Introduction and Overview

EC 421, Set 1

Edward Rubin

Prologue

Why?

Motivation

Let's start with a few **basic, general questions:**

Why?

Motivation

Let's start with a few **basic, general questions:**

1. What is the goal of econometrics?

2. Why do economists (or other people) study or use econometrics?

Why?

Motivation

Let's start with a few **basic, general questions:**

1. What is the goal of econometrics?

2. Why do economists (or other people) study or use econometrics?

One simple answer: Learn about the world using data.

Why?

Motivation

Let's start with a few **basic, general questions:**

1. What is the goal of econometrics?
2. Why do economists (or other people) study or use econometrics?

One simple answer: Learn about the world using data.

- *Learn about the world* = Raise, answer, and challenge questions, theories, assumptions.
- *data* = Plural of datum.

Why?

Example

One might (reasonably) guess a company's sales are a function of its advertising spending, price, and intesity of competitors.

Why?

Example

One might (reasonably) guess a company's sales are a function of its advertising spending, price, and intesity of competitors.

So, one might hypothesize a model $\text{Sales} = f(\text{Ad}, \text{Price}, \text{Comp})$

where

- Ad represents dollars spent on advertising,
- Price is the product's price,
- Comp gives the product's competition.

Why?

Example

One might (reasonably) guess a company's sales are a function of its advertising spending, price, and intesity of competitors.

So, one might hypothesize a model $\text{Sales} = f(\text{Ad}, \text{Price}, \text{Comp})$

where

- Ad represents dollars spent on advertising,
- Price is the product's price,
- Comp gives the product's competition.

We expect that sales $\uparrow$ with advertising and $\downarrow$ with price and competition.

But who needs to *expect*?

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

But who needs to *expect*?

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

More importantly: Regression estimates the *size* of these effects

- *How much* does an additional dollar of advertising increase sales?
- *How much* does a one-dollar increase in price decrease sales?
- *How much* does an additional competitor reduce sales?

These (causal) questions are central to efficient decision-making
and are the bread and butter of econometrics.

Why?

Example, cont.

Regression model:

$$\text{Sales}_i = \beta_0 + \beta_1 \text{Ad}_i + \beta_2 \text{Price}_i + \beta_3 \text{Comp}_i + \varepsilon_i$$

With this basic regression model, we can test/estimate/quantify the (linear) relationship between sales and advertising, price, and competition.

Why?

Example, cont.

Regression model:

$$\text{Sales}_i = \beta_0 + \beta_1 \text{Ad}_i + \beta_2 \text{Price}_i + \beta_3 \text{Comp}_i + \varepsilon_i$$

(Review) Questions

Why?

Example, cont.

Regression model:

$$\text{Sales}_i = \beta_0 + \beta_1 \text{Ad}_i + \beta_2 \text{Price}_i + \beta_3 \text{Comp}_i + \varepsilon_i$$

(Review) Questions

- **Q:** How do we interpret β_1?

Why?

Example, cont.

Regression model:

$$\text{Sales}_i = \beta_0 + \beta_1 \text{Ad}_i + \beta_2 \text{Price}_i + \beta_3 \text{Comp}_i + \varepsilon_i$$

(Review) Questions

- **Q:** How do we interpret β_1?
- **A:** An additional dollar of advertising corresponds with a β_1-unit change in sales (holding price and competition fixed).

Why?

Example, cont.

Regression model:

$$\text{Sales}_i = \beta_0 + \beta_1 \text{Ad}_i + \beta_2 \text{Price}_i + \beta_3 \text{Comp}_i + \varepsilon_i$$

(Review) Questions

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

Why?

Example, cont.

Regression model:

$$\text{Sales}_i = \beta_0 + \beta_1 \text{Ad}_i + \beta_2 \text{Price}_i + \beta_3 \text{Comp}_i + \varepsilon_i$$

(Review) Questions

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

Why?

Example, cont.

Regression model:

$$\text{Sales}_i = \beta_0 + \beta_1 \text{Ad}_i + \beta_2 \text{Price}_i + \beta_3 \text{Comp}_i + \varepsilon_i$$

(Review) Questions

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

Why?

Example, cont.

Regression model:

$$\text{Sales}_i = \beta_0 + \beta_1 \text{Ad}_i + \beta_2 \text{Price}_i + \beta_3 \text{Comp}_i + \varepsilon_i$$

(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.

Why?

Example, cont.

