Correlation…and causality

EDUC 643: Unit 1

David D. Liebowitz

UNIVERSITY OF OREGON

Roadmap

Foundations

1. Introduction to regression

- The General Linear Model (GLM)
- Review of bivariate regression
- Coefficient- and model-level inference
- Correlation…and causality

2. Assumptions & Diagnostics

- Measurement error, normality, linearity, homoscedasticity, and independence
- Residuals: raw, studentized & standardized
- Outliers
- Diagnostics and solutions

Adding more and different predictors

3. Multiple regression

- Statistical adjustments (“controls”)
- Statistical inference
- Multi-collinearity

4. Categorical predictors

- Two-sample *t*-tests
- Regression with dummy variables
- ANOVA
- ANCOVA
- Variance decomposition

5. Interactions & Non-linearity

- Interactions in MR models
- Categorical * continuous
- Continuous * continuous
- Transformations to achieve linearity

Putting it together!

6. Applied regression modeling

Goals for the unit

- Characterize a bivariate relationship along five dimensions (direction, linearity, outliers, strength and magnitude)
- Describe how statistical models differ from deterministic models
- Mathematically represent the population model and interpret its deterministic and stochastic components
- Formulate a linear regression model to hypothesize a population relationship
- Estimated a fitted regression line using Ordinary-Least Squares regression
- Describe residuals and how they can describe the degree of our OLS model fit
- Conduct an inference test for a regression coefficient and our regression model

- Explain R^2, both in terms of what it tells us and what it does not
- Calculate a correlation coefficient (r) and describe its relationship to R^2
- Distinguish between research designs that permit correlational associations and those that permit causal inferences

Correlation ...and causality

Correlations

- Correlation coefficients (r) describe the **strength** of a linear relationship between two variables.
- The concept was first developed by Karl Pearson a eugenics professor at the University College of London. As we discussed last term, he held many despicable views.
- He (along with Francis Galton and RA Fisher) also pioneered many of the basic tools of modern statistics, including the concepts of standard deviation, χ^2, goodness of fit and the correlation coefficient
- Correlations are dimensionless measures that eliminate the metrics of any particular scale.
- To construct these dimensionless measures requires **standardizing** each variable.

Standardizing variables

- Any variable can be standardized using a simple algorithm.

Each observation (i) is transformed into standardized form using the following formula:

$$z_i = \frac{X_i - \mu}{\sigma}$$

- The standardized value is calculated calculated by **subtracting the mean** from each value and **dividing by the standard deviation**.

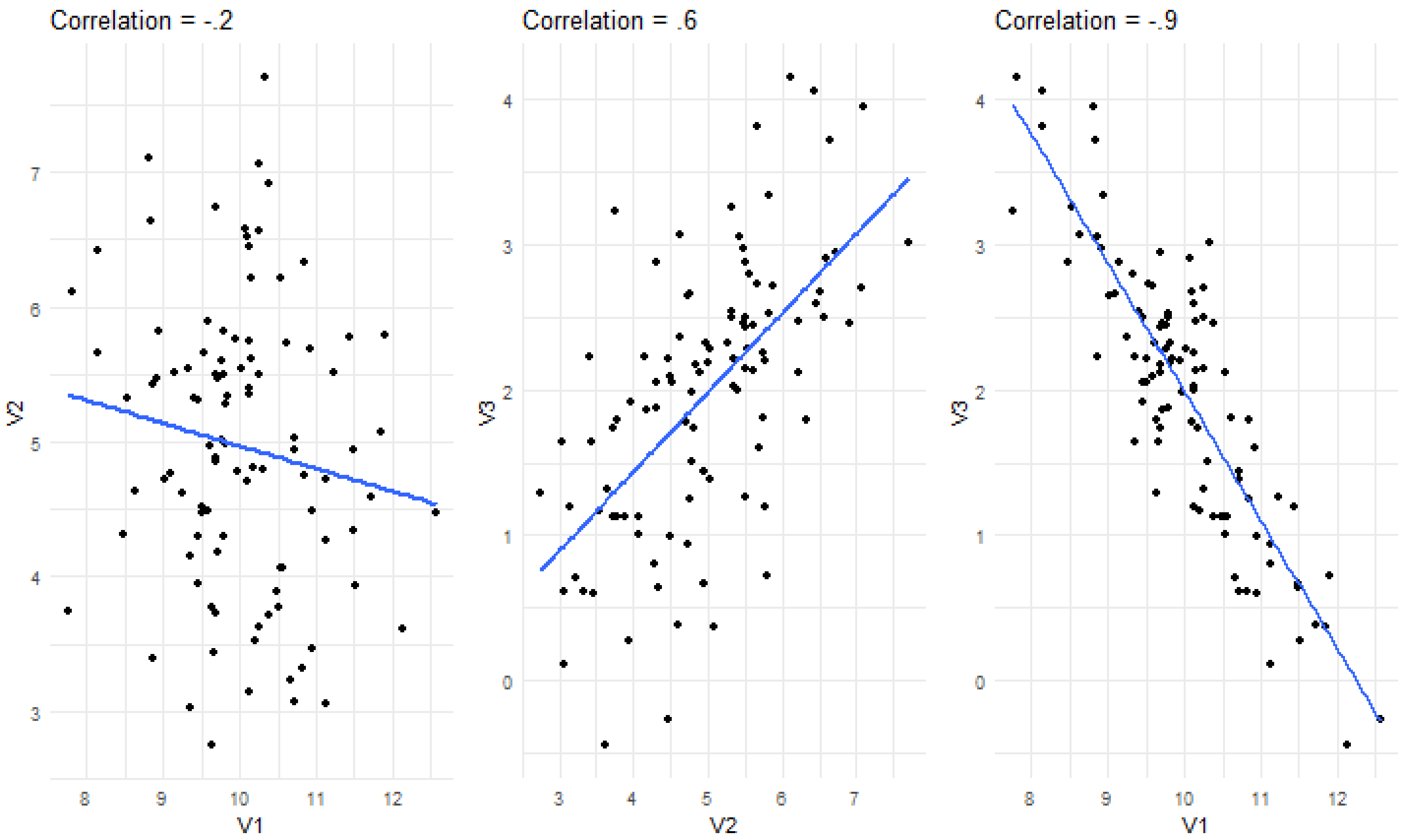
- The sample mean of the new variable is 0 and its standard deviation is 1

