

Week 1

Intro to Practical Econometrics

Dragos Ailoae

dragos@nyu.edu

Applied Statistics and Econometrics 2

ECON GA-1102

Lab Sections 031 and 032

New York University

January 22, 2024

Today

1. This class
2. Research project
3. Next steps

This Class

Lab = Enhanced Recitation

- Recitation
 - Review theory
 - Solve exercises to complement theory, homework
- Empirical work in R
 - Complement lecture
 - Research project
- Answer your questions – resuscitation

Office hours

- No formal office hours but available by email
- Last half hour of class for 1 on 1 discussion or research group meetings
 - by appointment or “walk in”

Research Project

Intro: Motivation

One of the priorities of this course is to guide you into producing your own research by the end of the semester

- Great opportunity to explore a topic of interest
- Apply the econometric methods you learned
- Excellent topic of discussion for job interviews

Intro: Logistics

- Detailed in the “Project Outline” handout
- Groups of 3-5 students (ideally from same lab section)
- Important dates:

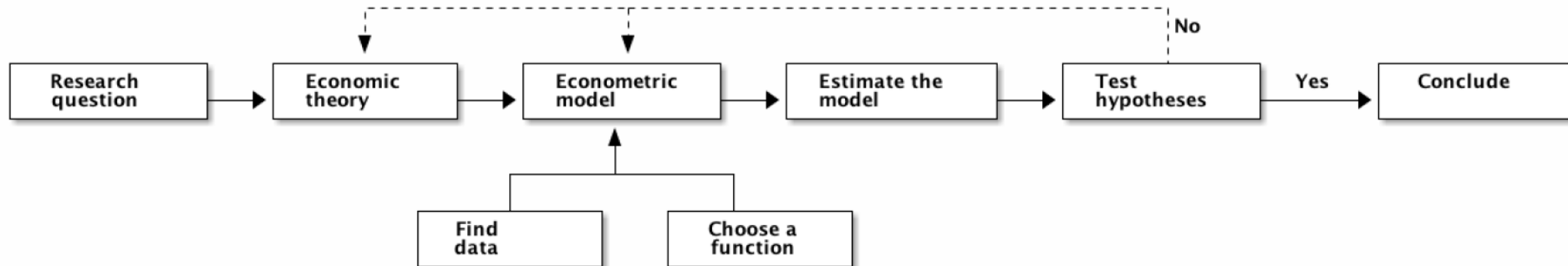
Project Requirement	Date Due
Group Signup	Feb 5
Problem Statement	Feb 26
Model Description	Apr 1
Presentation	Apr 29
Final Report	May 10

Research Project: Guidelines

Three ingredients of a successful research project

1. Academic rigor
 - a) Understand and encompass the existing literature
 - b) Innovative, yet appropriate, use of data
 - c) Appropriate causal inference
2. Policy relevance
 - a) Tied to new facts or trends
 - b) Framed in terms of policy levers
 - c) Timely
3. Broadly communicated
 - a) Accessible to a wide range of audiences
 - b) High potential for media coverage
 - c) Partnered with policy makers

Econometric research workflow



Model should be anchored in established economic theory

Avoid data mining! Put the Econ in the Econometrics

Some (broad) theoretical frameworks:

- Supply / demand
- Consumption smoothing
- Monopolistic competition

Keep your eyes open for empirical examples in your textbooks
(Chiang book, Greene book)