Regression model:

$$\text{Sales}_i = \beta_0 + \beta_1 \text{Ad}_i + \beta_2 \text{Price}_i + \beta_3 \text{Comp}_i + \varepsilon_i$$

(Review) Questions

- **Q:** What is ε_i?

Why?

Example, cont.

Regression model:

$$\text{Sales}_i = \beta_0 + \beta_1 \text{Ad}_i + \beta_2 \text{Price}_i + \beta_3 \text{Comp}_i + \varepsilon_i$$

(Review) Questions

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

Population parameters are averages; individuals are rarely average.

Why?

Example, cont.

Regression model:

$$\text{Sales}_i = \beta_0 + \beta_1 \text{Ad}_i + \beta_2 \text{Price}_i + \beta_3 \text{Comp}_i + \varepsilon_i$$

(Review) Questions

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

Example, cont.

Regression model:

$$\text{Sales}_i = \beta_0 + \beta_1 \text{Ad}_i + \beta_2 \text{Price}_i + \beta_3 \text{Comp}_i + \varepsilon_i$$

(Review) Questions

- **Q:** Which assumptions do we impose when estimating with OLS?
- **A:**
  - The relationship between the sales 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.

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.

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.

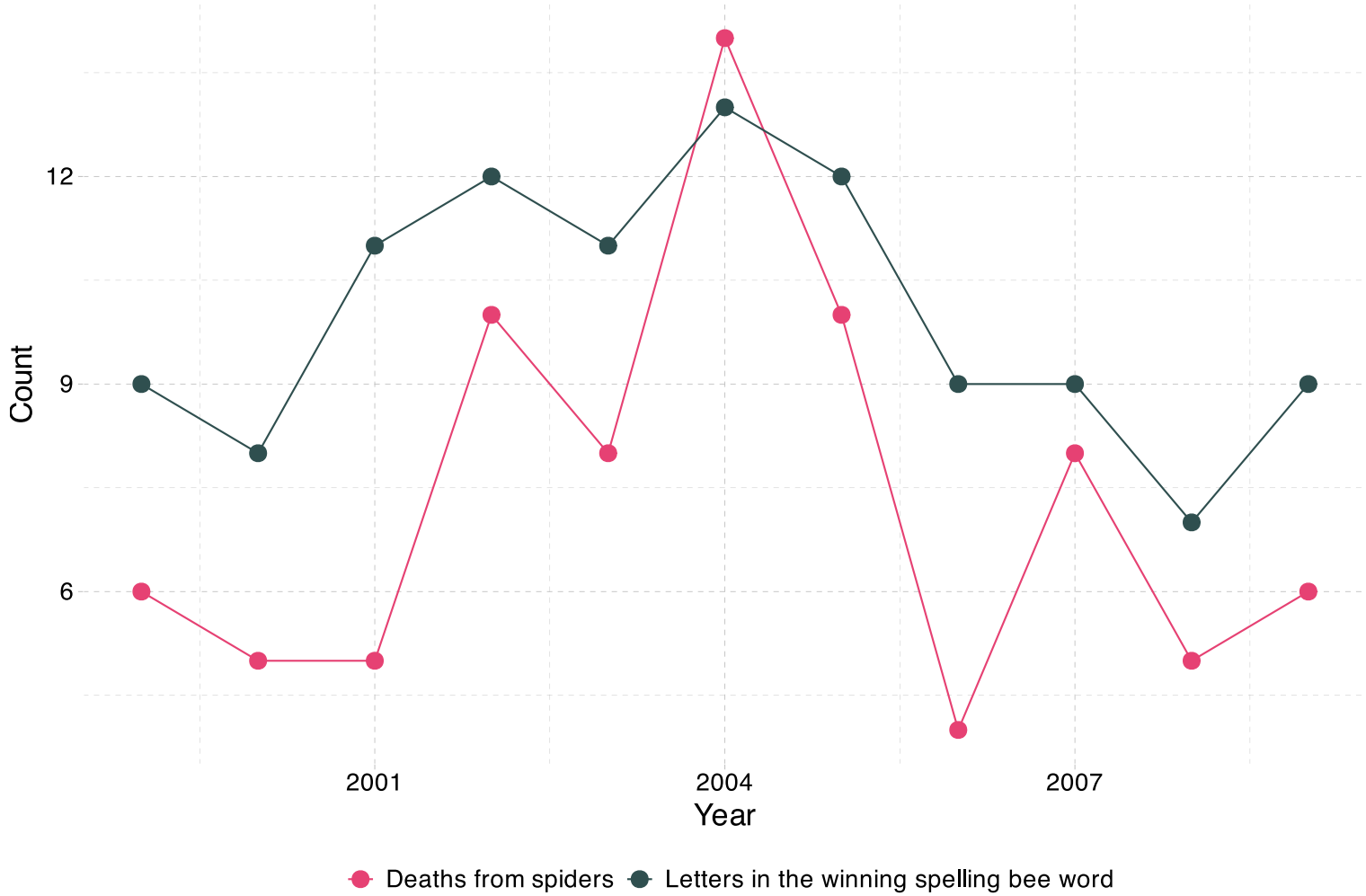
Real life often violates these assumptions.

