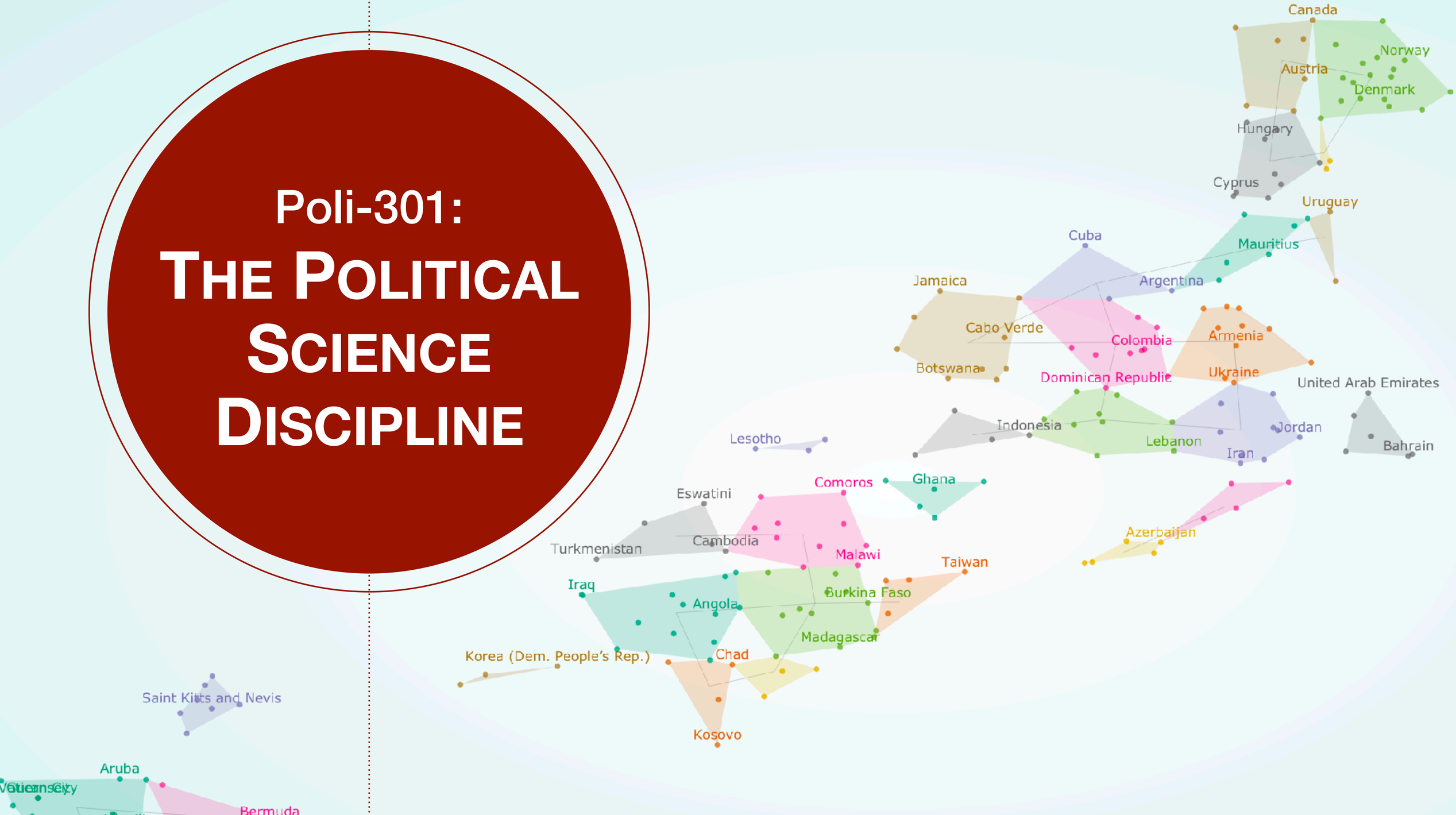
Poli-301:
The Political Science Discipline