Research Project: Data

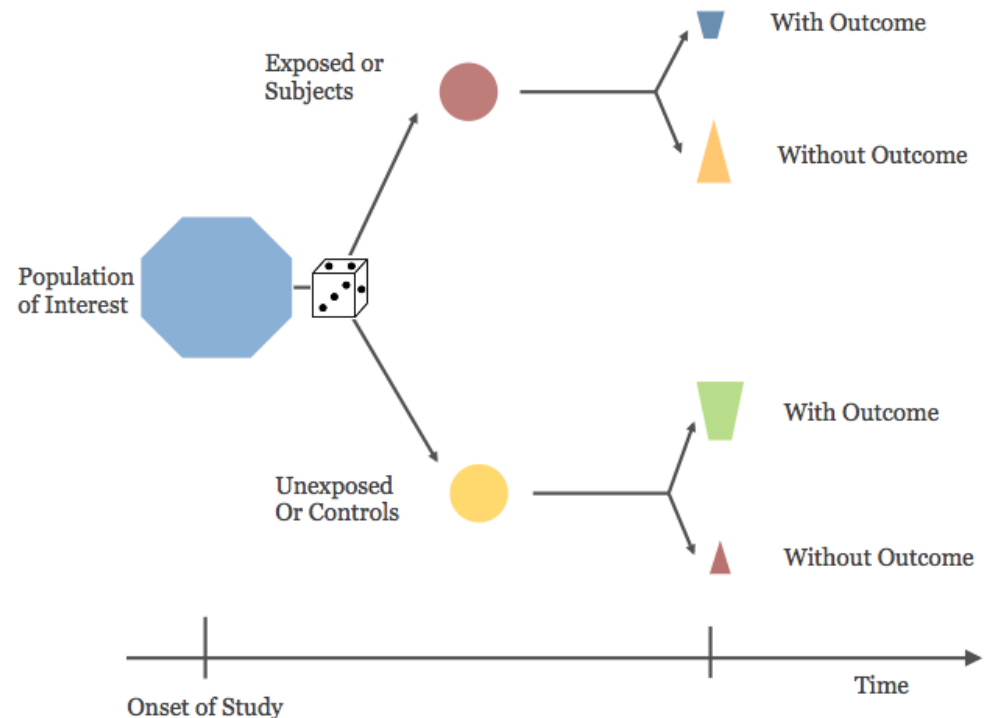
Experimental vs. observational data

- Experimental data come from experiments designed to evaluate a treatment or policy or to investigate a causal effect
- Observational (nonexperimental) data are collected using surveys and administrative records

Experimental data: RCTs

Randomized Control Trials:

- All participants are randomly assigned into two groups.
- The control group receives no treatment (or placebo)
- The experimental group receives the treatment.
- After a follow-up period, compare the two groups



RCTs: advantages

- The gold standard for causal inference
- Randomization minimizes selection bias
 - Ensures that the only systematic difference between the control treatment group is the treatment itself, with the effects from other confounding factors eliminated

RCTs: disadvantages

- **Cost:** Called “the gold standard” because expensive (in money and time)
- **Ethics:** Especially in social science, we cannot impose some treatment due to ethic concerns

Observational data: advantage

Readily available:

Public databases

- Federal Reserve Economic Data <https://fred.stlouisfed.org/>
- US Census <https://www.census.gov/en.html>
- US Bureau of Labor Statistics <https://www.bls.gov/>
- US Economic Accounts <https://www.bea.gov/data>
- Penn World Tables <https://cid.econ.ucdavis.edu/pwt.html>
- IMF <https://www.imf.org/en/Data> OECD: <https://data.oecd.org/>

Replication data sets

- openICPSR <https://www.openicpsr.org/openicpsr/repository/>
- Harvard Dataverse <https://dataverse.harvard.edu/>

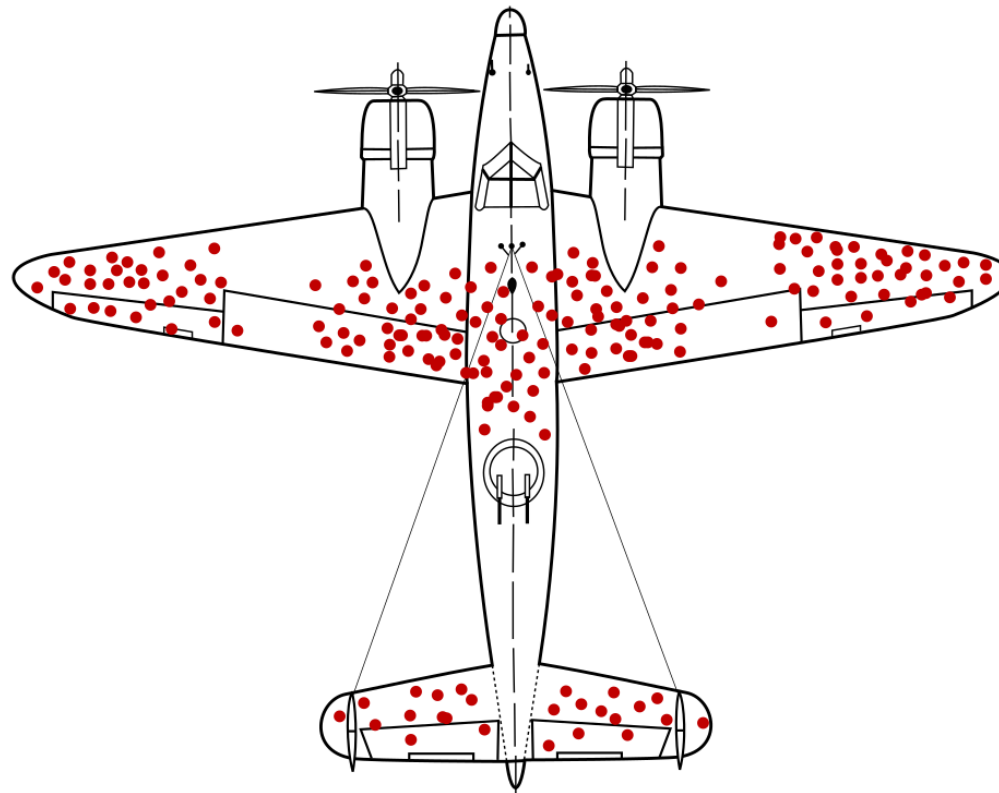
Author personal website

Paid Haver Analytics, Bloomberg, FactSet, Markit, CapitalIQ

Curated datasets

- R datasets <https://vincentarelbundock.github.io/Rdatasets/articles/data.html>
- Data and Story Library <https://dasl.datadescription.com/datafiles/>

