Applied Microeconometrics I

Lecture 1: Introduction

Tuomas Pekkarinen

Aalto University

September 14, 2021

`Lecture Slides`

Goals

- Main aim: to provide the basic tools to do empirical analysis with the aim of causal inference.
- This course has a practical flavor $\Rightarrow$ emphasis is not on proofs but on intuitions and on applications
- I will presume that you already know basic Econometrics
- We will be more concerned in general with consistency that with efficiency.

Schedule

- The course begins on September 14 and lasts until October 21.
- Lectures will be held remotely via Zoom
  - Tue: 13:15-14:45
  - Wed: 16:15-17:45 (typically discussion of exercises)
  - Thu: 13:15-14:45
- Software tutorial also remotely via Zoom
  - Wed 10/9: 16:15-17:45 Stata
  - Wed 17/9: 16.15-17.35 R (to be confirmed)

Schedule

- Exceptions
  - On Wed 20/10: 16:15-17:45 Lecture instead of exercises
  - On Thu 21/10: 13:15-14:45 Exercises instead of lectures

A few things you need to know

- Office hours:
  - Please, send an email (tuomas.pekkarinen@aalto.fi) to fix an appointment.
  - For STATA related issues please contact the teaching assistant Lassi Tervonen: lassi.tervonen@aalto.fi
- Please, provide feedback also during or after the lectures
- Because interaction may be difficult during lectures, I will try to stick around in Zoom after the lecture for a while to answer questions
- Lectures will be recorded but not edited. Recordings are available after the lecture at the course website

Course Requirements

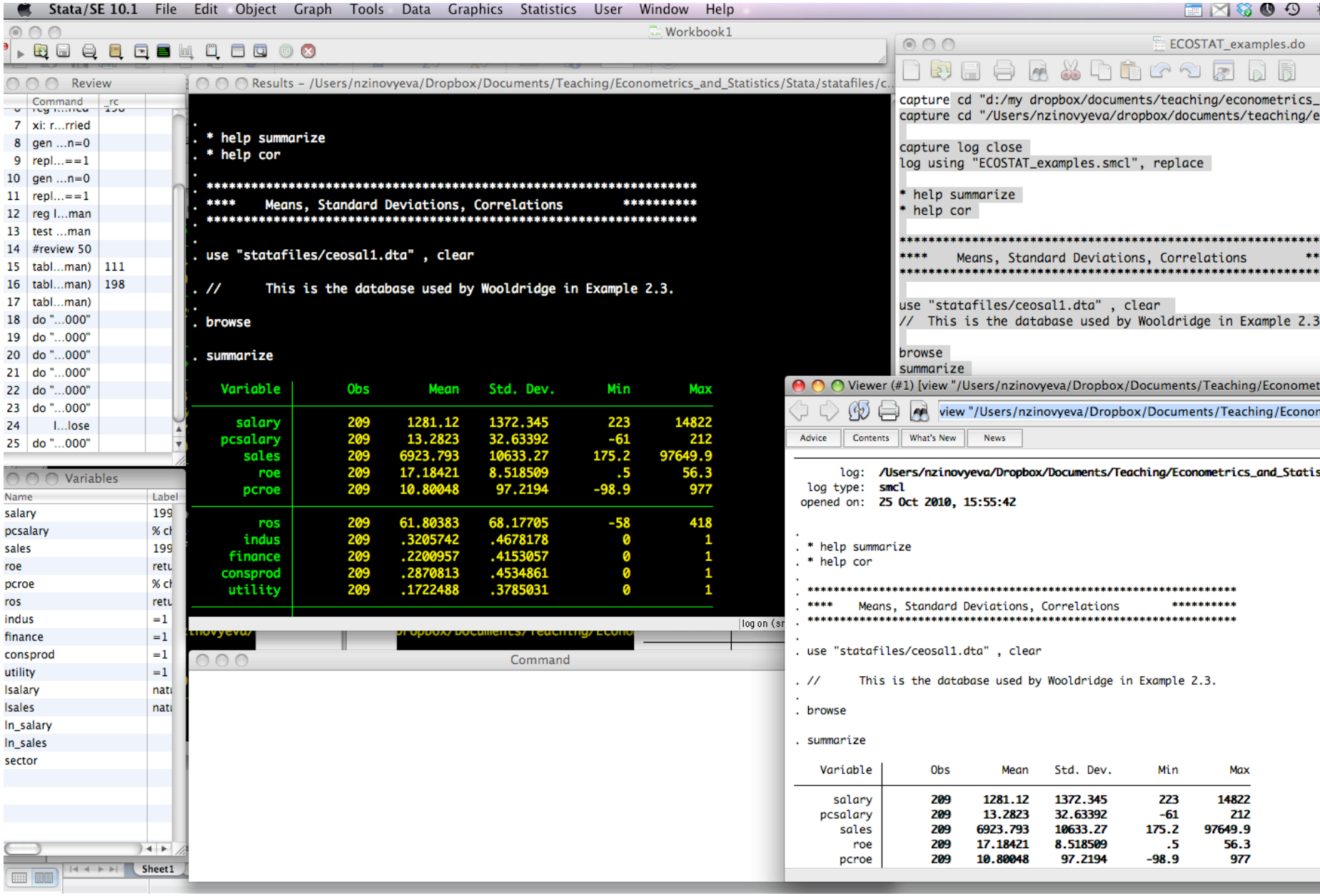
- Evaluation: five problem sets (50 per cent), final exam (50 per cent).
  - To pass the course a passing grade in the exam is required.
- Problem sets:
  - Available on Fridays at mycourses, due the following Friday at 22:00.
  - Deadlines are strict. The lowest graded problem sheet will not contribute to the grade.
  - If you enrol through Oodi, you are automatically enrolled in mycourses. Otherwise contact me!
  - Please include NAME and STUDENT NUMBER at the beginning of your problem set.
- Exam:
  - Date: 26/10 (presumably)
  - Retake: 21/12
  - Check, if you need to register separately for the exam
- PhD students, please talk with me after class