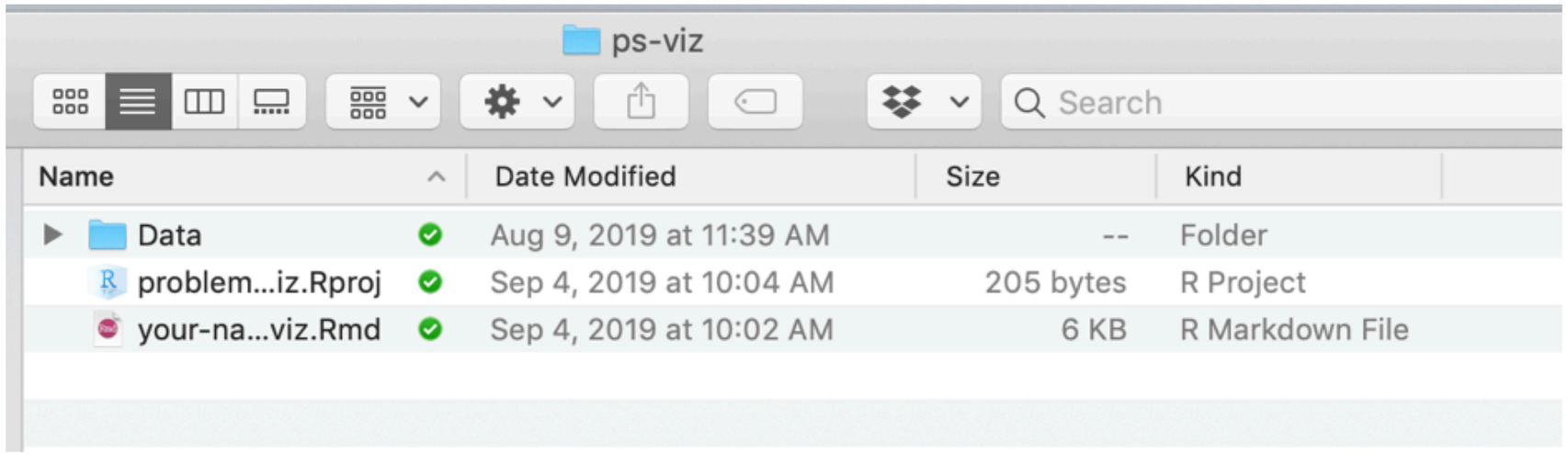
TODAY'S AGENDA

1. Regression Recap
2. Regression with categorical variables
3. "Explaining" outcomes

Loading data