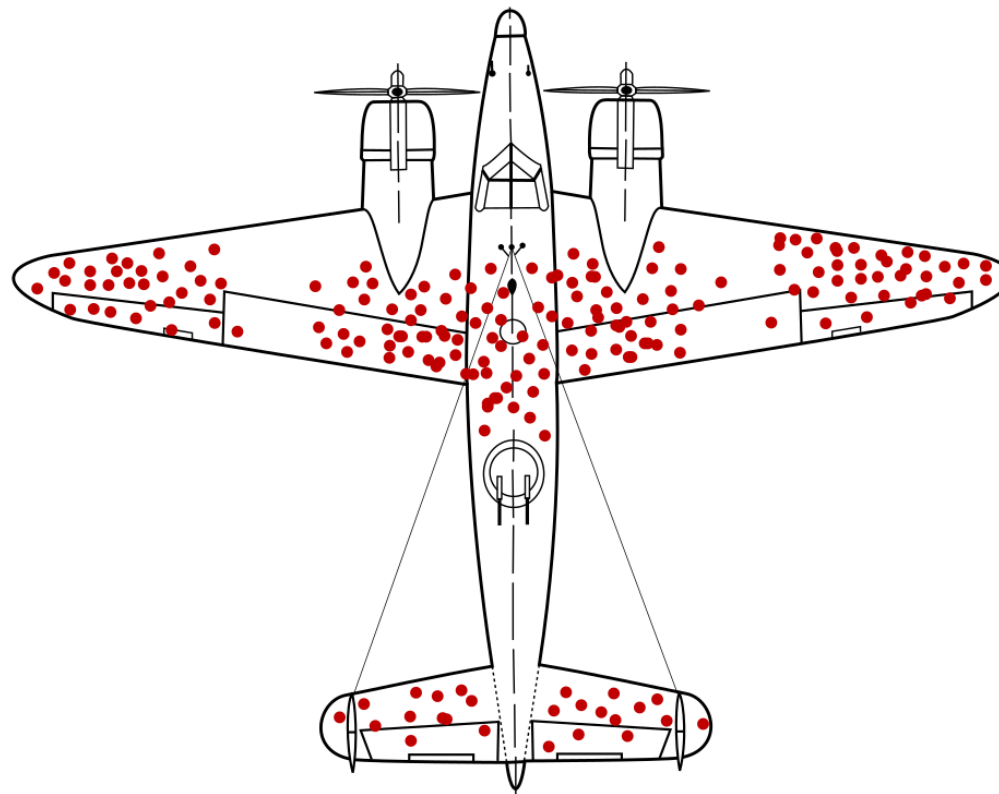
Observational data: disadvantage 1



Source: Martin Grandjean

Observational data: disadvantage 1

Choices already baked in: Know your data collection methodology!
(see Abraham Wald, survivorship bias, selection bias, truncation, censoring)



