Introduction and Overview

ECO 6973, Set 1

Jonathan Moreno-Medina
Fall 2021

Prologue

Acknowledgements

First things first! Huge thanks to Ed Rubin for publicly sharing his slides and templates.

=)

About me

- Parents first gen (P.E. and Communications)
- Undergrad in Colombia, MA in Belgium, PhD in North Carolina
- Applied microeconomist (i.e. data to study well-defined markets and policies)
- Focus on Public, Urban and Media Economics
- Moved recently to San Antonio -- Really like it here! (I'm still getting used to the heat, though)

Why the course?

Motivation

This course:

Why the course?

Motivation

This course:

1. What is the goal of causal inference?

Why the course?

Motivation

This course:

1. What is the goal of causal inference?

One simple answer: "Leverage theory and deep knowledge of institutional details to estimate the **impact** of events and choices on a given outcome of interest." - Cunningham, 2021

1. What is the goal of machine learning?

Why the course?

Motivation

This course:

1. What is the goal of causal inference?

One simple answer: "Leverage theory and deep knowledge of institutional details to estimate the **impact** of events and choices on a given outcome of interest." - Cunningham, 2021

1. What is the goal of machine learning?

One simple answer: "To automatically build robust **predictions** from complex data." - Taddy, 2021

Why?

Related concept: what is big data?

Could mean **long** (many observations - think IRS and income), or **wide** (many variables - think Amazon or Google)

Why?

Related concept: what is big data?

Could mean **long** (many observations - think IRS and income), or **wide** (many variables - think Amazon or Google)

1. Causal inference requires **long** data
2. Machine learning developed in a framework with **wide** data

Why?

Topics covered:

Why?

Topics covered:

1. Causal inference: Potential outcomes framework, DAGS, RCTs, PS, RD, IV, event study, dif-in-dif, synthetic control
2. Machine learning: LASSO, Decision trees, Random Forest, Neural Networks, NLP

Small Digression

This class:

1. Presentation: 30 minutes on chosen material. Select the presentation you would like to give; ties will be sorted out randomly

Small Digression

This class:

1. Presentation: 30 minutes on chosen material. Select the presentation you would like to give; ties will be sorted out randomly
2. Homework: hands-on working with data implementing and interpreting code

Small Digression

This class:

1. Presentation: 30 minutes on chosen material. Select the presentation you would like to give; ties will be sorted out randomly
2. Homework: hands-on working with data implementing and interpreting code
3. Participation: Key to engage with presenters, and when discussing papers

Why?

Example

Before we dive into those, consider this:

GPA is an output from endowments (ability) and hours studied (inputs). So, one might hypothesize a model

$$\text{GPA} = f(H, \text{SAT}, \text{PCT})$$

where H is hours studied, SAT is SAT score and PCT is the percentage of classes an individual attended. We expect that GPA will rise with each of these variables (H, SAT, and PCT).

But who needs to *expect*?

We can test these hypotheses **using a regression model**.

Why?

Example, cont.

Regression model:

$$\text{GPA}_i = \beta_0 + \beta_1 H_i + \beta_2 \text{SAT}_i + \beta_3 \text{PCT}_i + \varepsilon_i$$

We want to test estimate/test the relationship $\text{GPA} = f(H, \text{SAT}, \text{PCT})$.

Why?

Example, cont.

Regression model:

$$\text{GPA}_i = \beta_0 + \beta_1 H_i + \beta_2 \text{SAT}_i + \beta_3 \text{PCT}_i + \varepsilon_i$$

(Review) Questions

Example, cont.

Regression model:

$$\text{GPA}_i = \beta_0 + \beta_1 H_i + \beta_2 \text{SAT}_i + \beta_3 \text{PCT}_i + \varepsilon_i$$

(Review) Questions

- **Q:** How do we interpret β_1?

Example, cont.

Regression model:

$$\text{GPA}_i = \beta_0 + \beta_1 H_i + \beta_2 \text{SAT}_i + \beta_3 \text{PCT}_i + \varepsilon_i$$

(Review) Questions

- **Q:** How do we interpret β_1?
- **A:** An additional hour in class correlates with a β_1 unit increase in an individual's GPA (controlling for SAT and PCT).

Example, cont.

Regression model:

$$\text{GPA}_i = \beta_0 + \beta_1 H_i + \beta_2 \text{SAT}_i + \beta_3 \text{PCT}_i + \varepsilon_i$$

(Review) Questions

- **Q:** How do we interpret β_1?
- **A:** An additional hour in class correlates with a β_1 unit increase in an individual's GPA (controlling for SAT and PCT).

- **Q:** Are the β_k terms population parameters or sample statistics?

Why?

Example, cont.

Regression model:

$$\text{GPA}_i = \beta_0 + \beta_1 H_i + \beta_2 \text{SAT}_i + \beta_3 \text{PCT}_i + \varepsilon_i$$

(Review) Questions

- **Q:** How do we interpret β_1?
- **A:** An additional hour in class correlates with a β_1 unit increase in an individual's GPA (controlling for SAT and PCT).

