Applied Microeconometrics I

Lecture 1: Introduction

Stefano Lombardi

Aalto University

September 6, 2023
Lecture Slides

Goals

- **Aim**: to provide the basic tools to do causal inference in empirical analysis.
- This course has a **practical flavor**
  - emphasis is not on proofs but on intuitions and on applications
  - however, we will also go through key derivations
- I will **presume** that you already know **basic Econometrics**
  - to a minimum, you should be (very) familiar with OLS
  - basic expected value operations
- We will be more concerned in general with **consistency** (convergence in probability to true parameter value) that with efficiency (standard errors).
- The **next course** (Applied Microeconometrics II) by Ciprian Domnisoru will touch upon advanced topics

Schedule

- The course begins on September 6 and lasts until October 12
  - important to review materials between lectures
  - exercise sessions will help
  - I will start each lecture with a recap of what was done last time
- **Lectures** will be held on site:
  - Tue: 12:15-13:45
  - Wed: 16:15-18:00 (discussion of exercises)
  - Thu: 12:15-13:45
- **Software tutorials**, also on site:
  - Wed 6/9: 16:15-18:00 Stata (with Atte Pudas)
  - Wed 13/9: 16.15-18:00 R (with Ramin Izadi)
- **Exceptions**:
  - This first lecture is on Wed, not Tue
  - Thu 14 and Tue 19 will be taught by Cristina Bratu

A few things you need to know

- **Office hours**:
  - Send an email (stefano.lombardi@vatt.fi) to fix an appointment
  - For issues related with problem sets/STATA please contact the teaching assistant Atte Pudas: atte.pudas@aalto.fi
- Please, **provide feedback** during the lectures!
- Please provide feedback via the **course evaluation**
  - Things that worked
  - Things that did not work
  - Things that you would like to be covered

Course Requirements

- **Evaluation**:
  - Five problem sets (50%), final exam (50%).
  - To pass the course a passing grade in the exam is required.
- **Exam** (Please check if you need to register separately):
  - Date: 17/10, 13:00-16:00
  - Retake: 12/12, 13:00-16:00
- **PhD students**, please talk with me after class

Course Requirements

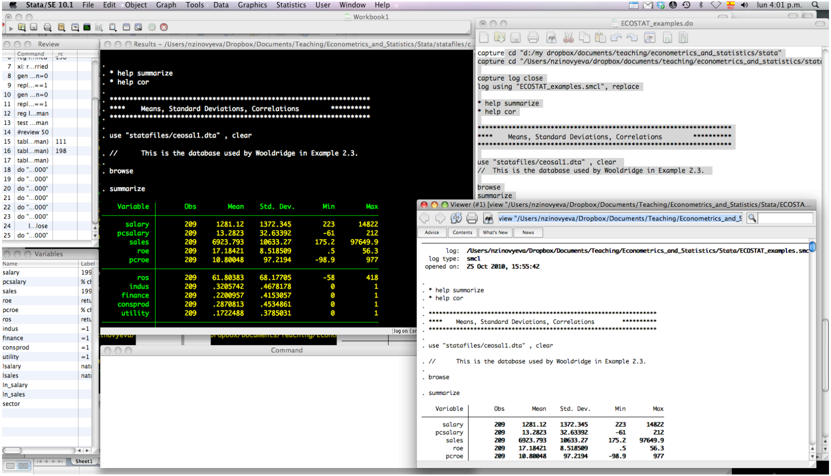
- Problem sets:
  - Available on **Fridays** (MyCourses).
  - **Due following Friday** at 23:59 and discussed on Wednesdays.
  - Please note that **deadlines are strict**.
  - However, lowest graded problem set will not be counted.
- What do I expect from the problem sets?
  - NAME and STUDENT NUMBER at the beginning.
  - Submit a **pdf** (compiled from Word or Latex)
  - For STATA/R questions, attach log file at end of document.
  - Use whichever software you are more comfortable with.
  - Be **concise**. Use examples to make us understand
  - **Effort will be rewarded**
  - **Cheating** will not be tolerated

