

# 1.1 — Introduction to Econometrics

ECON 480 • Econometrics • Fall 2020

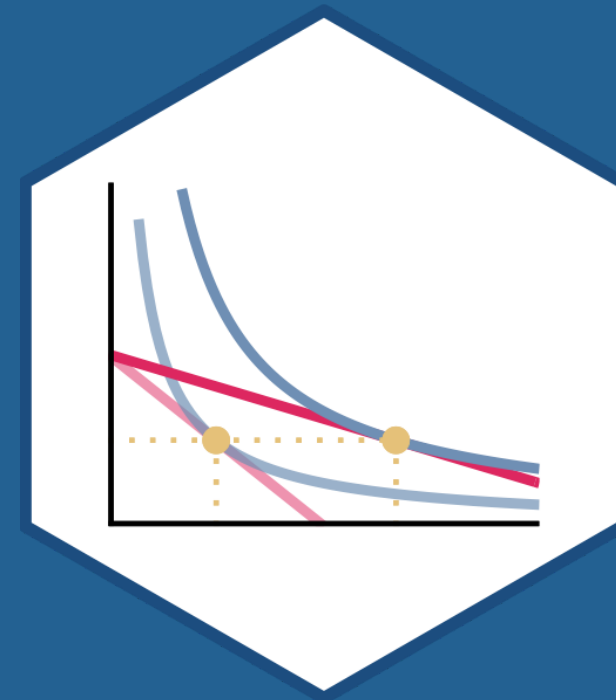
Ryan Safner

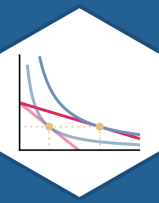
Assistant Professor of Economics

✉ [safner@hood.edu](mailto:safner@hood.edu)

🔗 [ryansafner/metricsF20](https://ryansafner/metricsF20)

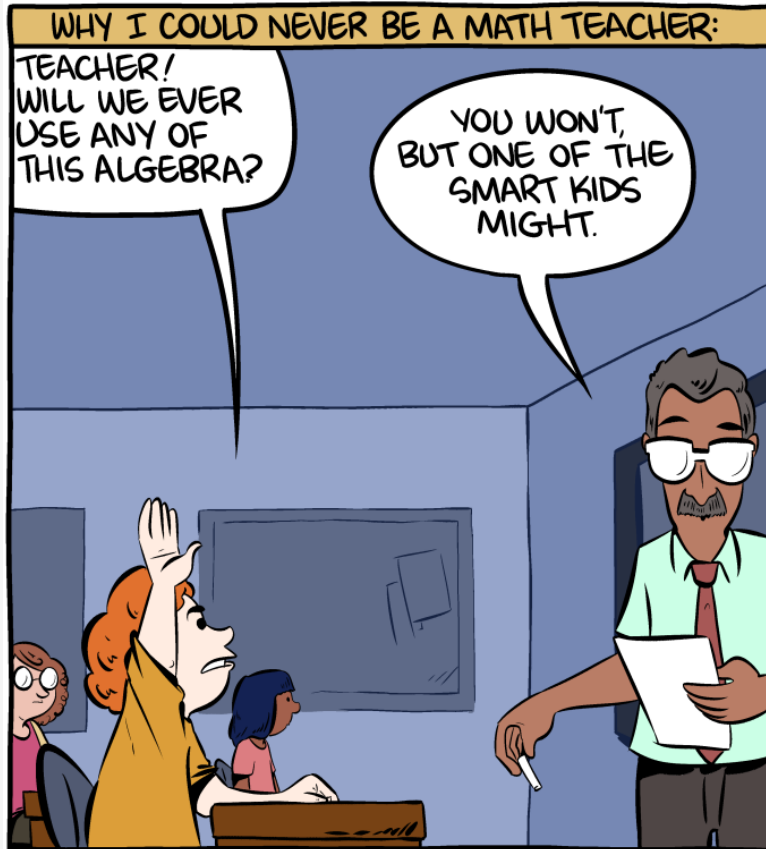
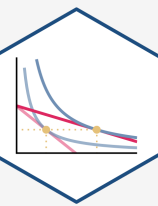
🌐 [metricsF20.classes.ryansafner.com](https://metricsF20.classes.ryansafner.com)





# What is Econometrics?

# Why Everyone, Yes *Everyone*, Should Learn Statistics

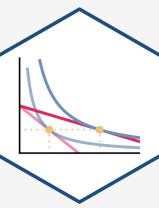


SMBC



SMBC

# We're Not so Good at Statistics: Votes I

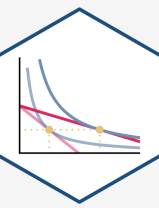


- Votes in the U.S. House of Representatives in favor of **passing** the *Civil Rights Act of 1964*:

Democrat	Republican
61%	80%

- Simple enough: "on average, Republicans tended to vote for passage more than Democrats"

# We're Not so Good at Statistics: Votes II

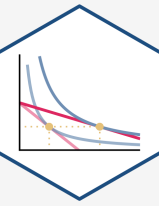


- Broken down further by Northern vs. Southern states:

	Democrat	Republican
<b>North</b>	<b>94%</b> (145/154)	85% (138/162)
<b>South</b>	<b>7%</b> (7/94)	0% (0/10)
<b>Overall</b>	61% (152/248)	<b>80%</b> (138/172)

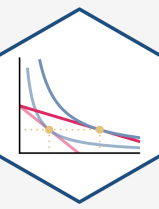
- Larger proportion of Democrats ( $\frac{94}{248}$ , 38%) than Republicans ( $\frac{10}{172}$ , 6%) were from South
- The 7% of southern Democrats voting *for* the Act dragged down the Democrats' *overall* percentage more than the 0% of southern Republicans

# We're Not So Good at Statistics: Kidney Stones I



- Suppose you suffer from kidney stones, your doctor offers you **treatment A** or **treatment B**
- In clinical trials, **Treatment A** was effective for a higher percentage of patients with *large* stones and a higher percentage of patients with *small* stones
- **Treatment B** was effective for a larger percentage of patients overall than **treatment A**
- Wait, what?

# We're Not So Good at Statistics: Kidney Stones II

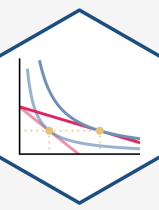


From a real medical study:

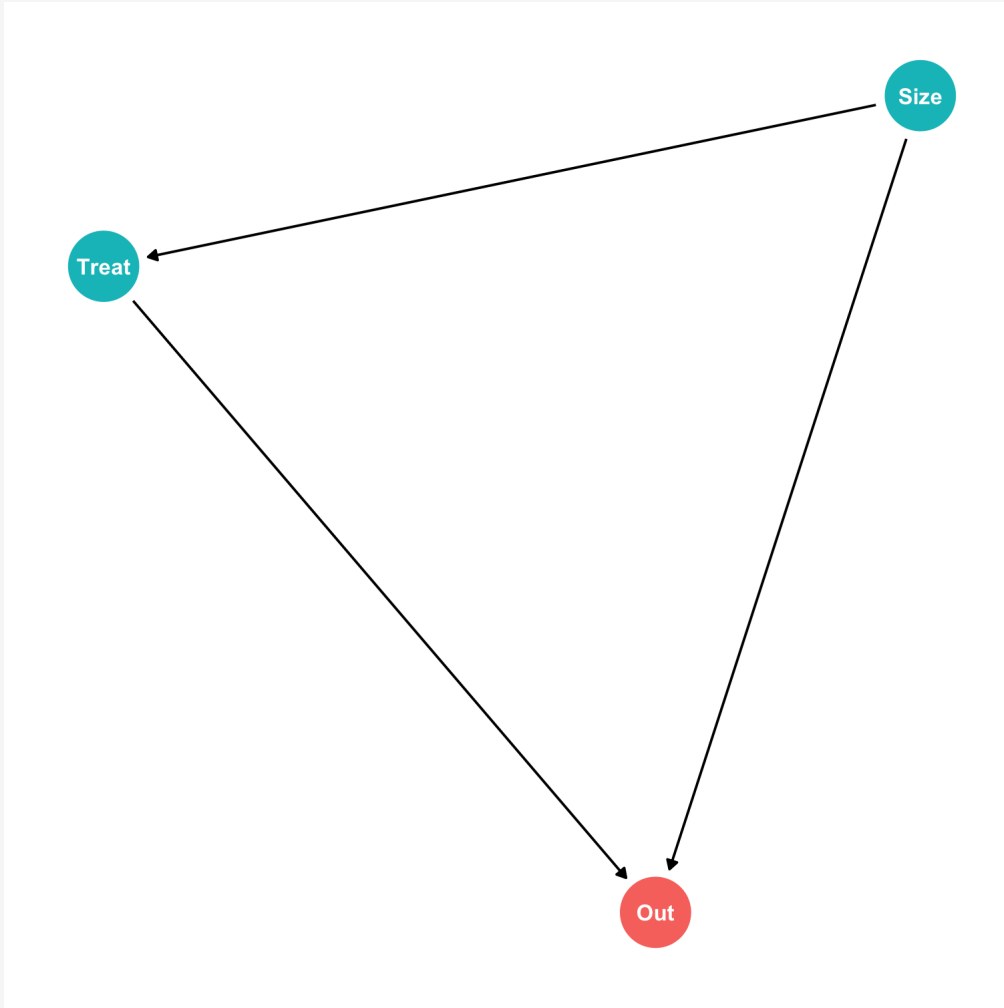
- The *sizes* of the two groups (i.e. who gets A vs B) are *very* different