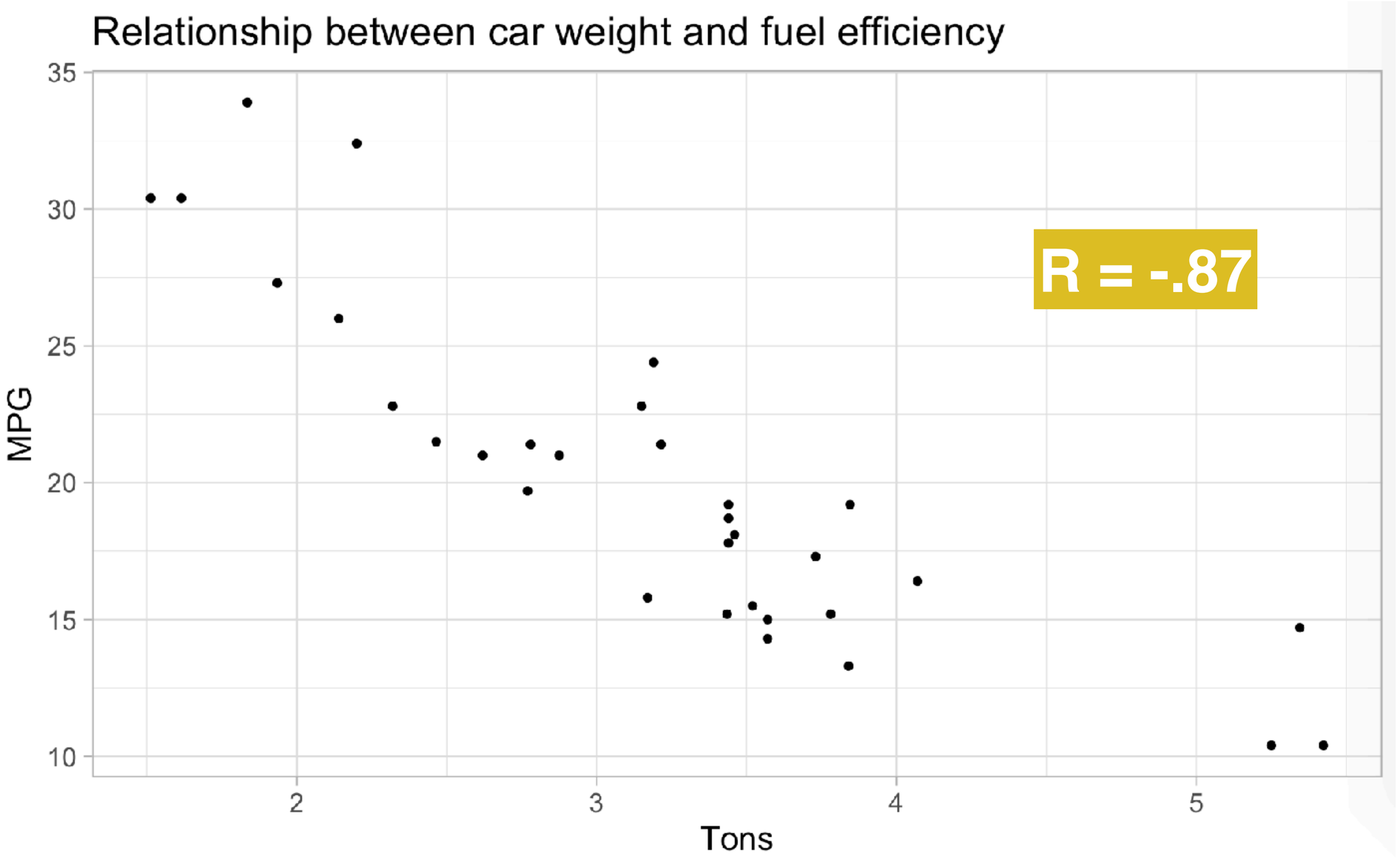
Need to be in RProject

Need file path for data and right extension




```
# read data
movies = read_csv("Data/movies.csv")
```