Source: Martin Grandjean

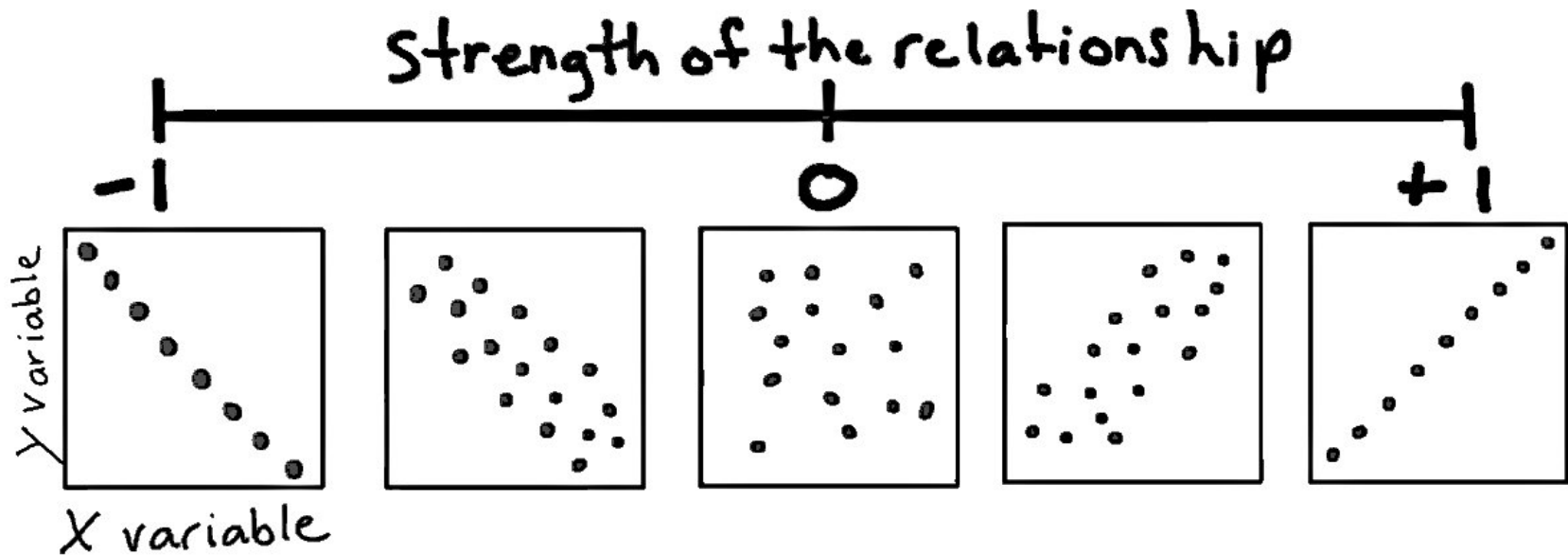
Observational data: disadvantage 2

“Treatment” is not randomly assigned so
difficult to estimate causal effects

**Much of econometrics dedicated to dealing
with causality using observational data**

Causal Inference

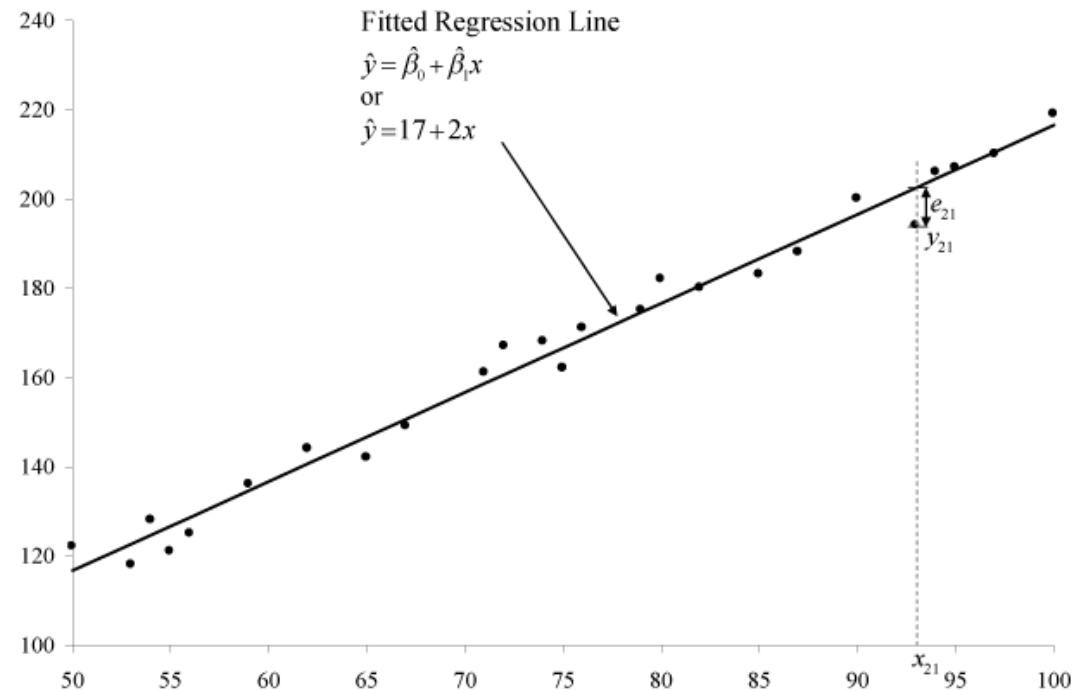
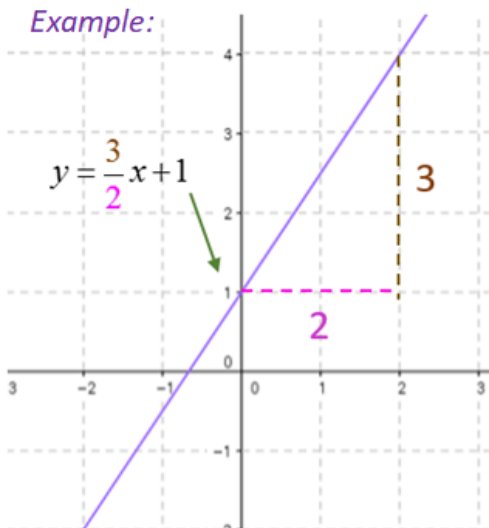
Measure of association: correlation coefficient



Regression

$$y = mx + b$$

slope of line y intercept, where the line crosses the y-axis at (0,b)



Regression

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Y = dependent variable

X = independent variable

ε = other factors (aka "error term")

$$\text{Lifespan} = \beta_0 + \beta_1 \text{RedWineConsumption} + \varepsilon$$

Wealth as possible confounder (wealthy people likely to drink wine but also likely to get better health care)

$$\text{LungCancer} = \beta_0 + \beta_1 \text{SmokingTobacco} + \varepsilon$$

Ronald Fisher (a smoker himself) argued on the side of tobacco companies about possible confounders (genetics etc)

