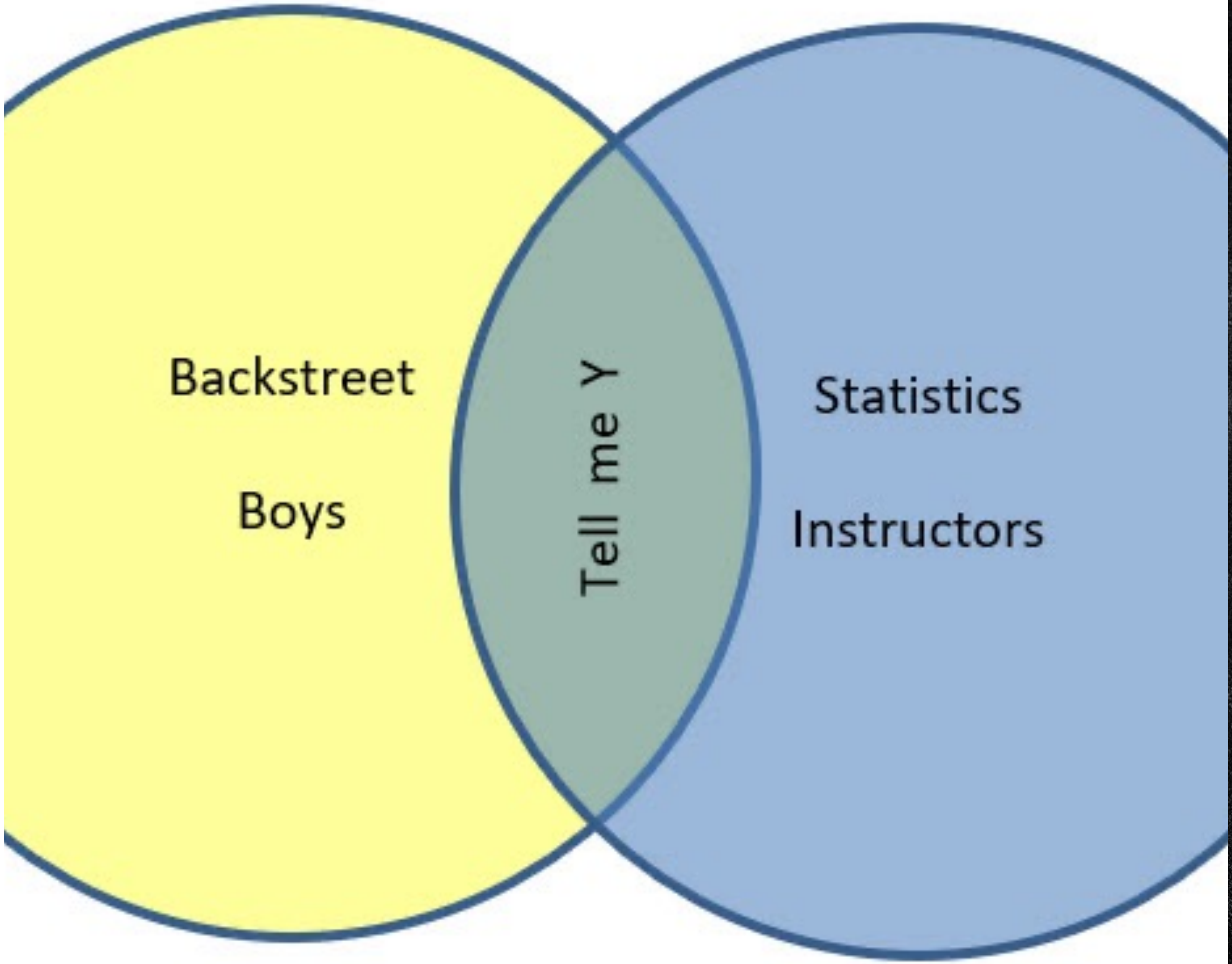


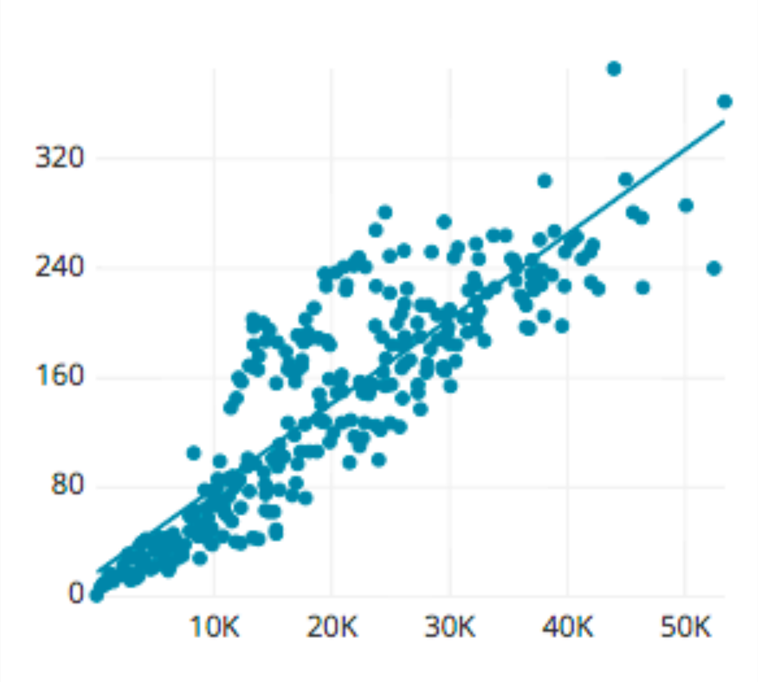
Regression

April 26, 2022

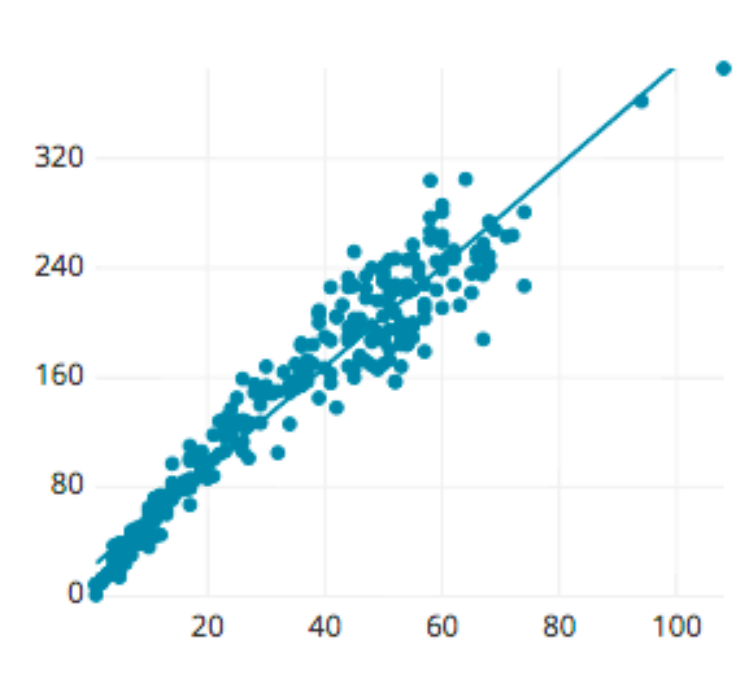
POLS 095

A "LINEAR RELATIONSHIP"