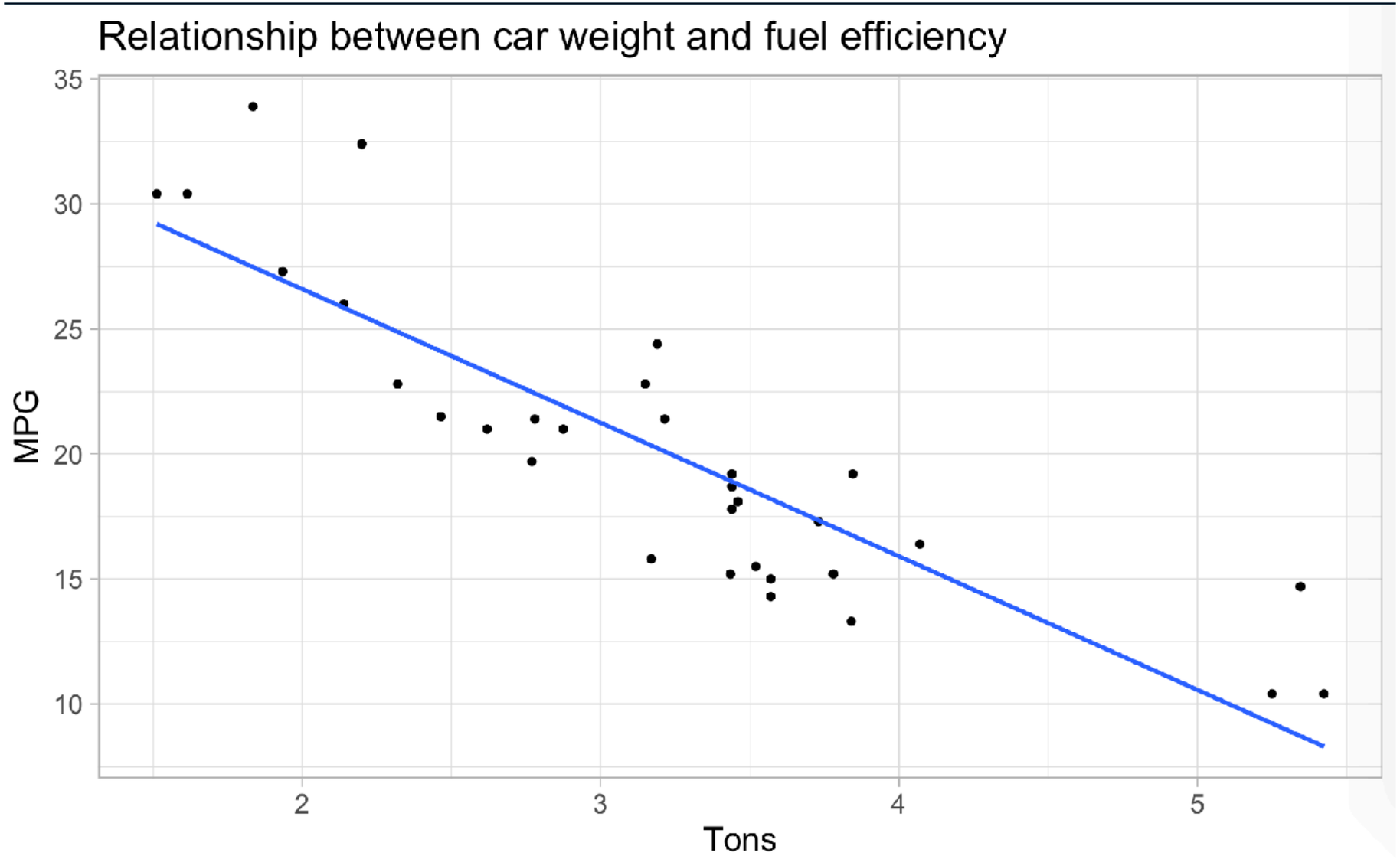
Cars


Relationship between car weight and fuel efficiency
MPG
35
30
25
20
15
10
2
3
4
5
Tons
R = -.87