- The new values represent an observation's distance from the mean in standard deviation units.

- **Doesn't change anyone's relative rank**

- **Doesn't create a normally distributed variable**

Correlations visualized



Visualize in our data

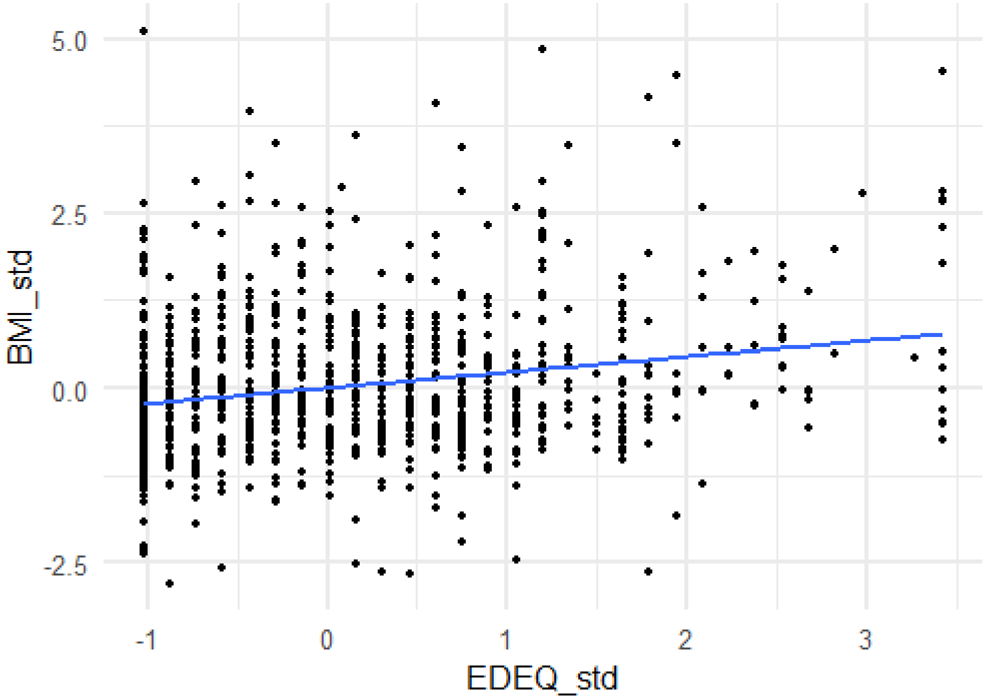
Let's transform *BMI* and *EDEQ_RESTRAINT* into standardized versions:

```
# Read in the data
do <- read_spss(here("data/male_do_eating.sav")) %>%
    select(OE_frequency, EDEQ_restraint, EDS_total,
           BMI, age_year, income_group) %>%
    mutate(EDS_total = ifelse(EDS_total==-99, NA, EDS_total)) %>%
    drop_na()

# Standardize the variables
do <- do  %>%
      mutate(BMI_std = (BMI - mean(BMI)) / sd(BMI))
do <- do %>%
      mutate(EDEQ_std =
      (EDEQ_restraint - mean(EDEQ_restraint)) / sd(EDEQ_restraint))
```