EC421 asks "**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



Not everything is causal

More seriously

Suppose you estimate our sales model for your boss.

$$\text{Sales}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{Ad}_i + \hat{\beta}_2 \text{Price}_i + \hat{\beta}_3 \text{Comp}_i + e_i$$

Can you trust that $\hat{\beta}_2$ gives you the actual effect of price on sales?

Econometrics

Applied econometrics, data science, analytics require:

1. Intuition for the **theory** behind statistics/econometrics
   (assumptions, results, strengths, weaknesses).

2. Practical knowledge of how to **apply theoretical methods** to data.

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

Econometrics

Applied econometrics, data science, analytics require:

1. Intuition for the **theory** behind statistics/econometrics (assumptions, results, strengths, weaknesses).
2. Practical knowledge of how to **apply theoretical methods** to 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.

Econometrics

Applied econometrics, data science, analytics require:

1. Intuition for the **theory** behind statistics/econometrics (assumptions, results, strengths, weaknesses).
2. Practical knowledge of how to **apply theoretical methods** to 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**

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?

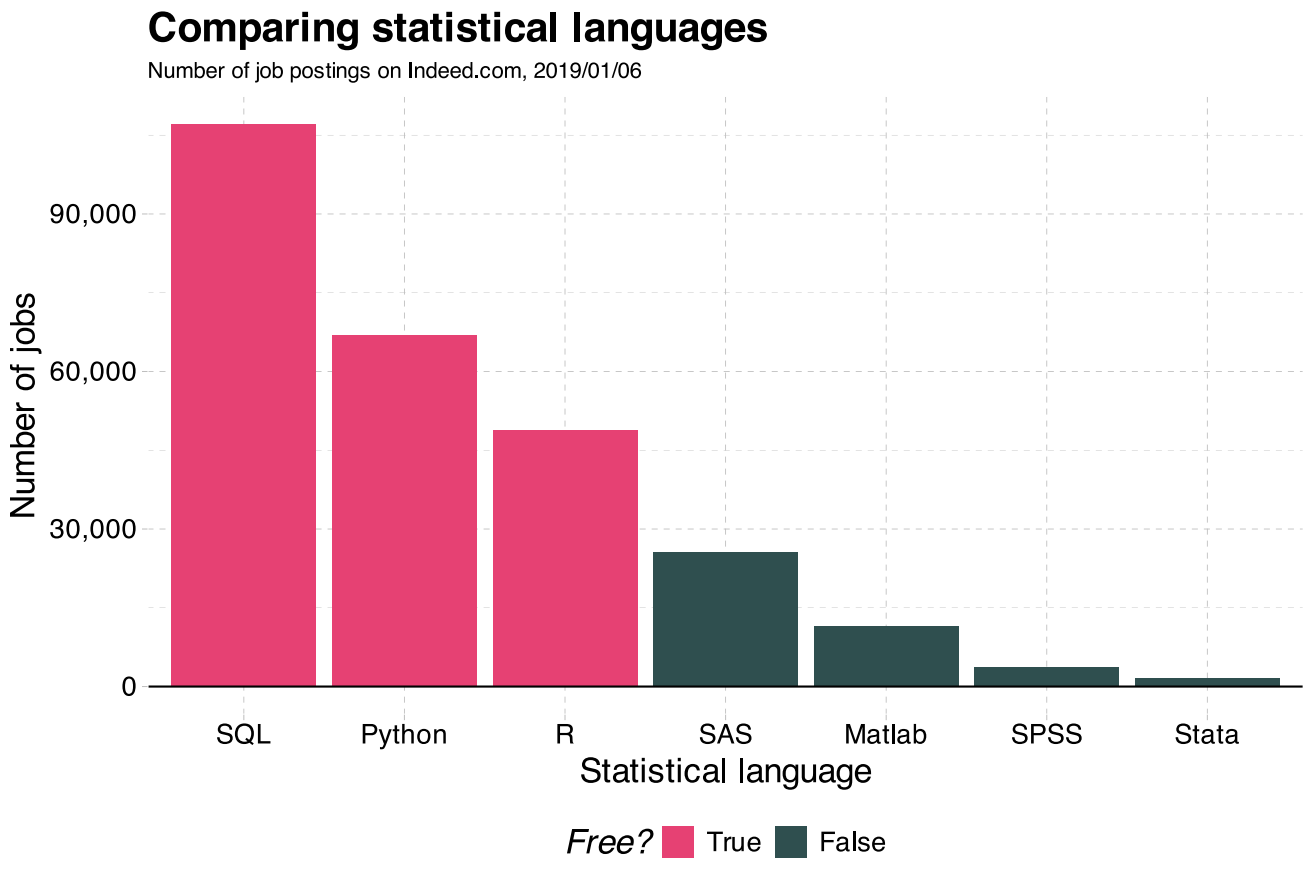
4. *Related:* R imposes **no limitations** on your amount of observations, variables, memory, or processing power. (I'm looking at **you**, `Stata`.)