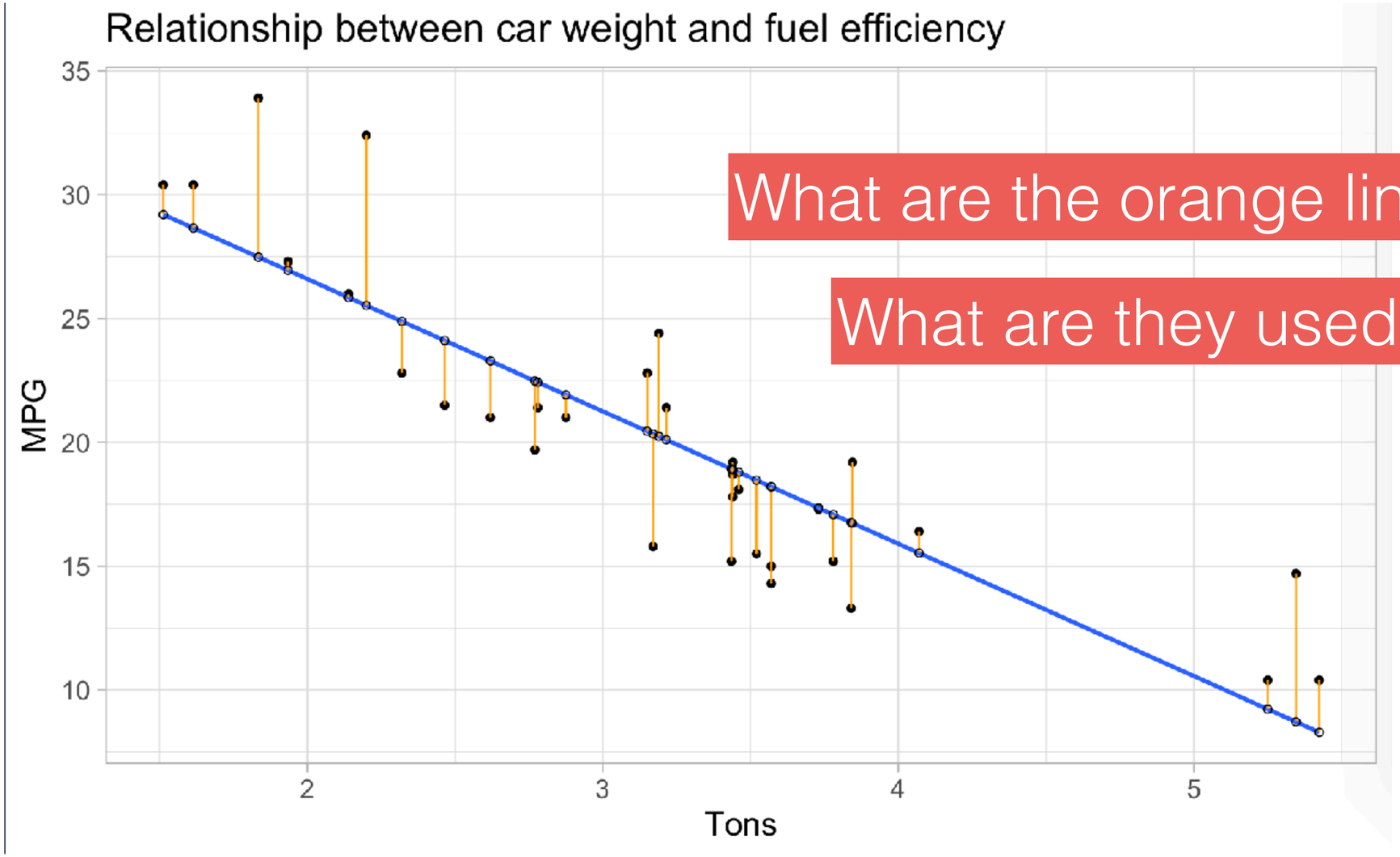
Linear models

Y = mx + b	$\hat{y} = \beta_0 + \beta_1 x_1 + \epsilon$	
Y	$\hat{y}$	Outcome variable
x	x_1	Explanatory variable
b	β_0	Y intercept
m	β_1	Slope
	ϵ	Error (residual)

Drawing lines



Drawing lines



What are the orange lines?

What are they used for?

Residuals

$$\hat{mpg} = \beta_0 + \beta_1 weight + \epsilon$$

$$\text{mpg} = 37.29 - 5.34(\text{wt}) + \epsilon$$

What is mpg_hat?

What are residual units in?

```
model_cars = lm(mpg ~ wt, data = mtcars)


## using moderndive
fitted = get_regression_points(model_cars)
```

	ID	mpg	wt	mpg_hat	residual
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
1	1	21	2.62	23.3	-2.28
2	2	21	2.88	21.9	-0.92
3	3	22.8	2.32	24.9	-2.09
4	4	21.4	3.22	20.1	1.30
5	5	18.7	3.44	18.9	-0.2
6	6	18.1	3.46	18.8	-0.693
7	7	14.3	3.57	18.2	-3.90
8	8	24.4	3.19	20.2	4.16
9	9	22.8	3.15	20.4	2.35
10	10	19.2	3.44	18.9	0.3