Course material

- Lectures
- Slides
  - Available at MyCourses a few hours before each lecture
- Main textbook:
  - Angrist, J. and J.S. Pischke (2014), *Mastering Metrics: The Path from Cause to Effect*, Princeton University Press.
  - Or the earlier version (for PhD students): Angrist, J. and J.S. Pischke (2009), *Mostly Harmless Econometrics*, Princeton University Press.
- To refresh basics of Econometrics:
  - Wooldridge, J. (2003), *Introductory Econometrics: A Modern Approach*. South-Western College Publishing.
- Papers discussed (on MyCourses, updated continuously)

Statistical software

- Problems will require the use of some statistical package, but this is **not** a course on programming.
- I use STATA for some things, R and Matlab for others.
- Why **STATA**?
    - Easy to start with
    - Economists/social scientists mostly use STATA
    - Drawback: proprietary software
- What about **R**?
    - Steeper learning curve
    - powerful tools for exploratory data analysis (e.g., *ggplot*)
    - Matrix operations/coding packages *much* easier
    - Huge community to solve coding issues (*Stack Overflow*)
- Software tutorials
    - Wed 6/9: 16:15-18:00 **Stata**
    - Wed 13/9: 16.15-18:00 **R**



```
     Command      _rc
  7  xi: r...rried
  8  gen ...n=0
  9  repl...==1
 10  gen ...n=0
 11  repl...==1
 12  reg l...man
 13  test ...man
 14  #review 50
 15  tabl...man)  111
 16  tabl...man)  198
 17  tabl...man)
 18  do "...000"
 19  do "...000"
 20  do "...000"
 21  do "...000"
 22  do "...000"
 23  do "...000"
 24       l...lose
 25  do "...000"
```

```
Name        Label
salary      199
pcsalary    % ch
sales       199
roe         retu
pcroe       % ch
ros         retu
indus       =1
finance     =1
consprod    =1
utility     =1
lsalary     natu
lsales      natu
ln_salary
ln_sales
sector
```

Results – /Users/nzinovyeva/Dropbox/Documents/Teaching/Econometrics_and_Statistics/Stata/statafiles/c...

```
.
. * help summarize
. * help cor
.
. *****************************************************************
. ****     Means, Standard Deviations, Correlations        **********
. *****************************************************************
.
. use "statafiles/ceosal1.dta" , clear

. //      This is the database used by Wooldridge in Example 2.3.
.
. browse

. summarize

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      salary |       209     1281.12    1372.345        223      14822
    pcsalary |       209     13.2823    32.63392        -61        212
       sales |       209    6923.793    10633.27      175.2    97649.9
         roe |       209    17.18421    8.518509         .5       56.3
       pcroe |       209    10.80048     97.2194      -98.9        977
-------------+--------------------------------------------------------
         ros |       209    61.80383    68.17705        -58        418
       indus |       209    .3205742    .4678178          0          1
     finance |       209    .2200957    .4153057          0          1
    consprod |       209    .2870813    .4534861          0          1
     utility |       209    .1722488    .3785031          0          1
```

ECOSTAT_examples.do

```
capture cd "d:/my dropbox/documents/teaching/econometrics_and_statistics/stata"
capture cd "/Users/nzinovyeva/dropbox/documents/teaching/econometrics_and_statistics/stat

capture log close
log using "ECOSTAT_examples.smcl", replace

* help summarize
* help cor

*****************************************************************
****     Means, Standard Deviations, Correlations        **********
*****************************************************************

use "statafiles/ceosal1.dta" , clear
//  This is the database used by Wooldridge in Example 2.3.

browse
summarize
```

Viewer (#1) [view "/Users/nzinovyeva/Dropbox/Documents/Teaching/Econometrics_and_Statistics/Stata/ECOSTA...

view "/Users/nzinovyeva/Dropbox/Documents/Teaching/Econometrics_and_S

```
       log:  /Users/nzinovyeva/Dropbox/Documents/Teaching/Econometrics_and_Statistics/Stata/ECOSTAT_examples.smc
  log type:  smcl
 opened on:  25 Oct 2010, 15:55:42

.
. * help summarize
. * help cor
.
. *****************************************************************
. ****     Means, Standard Deviations, Correlations        **********
. *****************************************************************
.
. use "statafiles/ceosal1.dta" , clear

. //      This is the database used by Wooldridge in Example 2.3.
.
. browse

. summarize

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      salary |       209     1281.12    1372.345        223      14822
    pcsalary |       209     13.2823    32.63392        -61        212
       sales |       209    6923.793    10633.27      175.2    97649.9
         roe |       209    17.18421    8.518509         .5       56.3
       pcroe |       209    10.80048     97.2194      -98.9        977
```