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

6. I 💖 **R**

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

R


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

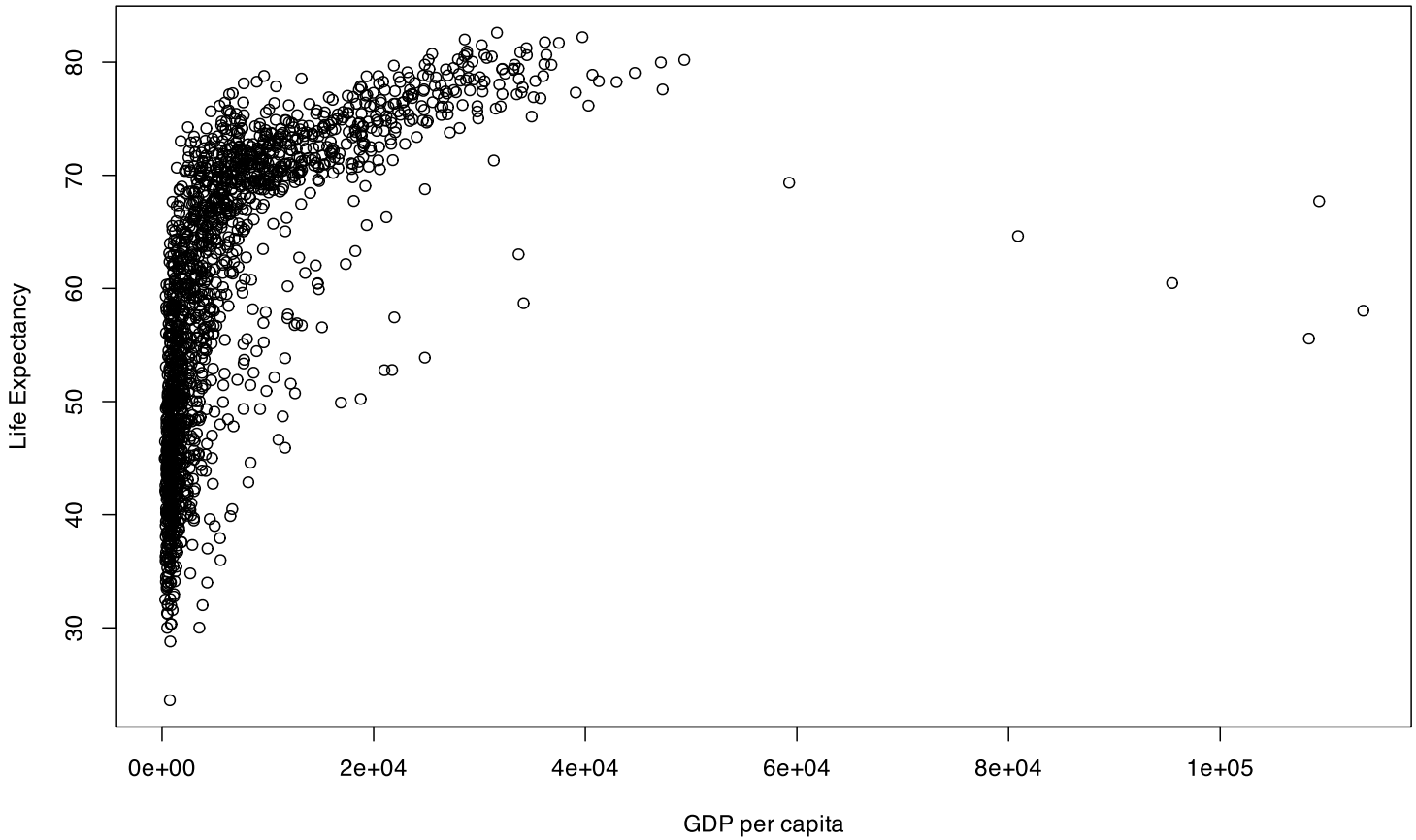
```r
# A simple regression
fit ← lm(dist ~ 1 + speed, data = cars)