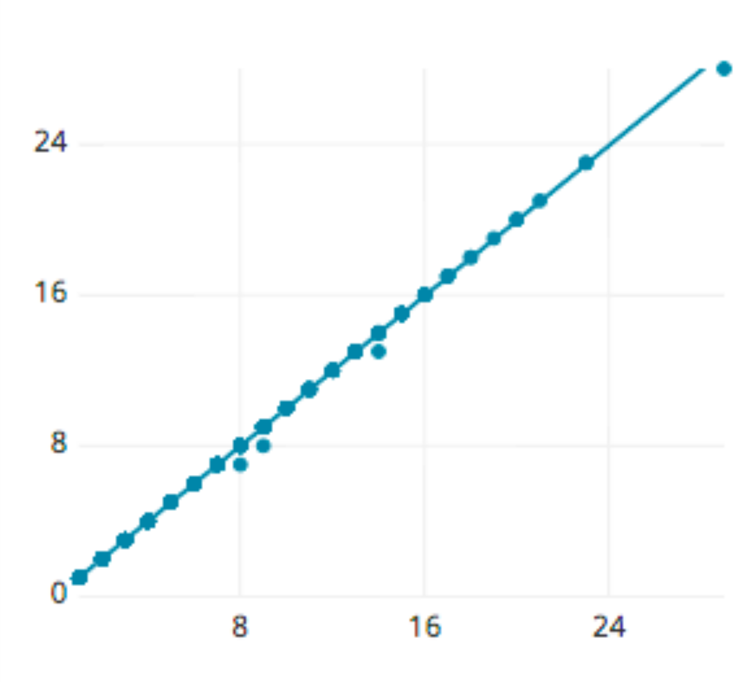
- Importantly, Pearson's *r* assumes a linear relationship. One good reason to look at your scatterplot first is to ensure that this is a good assumption.