Visualize in our data

Let's transform *BMI* and *EDEQ_RESTRAINT* into standardized versions:

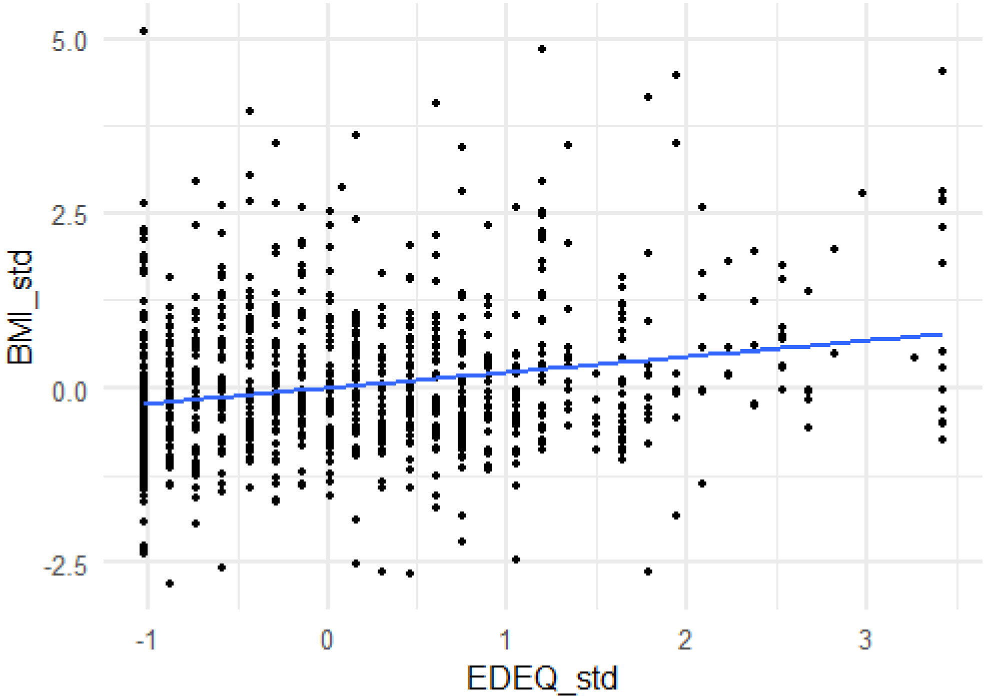


Note that the scale of our variables have changed.

- The standardized regression line goes through the origin (0, 0)

Visualize in our data

Let's transform *BMI* and *EDEQ_RESTRAINT* into standardized versions:



The new fitted regression line is:

$$\hat{BMI}_{std} = 0.000 + 0.2241 * DietaryRestraint_{std}$$

For fun, multiply that 0.2241 by itself: $(0.2241)^2 = 0.0502$. **Anything familiar about 0.05?**

r and R^2

$$r = \sqrt{R^2}$$

The coefficient on the regression of two standardized variables is called the Pearson product-moment coefficient. It is the same as the Pearson product-moment correlation (otherwise known as Pearson correlation). And it is the square root of the R^2.

Correlation coefficient values range from -1 to 1

- Positive Values: higher values of Y $\rightarrow$ higher values of X (and vice-versa)
- Negative Values: higher values of Y $\rightarrow$ lower values of X (and vice-versa)

Calculate correlation coefficient in R

```
cor(do$BMI, do$EDEQ_restraint)
```

```
## [1] 0.2240726
```

Formal correlation coefficient

Covariance:

$$\text{cov}_{XY} = \sigma_{XY} = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})$$

- However, units of covariance are hard to interpret, so...