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 that we discuss in the lectures and that are listed in the syllabus (on Mycourses, updated continuously)

Statistical software

- Problems will require the use of some statistical package
- I use STATA, but you are free to use any software
- Why STATA?
  - Easy to start with
  - Many other people use it
  - Drawback: proprietary software
- Software tutorial
  - Stata: Wed 15/09, 16:15-17:45
  - R (to be confirmed): Wed 22/09: 16:15-17.45
- We may use some STATA in the class


Stata/SE 10.1 File Edit Object Graph Tools Data Graphics Statistics User Window Help
Workbook1
ECOSTAT_examples.do
Review
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"
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
Variables
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
Command
Viewer (#1) [view "/Users/nzinovyeva/Dropbox/Documents/Teaching/Econometr
Advice Contents What's New News
log: /Users/nzinovyeva/Dropbox/Documents/Teaching/Econometrics_and_Statis
log type: smcl
opened on: 25 Oct 2010, 15:55:42
Sheet1

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):
   - Structural analysis
   - Clustering standard errors
   - Small samples
   - Multiple testing

- Ciprian Domnisoru will be teaching Applied Microeconometrics II in the spring!

Introduction

Economics is increasingly an empirical discipline

- 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*

	Type of study				
Year	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

- 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 mobility when accounting for measurement error
  - Björklund and Jäntti (1997): Mobility is higher in Scandinavia 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:

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

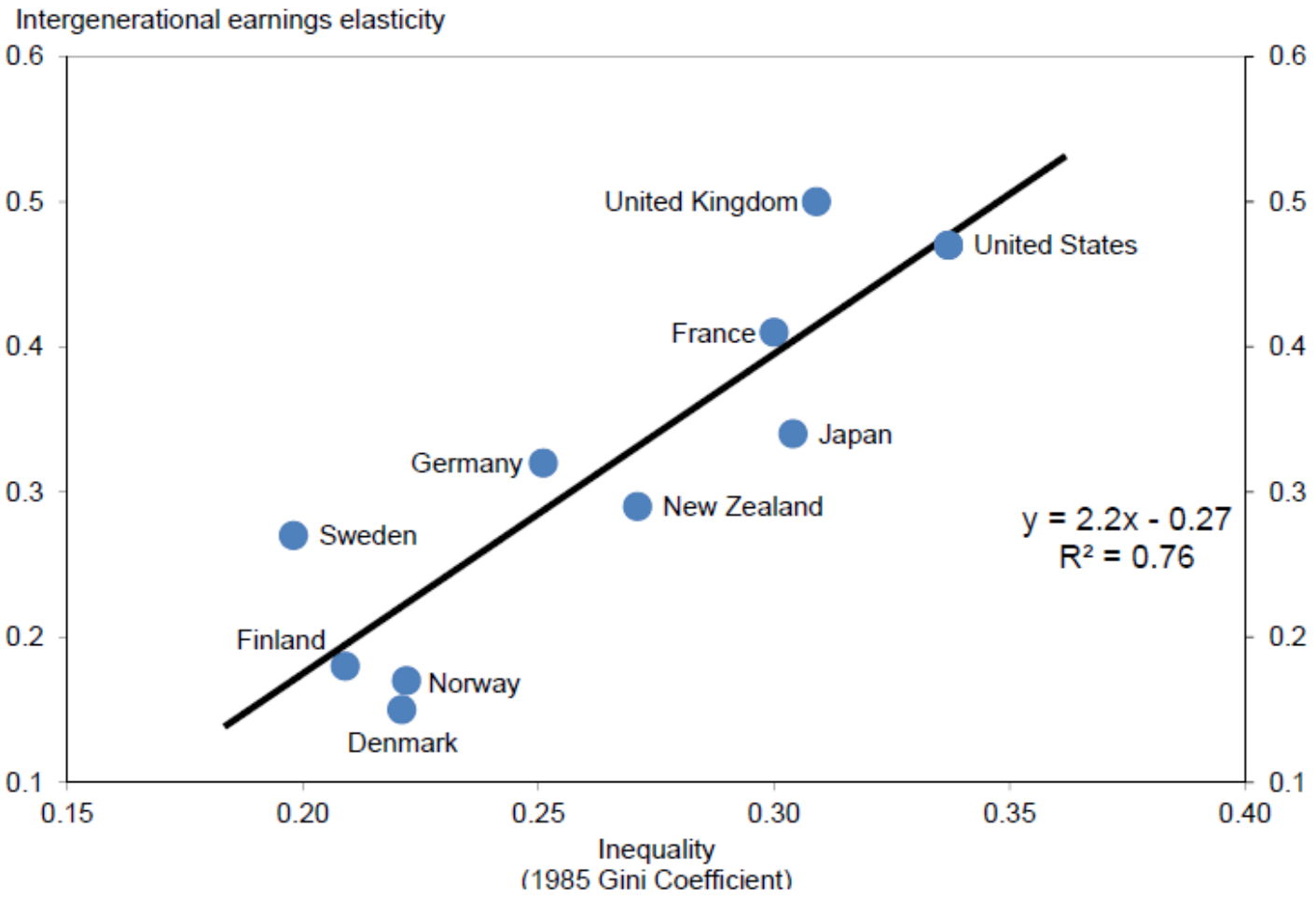
  where $y_{son,i}$ is son's and $y_{father,i}$ father's lifetime earnings, respectively
- Early results suffered from two sources of bias:
  1. Measurement error: $y_{father,i}$ proxied by one year of earnings
  2. Homogeneous samples: Lack of variation in $y_{father,i}$
- Both problems introduce negative 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): More mobility in Sweden 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 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 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