	Treatment A	Treatment B
<b>Small Stones</b>	<b>93%</b> (81/87)	87% (234/270)
<b>Large Stones</b>	<b>73%</b> (192/263)	69% (55/80)
<b>Overall</b>	78% (273/350)	<b>83%</b> (289/350)

# We're Not So Good at Statistics: Kidney Stones III

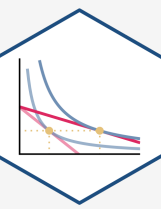


- The *sizes* of the two groups (i.e. who gets A vs B) are *very* different
- A **lurking variable** in the study is the severity of the case: doctors tended to give treatment B for less severe cases





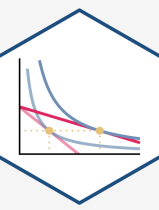
# Simpson's Paradox



**Simpson's Paradox:** The correlation between two variables can change (even reverse!) when additional variables are considered



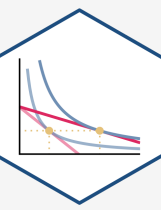
# We're Not so Good at Statistics: Smoking I



- 1964: U.S. Surgeon General issued a report claiming that cigarette smoking causes lung cancer
- Evidence based primarily on *correlations* between cigarette smoking and lung cancer



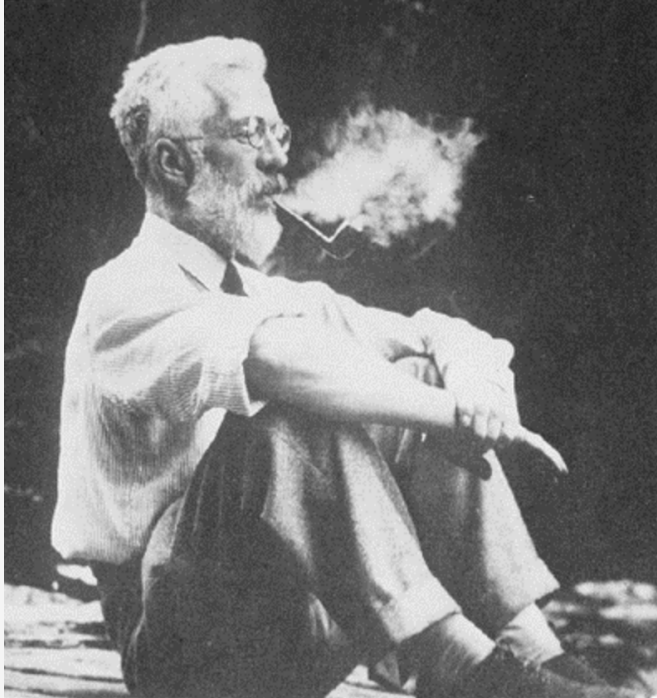
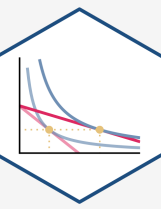
# We're Not so Good at Statistics: Smoking II



- Tobacco companies attacked the report, naturally



# We're Not so Good at Statistics: Smoking III

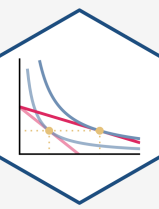


Ronald A. Fisher

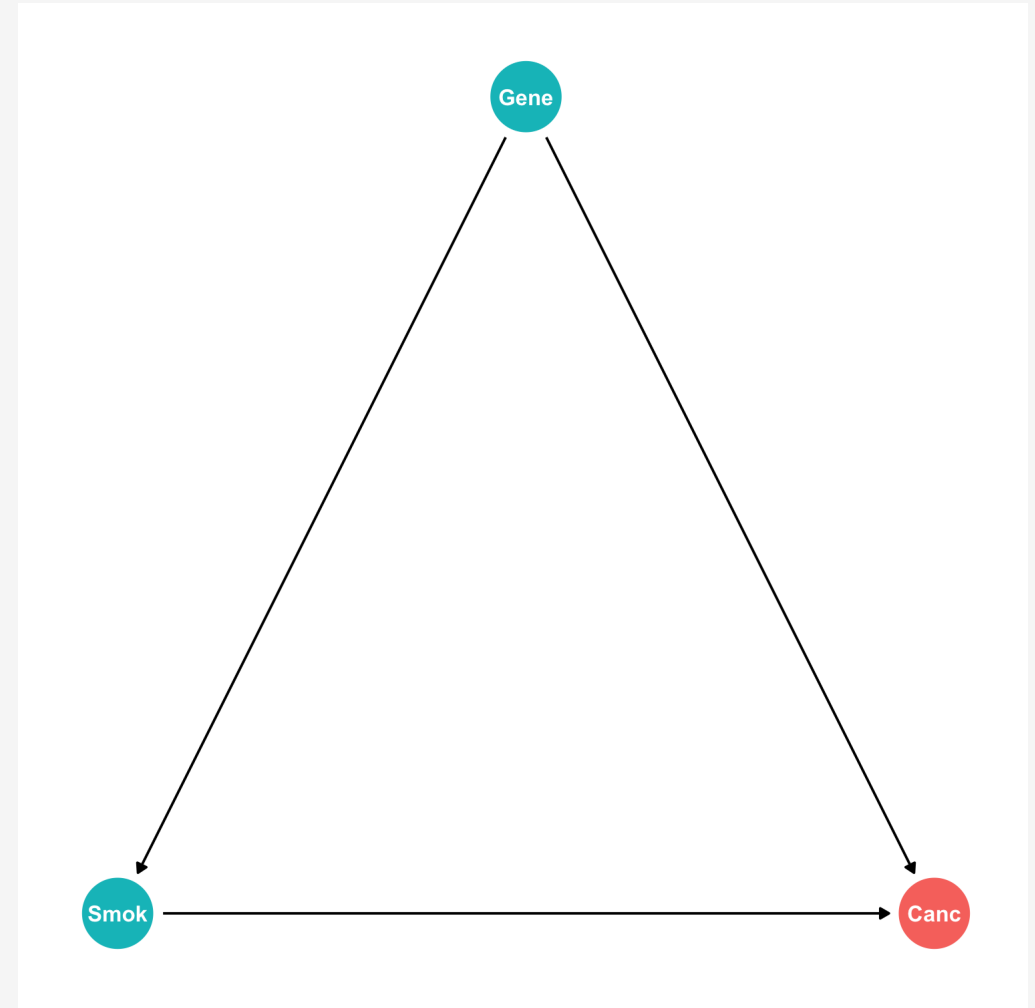
1890--1924

- But so did R. A. Fisher, the "father of modern statistics"

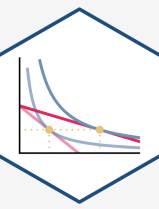
# We're Not so Good at Statistics: Smoking IV



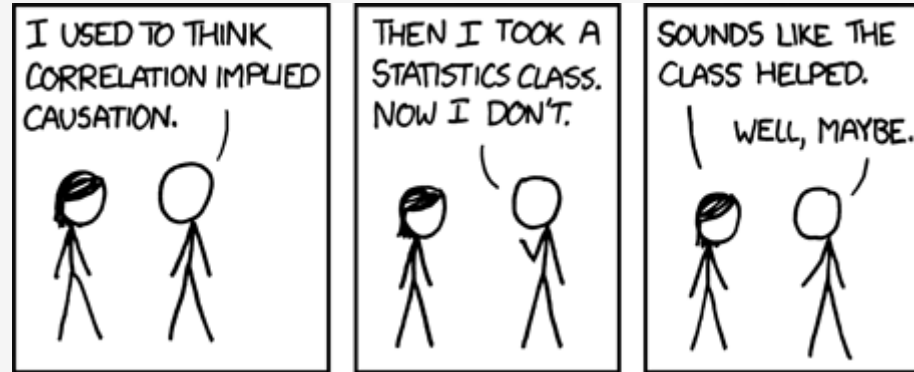
- There could be a confounding variable ("smoking gene") that causes *both* lung cancer *and* the urge to smoke
- Would imply: decision to smoke or not would have *no impact* on lung cancer!
- Correlation between smoking and cancer is spurious!



# Correlation Does Not Imply Causation I

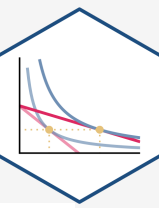


- The goal of every intro statistics class ever

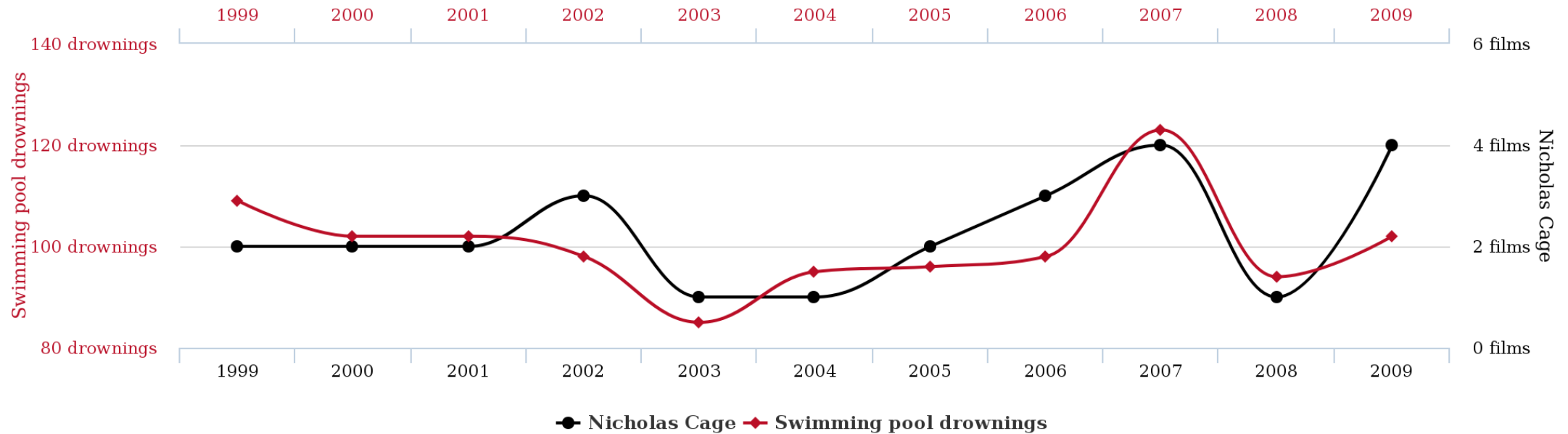


[XKCD: Correlation](#)

