Econometrics A

Handout 1: Introduction

Laura Mayoral

IAE and BGSE

Singapore, January 2019

Roadmap Session 1

1. Overview of the course

1.1 Goal

1.2 Requirements

1.3 Books

2. This class: 3 Important questions before you start an empirical analysis

2. 1. What is the causal relationship of interest? correlation versus causation

2. 2 What is your identification strategy?

2.3. What is your mode of statistical inference?

2.4. Quick preview of the course

1. This course

- Introduction to econometric analysis
- Goal: by the end of this course you should be able to
  - formulate a question
  - answer it using econometric techniques
  - use stata to prepare, describe and analyse a dataset
  - critically analyse other people's studies, assessing their weaknesses and their strengths

1.2. Requirements

- I assume you are familiar with

- Basic notions of probability and statistics:

  - Random variables;

  - Distributions (marginal, joint and conditional); Moments of random variables (mean, variance, correlations, conditional expectations).

  - Convergence of random variables: convergence in probability and distribution. LLN and CLT

  - Hypothesis testing: t-tests and F-tests

- Matrix algebra
  - Operations with matrices: sums, multiplications, inverse,
  - Basic concepts: transpose, rank of a matrix.

- Basic notions of real analysis
  - Integral of functions, derivatives.

- If you are not familiar with the topics above:
  - Greene, appendices A-D

1.3. Textbooks

*Jeffrey Wooldridge, Econometrics of Cross Section and Panel Data.

*Applied Econometrics using Stata, Cameron et al.

William H. Greene, Econometrics, 6th edition.

Joshua Angrist and Jorn Steffen Pischke, Mostly Harmless Econometrics.

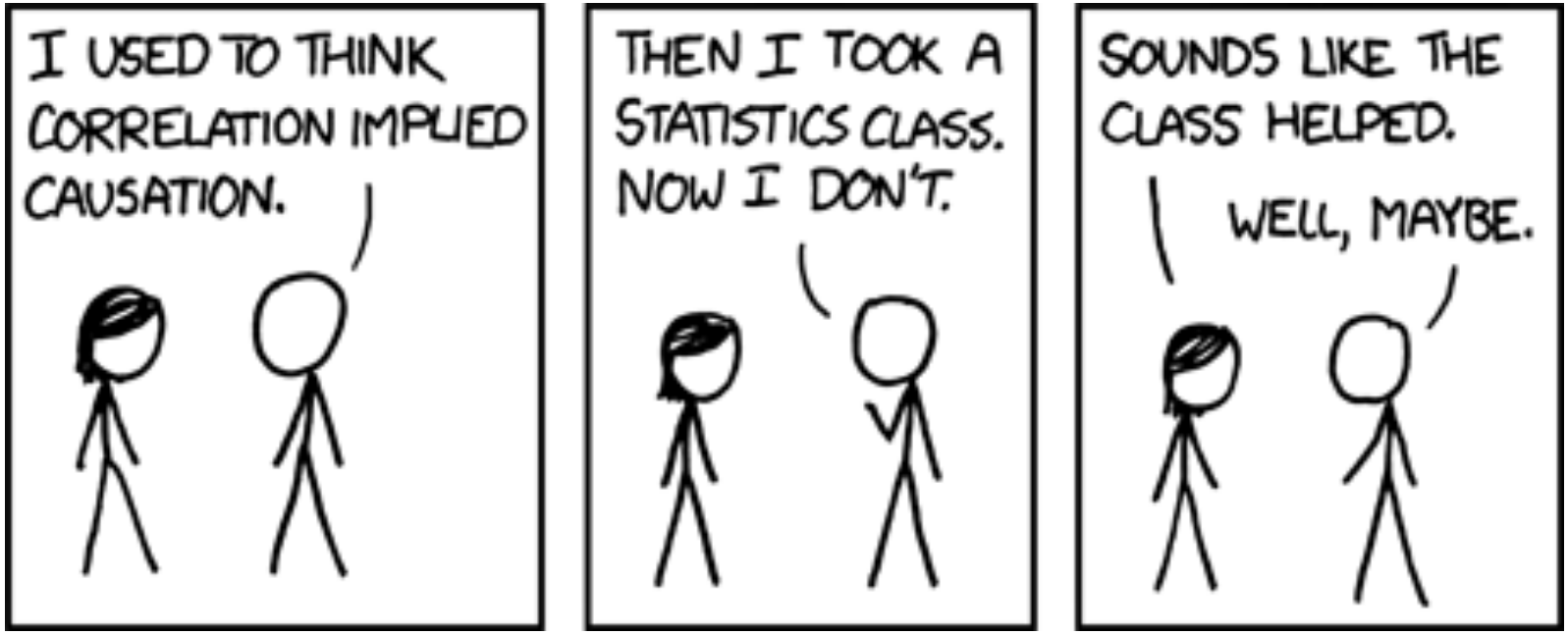
2. This class:

3 important questions to think about before you start an empirical analysis

2.1. What is the causal relationship of interest?

2.2 What is your identification strategy?

2.3. What is your mode of statistical inference?


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

2. 1. What is the causal relationship of interest?

- cause-effect relationships versus descriptive relationships (association among variables)

- Examples:

- What is the causal impact of class size on children's tests scores?

- What is the causal effect of an additional year of schooling on wages?

- What is the causal effect of democracy on development?

2. 1. What is the causal relationship of interest?

- cause-effect relationships versus descriptive relationships (association among variables)

- Examples:

- What is the causal impact of class size on children's tests scores?

- What is the causal effect of an additional year of schooling on wages?

- What is the causal effect of democracy on development?

- Causal relationship: involves to know what would happen in an alternative (or conterfactual) world.

- Correlation doesn't imply causality.

Statistical association, correlation and causality

- Association simply refers to statistical regularities among variables.

This means that knowing how a variable moves gives us information about the behaviour of another variable.

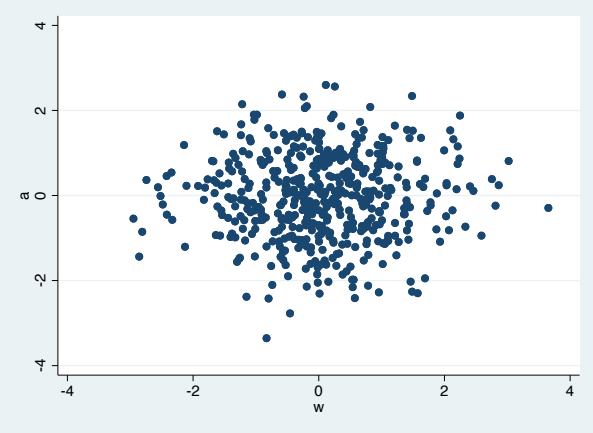
- Correlation is just one way two variables can be associated: linear association

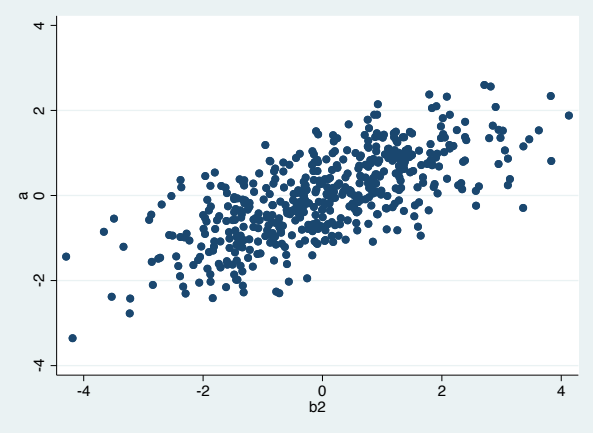
Statistical association, correlation and causality