Structure of the course

1. Introduction
2. Randomized control trials
3. Regression based on observables
4. Instrumental variables in action
5. Differences-in-differences
6. Regression discontinuity design
7. Other topics (time permitting):
    - Machine learning
    - Structural vs. reduced form analysis
    - Non-parametric methods
    - Clustering standard errors and multiple Hp. testing

- Ciprian Domnisoru will teach Applied Microeconometrics II

Introduction

Economics is increasingly an empirical discipline

- The Lindau Nobel Laureate Meetings in Economics
- Publications in top economic journals
- The credibility revolution

Introduction

- The Lindau Nobel Laureate Meetings in Economics
  - Only 5 out 150 attendants was doing theory
- Publications in top journals
  - Evidence from articles published in AER, JPE and QJE (Hamermesh 2013)

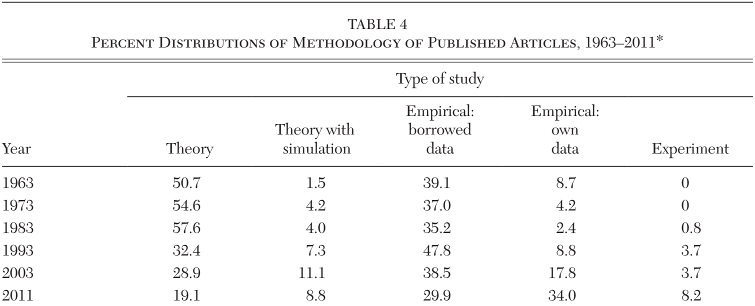


TABLE 4
Percent Distributions of Methodology of Published Articles, 1963–2011*

Year	Type of study Theory	 Theory with simulation	 Empirical: borrowed data	 Empirical: own data	 Experiment
1963	50.7	1.5	39.1	8.7	0
1973	54.6	4.2	37.0	4.2	0
1983	57.6	4.0	35.2	2.4	0.8
1993	32.4	7.3	47.8	8.8	3.7
2003	28.9	11.1	38.5	17.8	3.7
2011	19.1	8.8	29.9	34.0	8.2

Introduction

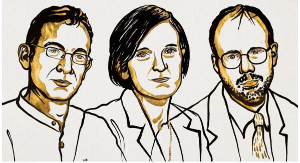
- The **credibility revolution** (Angrist and Pischke, 2010)
  - In early 1980s, the state of empirical economics had reached its lowest point (Leamer, 1983)

    "Hardly anyone takes data analysis seriously. Or perhaps more accurately, hardly anyone takes anyone data analysis seriously."

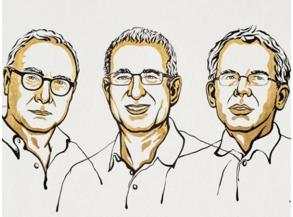
  - The revolution put emphasis on the quality of **empirical research designs**. This immensely improved both scientific and policy relevance of research.
- Over the next decades, Economics as whole saw a **disruptive change** in how we do empirical research:
  - The reaction from *structuralists* was fierce. More recently the discussion got civil (e.g., Deaton, 2010).
  - The result of this change is a **new paradigm** that puts empirical research designs and better data at the forefront

Introduction

- **2019 Nobel**: Banerjee, Duflo and Kremer, *For their experimental approach to alleviating global poverty*



- **2021 Nobel**: Card (1/2), Angrist (1/4), Imbens (1/4), *They have shown that natural experiments can be used to answer central questions for society [...]. Together, they have revolutionised empirical research in the economic sciences.*



Introduction

Three types of empirical questions

What kind of empirical research question to we have in mind? Three broad possibilities:

- Descriptive
- Forecasting
- Causal

Descriptive: Intergenerational mobility

- How are one's lifetime earnings correlated with one's parents' lifetime earnings?
- Equality of opportunity debate
- Development of views within economics:
    - Becker (1988): High mobility in the Anglo-Saxon countries
    - Solon (1992): Low mob. if we account for measurement error
    - Björklund and Jäntti (1997): Mobility in Scandinavia higher than in the US
    - Krueger (2012): Cross-sectional inequality and mobility are negatively correlated