- Regression can be useful but be careful not to interpret causally
- Safer to say “ X is associated with Y ” or, if you want to be more specific, “a one unit increase in X is associated with a β_1 increase/decrease in Y ”

Causal effect

- Causal effect - the effect on an outcome of a given action or treatment as measured in an ideal RCT
- The concept of the **ideal randomized controlled experiment** does provide a theoretical benchmark to define causal effects in research design
- Sometimes nature helps - natural experiments (quasi-experiments) provide randomization

Methods

- Difference in Differences – Greene Ch. 6
- Instrumental Variables – Greene Ch. 8

Difference in Differences

Jon Snow

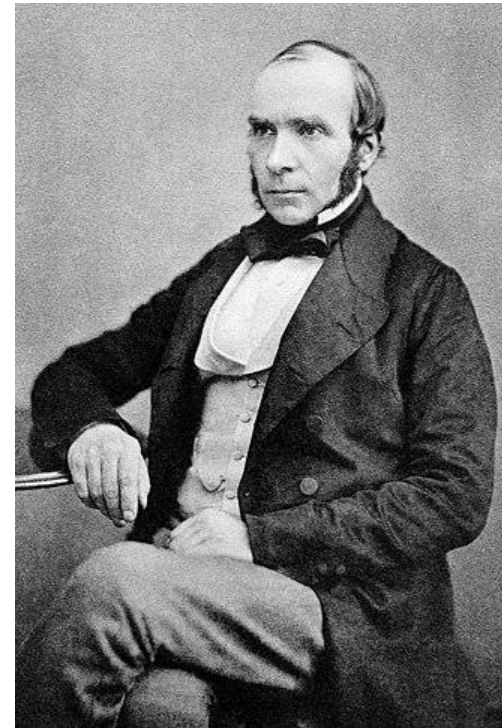
(“Game of Thrones” character)



VS

John Snow

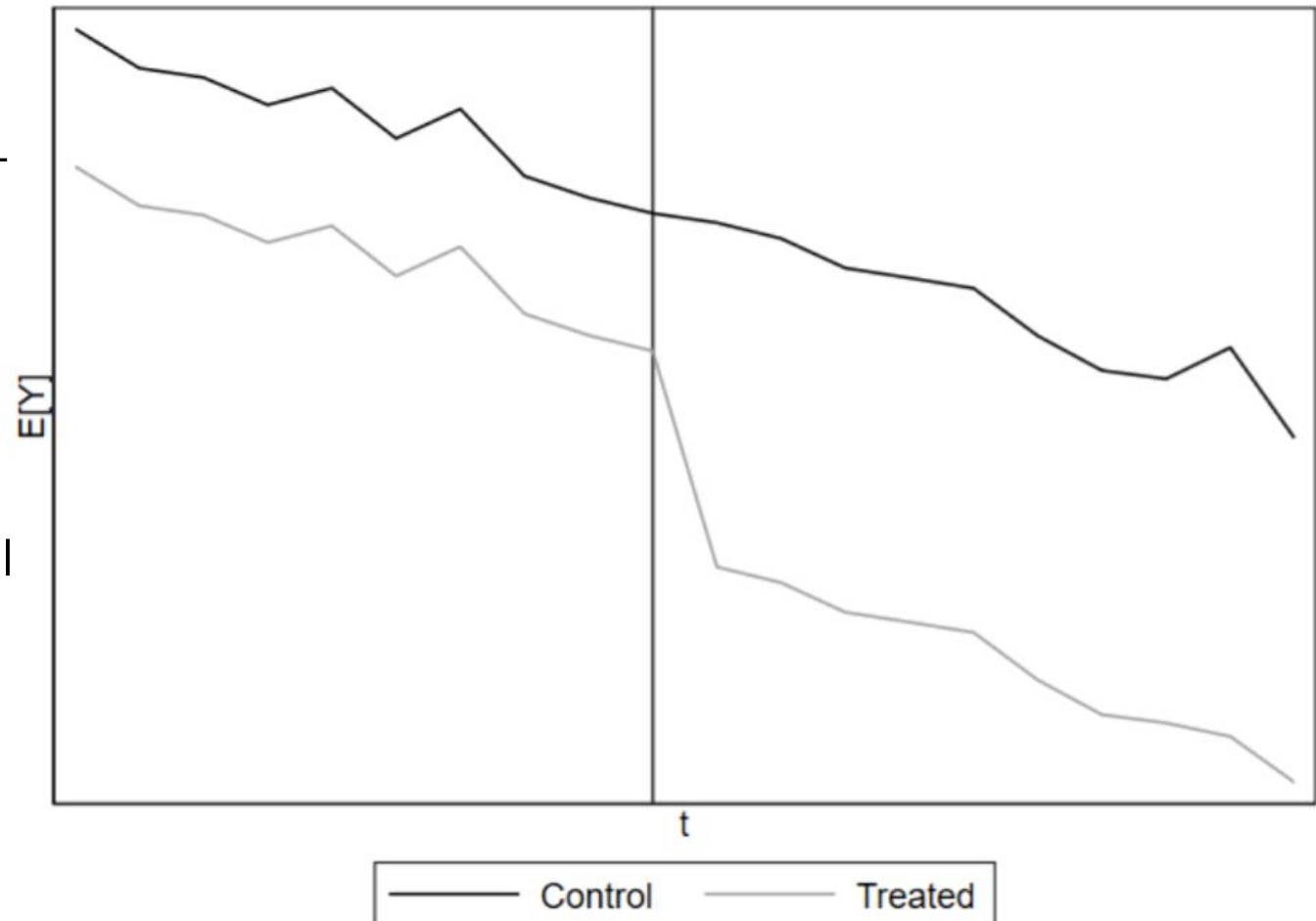
(Father of epidemiology)