0.9



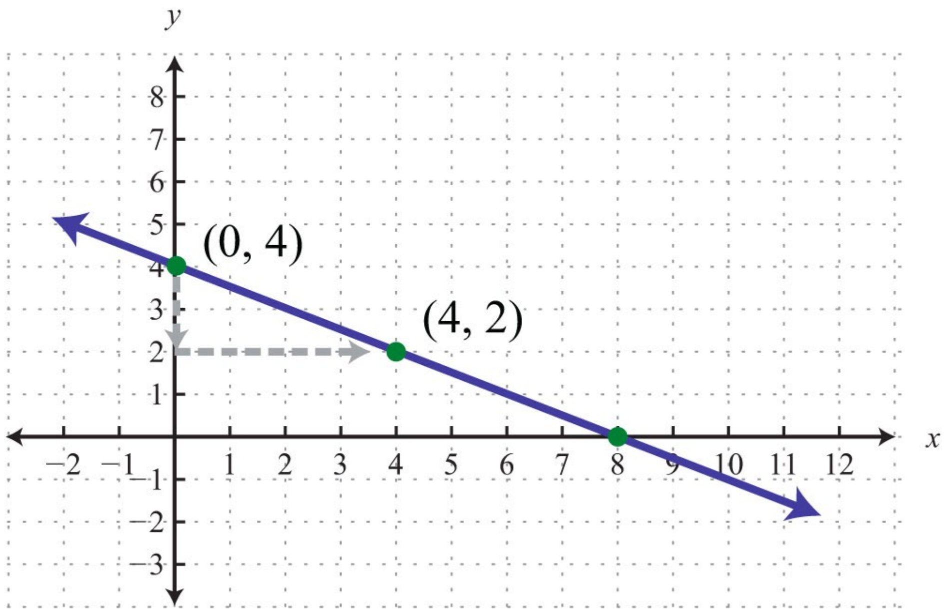
1.0



1.0

WHAT DOES THAT MEAN?

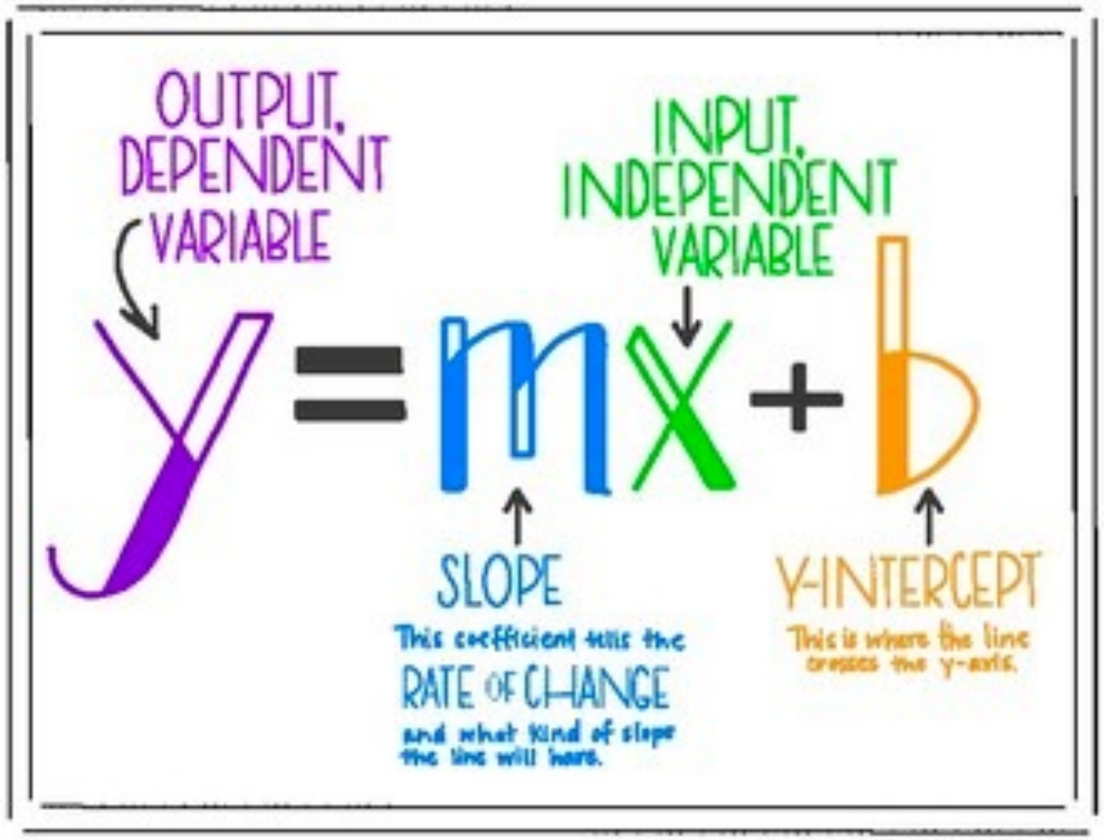
- A linear relationship assumes that the data can be approximated by a line.
- Implies that every unit increase in your independent variable, there is a certain change in your dependent variable, and that change is *constant* across the whole range of data.
- The change is the “slope” (“rise over run”).