- **Q:** Are the β_k terms population parameters or sample statistics?
- **A:** Greek letters denote **population parameters**. Their estimates get hats, *e.g.*, $\hat{\beta}_k$.

Why?

Example, cont.

Regression model:

$$\text{GPA}_i = \beta_0 + \beta_1 H_i + \beta_2 \text{SAT}_i + \beta_3 \text{PCT}_i + \varepsilon_i$$

(Review) Questions

- **Q:** Can we interpret the estimates for β_2 as causal?

Why?

Example, cont.

Regression model:

$$\text{GPA}_i = \beta_0 + \beta_1 H_i + \beta_2 \text{SAT}_i + \beta_3 \text{PCT}_i + \varepsilon_i$$

(Review) Questions

- **Q:** Can we interpret the estimates for β_2 as causal?
- **A:** Not without making more assumptions and/or knowing more about the data-generating process.

Example, cont.

Regression model:

$$\text{GPA}_i = \beta_0 + \beta_1 H_i + \beta_2 \text{SAT}_i + \beta_3 \text{PCT}_i + \varepsilon_i$$

(Review) Questions

- **Q:** Can we interpret the estimates for β_2 as causal?
- **A:** Not without making more assumptions and/or knowing more about the data-generating process.

- **Q:** What is ε_i?

Why?

Example, cont.

Regression model:

$$\text{GPA}_i = \beta_0 + \beta_1 H_i + \beta_2 \text{SAT}_i + \beta_3 \text{PCT}_i + \varepsilon_i$$

(Review) Questions

- **Q:** Can we interpret the estimates for β_2 as causal?
- **A:** Not without making more assumptions and/or knowing more about the data-generating process.

- **Q:** What is ε_i?
- **A:** An individual's random deviation/disturbance from the population parameters.

Why?

Example, cont.

Regression model:

$$\text{GPA}_i = \beta_0 + \beta_1 H_i + \beta_2 \text{SAT}_i + \beta_3 \text{PCT}_i + \varepsilon_i$$

(Review) Questions

- **Q:** Which assumptions do we impose when estimating with OLS?

Why?

Example, cont.

Regression model:

$$\text{GPA}_i = \beta_0 + \beta_1 H_i + \beta_2 \text{SAT}_i + \beta_3 \text{PCT}_i + \varepsilon_i$$

(Review) Questions

- **Q:** Which assumptions do we impose when estimating with OLS?
- **A:**
  - The relationship between the GPA and the explanatory variables is linear in parameters, and ε enters additively.
  - The explanatory variables are **exogenous**, *i.e.*, $E[\varepsilon|X] = 0$.
  - You've also typically assumed something along the lines of: $E[\varepsilon_i] = 0$, $E[\varepsilon_i^2] = \sigma^2$, $E[\varepsilon_i \varepsilon_j] = 0$ for $i \neq j$.
  - And (maybe) ε_i is distributed normally.

Assumptions

How important can they be?

You've learned how **powerful and flexible** ordinary least squares (**OLS**) regression can be.

However, the results you learned required assumptions.

Real life often violates these assumptions.

Assumptions

How important can they be?

You've learned how **powerful and flexible** ordinary least squares (**OLS**) regression can be.

However, the results you learned required assumptions.