2

Regression with categories

Regression with categorical variables

Numeric variables	Categorical variables
(continuous)	(Factors)
Numbers	**Words**

You tell me

Income

Weight

18-25
26-34
…
65+

Gender

GDP per capita

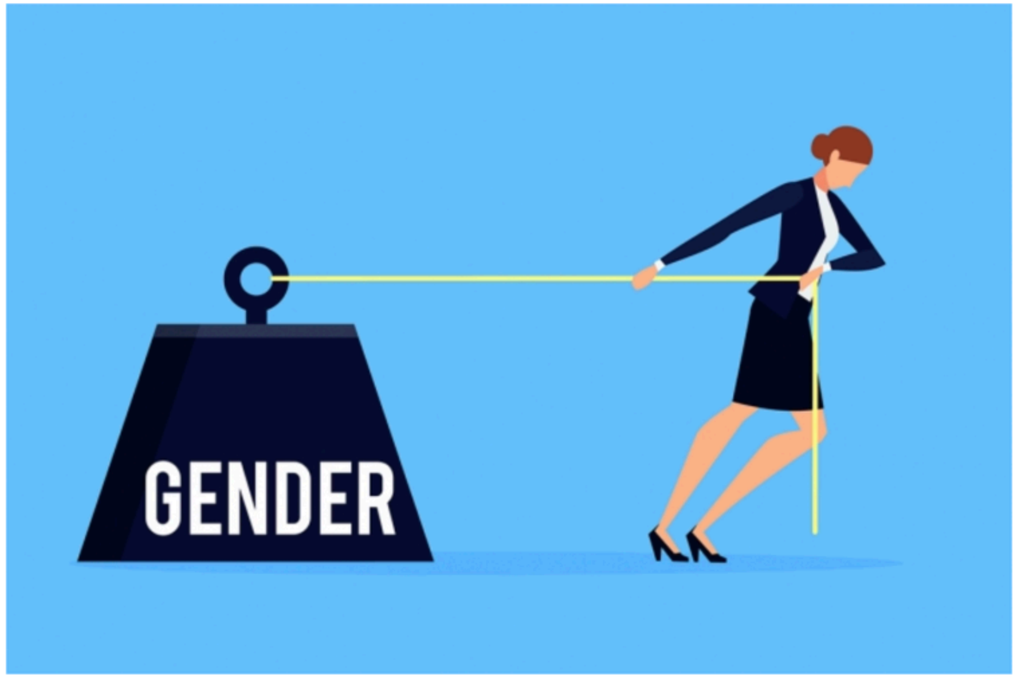
Life Expectancy

Strongly Agree
…
Strongly disagree

Bechdel test

Bias in student evaluations

Why We Must Stop Relying on Student Ratings of Teaching



iStock

