigma_v^2$$

RE Estimator (2/4)

- Let $w_i = (w_{i1}, w_{i2}, \dots, w_{iT})'$ be the $(T \times 1)$ vector of disturbances for the i th cross section.
- The $(T \times T)$ covariance matrix of w_i is:

$$\Omega_i = E(w_i w_i') = \begin{bmatrix} \sigma_u^2 + \sigma_v^2 & \sigma_v^2 & \sigma_v^2 & \dots & \sigma_v^2 \\ \sigma_v^2 & \sigma_u^2 + \sigma_v^2 & \sigma_v^2 & \dots & \sigma_v^2 \\ \sigma_v^2 & \sigma_v^2 & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ \sigma_v^2 & \sigma_v^2 & \dots & \dots & \sigma_u^2 + \sigma_v^2 \end{bmatrix}$$

RE Estimator (3/4)

- The correlation structure of the matrix remains constant over time:

$$\rho_{ts} = \frac{\text{cov}(w_{it} w_{is})}{\sqrt{\text{var}(w_{it}) \cdot \text{var}(w_{is})}} = \frac{\sigma_v^2}{\sigma_u^2 + \sigma_v^2}$$

- Combining all disturbances into the $(NT \times NT)$ composite disturbance vector $w = (w_1, w_2, \dots, w_N)'$, the full covariance matrix of w has the form:

$$\Omega = E(ww') = \begin{bmatrix} \Omega_1 & 0 & 0 & 0 \\ 0 & \Omega_2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \Omega_N \end{bmatrix}$$

RE Estimator (4/4)

Steps to estimate the parameters of the RE model:

1. Choose starting values for σ_u^2 and σ_v^2 .
2. Define the quasi-differenced parameter:

$$\lambda = 1 - \frac{\sigma_u^2}{\sqrt{\sigma_u^2 + T\sigma_v^2}}$$

and compute the quasi-differenced data:

$$\tilde{y}_{it} = y_{it} - \lambda \bar{y}_i$$

$$\tilde{x}_{it} = x_{it} - \lambda \bar{x}_i$$

where \bar{y}_i and \bar{x}_i are averaged over time.

3. The GLS estimator based on the starting values for (σ_u^2, σ_v^2) is obtained by estimating the following transformed equation by OLS:

$$\tilde{y}_{it} = \alpha(1 - \lambda) + \tilde{x}'_{it}\beta + \tilde{w}_{it}$$

where \tilde{w}_{it} is a disturbance term.

References

Mundlak, Yair. 1978. "On the Pooling of Time Series and Cross Section Data." *Econometrica: Journal of the Econometric Society*, 69–85.