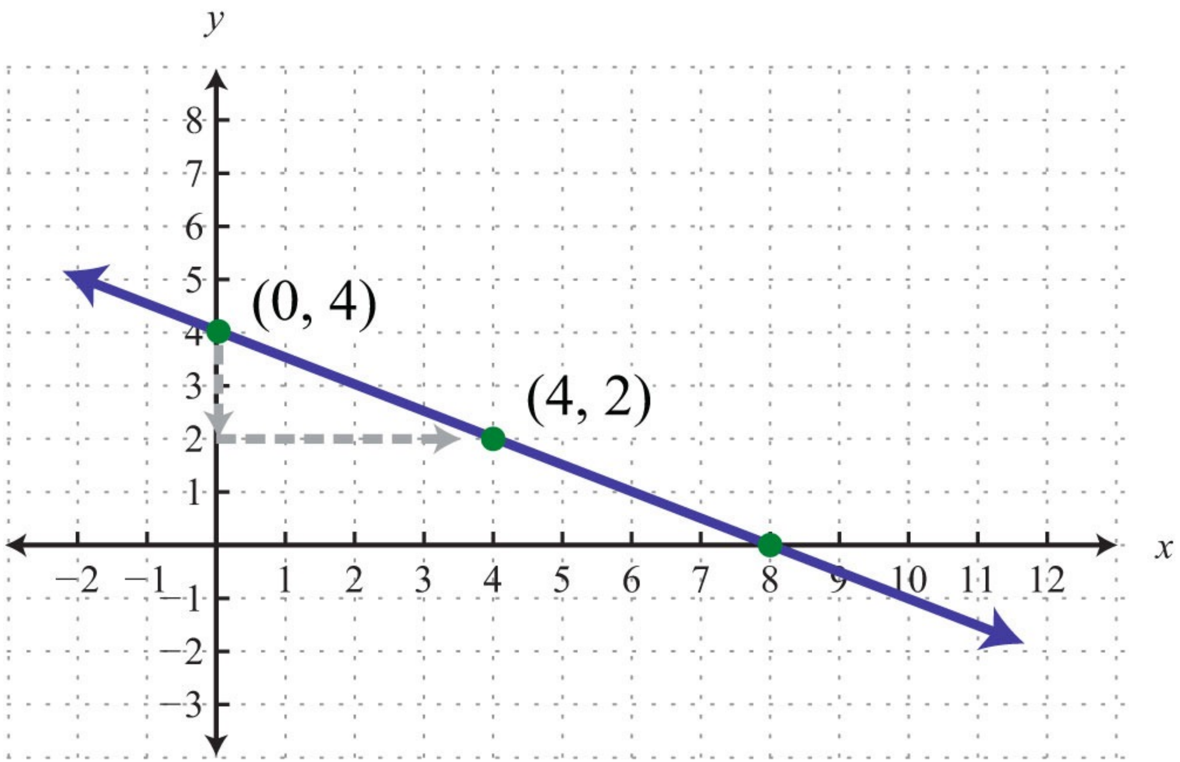
THE EQUATION FOR A LINE

Slope: $y = mx + b$

- If we identify x as our independent variable and y as our dependent variable (as we've been doing all semester), the linear relationship between x and y is right here!
- Every time x increases by 1 unit (whatever unit that is), we should expect y to change in a predictable way – the slope (m) tells us how.
- Can be positive or negative, big or small.



WHAT IF THE SLOPE = 4, -.3, OR 8,099?



Slope: y = mx + b

If m = -2, then every time we increase x by 1, y will decrease by 2.

If m = -.3, then every time we increase x by 1, y will decrease by .3.

If m = 8,099, then every time we increase x by 1, y will increase by 8,099.

HOW DO WE FIND THAT LINE?

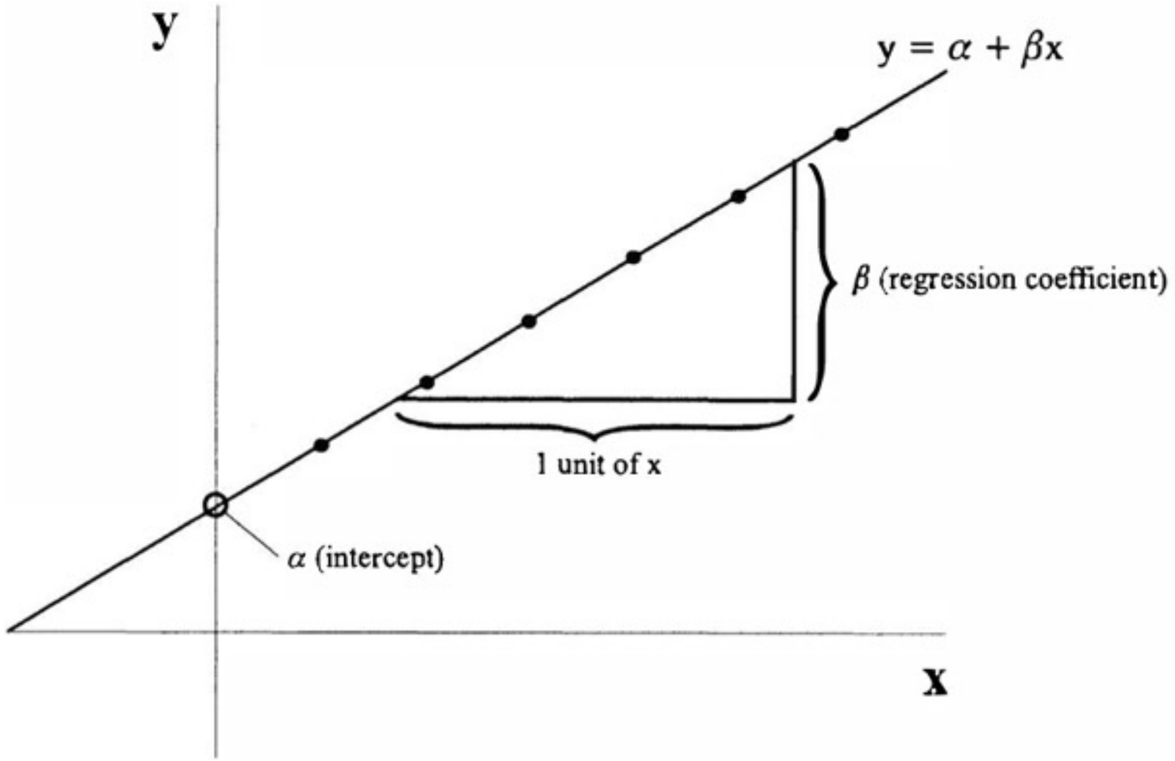
- The principle is the same as for correlation:
  - how far is the value of x from the mean value of x across all the data?
  - how far is the value of y from the mean value of y across all the data?