# Show the coefficients
coef(summary(fit))
```

```
#>               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 × 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`)

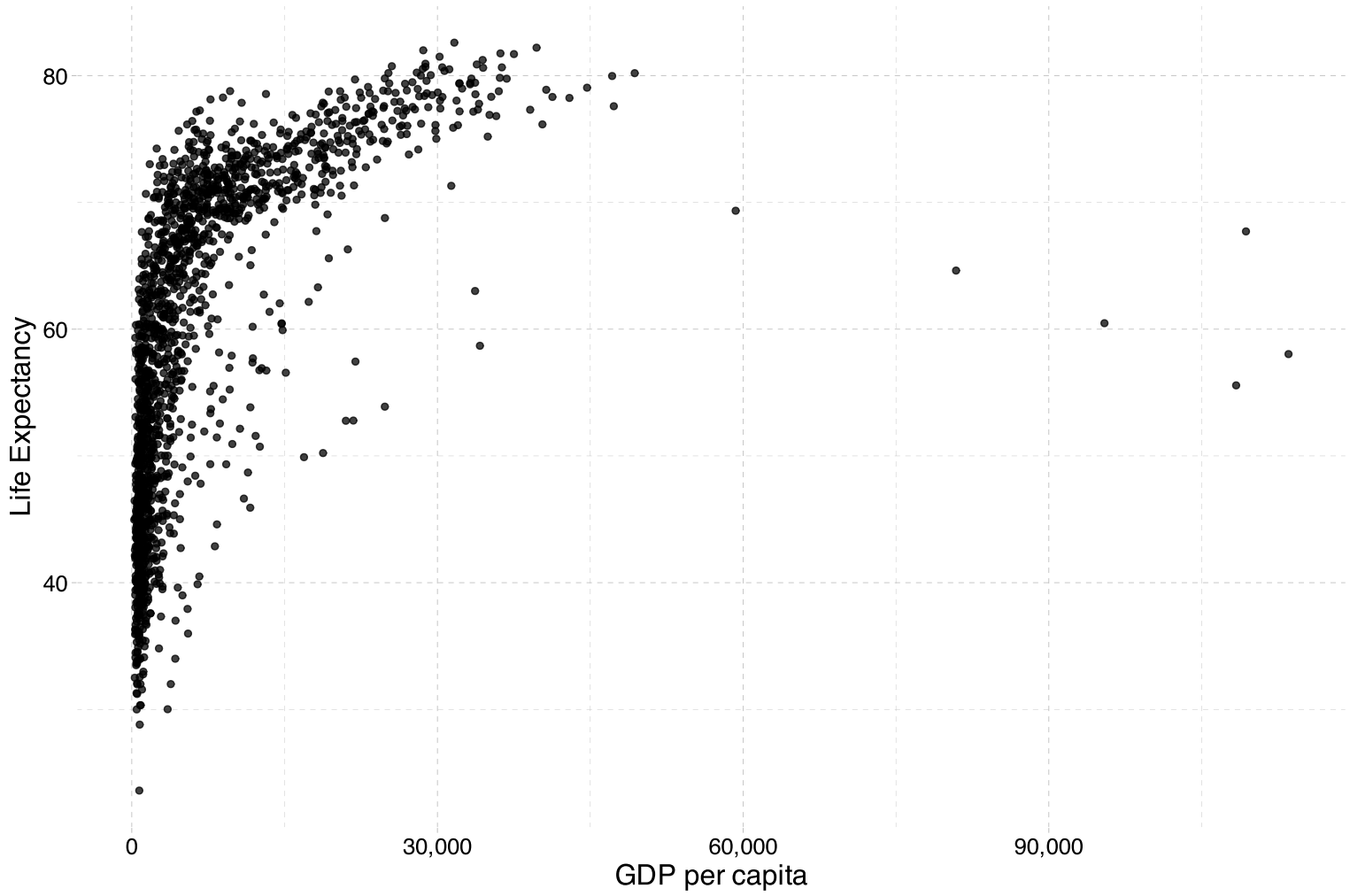


R + Plotting (w/ `plot`)