# Correlation Does Not Imply Causation II



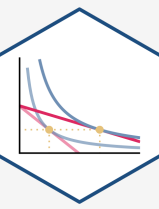
**Number of people who drowned by falling into a pool**  
correlates with  
**Films Nicolas Cage appeared in**



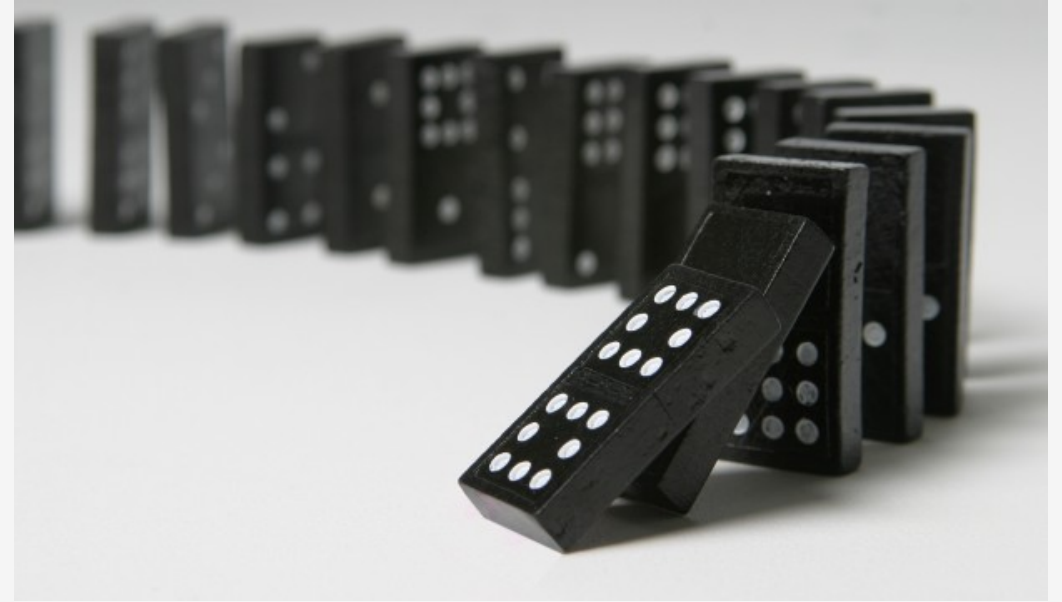
tylervigen.com

Spurious Correlations

# Correlation *Can* Imply Causation..

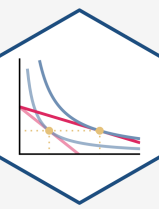


- It's always good to be skeptical of causal claims
- But this is actually where **econometrics** shines





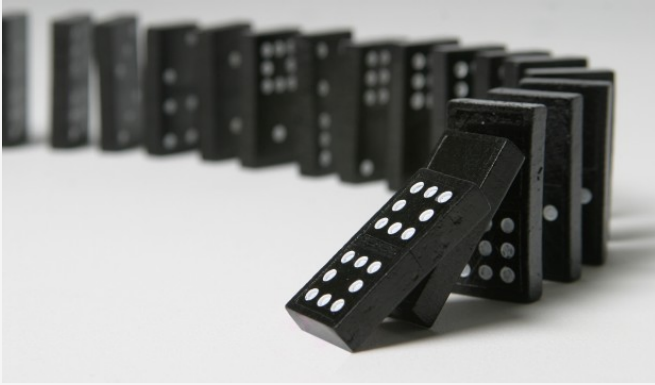
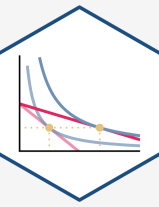
# ...With the Right Tools



- **Econometrics** is the application of statistical tools to *quantify* economic relationships in the real world
- Uses real data to
  - test economic hypotheses
  - quantitatively estimate the magnitude of relationships between economic variables
  - forecast future events

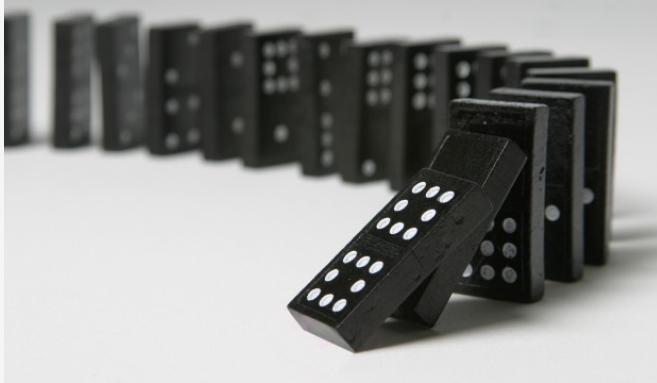
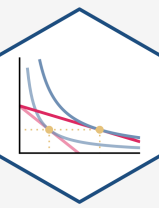


# Causal Inference I



- What sets econometrics apart from mere statistics (or uses of statistics in other disciplines) is its role in **causal inference**
- We can, with proper tools and interpretations, make *quantitative causal* claims
  - about the effects of individual choices
  - about the effects of policy interventions
  - about the impact of political institutions
  - about economic history and economic development
  - etc...

# Causal Inference II



A 50% increase in police presence in a metropolitan area lowers crime rates by 15%, on average<sup>1</sup>

Being an incumbent in office raises the probability of re-election by 40-45 percentage points<sup>2</sup>

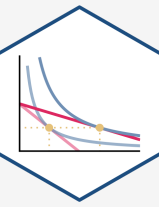
European cities with at least one printing press in 1500 were at least 29% more likely to become Protestant by 1600<sup>3</sup>

<sup>1</sup> Klick, Jonathan and Alexander Tabarrok, 2005, "Using Terror Alert Levels to Estimate the Effect of Police on Crime," *Journal of Law and Economics* 48(1): 267-279

<sup>2</sup> Lee, David S, 2001, "The Electoral Advantage to Incumbency and Voters' Valuation of Politicians' Experience: A Regression Discontinuity Analysis of Elections to the U.S," *NBER Working Paper* 8441

<sup>3</sup> Rubin, Jared, 2014, "Printing and Protestants: An Empirical Test of the Role of Printing in the Reformation," *Review of Economics and Statistics* 96(2): 270-286

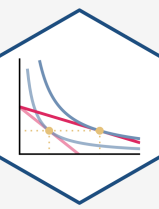
# Example 1: Education



Does reducing class sizes improve student performance?



# Example 1: Education

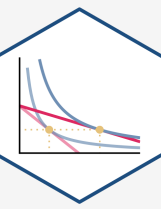


Does reducing class sizes improve student performance?

- A policy-relevant tradeoff with a budget constraint
- What is the *precise* effect of class size on performance?
- Is it worth hiring new teachers and building more schools over?



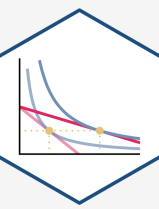
# Example 2: Discrimination in Lending



Is there racial discrimination in home mortgage lending?



# Example 2: Discrimination in Lending

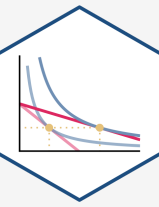


Is there racial discrimination in home mortgage lending?

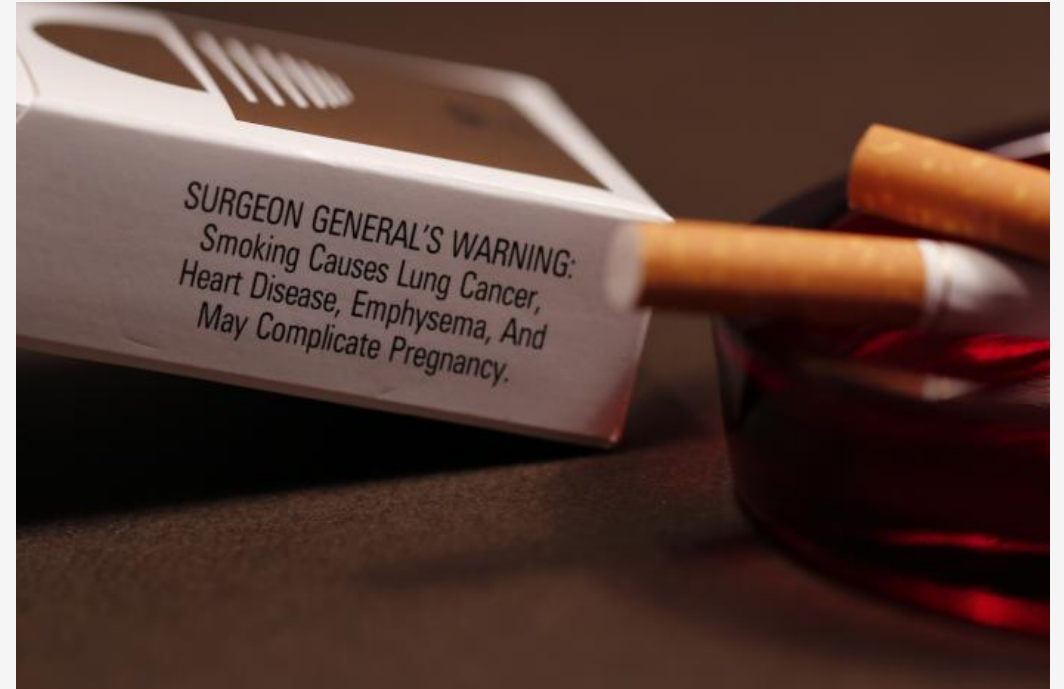
- Boston Fed: 28% of African-Americans are denied mortgages compared to only 9% of White Americans
- Is this due to factors such as credit history, income, or discrimination *purely* because of race?