FIRST…

THE EQUATION

- Re-state the equation of a line (forget $y = mx + b$)
  - $Y = \alpha + \beta x$
  - This is exactly the same as $y = mx + b$, just different notation.
  - β = slope; α = intercept

- β is our “coefficient” and just like any slope, it can be positive or negative, big or small.
  - This is the effect



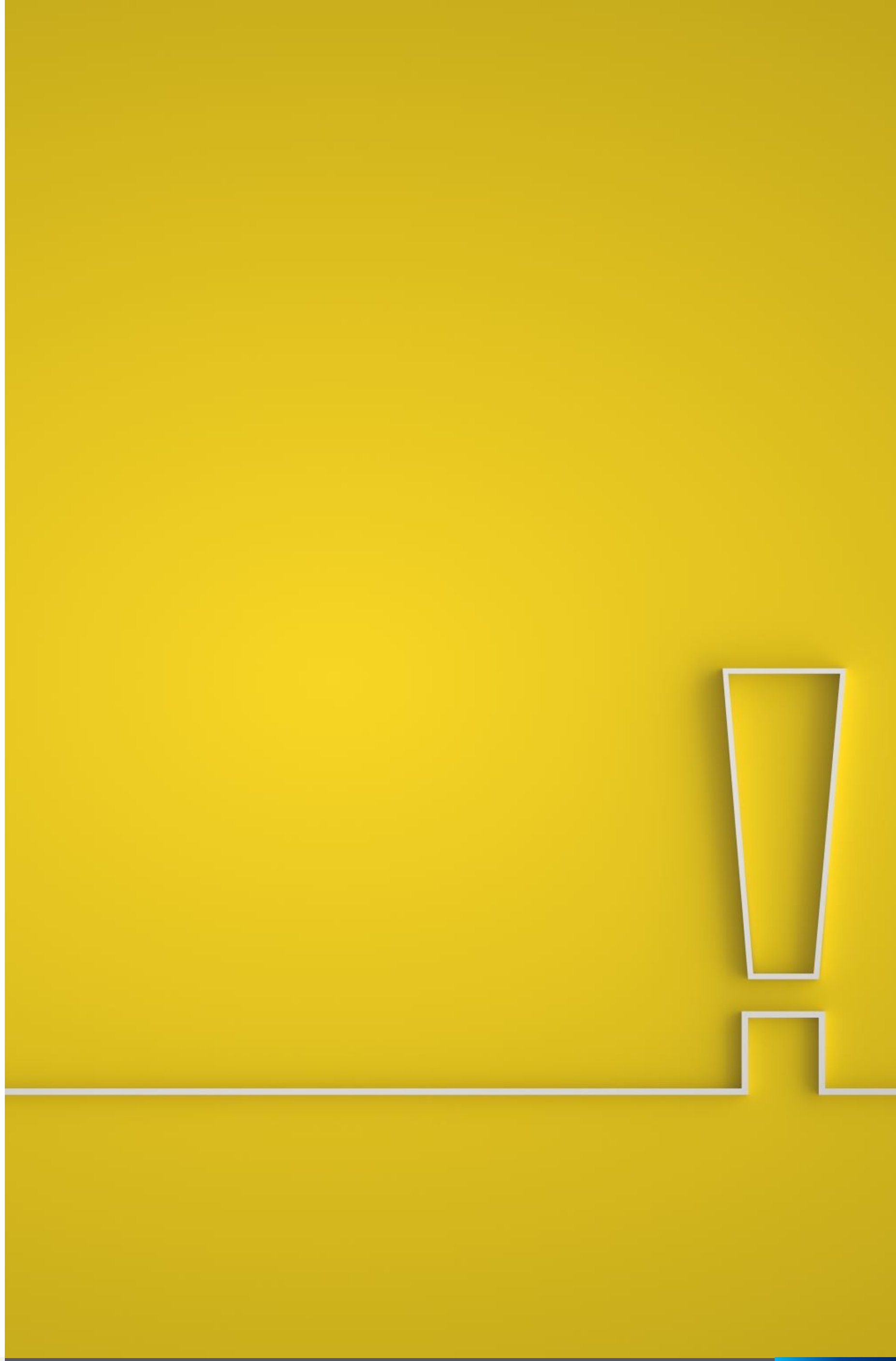
SOCIAL DATA ARE MESSY

Dependent variable = constant + independent variable + control variable 1 + control variable 2 + error

There will be error.