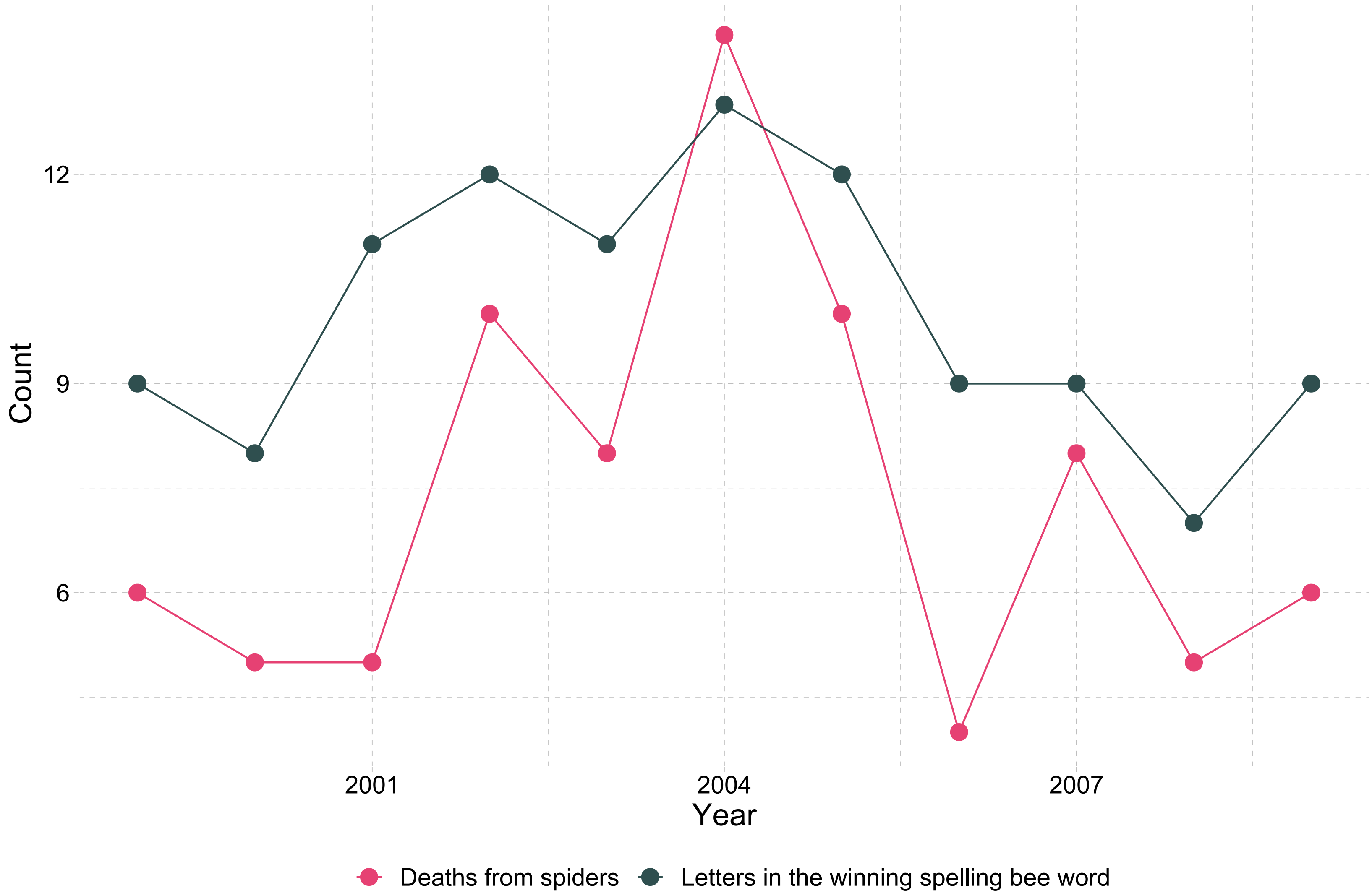
Real life often violates these assumptions.

"**what happens when we violate these assumptions?**"

- Can we find a fix? (Especially: How/when is β *causal*?)
- What happens if we don't (or can't) apply a fix?

OLS still does some amazing things—but you need to know when to be **cautious, confident, or dubious**.

Not everything is causal



Econometrics

An applied econometrician[†] needs a solid grasp on (at least) three areas:

1. The **theory** underlying econometrics (assumptions, results, strengths, weaknesses).

2. How to **apply theoretical methods** to actual data.

3. Efficient methods for **working with data**—cleaning, aggregating, joining, visualizing.

[†]: *Applied econometrician* = Practitioner of econometrics, *e.g.*, analyst, consultant, data scientist.

Econometrics

An applied econometrician[†] needs a solid grasp on (at least) three areas:

1. The **theory** underlying econometrics (assumptions, results, strengths, weaknesses).

2. How to **apply theoretical methods** to actual data.

3. Efficient methods for **working with data**—cleaning, aggregating, joining, visualizing.

This course aims to deepen your knowledge in each of these three areas.

[†]: *Applied econometrician* = Practitioner of econometrics, *e.g.*, analyst, consultant, data scientist.

Econometrics

An applied econometrician[†] needs a solid grasp on (at least) three areas:

1. The **theory** underlying econometrics (assumptions, results, strengths, weaknesses).

2. How to **apply theoretical methods** to actual data.

3. Efficient methods for **working with data**—cleaning, aggregating, joining, visualizing.

This course aims to deepen your knowledge in each of these three areas.

- 1: As before.
- 2–3: **R**

[†]: *Applied econometrician* = Practitioner of econometrics, *e.g.*, analyst, consultant, data scientist.

R

What is R?

To quote the R project website:

> R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.

What is R?

To quote the R project website:

> R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.

What does that mean?

- R was created for the statistical and graphical work required by econometrics.
- R has a vibrant, thriving online community. (stack overflow)
- Plus it's **free** and **open source**.

Why are we using R?

1. R is **free** and **open source**—saving both you and the university 💰💵💰.

2. *Related:* Outside of a small group of economists, private- and public-sector **employers favor R** over `Stata` and most competing softwares.

3. R is very **flexible and powerful**—adaptable to nearly any task, *e.g.*, 'metrics, spatial data analysis, machine learning, web scraping, data cleaning, website building, teaching. My website, the TWEEDS website, and these notes all came out of R.