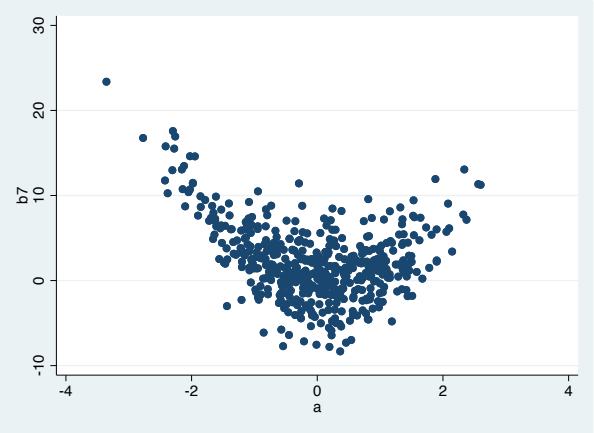
- Association simply refers to statistical regularities among variables.

This means that knowing how a variable moves gives us information about the behaviour of another variable.

- Correlation is just one way two variables can be associated: linear association







The perils of confounding correlation with causation

Correlation or causation?

- Health is negatively correlated with the number of days spent in hospital.
- Do hospitals kill patients?
- Shoe size is correlated with IQ.
- can we boost children reading scores using this information?
- Women earn less than men.
- Sign of discrimination?

More about causality: The Fundamental Problem of Econometrics, I

- Potential Outcomes Framework (Rubin Causality Model)

- Question: what is the effect of a treatment D on an outcome of interest Y?

- D=1: individual i is treated; D=0: individual i is not treated

- To measure the impact of D, treat individual i, measure Y if D=1, i.e., Y(D=1)

- Ideally: do not treat individual i, measure Y if D=0, i.e., Y(D=0)

- Ideally: compare Y(D=1) and Y(D=0) and you are done!

More about causality: The Fundamental Problem of Econometrics, I

- Potential Outcomes Framework (Rubin Causality Model)

- Question: what is the effect of a treatment D on an outcome of interest Y?

- D=1: individual i is treated; D=0: individual i is not treated

- To measure the impact of D, treat individual i, measure Y if D=1, i.e., Y(D=1)

- Ideally: do not treat individual i, measure Y if D=0, i.e., Y(D=0)

- Ideally: compare Y(D=1) and Y(D=0) and you are done!

- Obviously, this is not possible!

- Either Y(D=1) or Y(D=0) is unobserved!

Randomization as the Golden Benchmark

- Any way to solve this problem?

Randomization as the Golden Benchmark

- Any way to solve this problem?

- (Ideal) Solution: Randomization.

- select a sample of individuals (large enough).

- Randomly assign some individuals to treatment

- Compare the average value of Y among the treated and the non-treated.

- Average treatment effect: E[Y(D=1)-Y(D=0)]

Randomization as the Golden Benchmark

- Any way to solve this problem?

- (Ideal) Solution: Randomization.

- select a sample of individuals (large enough).

- Randomly assign some individuals to treatment

- Compare the average value of Y among the treated and the non-treated.

- Average treatment effect: E[Y(D=1)-Y(D=0)]

- Why is randomization so useful:

- It solves the selection problem

The Selection Problem

- Example:

- What is the impact on elderly people's health of using emergency rooms in hospitals ?

- In principle, we think that hospitals are useful!

- But, they also have a high concentration of bacteria, they can be dangerous for the elderly.

- To answer this question, we collect data on elderly patients (National Health Interview Survey), going and not going to the hospital (from Mostly Harmless Econometrics (MHE).

The Selection Problem,II

MHE, p. 13			
Group	Sample Size	Mean Health	Std. Error
Hospital	7.774	3.21	.014
No-hospital	90.049	3.93	0.003

- 3.21-3.93=-.72
- What can you conclude from here?
- not much
- People that go to the hospital are less healthy to start with
- Selection!

Selection bias

- The problem that selection generates is called selection bias

- It is the bias introduced by the selection of individuals, groups or data for analysis in such a way that proper randomization is not achieved.