Descriptive: Intergenerational mobility

Gary Becker, 1988

In all these countries (US, UK, and Canada), low earnings as well as high earnings are not strongly transmitted from fathers to sons, and Knight's claim about family life causing growing inequality is inconsistent with the evidence

Descriptive: Intergenerational mobility

Gary Solon, 1992

- Evidence on intergenerational mobility is based on simple regressions:

  $$y_{son,i} = \rho y_{father,i} + \epsilon_i$$

  where $y_{son,i}$ is son's and $y_{father,i}$ father's lifetime earnings, respectively

- Early estimates (mentioned by Becker) biased by:
  1. Measurement error: $y_{father,i}$ proxied by one year of earnings
  2. Homogeneous samples: Lack of variation in $y_{father,i}$

- Both problems introduce negative bias to estimates of ρ ▸

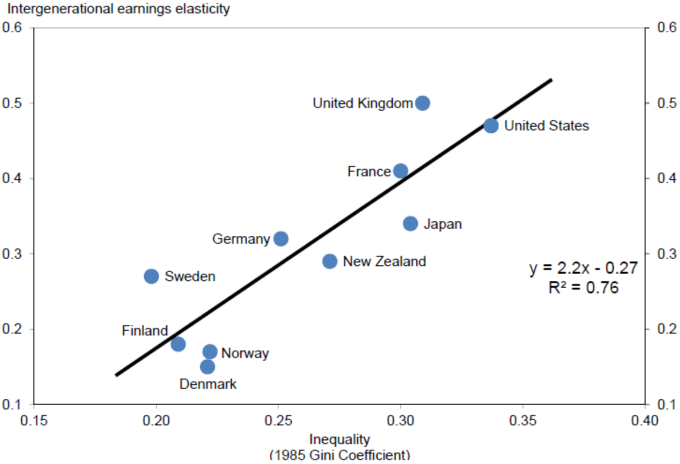
- Solon (1992) uses representative samples and several years of earnings data and obtains much higher estimates

Descriptive: Intergenerational mobility

Further evidence

- Björklund and Jäntti (1997): Mobility in Sweden is higher than in the United States
- Are equality of opportunity and equality of outcomes independent?

Descriptive: The Great Gatsby Curve (Krueger, 2012)



Source: Corak (2011), OECD, CEA estimates

Descriptive

In some ways descriptive questions are the easiest to answer in the sense that if we had enough data we would know the answer.

There are at least three challenges for the econometrician here:

- **Sampling**
  - We typically observe a sample from the full population, and we want to make inferences about the population based on it.
    - Example: Survey on labour market outcomes of university graduates
- **Measurement**
  - Example: measure the 'sentiment' of Facebook posts
- **Summary Statistics**
  - Often the data for some of these questions is complicated and we need to find a nice way to summarize it

Forecasting

Goal: to predict future events

Examples:

- Future GDP growth/unemployment?
- Prediction of election results based on exit polls
- Predict group membership based on choices

Prediction of group membership - partisanship in the U.S. congress 1873-2016

Gentzkow et al, 2017

- Can one predict group membership with observable choices?
- Examples:
  - Segregation in residential choices
  - Partisanship in media consumption
- Getnzkow et al study whether partisanship has increased in the U.S congress

Prediction of group membership - partisanship in the U.S. congress 1873-2016

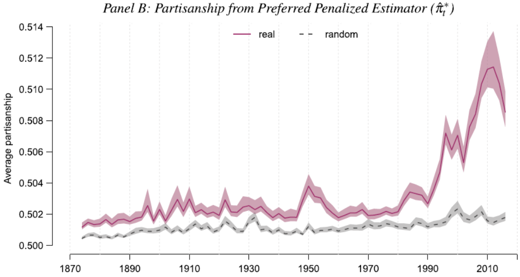
Gentzkow et al, 2017

- **Research question**: Can you tell to which party the representative belongs just by observing his or her speeches?
- If this has become easier over time, partisanship has increased
- Difficult econometric problems:
  - Dimensionality: the way we talk may differ just randomly
  - Computational burden

Prediction of group membership - partisanship in the U.S. congress 1873-2016

- Solve these problems with modern machine learning methods
  - Powerful tool in prediction
  - However, not a topic of this course (for a good reason)
- Use the choice of words to predict party membership in the U.S. congress between 1873-2016
- PhD thesis from Aalto University by Salla Simola does the same analysis for the Finnish parliament between 1907-2018

Partisanship in the U.S congress 1873-2016



Forecasting

Goal: to predict future events

Other examples:

- What will be grades obtain in the master thesis by students taking this course?
- Can you predict who are the pickpocketers in London's subway?
- Can you predict who is going to suffer a certain illness?