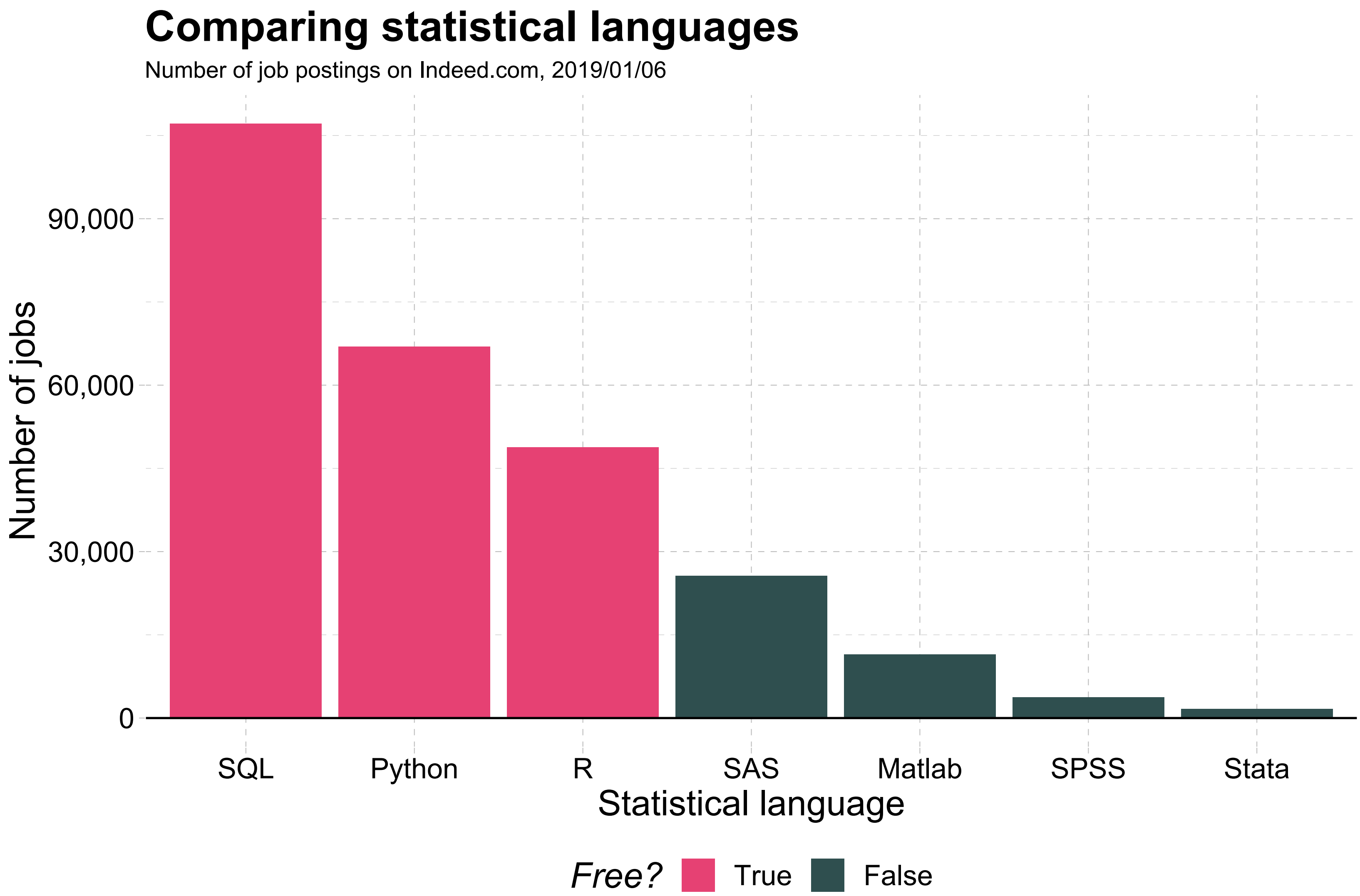
Why are we using R?

4. *Related:* R imposes **no limitations** on your amount of observations, variables, memory, or processing power. (mono[Stata])

5. If you put in the work,[†] you will come away with a **valuable and marketable** tool.

6. Although I sometimes use mono[Stata] (somethings can be easier), **R** is usually more flexible: higher benefit-cost in my opinion.

[†]: Learning R definitely requires time and effort.