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

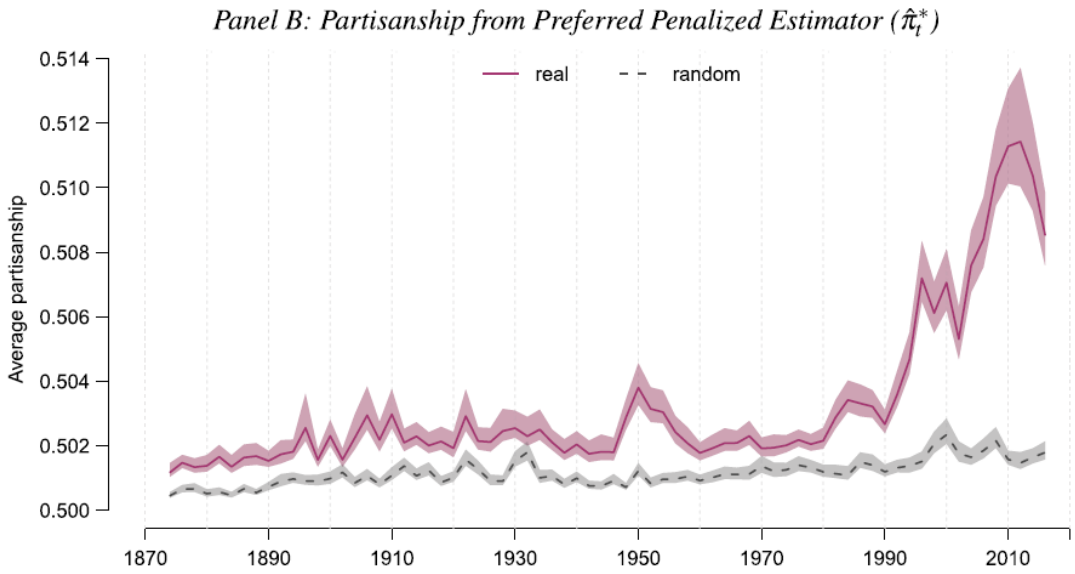
Gentzkow et al, 2017

- The 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 of speech: the way we talk may differ just randomly
  - Computational burden

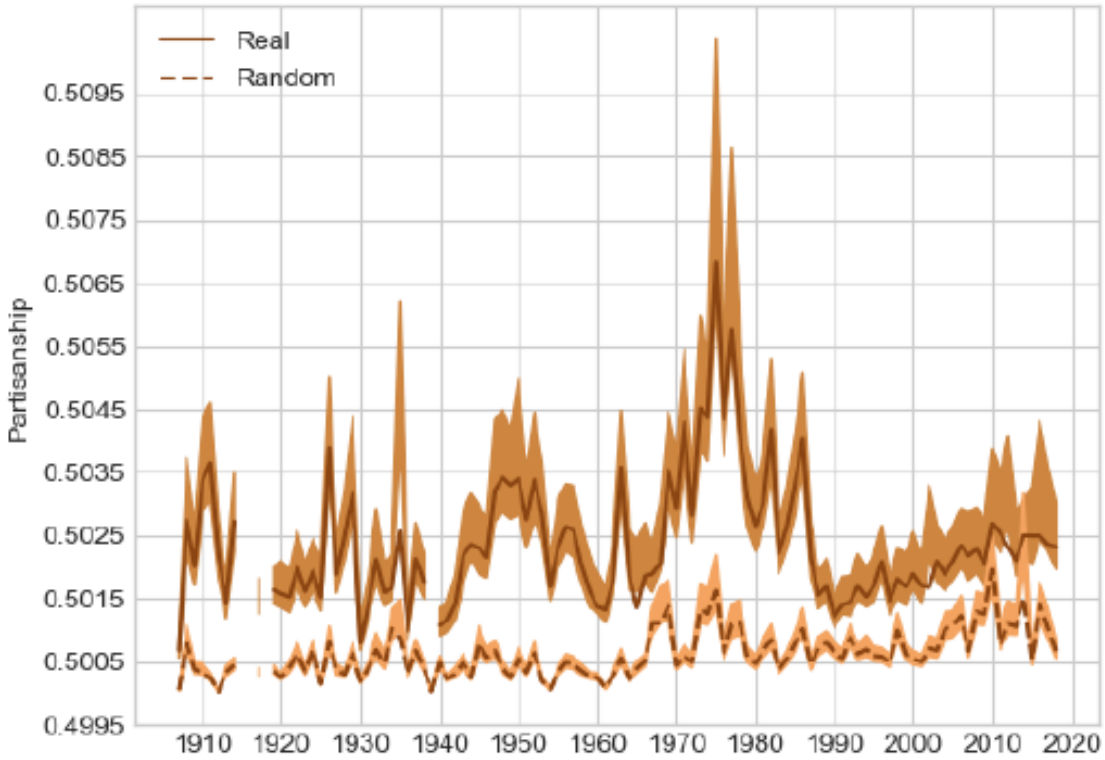
Prediction of group membership - partisanship in the U.S. congress 1873-2016

- 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
- Recent 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



Left-right partisanship in the Finnish parliament 1907-2018



Forecasting

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? Same teacher?

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 course affects the grade that you will obtain in your master thesis?
  - Note that this is different from the predictive question: “What is the grade that students taking this course will obtain with their master thesis?"
- How does a positive or a negative 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, Mas, Moretti and Saez 2012)

Causal effects

- Generally, we will never know the answer to these questions. Or, more precisely, we 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
  - 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

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 $y_{fit} = y_{fi} + \upsilon_{fit}$ 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

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

Mesurement error in father's lifetime earnings

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