```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)

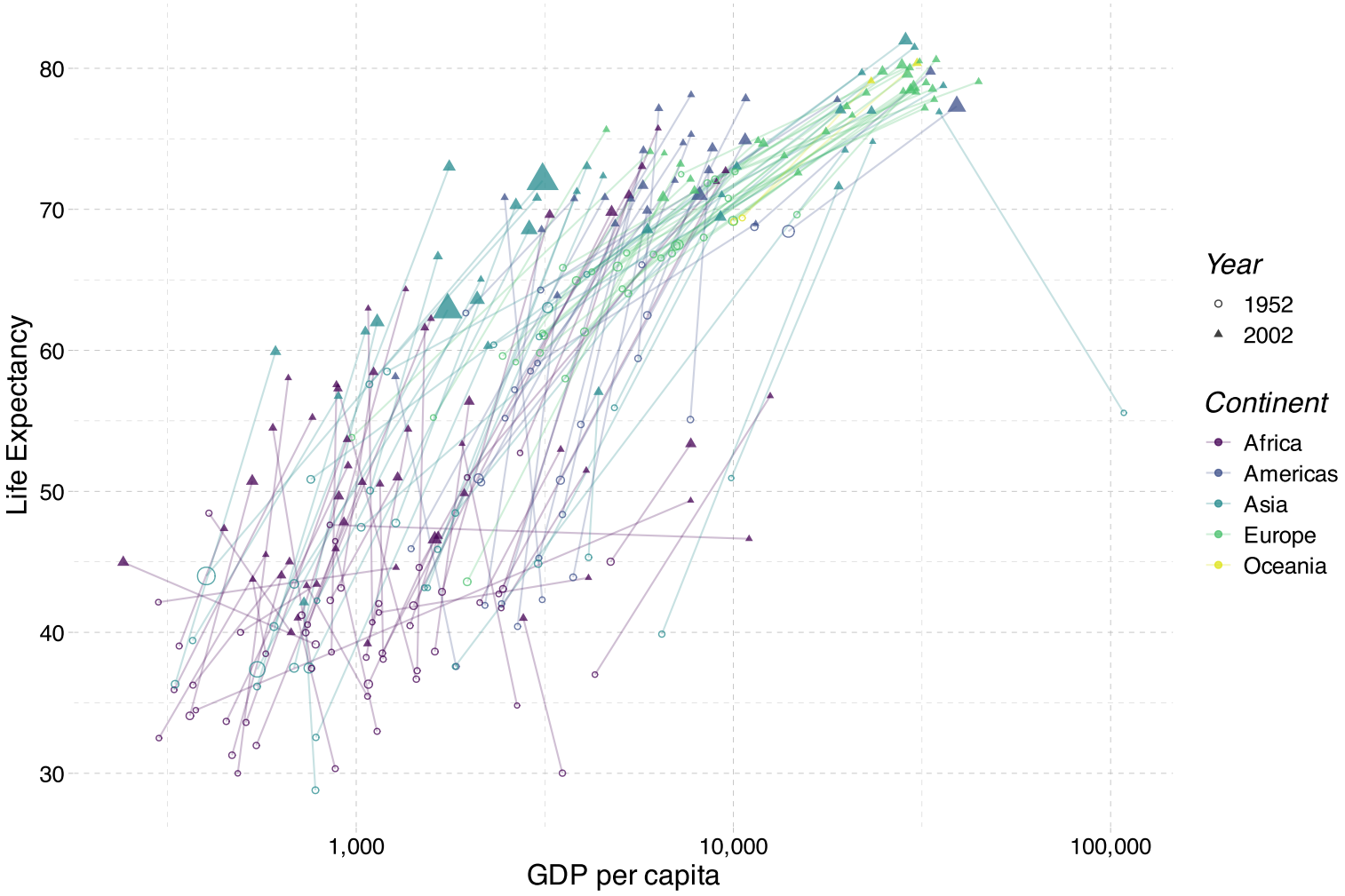


R + Plotting (w/ ggplot2)

```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)