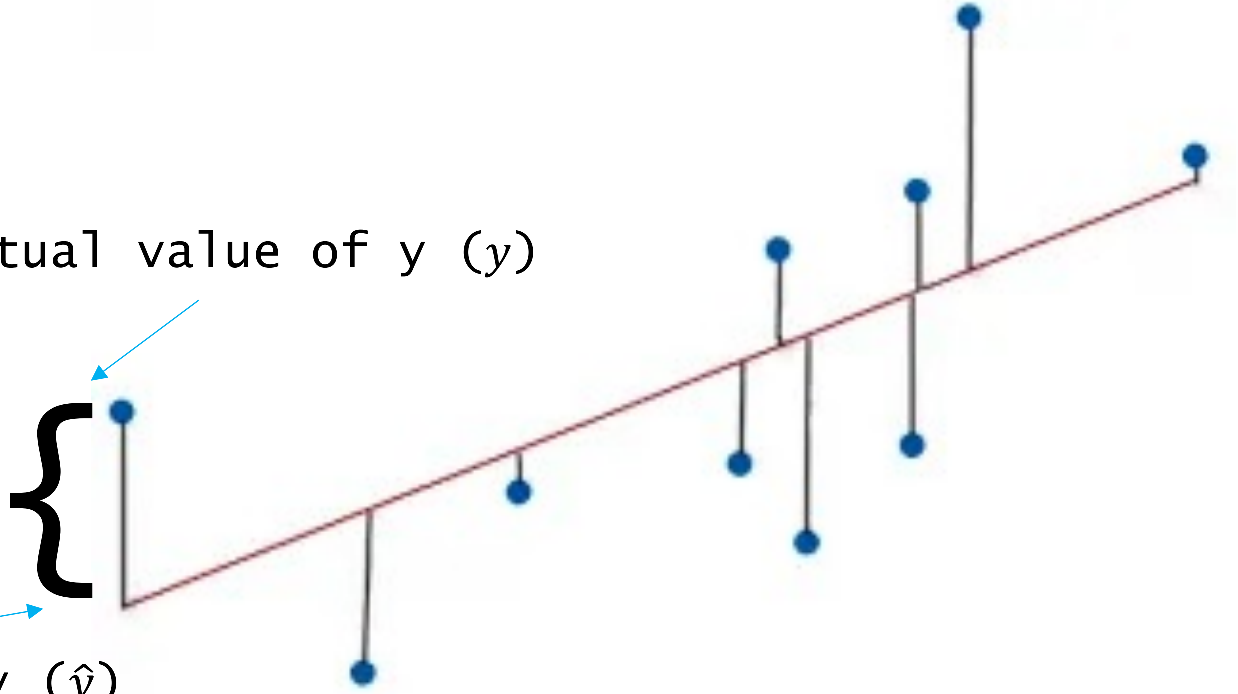
- I.e., not every case will fit the relationship (e.g., not all Democrats voted for Joe Biden, or Republicans for Trump; not every country with high GDP has a free press)

- $\hat{y} = \hat{a} + \hat{b}\,(x) + e$ where $\hat{y}$ is the estimated value of our dependent variable, $\hat{a}$ is our estimated intercept (α), and $\hat{b}$ is our estimated slope (β).
  - Hats = estimates



THE LINE OF BEST-FIT

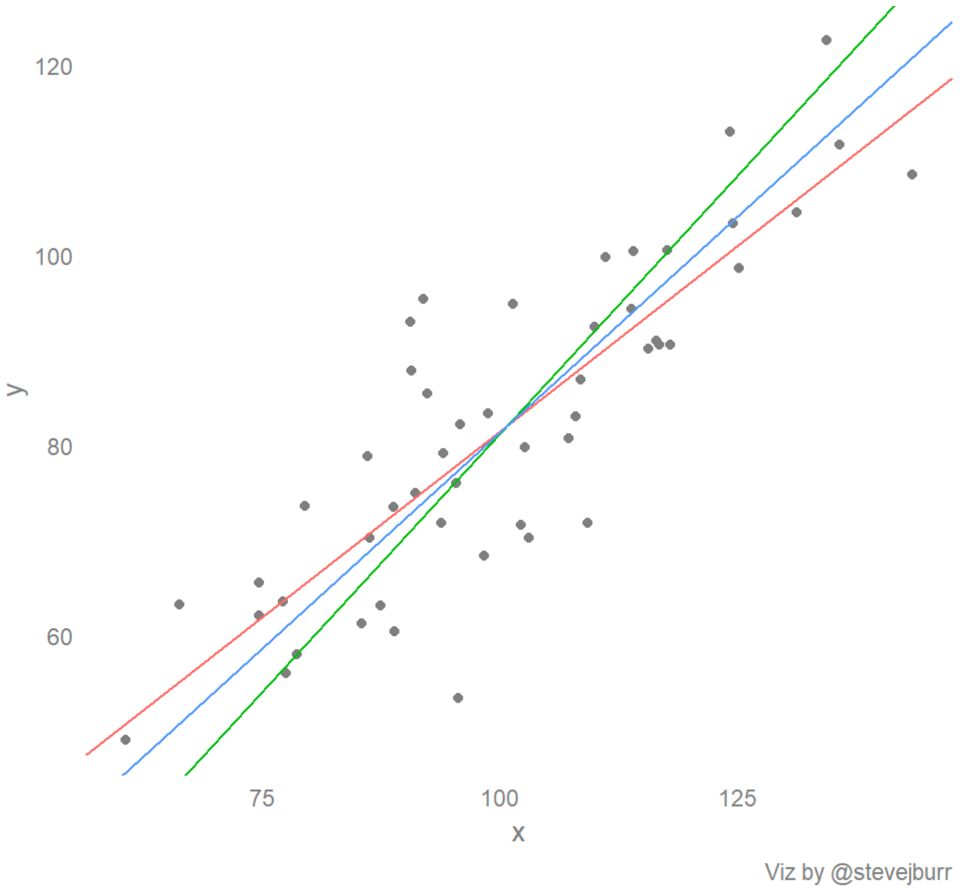
- How to find that line?
- For each data point and for any line you draw, there will be error in prediction, so we look at the actual value of y and the predicted value of y, the error in prediction.