- Thus, the sample obtained is not representative of the population intended to be analysed

More examples (what's wrong with them?)

- impact of government-subsidized training programs for disadvantage people. Then, you compare the average long-term employment status of the individuals that participated in those programs with individuals that didn't.

More examples (what's wrong with them?)

- impact of government-subsidized training programs for disadvantage people. Then, you compare the average long-term employment status of the individuals that participated in those programs with individuals that didn't.

- Non-experimental studies: worse effects for the trainees

More examples (what's wrong with them?)

- impact of government-subsidized training programs for disadvantage people. Then, you compare the average long-term employment status of the individuals that participated in those programs with individuals that didn't.

- Non-experimental studies: worse effects for the trainees

- impact of a training program within a firm. Then, you compare average productivity of individuals that followed the program with that of individuals that didn't follow the program.

More examples (what's wrong with them?)

- impact of government-subsidized training programs for disadvantage people. Then, you compare the average long-term employment status of the individuals that participated in those programs with individuals that didn't.

- Non-experimental studies: worse effects for the trainees

- impact of a training program within a firm. Then, you compare average productivity of individuals that followed the program with that of individuals that didn't follow the program.

- Non-experimental studies: might overestimate the impact of the program.

More examples (what's wrong with them?)

- impact of government-subsidized training programs for disadvantage people. Then, you compare the average long-term employment status of the individuals that participated in those programs with individuals that didn't.

- Non-experimental studies: worse effects for the trainees

- impact of a training program within a firm. Then, you compare average productivity of individuals that followed the program with that of individuals that didn't follow the program.

- Non-experimental studies: might overestimate the impact of the program.

- effect of class size on students' learning. To do that, you compare the average performance in an exam of kids coming from small/big classes.

- Non-experimental studies: no-effect of class size.

Randomization solves the selection problem

- The examples above share a common problem: individuals self-select themselves to the treatment.

- Randomization: individuals are assigned randomly to treatment

- By construction: the probability that individual i is treated is independent of potential outcomes!

Examples, II

- Randomly assign a training program to a group of disadvantaged people: since the treatment is randomly assign, it is no longer true that the people with the lowest earning potential self-select to the program

- Randomly assign students to classes of different size. Since the treatment is randomly assign, it is no longer true that kids assigned to smaller classes have learning difficulties.

- etc.

Experimental versus Observational studies

- Experimental studies: the researcher applies a treatment randomly to the experimental units

- Observational studies: the treatment that individuals receive is beyond the control of the researcher. She just observes and measures variables of interest.

- This course is about trying to estimate causal effects in observational studies.

2.2 What is your identification strategy?

Identification Strategy:

- the manner in which a researcher uses observational data (i.e., data not generated by a randomized trial) to approximate a real experiment.

- loosely speaking, identification looks at whether you could draw the right conclusions about the causal relationship of interest when using observational data (rather than randomized data, i.e., data generated in an experiment).

- Typically, a causal relationship between two variables involves the estimation of one (or more) parameter(s). Identification refers to the possibility of obtaining estimates with "good" properties [assuming that all the needed data is available]

- For each particular problem, we should design a specific identification strategy

2.3. What is your mode of statistical inference?

- This point refers to how closely we are able to approximate the causal effect of interest.

- Suppose that you are interested in one parameter (that captures a causal relationship between two variables). Assume further that your parameter is identified. Now we are considering how closely you can estimate this parameter given the available data.

- Issues we need to consider.

  - The population to be studied and the sample to be used

  - Assumptions made when constructing the standard errors of our estimators.

Population and Sample

- Population: a set of similar items or events which is of interest for some question or experiment.

- Typically, the population is very large, making a census or a complete enumeration of all the values in the population either impractical or impossible.

- Sample: is a set of data collected and/or selected from a statistical population by a defined procedure. The elements of a sample are known as sample points, sampling units or observations

- An unbiased or representative sample is a set of objects chosen from a population using a selection process that does not depend on the properties of the objects.