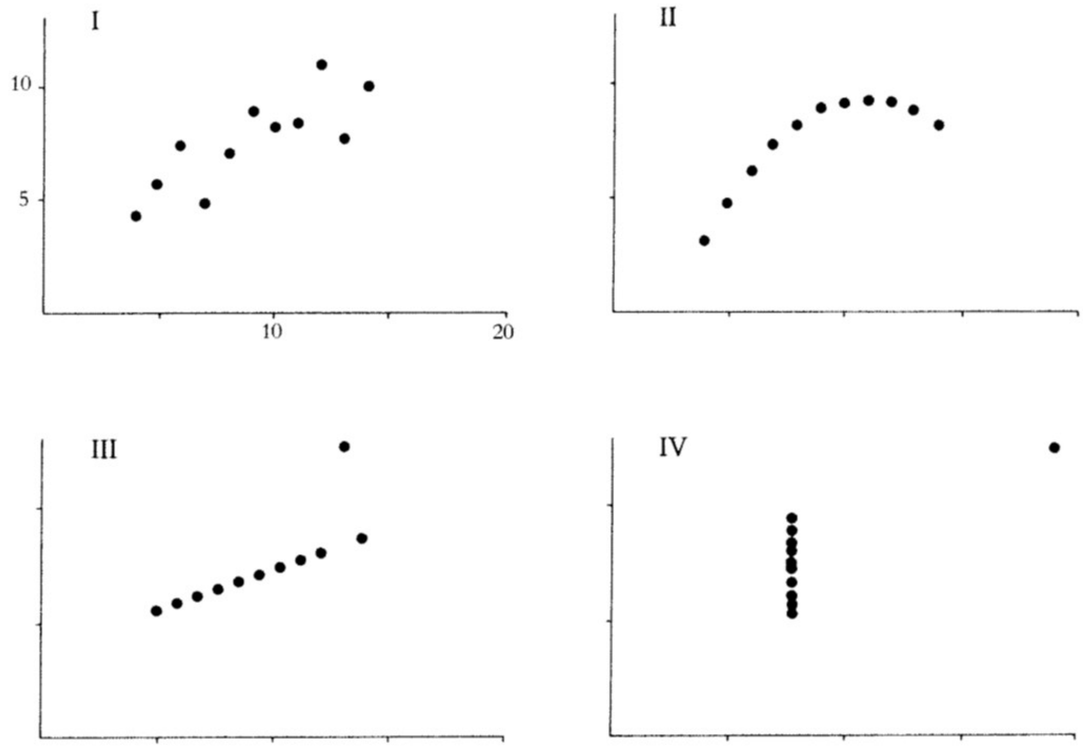
Correlation:

$$\text{corr}_{XY} = \rho_{XY} = \frac{\text{cov}_{XY}}{\hat{\sigma}_X \hat{\sigma}_Y}$$

- Pearson correlation divides by the standard deviation to put on a scale of -1 to 1

Anscombe's Quartet

...but, correlation is not everything. Frank Anscombe (1973) first highlighted the following set of distributions, all with correlations (r) of exactly 0.816.



What correlation does(n't) mean

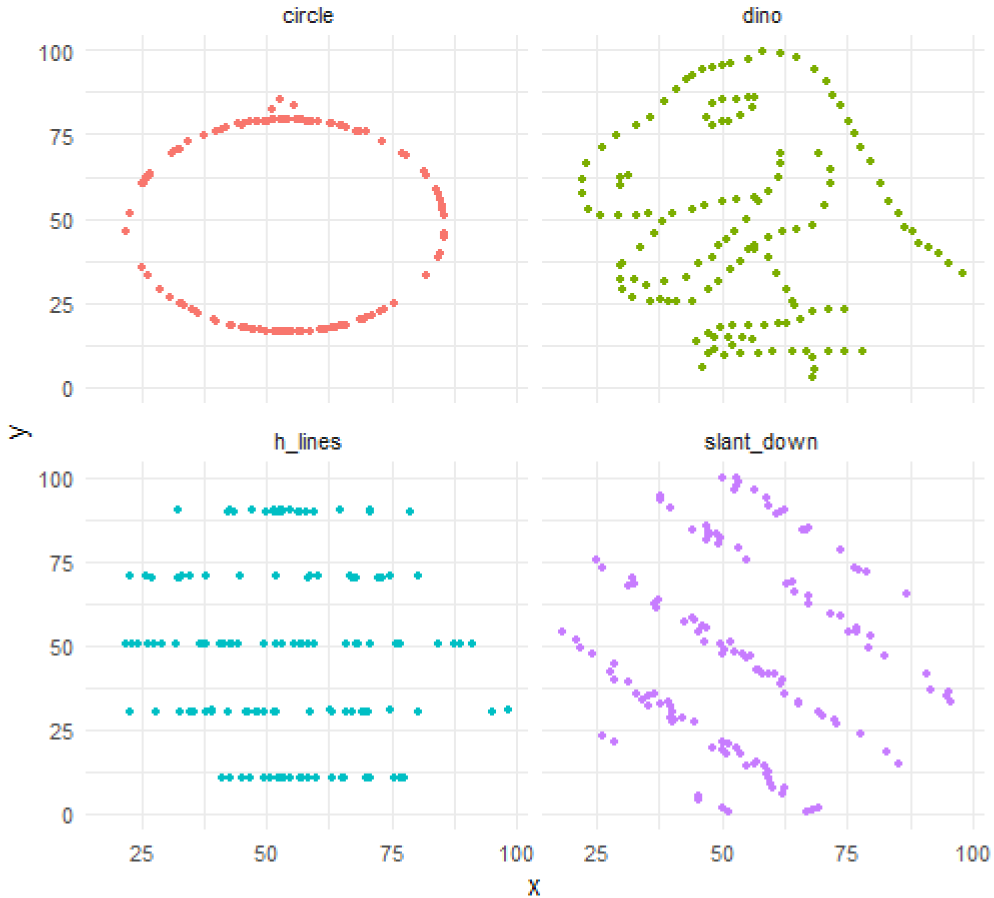
Here are four datasets, each with two variables with (nearly) identical means and correlations.

```
## # A tibble: 4 x 4
##   dataset    `mean(x)` `mean(y)` `cor(x, y)`
##   <chr>          <dbl>     <dbl>       <dbl>
## 1 circle          54.3      47.8     -0.0683
## 2 dino            54.3      47.8     -0.0645
## 3 h_lines         54.3      47.8     -0.0617
## 4 slant_down      54.3      47.8     -0.0690
```

What's the correlation between x and y across these four datasets?