THE LINE OF BEST-FIT

$$SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- The line of best fit will minimize those "residuals."
- Add up all residuals; seek the line that minimizes the total residuals
- Need to square the residuals because if we just add them up they'll cancel each other out
  - (just like with the standard deviation)
- Seeking the line with the smallest total squared residuals, or the line that is closest to the data.
- We call the method "Ordinary Least Squares" (OLS) regression.



HYPOTHESIS TESTING

- We are most interested in β, the true effect of a one unit increase in x on the value of y.
- H_o, the null hypothesis, is that $\beta = 0$. This would mean that a change in x has no effect on the value of y.
- We have an estimate of β, which is $\hat{b}$.
- Just like we can test the difference of means, we can use a *t*-test to evaluate the likelihood that the value of β is zero.

$$t = \frac{\hat{b} - \beta}{standard\ error\ of\ \hat{b}}$$

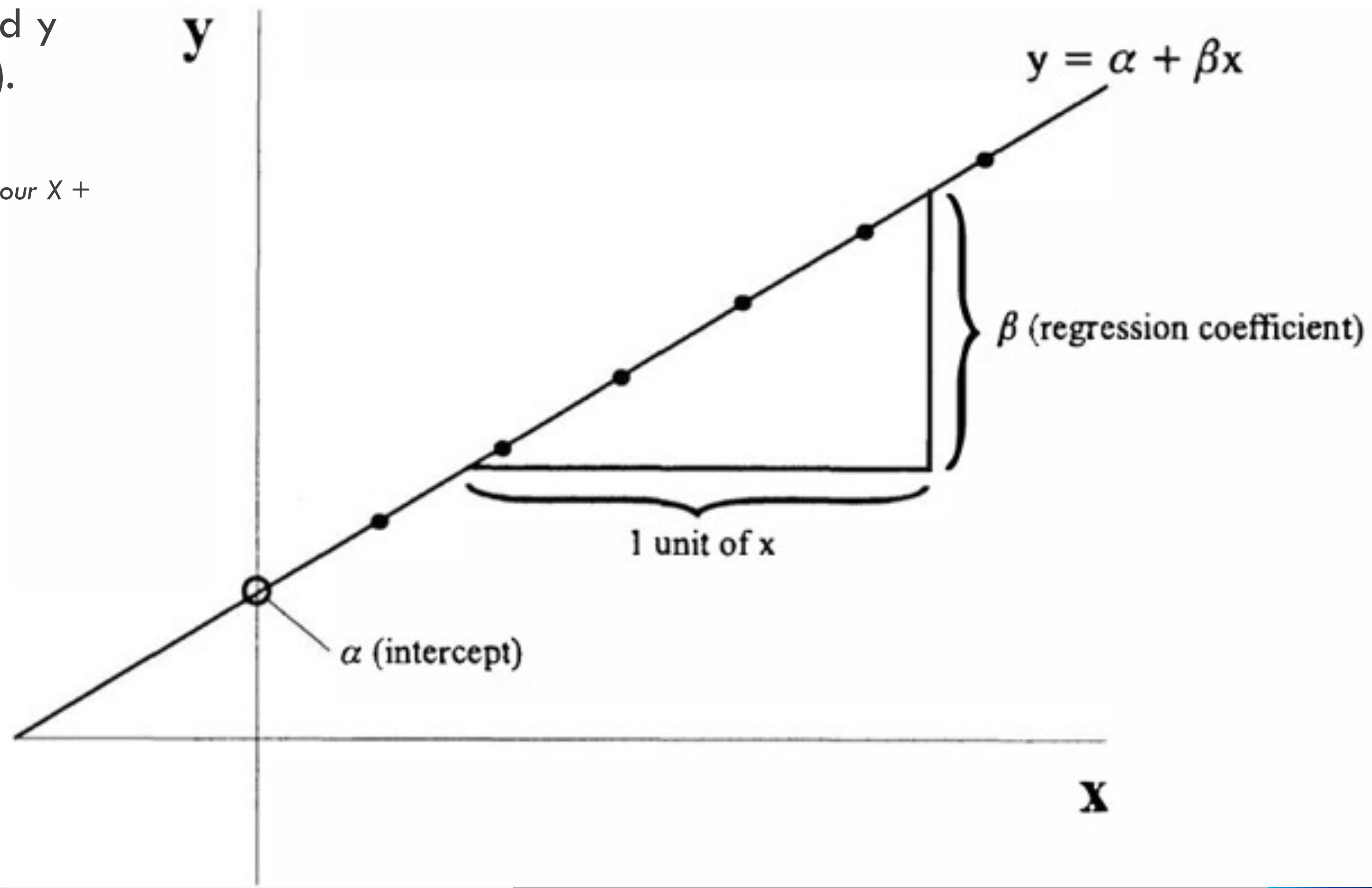
With d.f. = n - 2

😳

No…. You will not need to calculate this!