- For example, a representative sample of Australian men taller than 2m might consist of a randomly sampled subset of 1% of Australian males taller than 2m.

- If samples are representative of a population, one can make inferences or extrapolations from the sample to the population.

- Examples:

- Question: Impact of an educational reform in primary school.

  - population: all kids enrolled in primary school.

  - sample: a randomly sample subset containing 5% of all students.

- Question: Impact of a public polity consisting in giving mosquito nets for free in malaria affected areas.

  - population: everybody leaving in those areas

  - sample: individuals living in randomly selected villages in those areas.

Random sampling assumption.

Throughout the course we will assume that

- A population model has been specified.

- An independent and identically distributed (i.i.d.) sample can be drawn from this population.

Random sampling assumption.

Throughout the course we will assume that

- A population model has been specified.
- An independent and identically distributed (i.i.d.) sample can be drawn from this population.
- What is the advantage of this approach?

- The random sampling assumption allows us to separate the sampling assumptions from the assumptions made on the population model.

- In particular, it allows us stating all assumptions in terms of the population, (rather than the traditional approach of stating assumptions in terms of data matrices).

Summarizing (loosely speaking!)

- Identification is about whether you could obtain the right answer in the best case scenario (=when you are not limited by the fact that you are observing a sample, not the population)

- Studies of statistical inference seek to characterize the generally weaker conclusions that can be drawn from a finite number of observations.

- As the course develops, we'll see many examples of identification strategies (and learn to identify their threats) and we'll learn different ways to estimate the relevant parameters and to make inference (computing standard errors and for those parameters

Example 1

- Consider a lab experiment with 9 rats, you want to test the impact of some medication and you randomly assign the rats to a treatment group and a control group.

- Identification: Can you identify the effect of the medication on the rats using the random assignment?

- Inference: With 9 rats, can you say anything about the effectiveness of the medication?

Example 2

- you have data on class size and grades of the universe of students of singapore.

- Identification: can you conclude what is the relationship between class size and grades just by looking at this information?

- Inference: how big your confidence errors will be?

The rest of this course, I

- In econometrics there are three types of data:

- Cross-sectional data: data relative to n individuals gathered in the same point in time.

- Time series data: Data on one variable (or a few variables) observed over time, $t = 1, \ldots, T$

- Panel data: Data that combines both dimensions, n individuals observed over time $t = 1, \ldots, T$

  - Micro panels: n is very large with respect to T;

  - Macro panels: n and T are comparable in size;

The rest of this course, II

- Each type of data typically requires specific techniques.

- This course is an introduction to cross-sectional data and to panel data.

- Time series will be left for a more advance course

The rest of this course, II

- Each type of data typically requires specific techniques.

- This course is an introduction to cross-sectional data and to panel data.

- Time series will be left for a more advance course

- Contents

- Methods for cross-sectional data. Ordinary Least squares
  - Estimation, inference
  - problems, how to solve them (Instrumental variable estimation).

- Methods for Panel Data

Takeaway from this session

- This course is about estimating the causal effect of a treatment
- Do not confuse causality with correlation (or, more generally, statistical association among variables)
- Causality involves comparing the outcome with a counterfactual (what would happen if no treatment, everything else equal)
- Correlation simply implies co-movement, doesn't tell us anything about causality
- Randomisation of the treatment solves the selection problem
- Randomisation is not always possible: experimental vs. observational data.
- Selection bias: bias that arises when the sample is not representative of the underlying population

- Identification strategy: the way the researcher uses observational data as if it was coming from an experiment

- Statistical inference: assumptions we'll made to compute standard errors to our estimators

- Other key concepts

- Population

- Sample

- Representative samples; i.i.d samples (i.i.d.=independent and identically distributed)

- Readings:

Mostly Harmless, chapters 1 and 2

Writing tips (Cochrane, website).