Comparing statistical languages
Number of job postings on Indeed.com, 2019/01/06
Number of jobs
90,000
60,000
30,000
0
SQL
Python
R
SAS
Matlab
SPSS
Stata
Statistical language
Free?
True
False

R + Examples

R + Regression

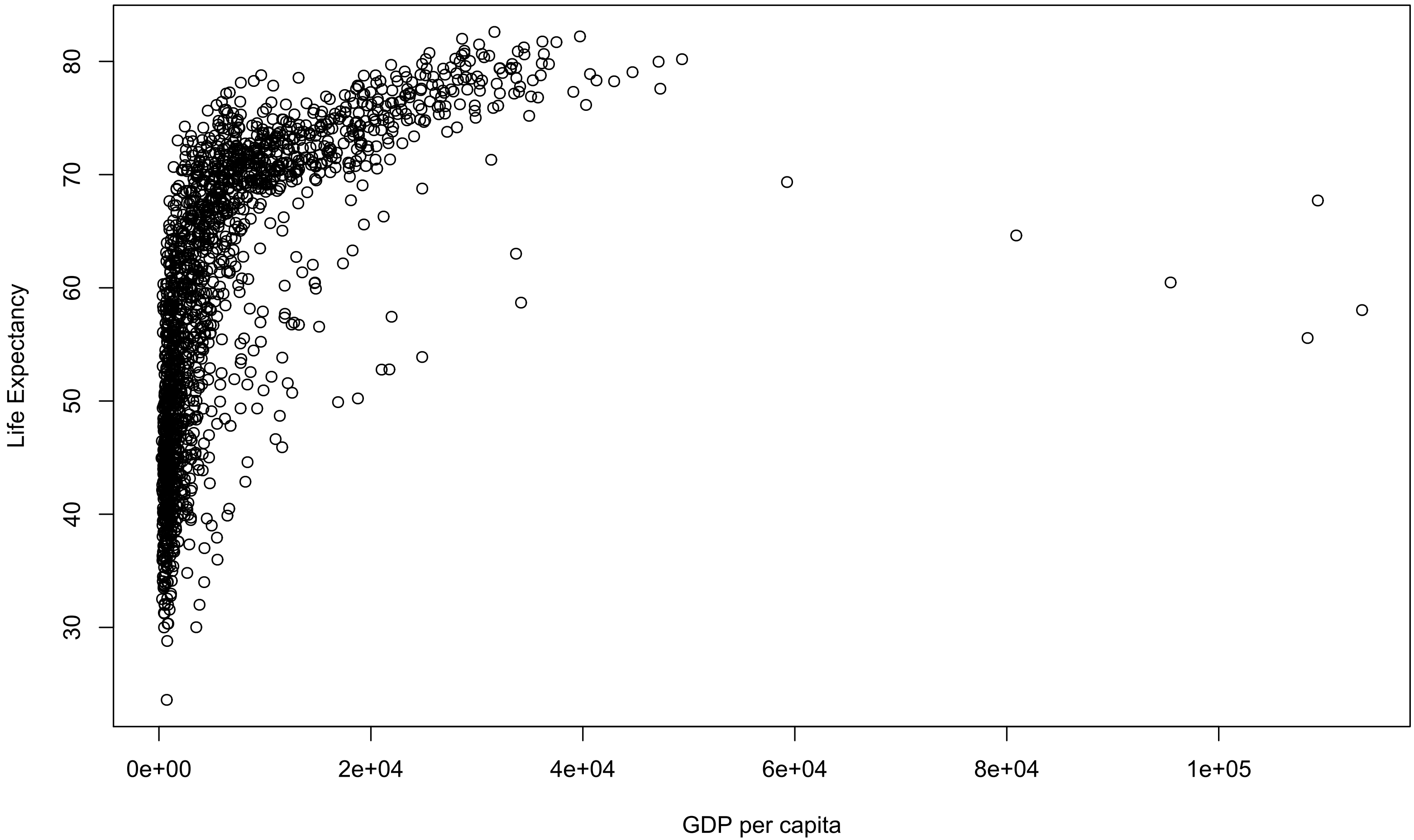
```r
# A simple regression
fit <- lm(dist ~ 1 + speed, data = cars)
# Show the coefficients
coef(summary(fit))
```

```
#>                Estimate Std. Error   t value     Pr(>|t|)
#> (Intercept) -17.579095  6.7584402 -2.601058 1.231882e-02
#> speed         3.932409  0.4155128  9.463990 1.489836e-12
```

```r
# A nice, clear table
library(broom)
tidy(fit)
```

```
#> # A tibble: 2 x 5
#>   term        estimate std.error statistic  p.value
#>   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)   -17.6      6.76      -2.60 1.23e- 2
#> 2 speed           3.93     0.416      9.46 1.49e-12
```

R + Plotting (w/ `plot`)