Forecasting

- We will not know the answer to these questions until the event happens (but when it happens, we will know)
- 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.

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 good in PISA exams?
- **Forward causal questions**: estimation of *effects of causes* (What if ...?).
    - Does teachers' IQ affect students performance?
    - Class size?

Causal effects

Economists often are motivated by *why* questions, but when they do research they tend by address *what if* questions.

Examples:

- How does taking this course affects the grade that you will obtain in your master thesis?
- How does a positive Facebook post affect your sentiment?
- If Uber increases prices, how would it affect demand?
- Does death penalty decrease crime rates?
- Would it be profitable for a firm to allow employees to work from home? (Yahoo 2013)
- Are employees more satisfied if they are informed about the salaries of their colleagues? (Card et al., 2012)

Causal effects

- Generally, will not know the answer to these questions unless a randomized controlled trial (RCT) is performed (or, somehow, we have an appropriate empirical strategy, more on this later!)
- 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 can run a RCT (or we have a plausible empirical strategy), we will never know the answer for sure.

Causal effects

- Economists usually think about causal questions
  - Economic theory leads to these 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 *predition/forecasting* questions
- Recently, machine learning has started to be used in contexts where causal questions are of interest:
  - predicting who to assign a given treatment (e.g., training)
  - data-driven heterogeneous effects

Mesurement error in father's lifetime earnings

- Write lifetime earnings of sons as: $y_{si} = y_{fi} + \epsilon_i$, where y_{fi} is father's lifetime earnings and $Cov(y_{fi}, \epsilon_i) = 0$.
- Suppose we only observe **proxy** $y_{fit} = y_{fi} + \upsilon_{fit}$ and let's further assume that $Cov(y_{fi}, \upsilon_{fit}) = 0$ and $Cov(y_{fit}, \epsilon_i) = 0$.
- It follows that:

$$\begin{aligned} Cov(y_{fit}, \upsilon_{fit}) &= Cov(y_{fi} + \upsilon_{fit}, \upsilon_{fit}) \\ &= Cov(y_{fi}, \upsilon_{fit}) + Cov(\upsilon_{fit}, \upsilon_{fit}) \\ &= Var(\upsilon_{fit}) \end{aligned}$$

Mesurement error in father's lifetime earnings

- Our regression of interest now becomes

$$y_{si} = \rho(y_{fit} - \upsilon_{fit}) + \epsilon_i$$

- OLS estimate of ρ is

$$\begin{aligned}
\hat{\rho} = \tfrac{Cov(y_{si}, y_{fit})}{Var(y_{fit})} \quad &= \quad \tfrac{Cov(\rho(y_{fit} - \upsilon_{fit}) + \epsilon_i), y_{fit})}{Var(y_{fit})} \\
&= \quad \rho\left(\tfrac{Cov(y_{fit}, y_{fit})}{Var(y_{fit})}\right) - \rho\left(\tfrac{Cov(\upsilon_{fit}, y_{fit})}{Var(y_{fit})}\right) + 0 \\
&= \quad \rho - \rho\left(\tfrac{Cov(\upsilon_{fit}, y_{fit})}{Var(y_{fit})}\right) \\
&= \quad \rho - \rho\left(\tfrac{Var(\upsilon_{fit})}{Var(y_{fi}) + Var(\upsilon_{fit})}\right) \\
&= \quad \rho\left(\tfrac{Var(y_{fi})}{Var(y_{fi}) + Var(\upsilon_{fit})}\right) < \rho
\end{aligned}$$

Mesurement error in father's lifetime earnings

- Short spells of y_{fit} usually mean that $Var(v_{fit}) \to \infty$ which implies that $\hat{\rho} \to 0$
- Homogeneous samples usually mean that $Var(y_{fi}) \to 0$ which again implies that $\hat{\rho} \to 0$