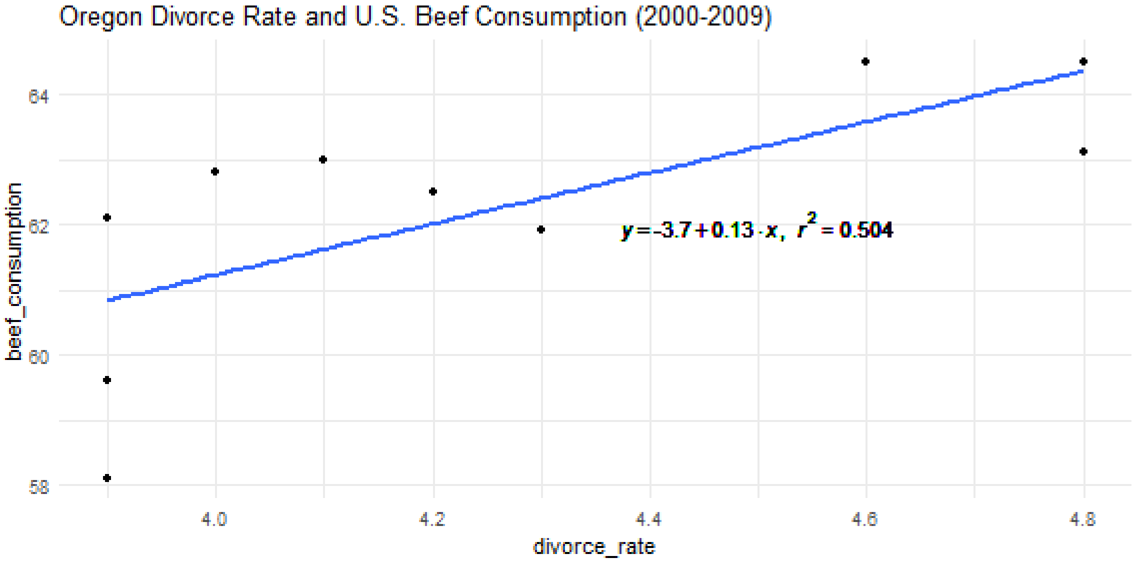
All seem pretty similar, right? Let's take a look at their bivariate relationship...

What correlation does(n't) mean



Correlation $\neq$ causation pt. 562

RQ: What is the relationship between Oregon's annual per capita divorce rate and the U.S. per capita annual beef consumption?



On the 10 o'clock news tonight: does U.S. beef consumption cause more "beefs" between Oregonians and their spouses?

Divorce and Beef

If we regress U.S. beef consumption on Oregon's divorce rate...

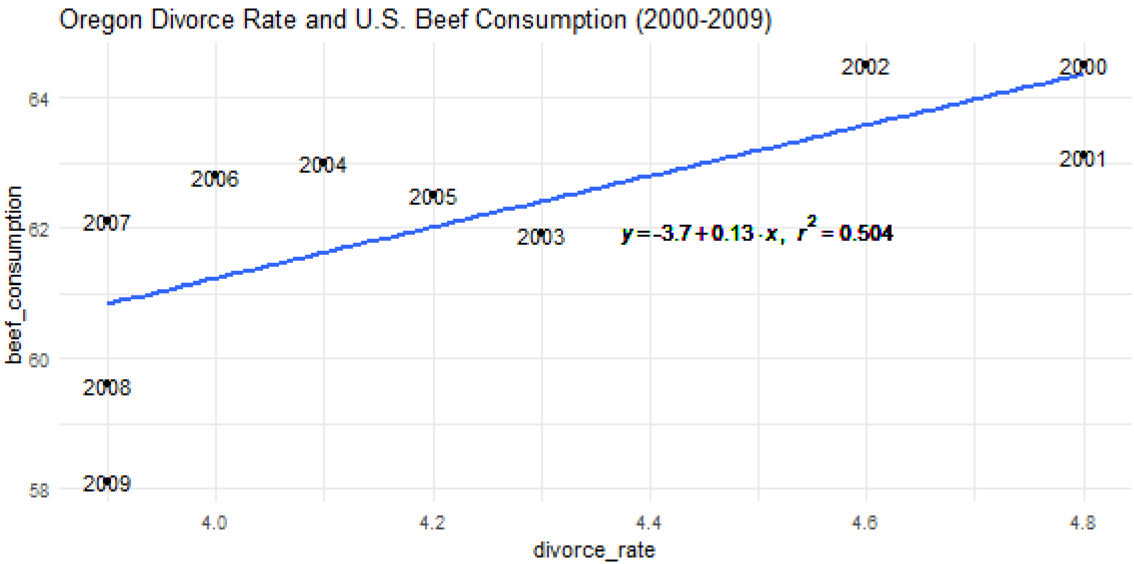
```
...
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    45.551      5.866   7.765 5.41e-05 ***
divorce_rate    3.920      1.376   2.849   0.0215 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.498 on 8 degrees of freedom
Multiple R-squared:  0.5037,	Adjusted R-squared:  0.4416
F-statistic: 8.119 on 1 and 8 DF,  p-value: 0.0215
...
```