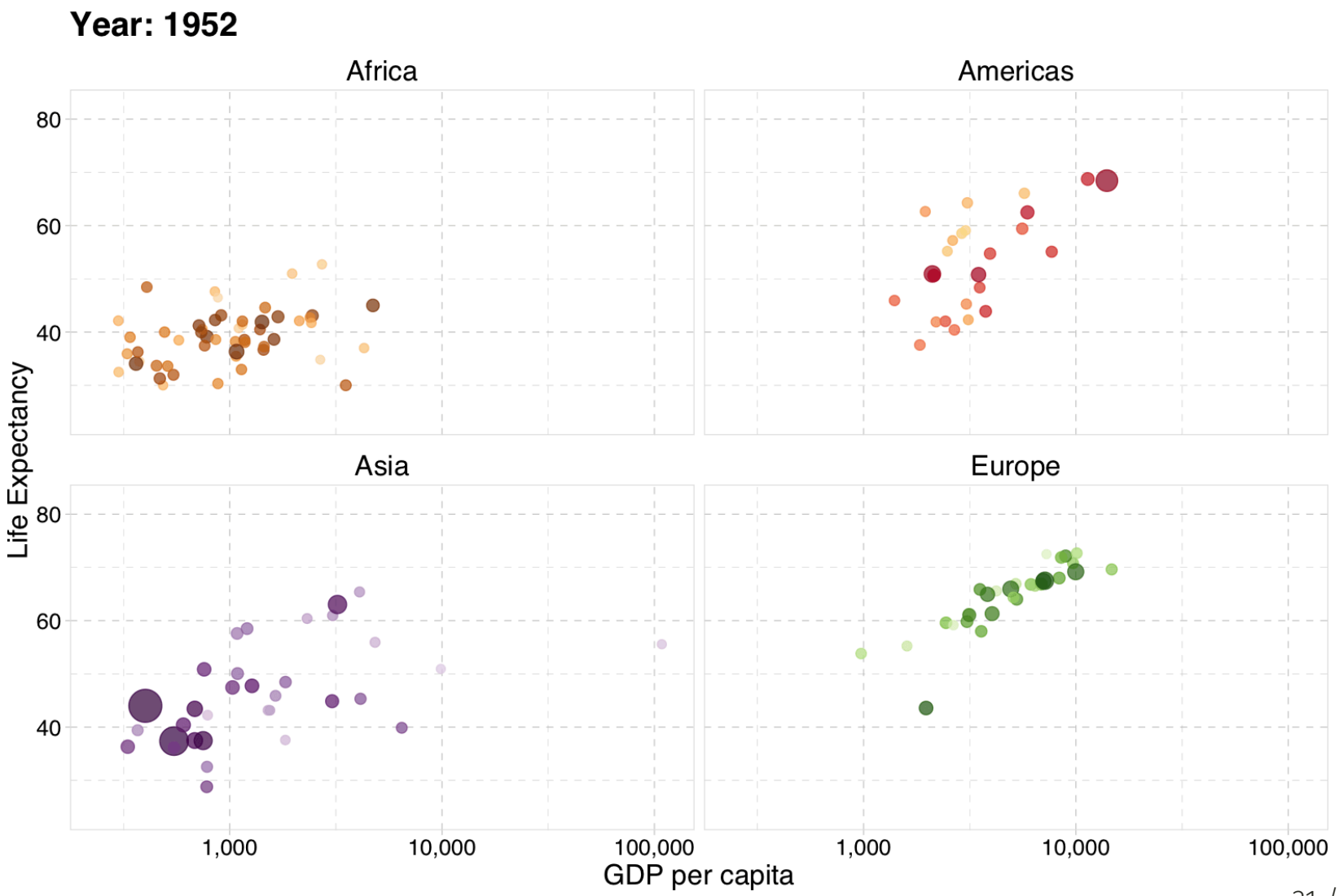


R + More plotting (w/ ggplot2)

```r
# 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 + Animated plots (w/ gganimate)