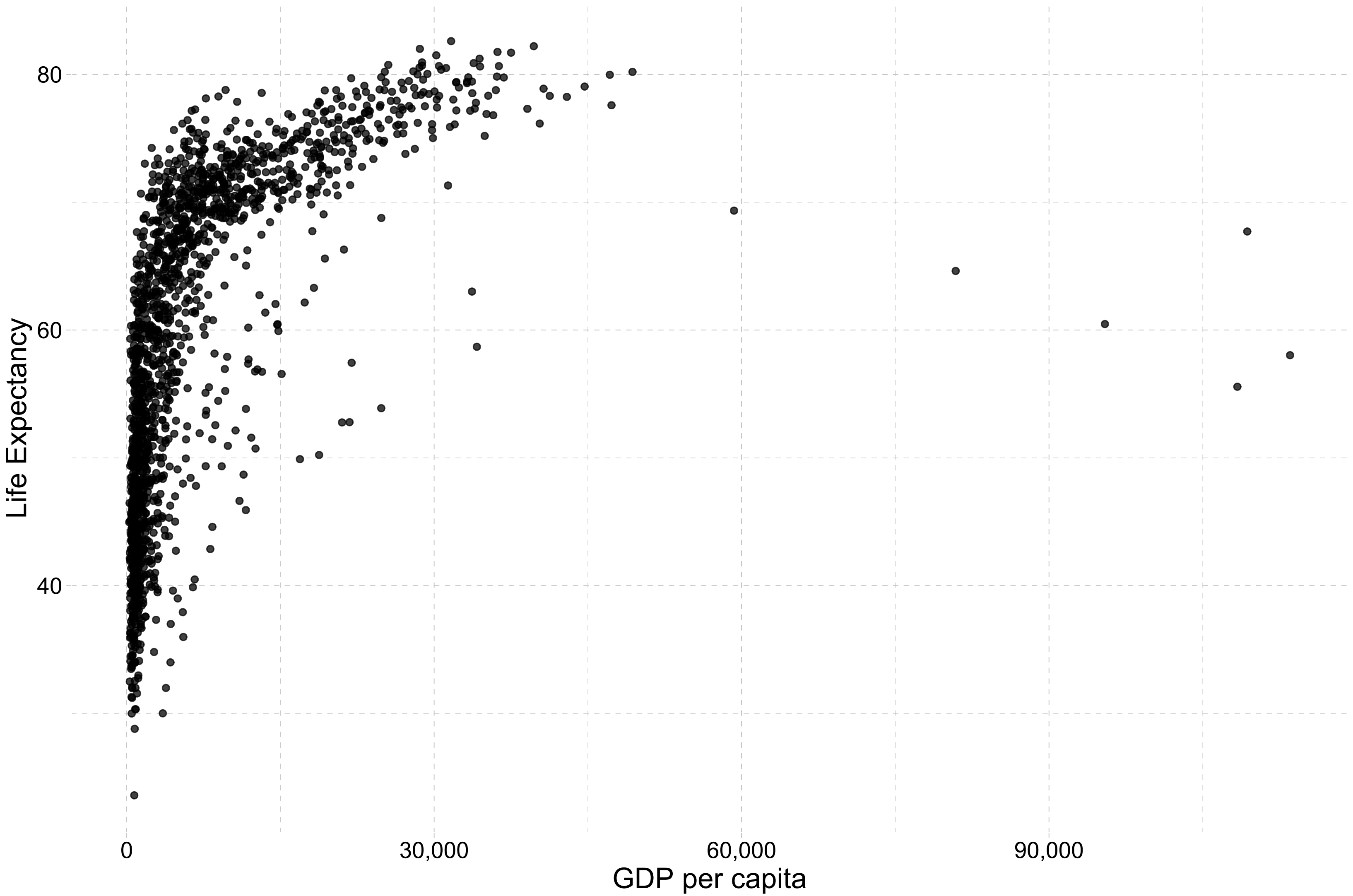


R + Plotting (w/ plot)

```r
# Load packages with dataset
library(gapminder)

# Create dataset
plot(
  x = gapminder$gdpPercap, y = gapminder$lifeExp,
  xlab = "GDP per capita", ylab = "Life Expectancy"
)
```

R + Plotting (w/ ggplot2)