The relationship between Oregon's divorce rate and U.S. beef consumption is statistically significant. In fact, Oregon's divorce rate accounts for 50% of the variance in U.S. beef consumption!

Divorce and Beef

Do increases in beef consumption in Oregon **cause** increases in the U.S. divorce rate?



This is a classic problem of a confounder![1]

[1] More fun with spurious correlations

Why correlation $\neq$ causation?

Common barriers in attributing causality to observed co-relationships include:

- Confounders: a third variable causes changes in X and also in Y
- Colliders: a third variable that is caused by both the predictor and outcome; controlling for this can make a true causal relationship disappear!
- Reverse causation: X may cause Y **or** Y may cause X
- Simpson's Paradox: a third variable may reverse the correlation
- Selection bias: due to the way in which the sample was constructed, observed relationships (or lack thereof) mask the true underlying relationship.
- Also, **lack** of correlation $\neq$ **lack** of causality

No correlation doesn't mean no causality.
CAUSAL INFERENCE: THE MIXTAPE, SCOTT CUNNINGHAM

h/t @causalinf

No racial bias in policing?

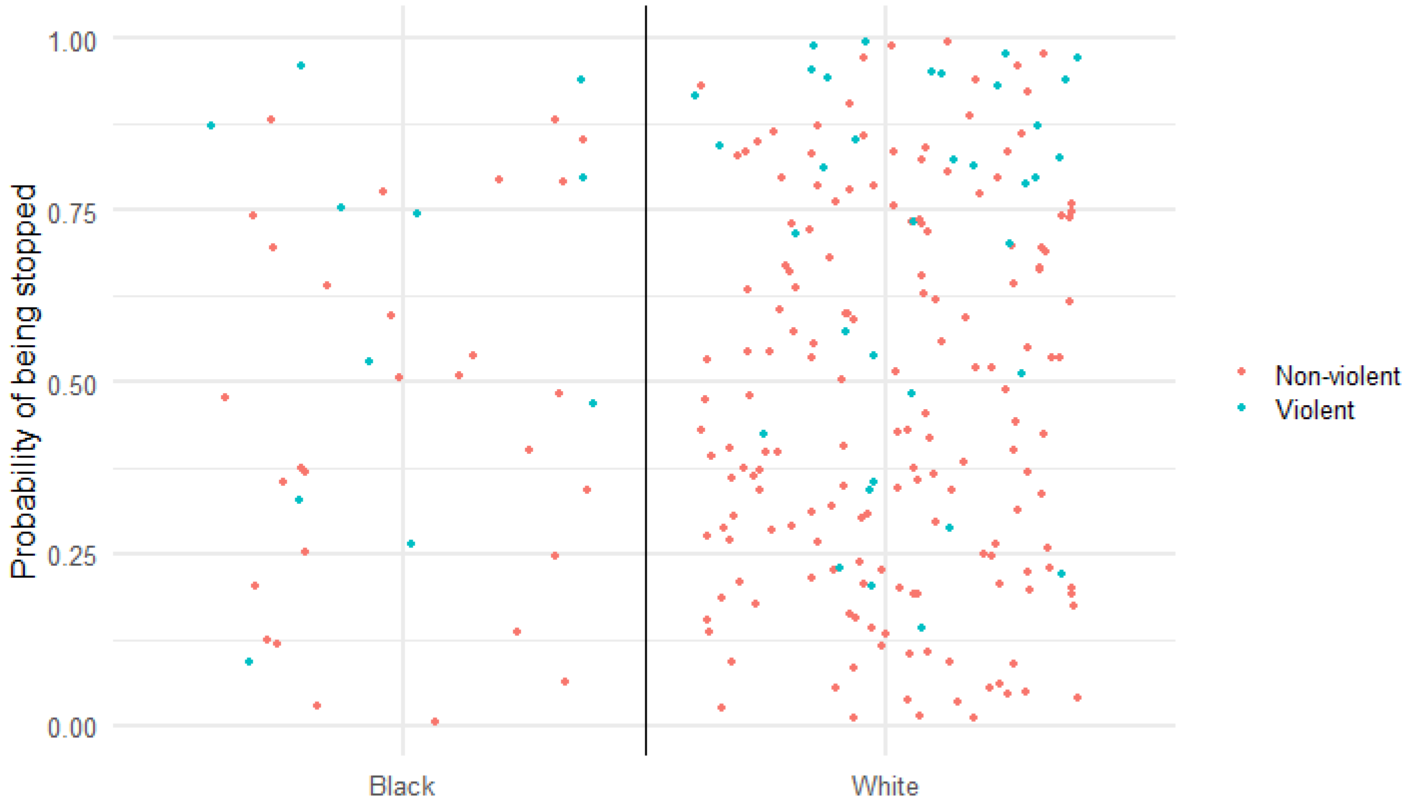
In 2016, Harvard economics professor Roland Fryer[1] released a study that claimed to find that police were ***equally likely*** to use force (i.e., officer-involved shootings) on whether the individuals they stopped were Black or White.