# Example 3: Public Health and Public Finance

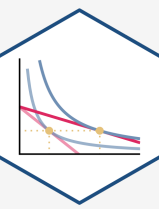


How much do state cigarette taxes reduce smoking rates?



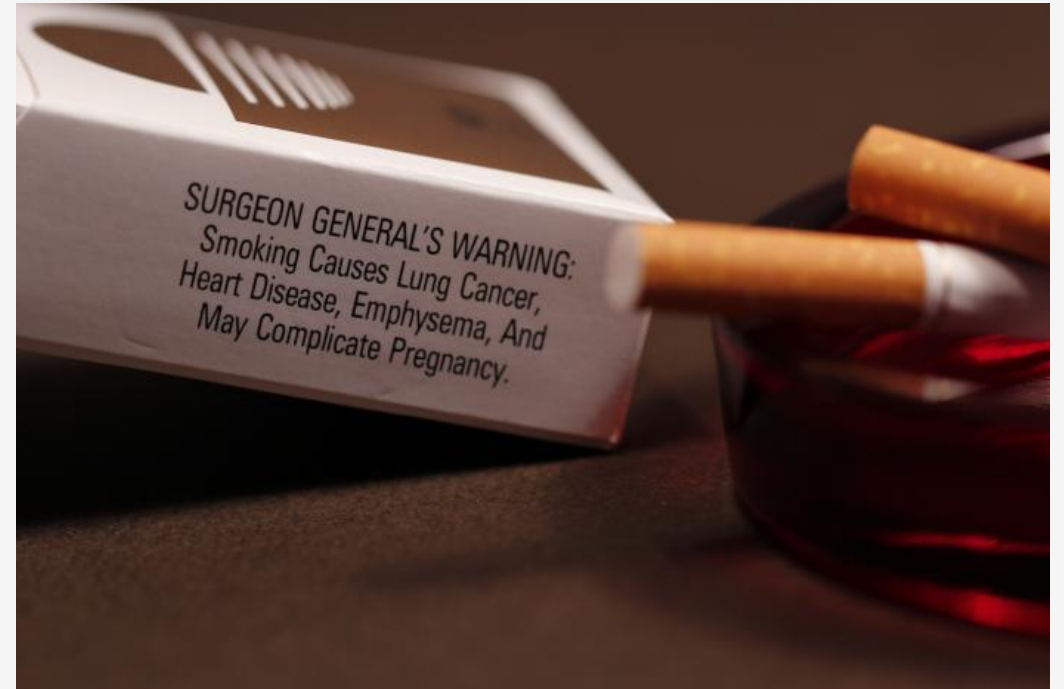


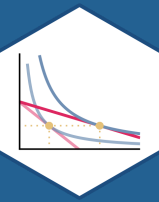
# Example 3: Public Health and Public Finance



How much do state cigarette taxes reduce smoking rates?

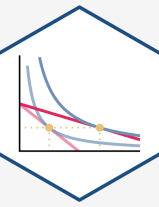
- Econ 101: raise price  $\implies$  lower quantity consumed
- What is the *price elasticity of demand* for smoking?
- How much tax revenue will this generate?
- Probably: *Taxes*  $\rightarrow$  *Smokers*
- Maybe?: *Taxes*  $\leftarrow$  *Smokers*



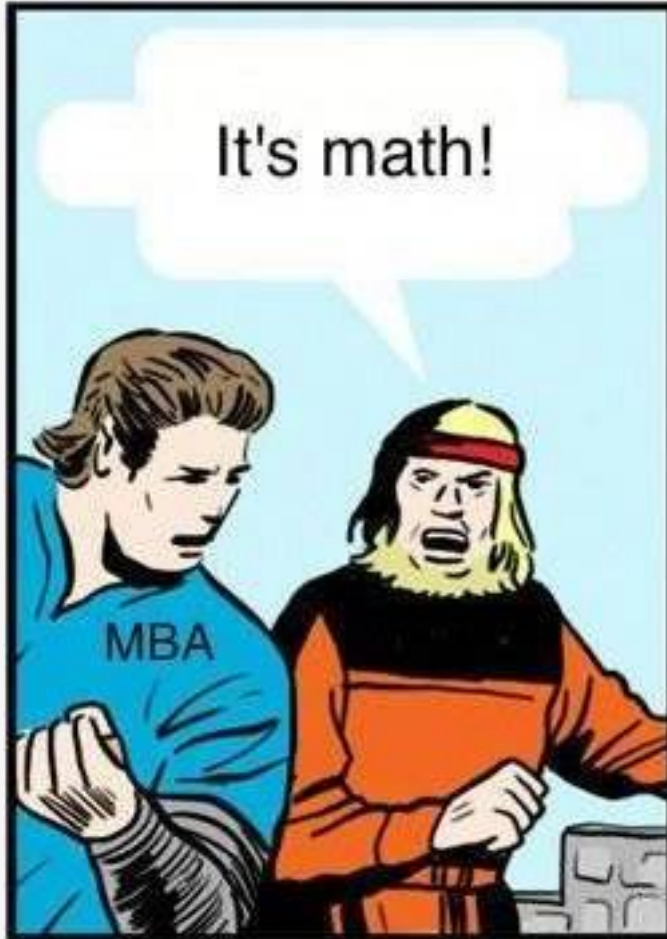
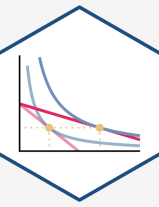


# About this Class

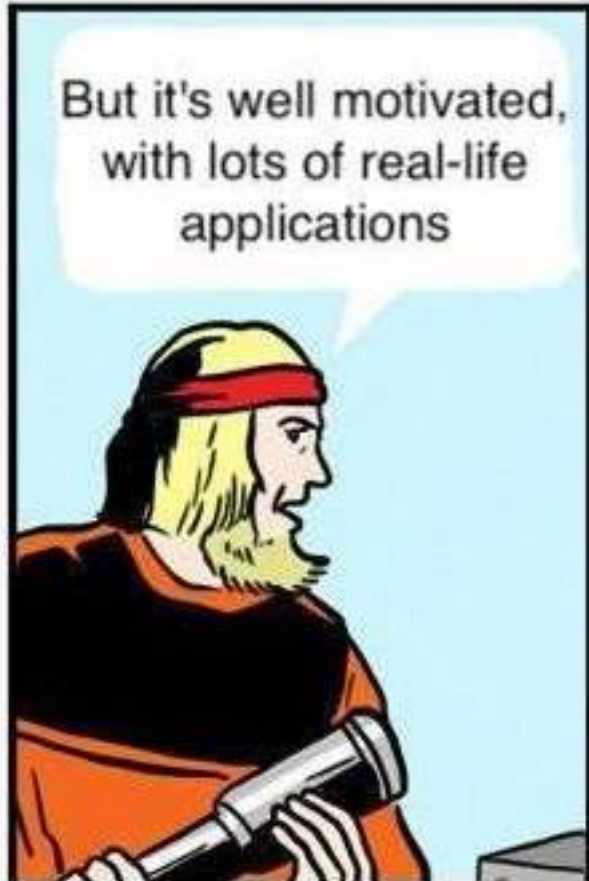
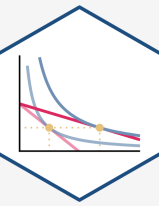
# Real Talk I