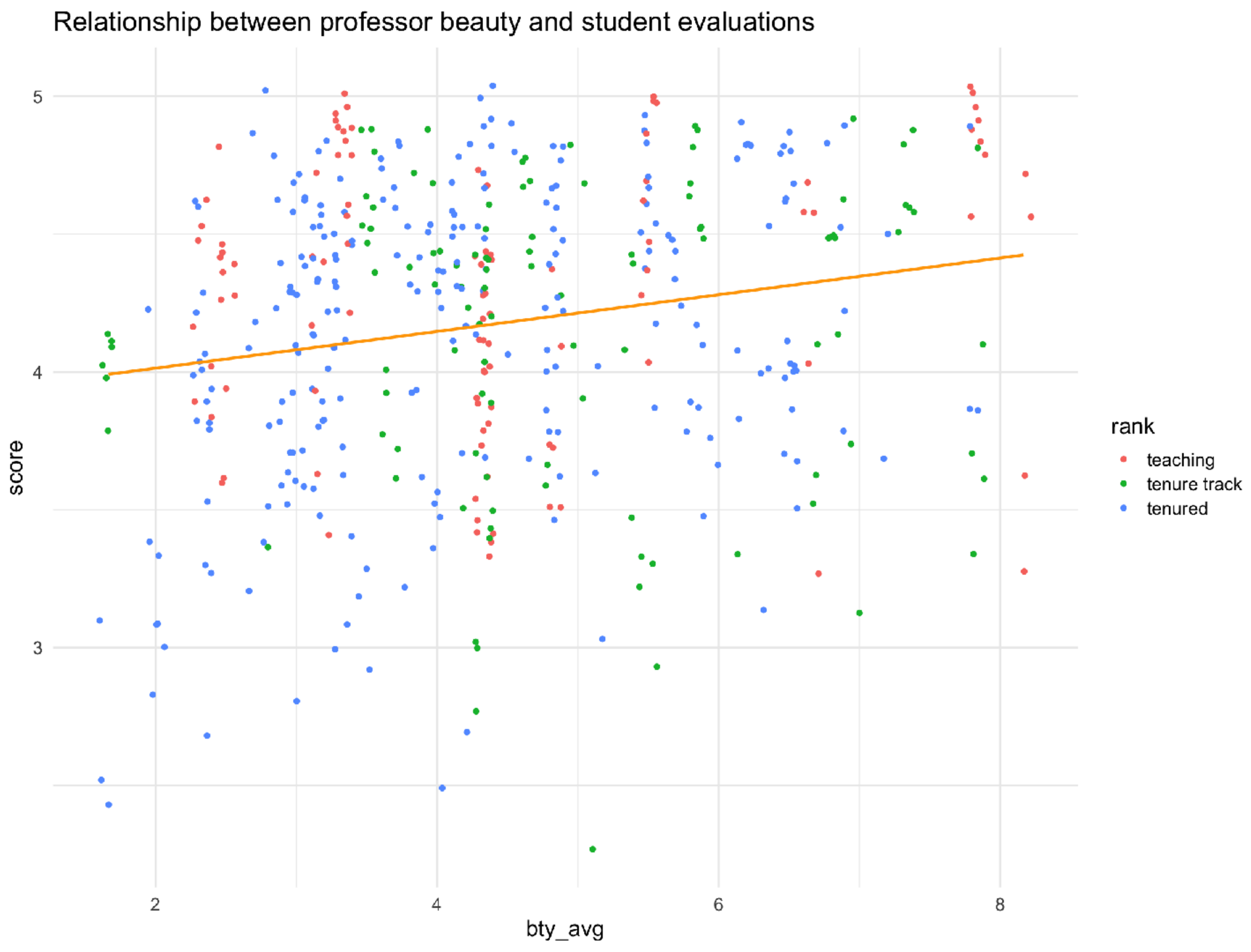
News & Views Careers Events Reports & Data

Trending: New Admissions Era? Winning Tenure, in Court

INSIDE HIGHER ED

#News #Teaching And Learning

Teaching Evals: Bias and Tenure


Relationship between professor beauty and student evaluations
score
bty_avg
5
4
3
2
4
6
8
rank
teaching
tenure track
tenured

Rank differences

```r
## average by tenure
evals %>%
  group_by(rank) %>%
  summarise(avg = mean(score))
```