Source: Wikipedia

Difference in Differences

- John Snow 1850s – cholera incidence vs. water provider
- Card and Krueger (1994) – NJ, PA unemployment level vs. min wage



Difference in Differences



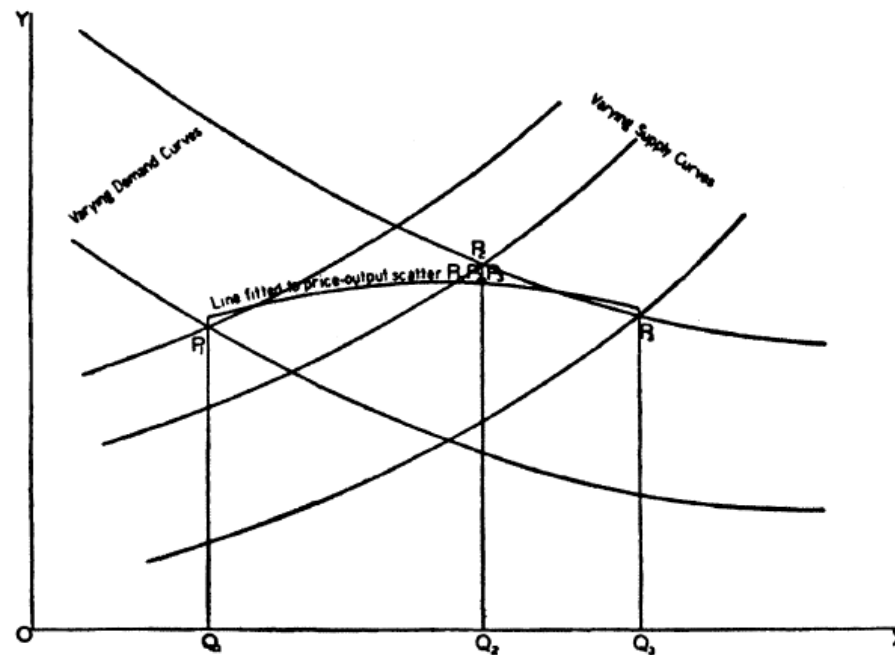
- Sources of randomization:
 - Local governments change policy (marijuana, pay-day loans, min. wage)
 - Jurisdictions hand down legal rulings (abortion)
 - Natural disasters (wildfires in California, hurricanes in Louisiana)
 - Firms lay off workers

Image source: Scott Cunningham, Causal Inference: The Mixtape(2020)