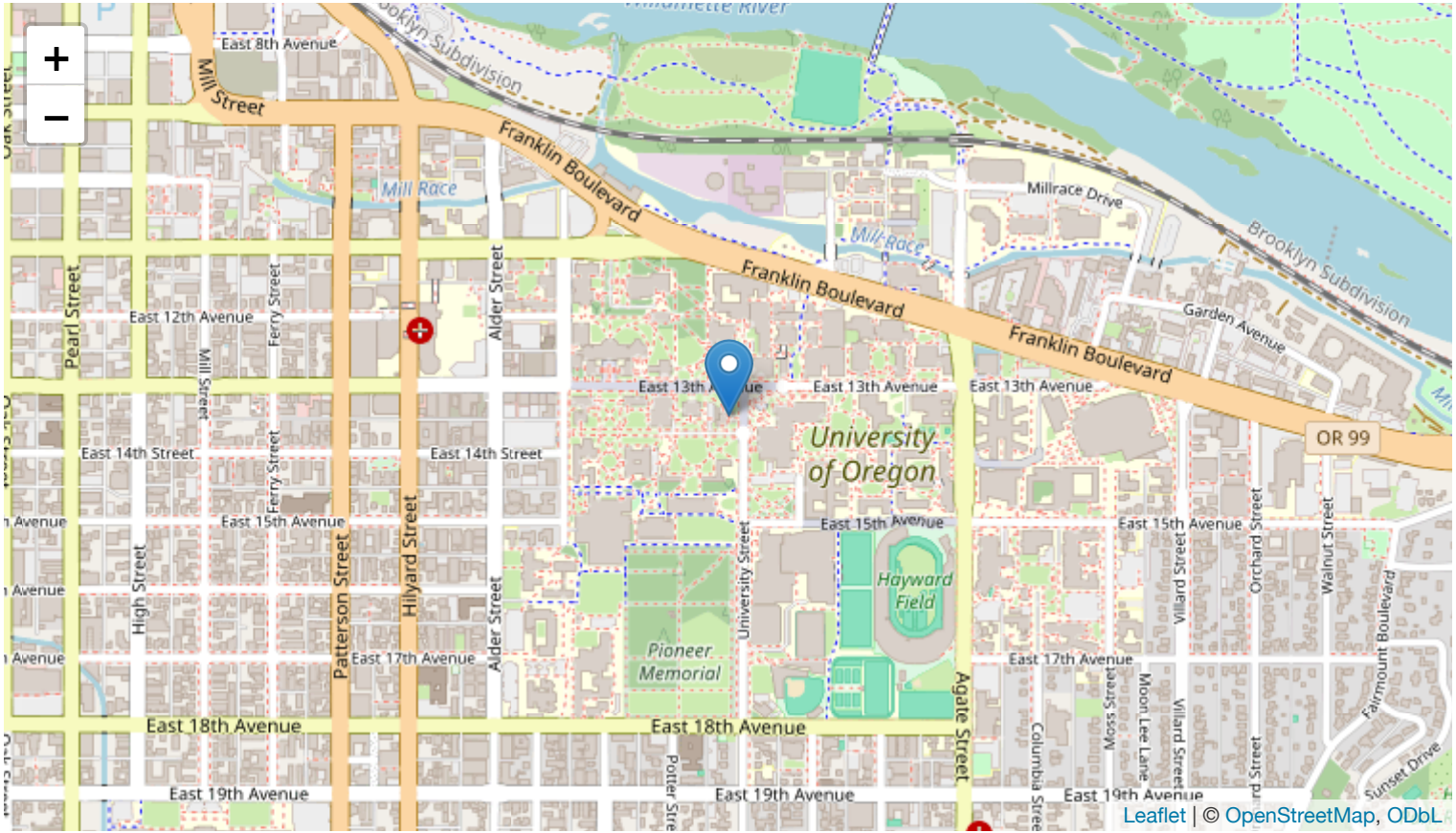


R + Animated plots (w/ gganimate)

```r
# The package for animating ggplot2
library(gganimate)
# As before
ggplot(
  data = gapminder %>% filter(continent != "Oceania"),
  aes(gdpPercap, lifeExp, size = pop, color = country)
) +
geom_point(alpha = 0.7, show.legend = FALSE) +
scale_colour_manual(values = country_colors) +
scale_size(range = c(2, 12)) +
scale_x_log10("GDP per capita", label = scales::comma) +
facet_wrap(~continent) +
theme_pander(base_size = 16) +
theme(panel.border = element_rect(color = "grey90", fill = NA)) +
# Here comes the gganimate-specific bits
labs(title = "Year: {frame_time}") +
ylab("Life Expectancy") +
transition_time(year) +
ease_aes("linear")
```

R + Maps

```
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:** Many UO labs have 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
- ChatGPT, Copilot, and other AI assistants
- 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`

Next: Metrics review(s)