# Real Talk I



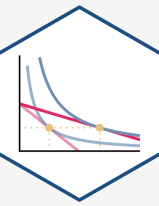
# Real Talk I



memecenter.com

MemeCenter

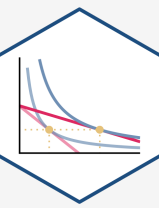
# Real Talk II



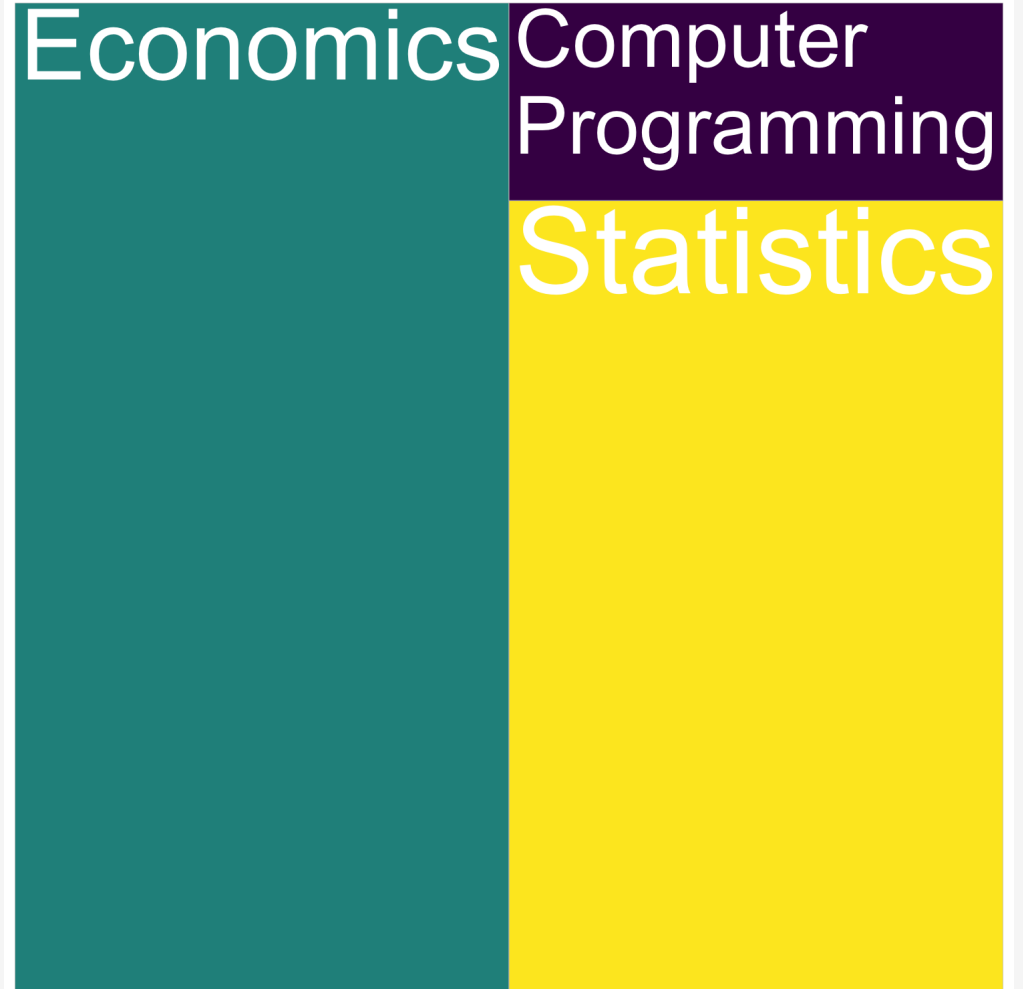
- This will be one of the hardest courses you take at Hood
- There will be moments where you have no idea WTF is going on (*this is normal*)
- Yes, you can still get an **A**



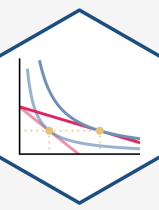
# This Class Is



- **Economics:** take your *preexisting* intuition and models for causal inference
- **Statistics:** add regression and statistical inference
- **Computer Programming:** using `R` and `R Studio` for analyzing and presenting data



# This Class Is



## Old School Statistics Courses

- $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$

- $\sigma_x = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$

- $r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$

- Use pre-cleaned "toy" data, if at all

## Hip New Data Science Courses

- `mean(x)`

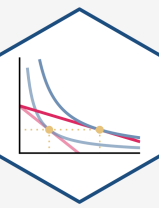
- `sd(x)`

- `cor(x, y)`

- Clean and manipulate raw data from scratch (like *real life!*)



# Prerequisites



- **Courses:**

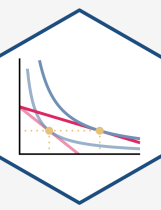
- ECON 205
- ECON 206
- ECON 305 or ECON 306
- MATH 112 or ECMG 212

- **Math Skills:**

- Basic algebra
- Probability-ish
- Statistics-ish

- **Computer Science Skills:**

# What You'll Get Out of This Class

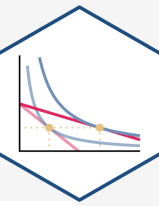


By the end of this semester, you will:

1. understand how to evaluate statistical and empirical claims;
2. use the fundamental models of causal inference and research design;
3. gather, analyze, and communicate with real data in R.



# This Class Opens Doors



Regressions!



[Redacted]@hood.edu>

Thu, Jun 6, 11:06 AM



to Ryan ▾

Hi Dr. Safner,

I hope all is well and you are enjoying the start to summer. I changed jobs in March to work as a researcher at an investment fund right outside of DC in Virginia.

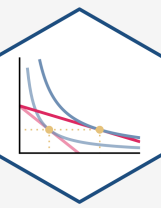
I often find myself running regressions (unfortunately my boss doesn't understand R so I have to use Python!) and using time series data; however I am in need of some notes from our courses together. Would you be willing to pass along past lecture presentations from Econometrics? I have hard copy notes to go along, but they are less useful than with the slides.

That is probably the single most influential class I have taken, especially given my new job function and I would certainly benefit from a refresher.

Best Regards,

[Redacted]  
Hood College, B.A. Economics  
Class of 2018

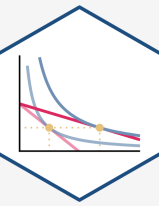
# This Class Gives You a Hybrid of Skills



- **"Data Science"**: ???
- **Causal Inference**: economists' comparative advantage!

Data Science	Causal Inference
-----------------	---------------------

# Data Science I



**Josh Wills**  
@josh\_wills

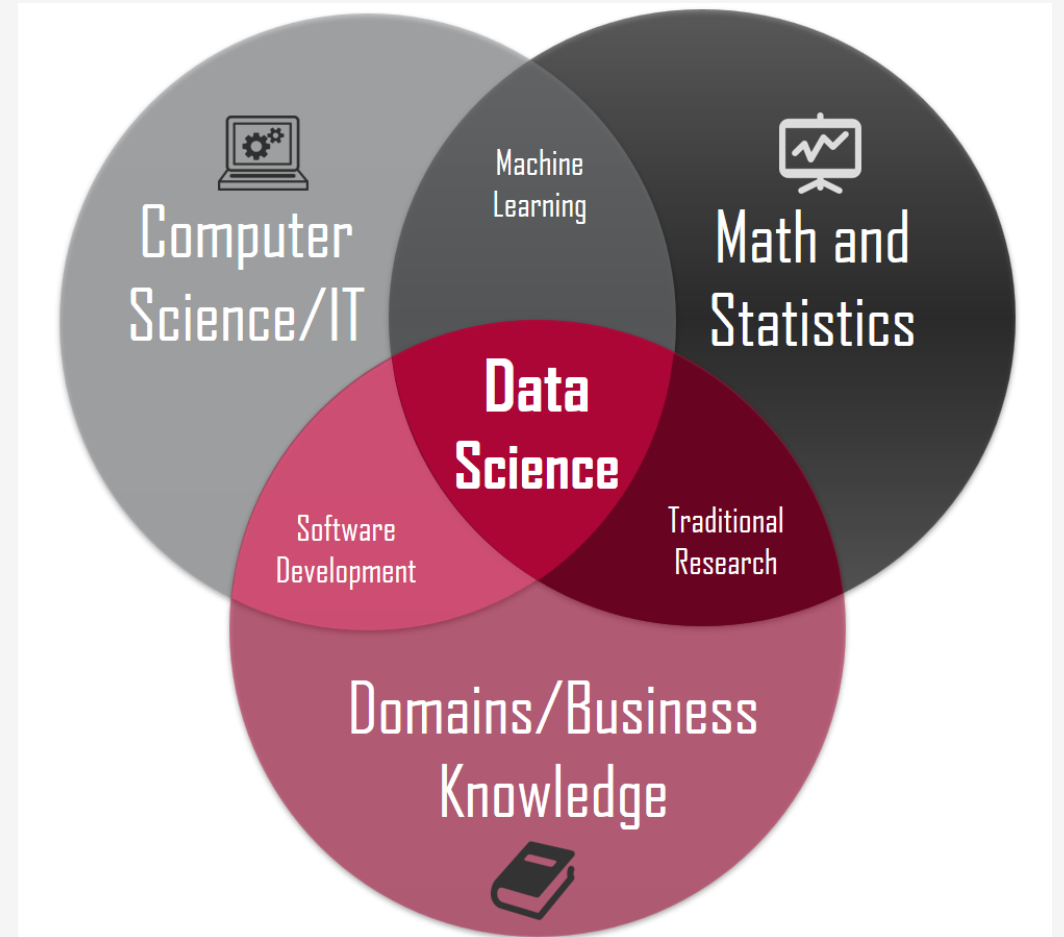


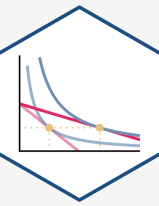
Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

12:55 PM · May 3, 2012



1.8K 1.9K people are Tweeting about this





## Harvard Business Review

Harvard Business Review

Subscribe

Latest Magazine Popular Topics Podcasts Video Store The Big Idea Visual Library Reading Lists Case Selections

ARTWORK: TAMAR COHEN, ANDREW J. BUBOLTZ, 2011, SILK SCREEN ON A PAGE FROM A HIGH SCHOOL YEARBOOK, 8.5" X 12"