Instrumental Variables

Phillip G Wright's original illustration of the identification problem

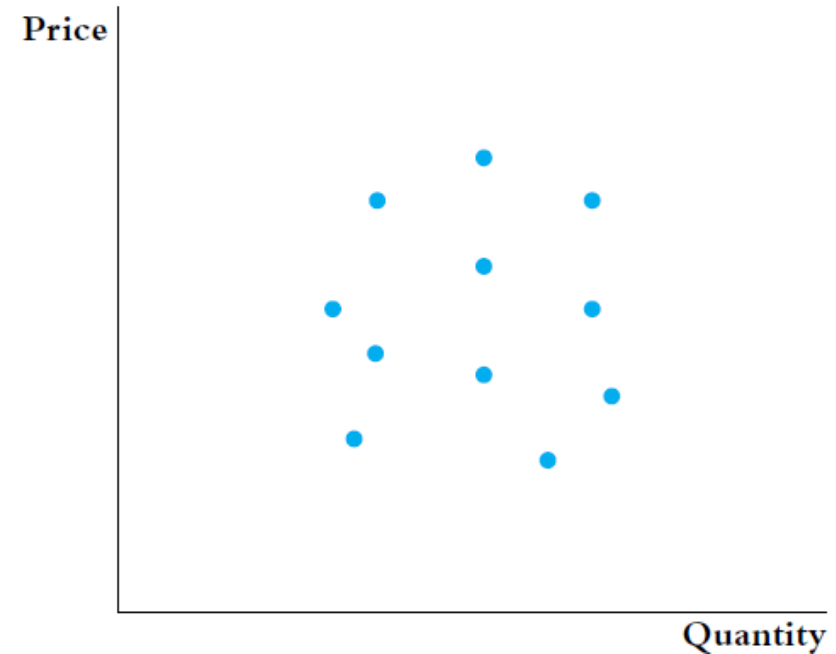
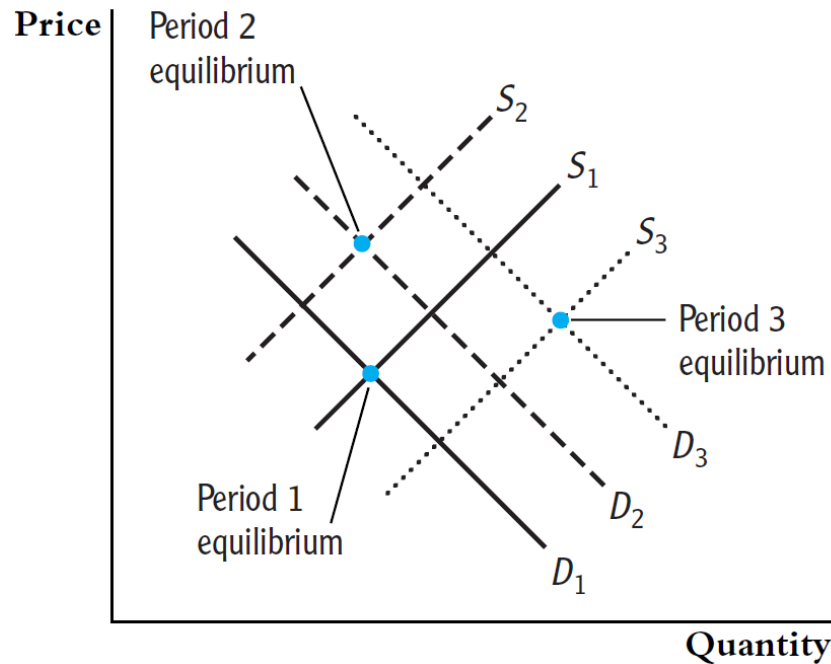
FIGURE 4. PRICE-OUTPUT DATA FAIL TO REVEAL EITHER SUPPLY OR DEMAND CURVE.



Source: PG Wright, The Tariff on Animal and Vegetable Oils (1928)

Instrumental Variables

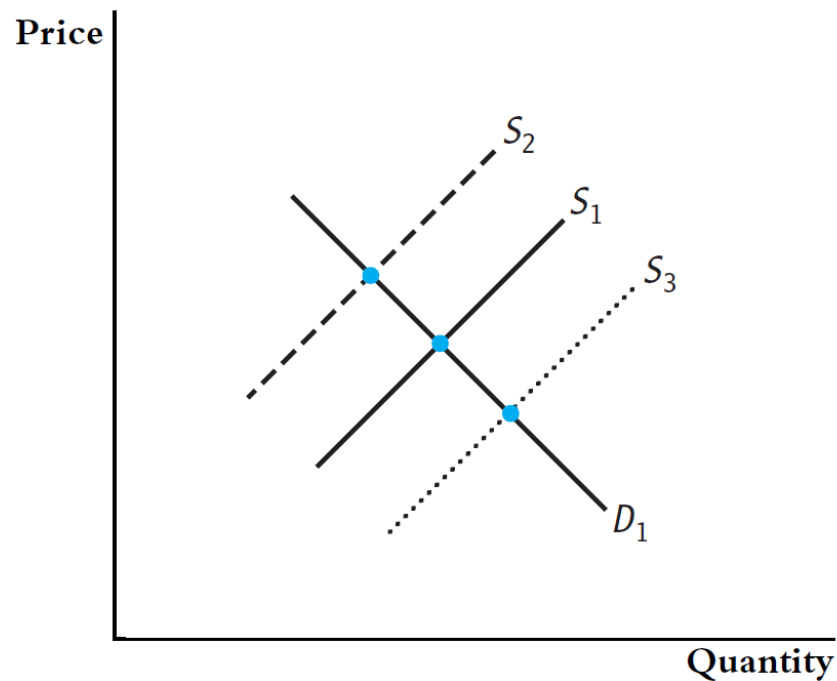
$$\ln(Q_i^{butter}) = \beta_0 + \beta_1 \ln(P_i^{butter}) + u_i$$



Source: Stock and Watson

Instrumental Variables

Using Rainfall as Instrumental Variable for Butter Supply



Source: Stock and Watson

Statistical Data Types

1. Cross-sectional data

- Data on different entities for a single time period are called cross-sectional data
- The sequence of each observation number is arbitrarily assigned
- Cross-sectional data can be experimental data or observational data

person	year	income	age	sex
1	2018	50	27	M
2	2018	80	38	F

2. Time series data

- Data for a single entity collected at multiple time periods
- The sequence of each record is based on the time period it happened

person	year	income	age	sex
1	2018	50	27	M
1	2019	55	28	M
1	2020	60	29	M

- Econometrics 2:
 - Serial correlation, nonstationarity
 - Spurious correlation <http://tylervigen.com/spurious-correlations>
 - Vector Autoregressive models (VAR), GARCH etc.