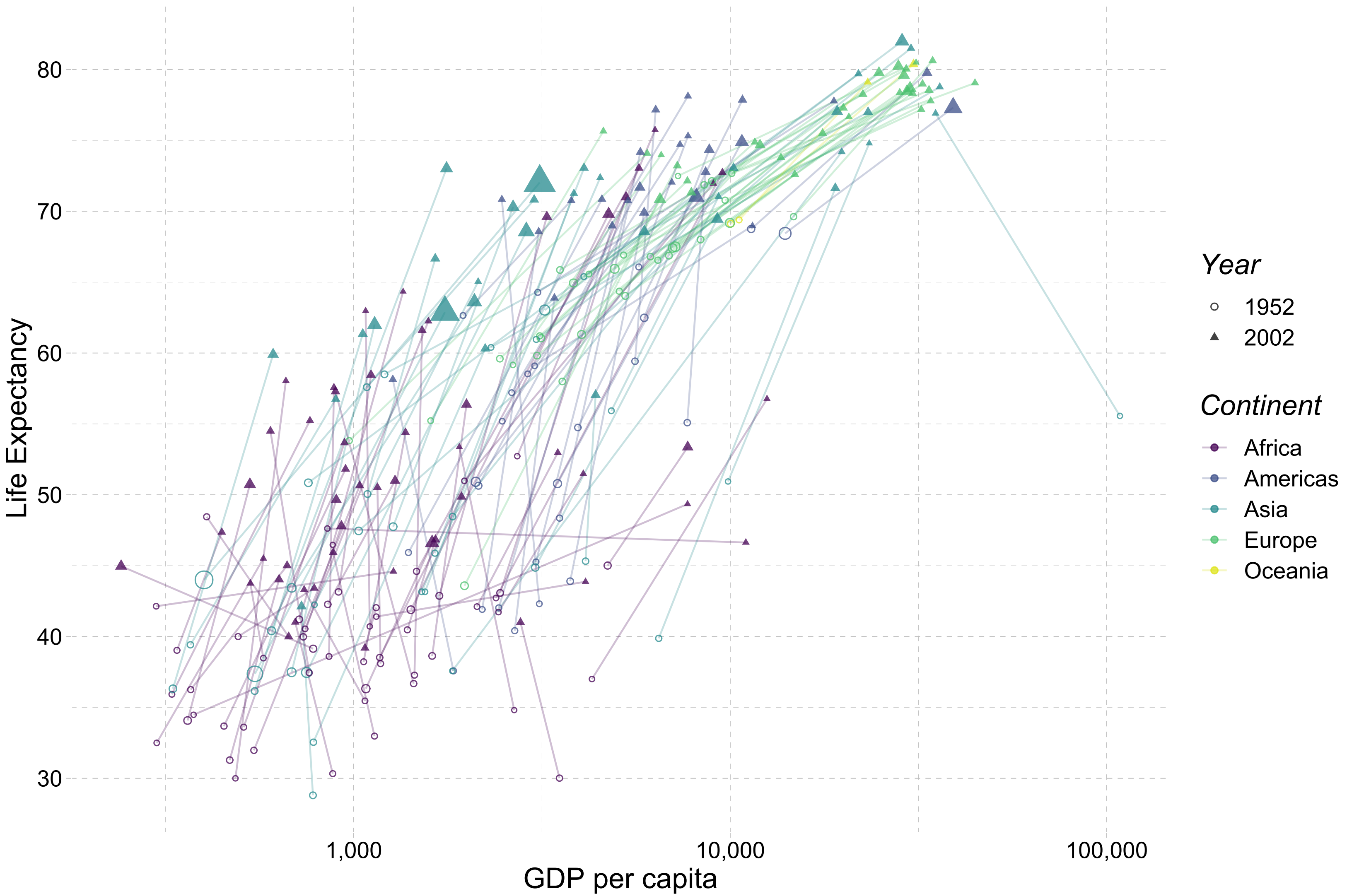


R + Plotting (w/ ggplot2)

```r
# Load packages
library(gapminder); library(dplyr)

# Create dataset
ggplot(data = gapminder, aes(x = gdpPercap, y = lifeExp)) +
geom_point(alpha = 0.75) +
scale_x_continuous("GDP per capita", label = scales::comma) +
ylab("Life Expectancy") +
theme_pander(base_size = 16)
```

R + More plotting (w/ ggplot2)