**Data**

### Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

Summary Save Share Comment 16 Text Size Print \$8.95 Buy

4/6 FREE ARTICLES LEFT > SUBSCRIBE TO ACCESS THE ARCHIVE

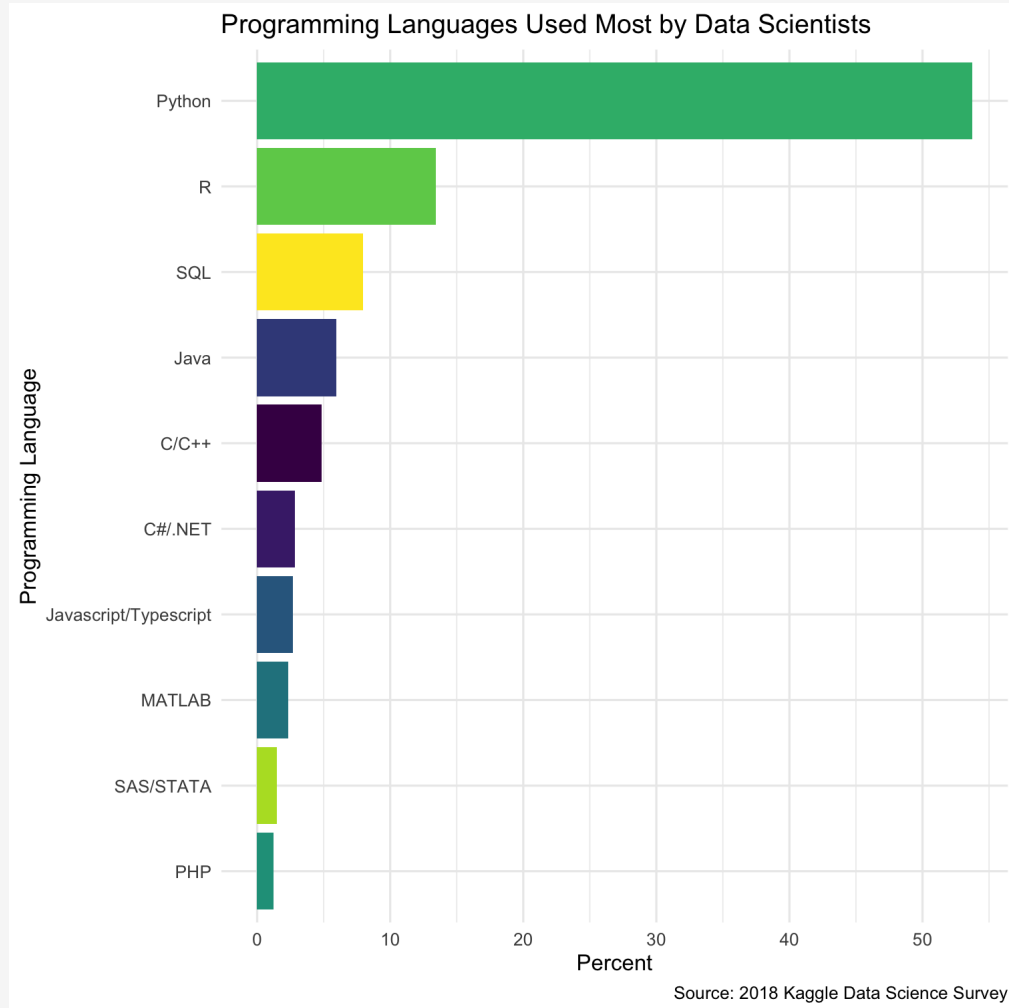
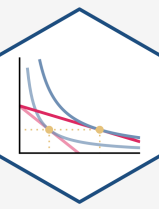
## LinkedIn 2018 Emerging Jobs Report

LinkedIn Economic Graph

Research Resources Blog About

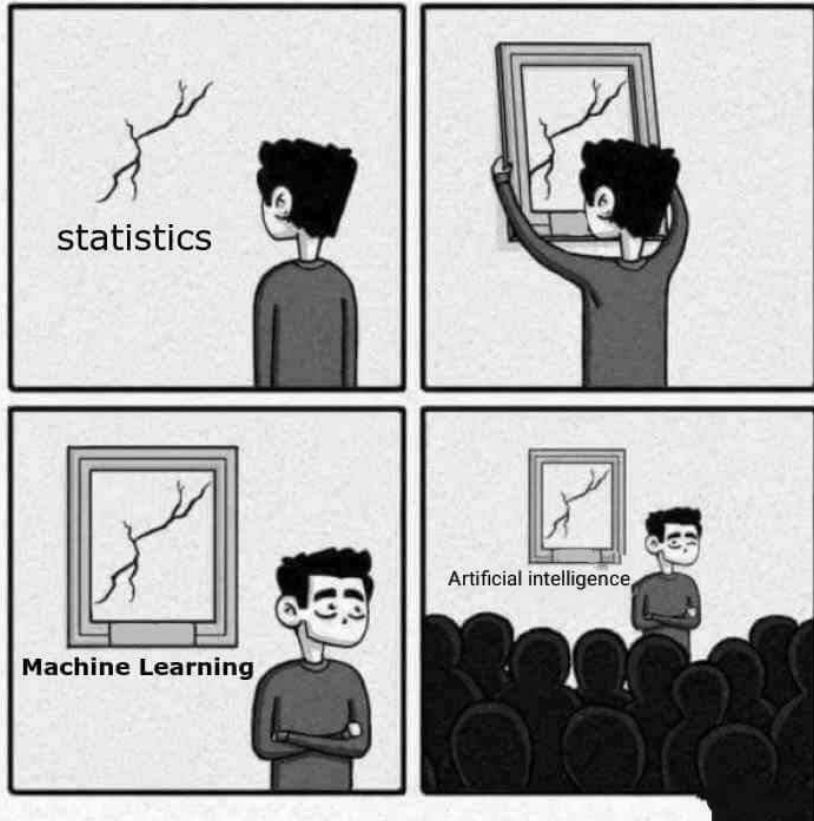
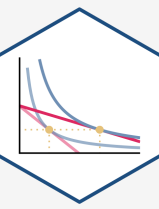
- 6. Relationship Consultant** (5.5X growth)
  - Top Skills:** Banking, Retail Banking, Loans, Consumer Lending, Credit
  - Where They Work:** Regions Bank, Merrill Edge, Vanguard
  - Top Industries:** Banking, Financial Services, Insurance
  - Cities Where Demand is High:** Jacksonville, New York City, St. Louis
- 7. Data Science Specialist** (5X growth)
  - Top Skills:** Machine Learning, Data Science, Python, R, Apache Spark
  - Where They Work:** IBM, Facebook, McKinsey & Company
  - Top Industries:** Higher Education, Information Technology & Services, Computer Software
  - Cities Where Demand is High:** New York City, San Francisco, Chicago
- 8. Assurance Staff** (5X growth)
  - Top Skills:** Auditing, Accounting, Financial Reporting, Internal Controls
  - Where They Work:** EY, Plante Moran, Moss Adams
  - Top Industries:** Accounting, Higher Education, Financial Services
  - Cities Where Demand is High:** Detroit, Philadelphia, Boston
- 9. Sales Development Representative** (4X growth)
  - Top Skills:** Salesforce, Cold Calling, Software-as-a-Service, Lead Generation, Sales Prospecting

# R Skills are In Demand



[Kaggle Data Science Survey 2018](#)

# Yada Yada Machine Learning

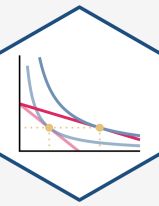


"When you're fundraising, it's AI. When you're hiring, it's ML. When you're implementing, it's logistic regression."

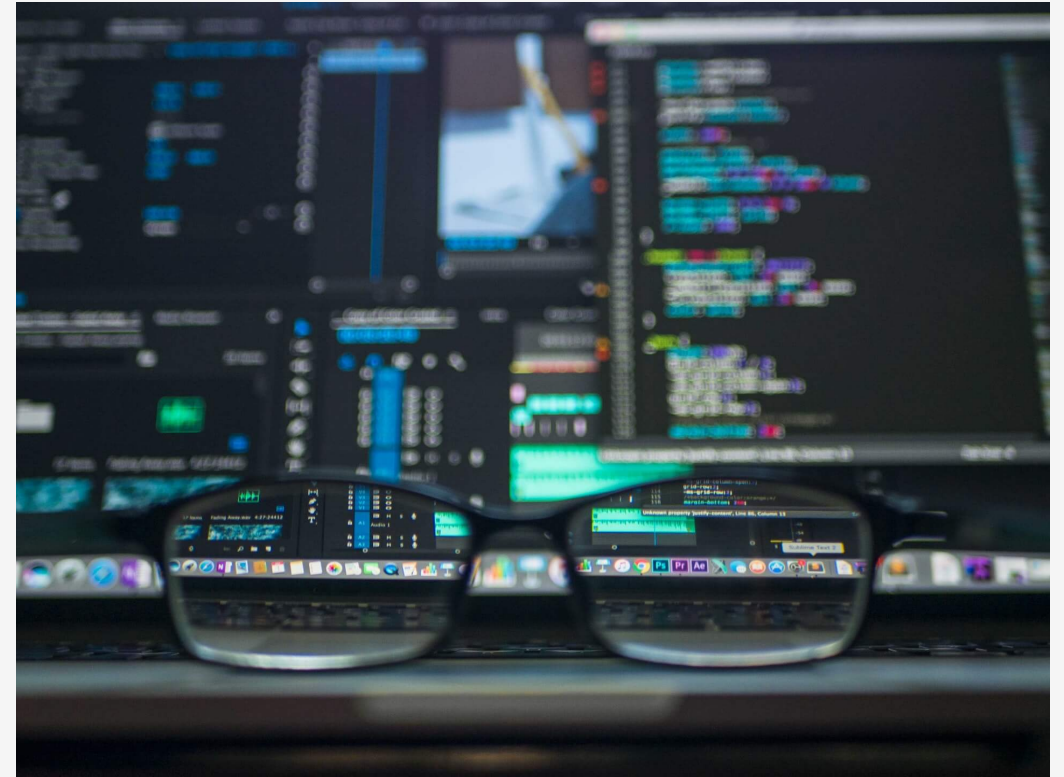
- everyone on Twitter ever ([Source](#))