rank	avg
<fct>	<dbl>
1 teaching	4.28
2 tenure track	4.15
3 tenured	4.14

```r
model_rank = lm(score ~ rank, data = evals)
```

```r
# get model output
get_regression_table(model_rank)
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1 intercept	4.28	0.054	79.9	0	4.18	4.39
2 ranktenure track	-0.13	0.075	-1.73	0.084	-0.277	0.017
3 ranktenured	-0.145	0.064	-2.28	0.023	-0.27	-0.02

$$\text{score} = \alpha + \beta_1(\text{rank}_{\text{tenure track}}) + \beta_2(\text{rank}_{\text{tenured}}) + \epsilon$$

$$\text{score} = 4.28 - 0.13(\text{rank}_{\text{tenure track}}) - 0.15(\text{rank}_{\text{tenured}}) + \epsilon$$

Score for teaching faculty

$$\text{score} = 4.28 - 0.13(\text{rank}_{\text{tenure track}}) - 0.15(\text{rank}_{\text{tenured}}) + \epsilon$$

$$\text{score} = 4.28 - 0.13 \times 0 - 0.15 \times 0 + \epsilon$$

$$\text{score} = 4.28$$

Score for tenured faculty

$$\text{score} = 4.28 - 0.13(\text{rank}_{\text{tenure track}}) - 0.15(\text{rank}_{\text{tenured}}) + \epsilon$$

$$\text{score} = 4.28 - 0.13 \times 0 - 0.15 \times 1 + \epsilon$$

$$\text{score} = 4.28 - 0.15$$

$$4.13$$

Regression coefficients

term	estimate	std_
<chr>	<dbl>	
intercept	4.28	
ranktenure track	−0.13	
ranktenured	−0.145	

Averages

	rank	avg
	<fct>	<dbl>
1	teaching	4.28
2	tenure track	4.15
3	tenured	4.14

Why do you think teaching = intercept?

Factors

Baseline category/intercept = lowest level in a factor variable

```
[457] tenure track tenure track tenure track
[463] tenure track
Levels: teaching tenure track tenured
```

When categorical variable *isn't* a factor, R picks one, typically first in alphabetical order

But you should set the factor a value that makes sense!

Interpretation

On average, y is βn units larger (or smaller) in xn, compared to x0

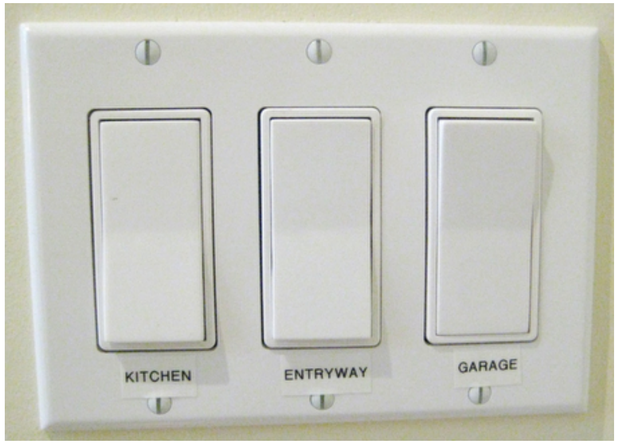
On average, student evaluations are .13 points lower for tenure-track faculty than for teaching faculty

Compared to teaching faculty, tenured faculty get, on average, .15 fewer points on teaching evals

A useful analogy: sliders and switches

$$\text{score} = 3.88 + 0.07(\text{bty\_avg}) + \epsilon$$

This is a **SLOPE**



$$\text{score} = 4.28 - 0.13(\text{rank}_{\text{tenure track}}) - 0.15(\text{rank}_{\text{tenured}}) + \epsilon$$

This is an **OFFSET**

Your turn

Regress score against a different categorical variable in *evals*

Interpret results for us

Most negative residual?

Most positive residual?

Interpreting residuals

What does Afghanistan having a residual of -27 tell us?

TABLE 5.9: Regression points (First 10 out of 142 countries)

country	lifeExp	continent	lifeExp_hat	residual
Afghanistan	43.8	Asia	70.7	-26.900
Albania	76.4	Europe	77.6	-1.226
Algeria	72.3	Africa	54.8	17.495
Angola	42.7	Africa	54.8	-12.075
Argentina	75.3	Americas	73.6	1.712
Australia	81.2	Oceania	80.7	0.515
Austria	79.8	Europe	77.6	2.180
Bahrain	75.6	Asia	70.7	4.907
Bangladesh	64.1	Asia	70.7	-6.666
Belgium	79.4	Europe	77.6	1.792