As you can imagine, this drew substantial media attention and controversy. We read about this in Lily Hu's article in the *Boston Review*.

[1] Since this study's release, Fryer has been accused of improper behavior towards members of his research lab.

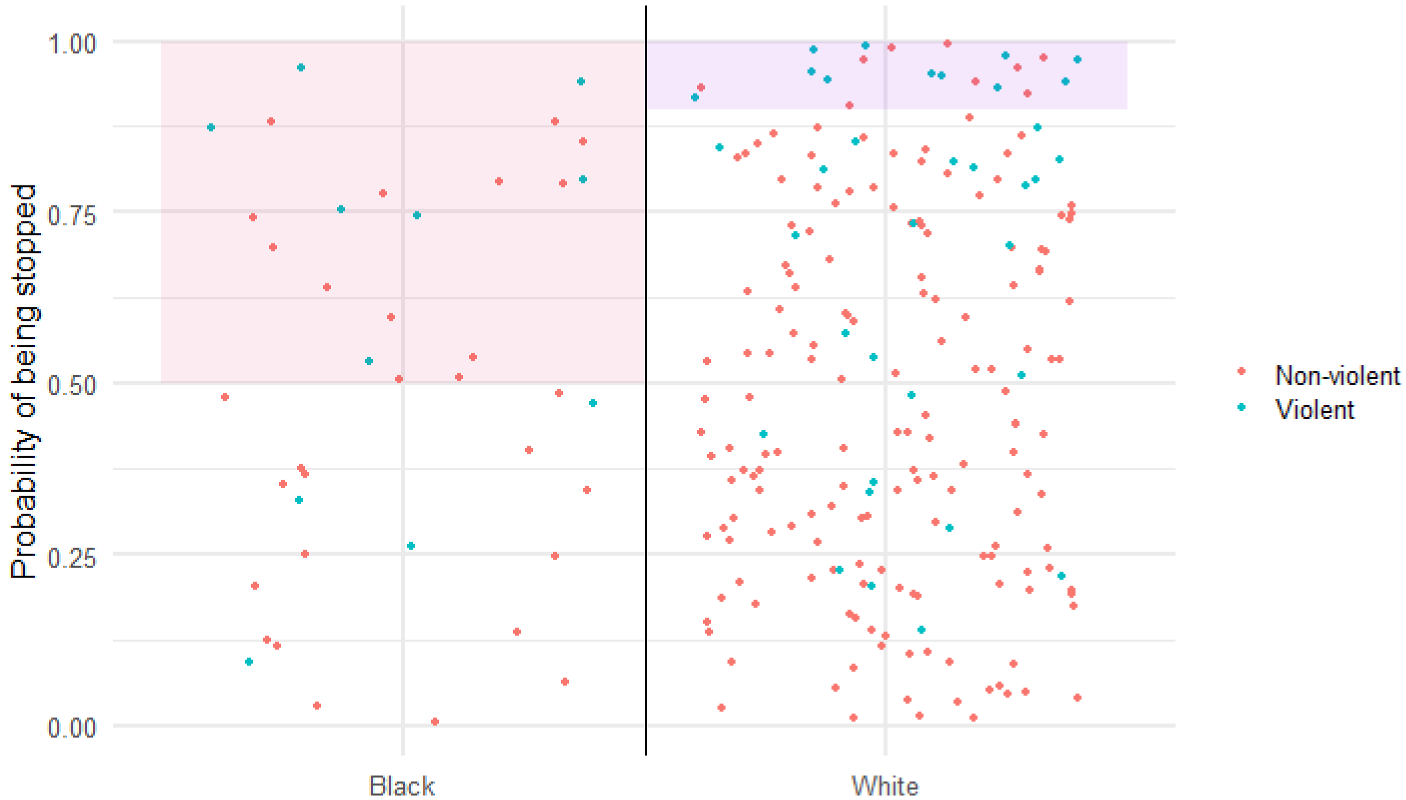
Selection bias

As detailed in Knox, Lowe and Mummolo (2020), Fryer's results are a product of a fundamental statistical error: **selecting on the outcome** (dependent variable)