# Causal Inference I

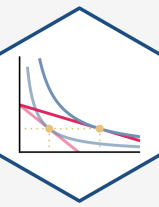


- Machine learning and artificial intelligence are "dumb"<sup>1</sup>
- With the right models and research designs, we *can* say "X causes Y" and quantify it!
- Economists are in a unique position to make *causal* claims that mere statistics cannot



<sup>1</sup> For more, see [my blog post](#), and Pearl & MacKenzie (2018), *The Book of Why*

# Causal Inference II

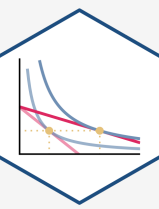


The screenshot shows the Harvard Business Review website. The main article is titled "Why Tech Companies Hire So Many Economists" by Susan Athey and Michael Luca, dated February 10, 2019. The article features a large, colorful abstract graphic with purple and yellow lines. Below the graphic, the text begins: "Over a recent coffee discussion, the COO of a multi billion-dollar tech company told us he was interested in building an internal team of economists. He had collaborated with economists before and was excited about the results. He now wanted to know how to...". To the right of the main article, there are several recommended items: "Using Experiments to Launch New Products", "The Latest Research: AI and Machine Learning" (a press toolkit for \$49.95), "Trader Joe's" (a case study for \$8.95), and "Artificial Intelligence Set: What You Need to...". At the bottom of the page, there is a red banner that says "5/6 FREE ARTICLES LEFT > SUBSCRIBE TO ACCESS THE ARCHIVE".

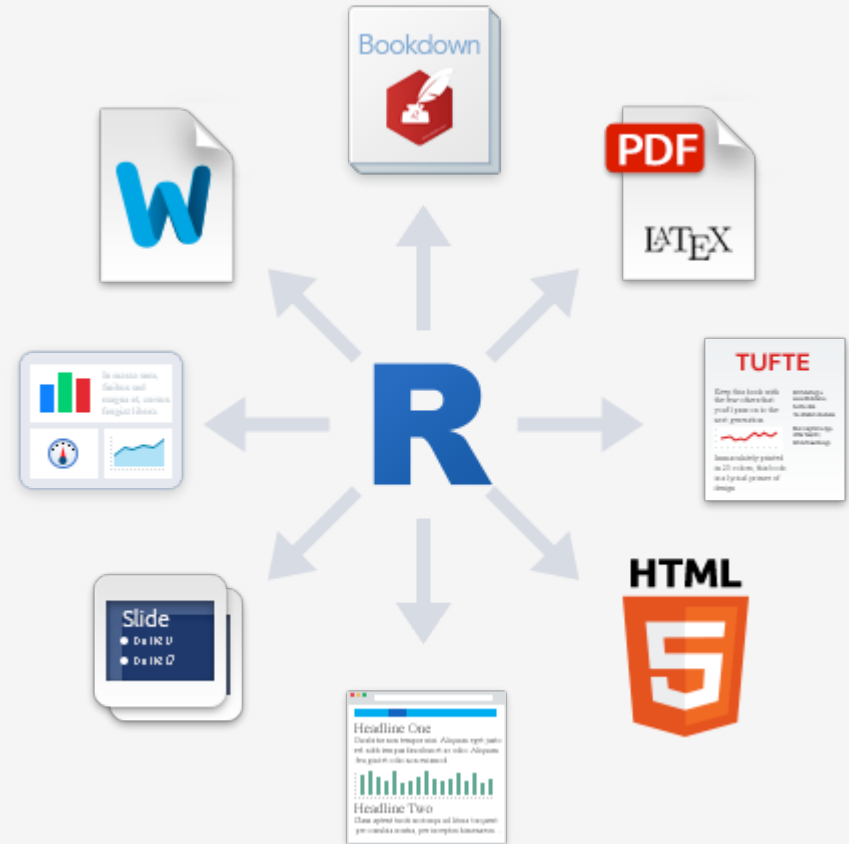
## Harvard Business Review

"First, the field of economics has spent decades developing a toolkit aimed at investigating empirical relationships, focusing on techniques to help understand which correlations speak to a causal relationship and which do not. This comes up all the time — does Uber Express Pool grow the full Uber user base, or simply draw in users from other Uber products? Should eBay advertise on Google, or does this simply syphon off people who would have come through organic search anyway? Are African-American Airbnb users rejected on the basis of their race? These are just a few of the countless questions that tech companies are grappling with, investing

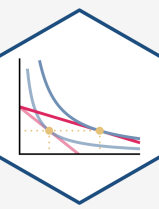
# Building Good Workflow Habits



- I will show you the tools to make your workflow:
  - Reproducible
  - Computer- and Human-Readable (!)
  - Automated
  - All in one program

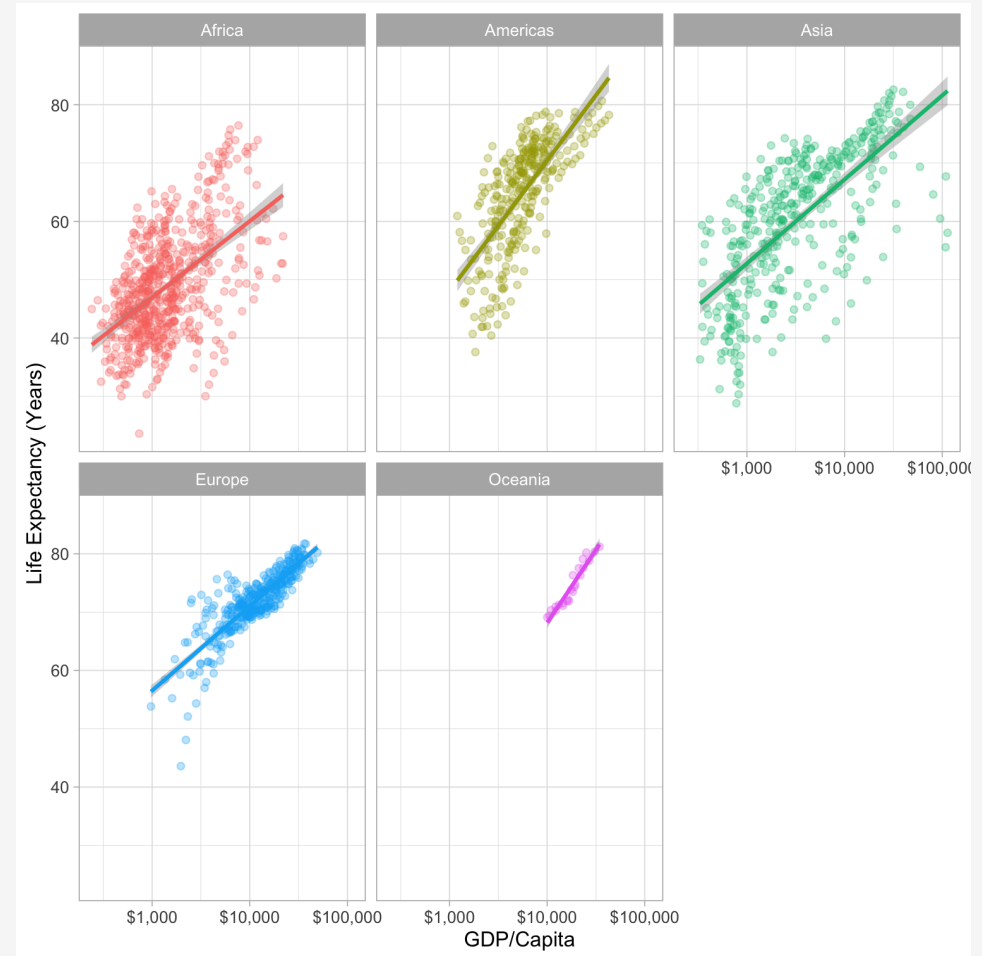


# A Quick Example

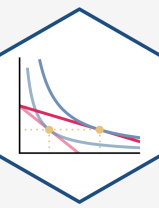


```
library("gapminder")

ggplot(data = gapminder,
       aes(x = gdpPercap,
           y = lifeExp,
           color = continent)) +
  geom_point(alpha=0.3) +
  geom_smooth(method = "lm") +
  scale_x_log10(breaks=c(1000,10000,
                        label=scales::dollar
  labs(x = "GDP/Capita",
       y = "Life Expectancy (Years)")
  facet_wrap(~continent) +
  guides(color = F) +
  theme_light()
```

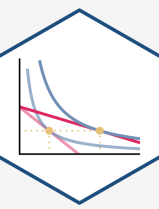


# Logistics: Hybrid Course



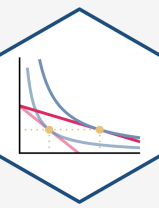
- **hybrid**: more **synchronous** material than **asynchronous** material
- I will always be teaching **remotely**
  - A classroom is available to you
  - I may make occasional visits to campus if you *need* something in person (TBD)
- Office hours: Tu/Th 3:30-5:00 PM on Zoom
  - Zoom link in Blackboard's **LIVE CLASS SESSIONS** link
  - Slack channels
- Teaching Assistant(s): TBD
  - grade HWs & hold (likely virtual) office hours

# Logistics: Hybrid Course



- We will have **synchronous** sessions Mon/Wed 2:00-3:15 PM on **Zoom**
- Lecture videos will be posted on **Blackboard** via Panopto for students unable to join synchronously
  - If you were present, you do not need to watch the video (again)!
  - You are not *required* to attend synchronously, but it will help you
- All graded assignments are **asynchronous**
  - (Probably) submitted on Blackboard by 11:59 PM Sundays
  - (Probably) timed exams on Blackboard