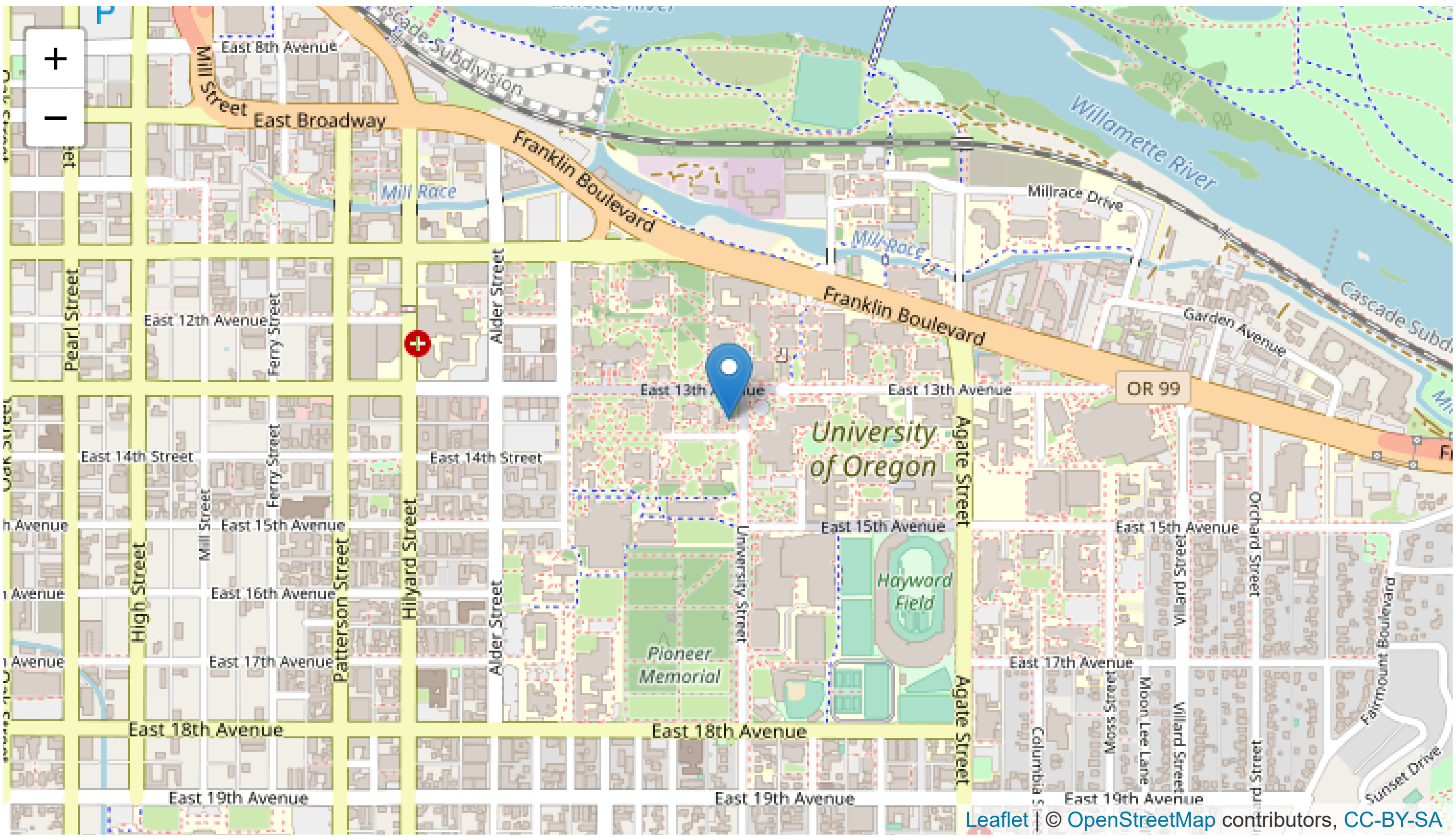
R + More plotting (w/ ggplot2)

```
# Load packages
library(gapminder); library(dplyr)

# Create dataset
ggplot(
  data = filter(gapminder, year %in% c(1952, 2002)),
  aes(x = gdpPercap, y = lifeExp, color = continent, group = country)
) +
geom_path(alpha = 0.25) +
geom_point(aes(shape = as.character(year), size = pop), alpha = 0.75) +
scale_x_log10("GDP per capita", label = scales::comma) +
ylab("Life Expectancy") +
scale_shape_manual("Year", values = c(1, 17)) +
scale_color_viridis("Continent", discrete = T, end = 0.95) +
guides(size = F) +
theme_pander(base_size = 16)
```

R + Maps

```r
library(leaflet)
leaflet() %>%
  addTiles() %>%
  addMarkers(lng = -123.075, lat = 44.045, popup = "The University of Oregon")
```



Getting started with R

Starting R

Installation

- Install R.
- Install `RStudio`.
- **Optional/Overkill:** Git
    - Create an account on GitHub
    - Register for a student/educator discount.
    - For installation guidance and troubleshooting, check out Jenny Bryan's website.
- **Note:** The lab in 442 McKenzie has R installed and ready. That said, having a copy of R on your own computer will likely be very convenient for homework, projects, *etc.*

Starting R

Resources

Free(-ish)

- Google (which inevitably leads to StackOverflow)
- Time
- Your classmates
- Your GEs
- Me
- R resources here and here

Money

- Book: *R for Stata Users*
- Short online course: DataCamp

Starting R

Some R basics

You will dive deeper into R in lab, but here six big points about R:

1. Everything is an **object**. `foo`
2. Every object has a **name** and **value**. `foo <- 2`
3. You use **functions** on these objects. `mean(foo)`
4. Functions come in **libraries** (**packages**) `library(dplyr)`
5. R will try to **help** you. `?dplyr`
6. R has its **quirks**. `NA; error; warning`

Starting R

R *vs.* `Stata`

Coming from **`Stata`**, here are a few important changes (benefits):

- Multiple objects and arrays (*e.g.*, data frames) can exist in the same workspace (in memory). No more `keep`, `preserve`, `restore`, `snapshot` nonsense!

- (Base) R comes with lots of useful built-in functions—and provides all the tools necessary for you to build your own functions. However, many of the *best* functions come from external libraries.

- You don't need to `tset` or `xtset` data (you can if you really want... `ts`).

Starting R

Google CoLab

If you rather not install R on your machine (slow, little space, etc) you can also use Cloud Computing options. The one I recommend for R is Colab.

- Similar to Jupyter Notebook if you are familiar with Python
- Can compile in both Python and **R**
- Great to share code (no need for anyone to install anything)
- Limitations:
    - Usage limits and hardware availability
    - Memory shortage
    - GPUs availability vary

Next: Metrics review + Potential outcomes + ML/CI distinction

Cunningham 1 and 2 (for today) + Mullainathan and Spiess (2017)