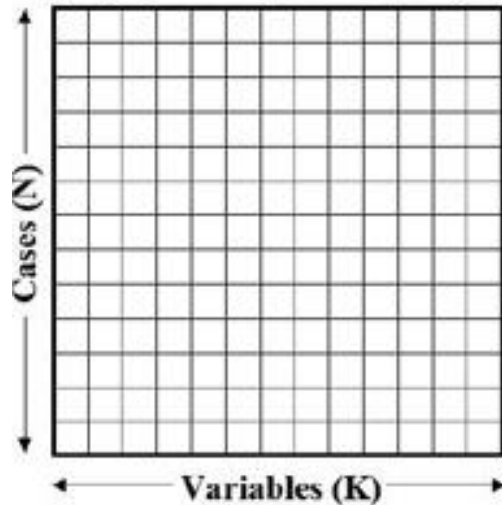
3. Panel data

- Also called longitudinal data - data for multiple entities in which each entity is observed at two or more time periods.
- Panel data are **very useful for estimating causal effects**

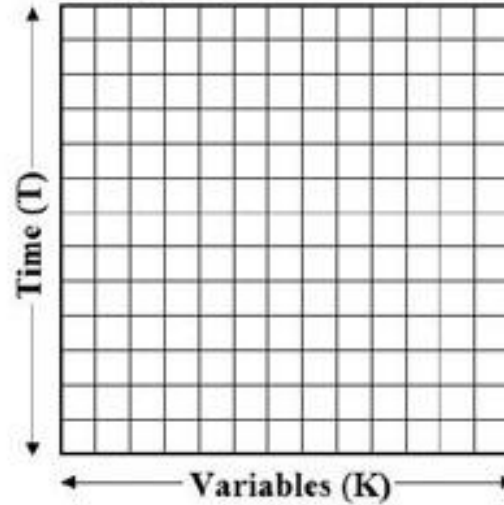
person	year	income	age	sex
1	2018	50	27	M
1	2019	55	28	M
1	2020	60	29	M
2	2018	80	38	F
2	2019	85	39	F
2	2020	90	40	F

Statistical data types visualization

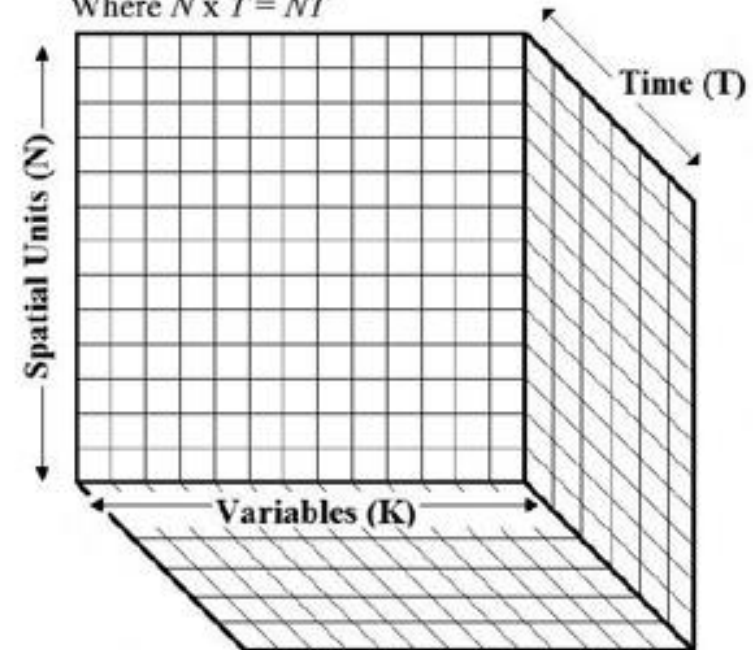
a) Cross-Sectional Analysis
(Single, or average time point)



b) Time-Series Analysis
(Single unit)



c) Time-Series Cross-Sectional Analysis
Where $N \times T = NT$



Summary

- Great topic for future job interviews
- Writeup is like an empirical final exam (i.e. show you've learned the material) but packaging matters (policy relevance)
- A topical research question has legs
- Econometric model should be anchored in economic theory (careful with data mining)

TLDR

- **Find a good reference paper**
- **Start now!**

Next Steps

Next Steps

- Start thinking about your project - group sign-up due Feb 5
- I will post homework 1 early next week
- Next week: Panel data analysis with R (may help with HW)
 - Install Rstudio: <https://www.datacamp.com/community/tutorials/installing-R-windows-mac-ubuntu>
 - Bring laptops (fully charged, few outlets in classroom)