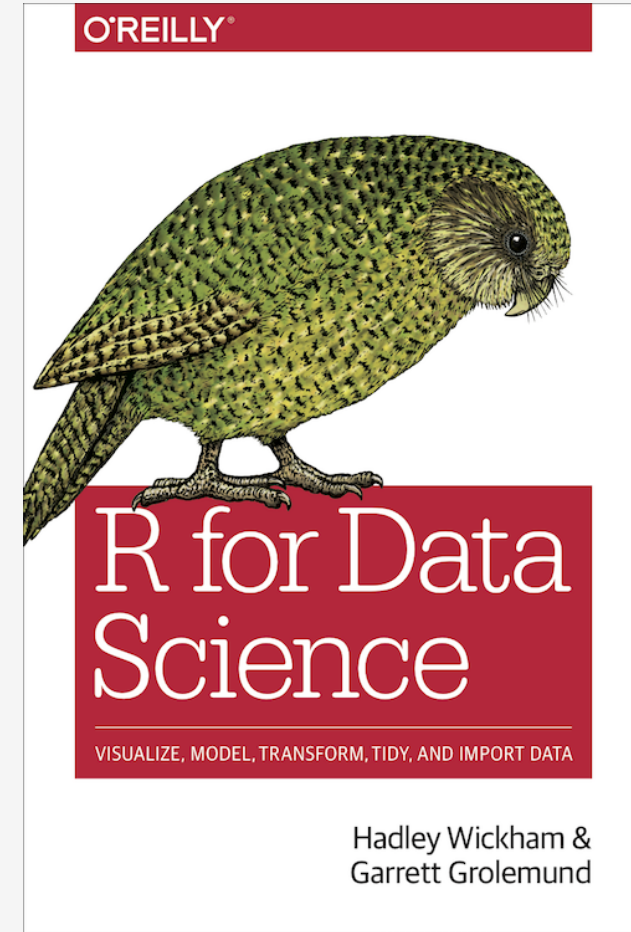
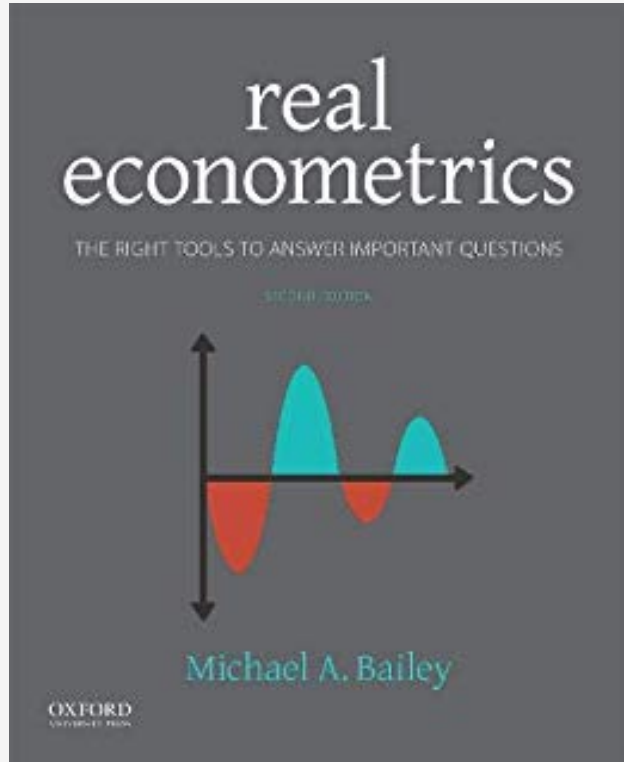
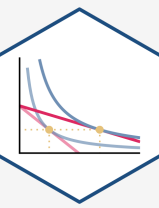
# Assignments



- Research project:
  - Come up with a testable research question
  - Find data
  - Analyze data
  - Present your results (in writing and verbally)
- HWs
- Midterm, Final exam (in-class, closed notes)

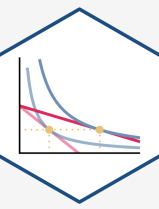
	<b>Assignment</b>	<b>Percent</b>
1	Research Project	30%
n	Homeworks (Average)	25%
1	Midterm	20%
1	Final	25%

# Your "Textbooks"

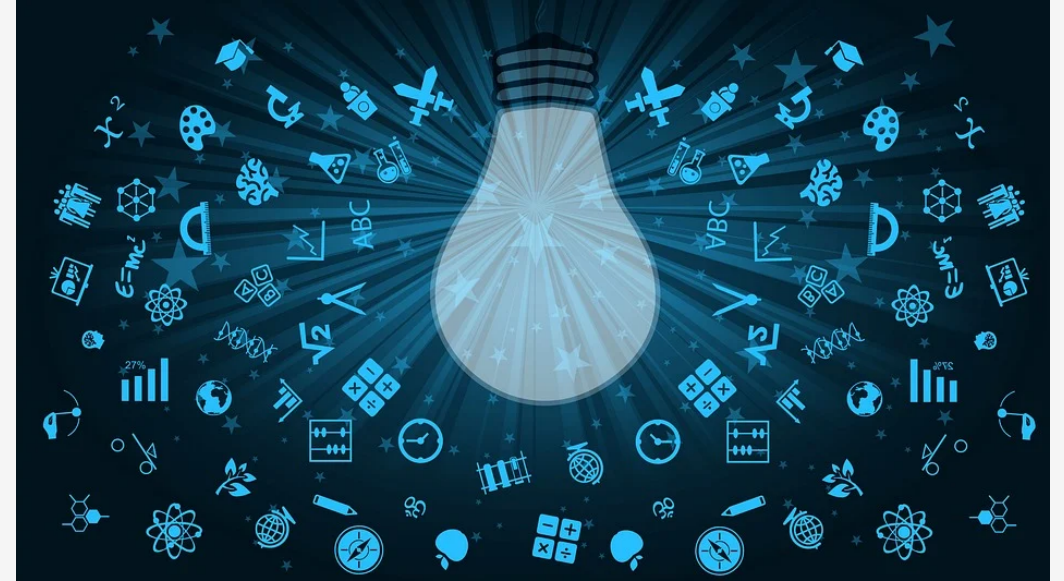




# Tips for Success In This Course



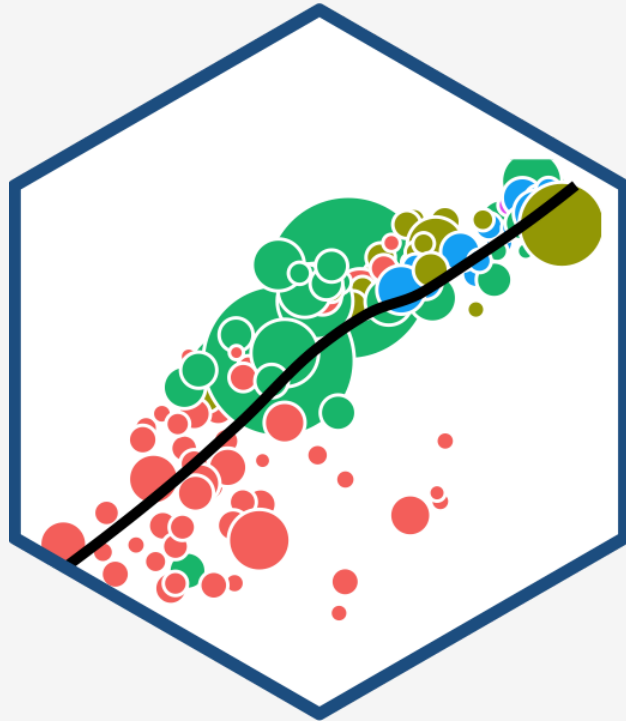
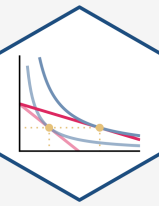
- *Take notes. On paper. Really.*
- **Work together** on assignments and study together.
- Ask questions, come to office hours.  
Don't struggle in silence, you are not alone!



- **You are learning how to learn**<sup>1</sup>
- See the [reference page](#) for more

<sup>1</sup> A properly worded Google search will become your secret weapon. Believe me. It's still mine.

# Course Website



metricsf20.classes.ryansafner.com

## ECON 480: ECONOMETRICS

SYLLABUS SCHEDULE ASSIGNMENTS REFERENCE RSTUDIO.CLOUD SLACK

### SCHEDULE

This page contains all of the following resources for each class meeting:

- Readings** include textbook chapters and occasional journal articles
- Assignments** are due by the beginning of class unless otherwise stated
- Class** materials contain more details, math appendices, and other helpful resources<sup>1</sup>
- Slides** are "Xaringan" presentations in html that can be opened in any browser<sup>2</sup>
- R** materials contain extra tutorials, videos, practice exercises for using R

Relevant materials (if applicable, icons will become links) will be posted before class meets.

**Last Update:** 21:30:59 Mon Aug 17 2020

I. DATA ANALYSIS IN R	READING	CLASS	SLIDES	ASSIGNMENT
Preliminary Survey				
1.1 Introduction to Econometrics				
1.2 Meet R				
1.3 Data Visualization with ggplot2				
1.4 Data Wrangling in the tidyverse				
1.5 Optimize Workflow: Projects, Markdown, and Git				
<b>Problem Set 1</b>				

[metricsf20.classes.ryansafner.com](https://metricsf20.classes.ryansafner.com) ]

# Roadmap for the Semester