Selection bias

As detailed in Knox, Lowe and Mummolo (2020), Fryer's results are a product of a fundamental statistical error: **selecting on the outcome** (dependent variable)

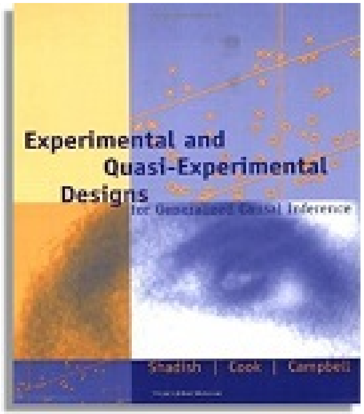


Helpful thread with a nice illustration of how selecting on the dependent variable and collider bias are interrelated.

From correlation to causality

Three criteria for establishing causality:[1]

1. Cause must precede effect in time
2. Systematic relationship between variation in cause and variation in effect
3. No plausible alternative explanation



Highest priority is establishing exogeneous variation in exposure to some "treatment" **OR** make an exceedingly convincing case that whether or not someone receives a "treatment" is a product of selection on observables (and not on any unobservables).

Research design is critical. So too can be Directed Acyclical Graphs (DAGs). We have whole classes dedicated to just this topic (EDLD 650, EDLD 679).

[1] Derived from Shadish, Cook and Cambpell (2002) and John Stuart Mill.

Turn and Talk

- Discuss science communication and how you would distinguish correlations from causal relationships to the average person.

- In what contexts might it be difficult to conduct an experimental study to establish causality?

- Consider the debate around the social science study of racial differences in officer-involved violence. Some of the conclusions analysts reach about critically important social issues diverge for highly technical and opaque reasons. What does this imply about the contributions of quantitative social science to public discourse and policy?

> *It is easy to prove that the wearing of tall hats and the carrying of umbrellas enlarges the chest, prolongs life, and confers comparative immunity from disease…A university degree, a daily bath, the owning of thirty pairs of trousers, a knowledge of Wagner's music, a pew in church, anything, in short, that implies more means and better nurture…can be statistically palmed off as a magic spell conferring all sorts of privileges…The mathematician whose correlations would fill a Newton with admiration, may, in collecting and accepting data and drawing conclusions from them, fall into quite crude errors by just such popular oversights.* –George Bernard Shaw (1906)

Color within the dots

aka, (mostly) don't predict beyond your data

Regression as a prediction

Regression equations can be used to evaluate the relationship between variables, and to predict expected values based on particular values of our predictors.

We can ask: What is the expected BMI value for a young male with a Dietary Restraint rating of **4**?

$$B\hat{M}I = 23.92 + 1.04 * (\mathbf{4}) = 28.1$$

The expected BMI for a Dietary Restraint of 4 is 28.1

Technically, there is no limit to what we can input!

Predicting beyond your data

Regression equations can be used to evaluate the relationship between variables, and to predict expected values based on particular values of our predictors.

We can ask: What is the expected BMI value for a young male with a Dietary Restraint rating of **400**?

Using our measure, this is not a possible value of Dietary Restraint but we can still estimate the predicted BMI using our regression equation.

$$B\hat{M}I = 23.92 + 1.04 * (\mathbf{400}) = 439.9$$

This is not a possible value for a human's BMI.

Only predict within the bounds of your data.

Synthesis and wrap-up

Goals for the unit

- Characterize a bivariate relationship along five dimensions (direction, linearity, outliers, strength and magnitude)
- Describe how statistical models differ from deterministic models
- Mathematically represent the population model and interpret its deterministic and stochastic components
- Formulate a linear regression model to hypothesize a population relationship
- Estimated a fitted regression line using Ordinary-Least Squares regression
- Describe residuals and how they can describe the degree of our OLS model fit
- Conduct an inference test for a regression coefficient and our regression model
- Explain R^2, both in terms of what it tells us and what it does not
- Calculate a correlation coefficient (r) and describe its relationship to R^2
- Distinguish between research designs that permit correlational associations and those that permit causal inferences

To-Dos

Reading:

- **By January 25**: LSWR Chapter 5.7

Quiz:

- Opens 3:45 Tuesday, Jan. 25 (closes 5pm on 1/26)

Assignment 1:

- Due February 5, 11:59pm

Next time: Regression assumptions