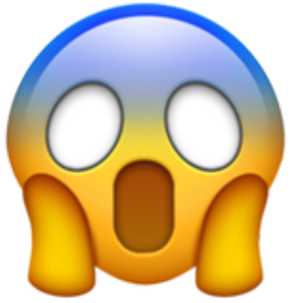
HOW TO INTERPRET

- The estimate of β is VERY valuable in that is allows for a precise statement of the relationship between x and y ("for every one unit increase in x, y changes by β").
- To learn how well the line fits the data, we use R^2
  - *It tells us what percentage of the variation in Y we can explain with X (or our X + controls)*



R^2

Still…. You do not need to compute any of this!



We call it a "goodness of fit" measure.

Given that we've squared all the errors, R^2 is often inflated, so we commonly adjust it and use a more conservative estimate, adjusted-R^2.

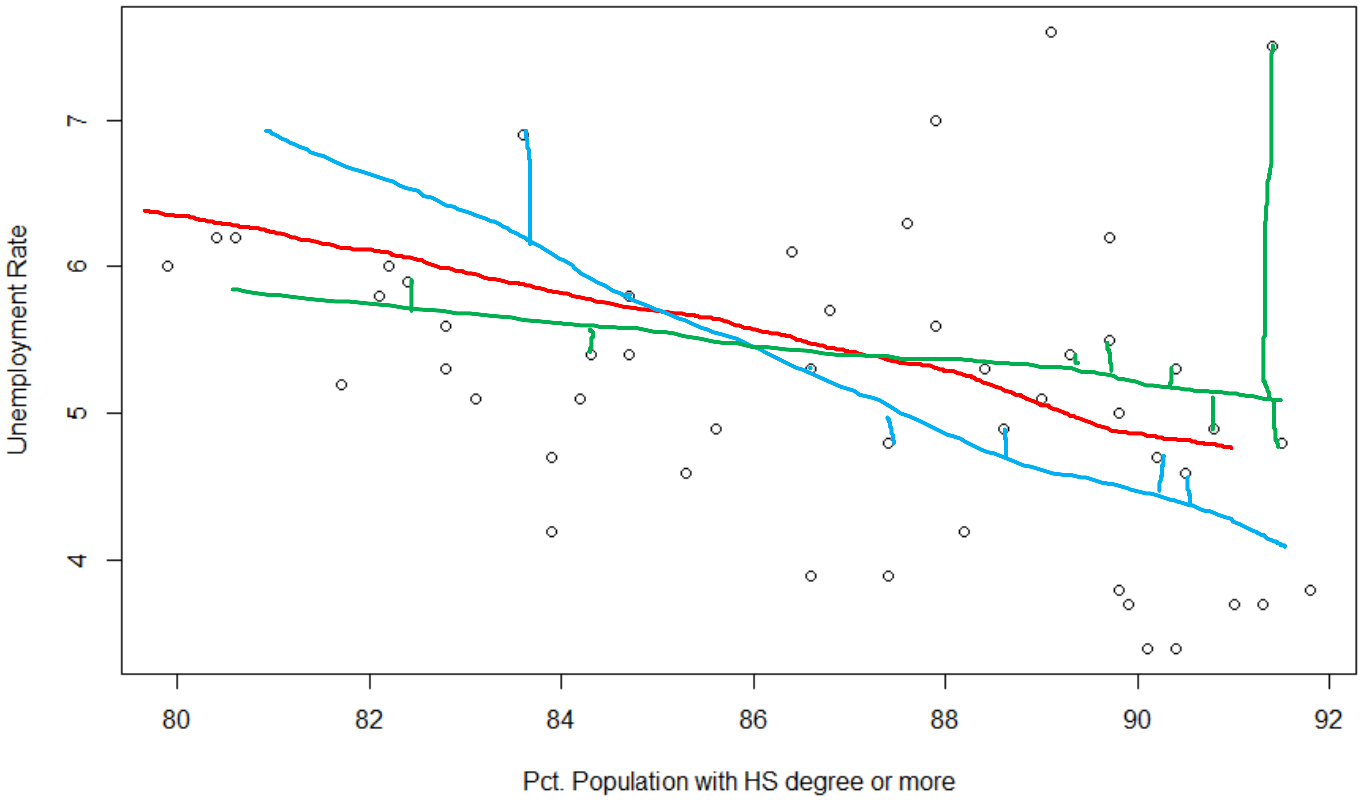
Practical interpretation, R^2 tells us the percent of variation in the dependent variable that can be explained by the independent variable.

EXAMPLE 1

Does lower education cause greater unemployment?

- H_A: In a comparison of states, those with lower education will have higher rates of unemployment than those with higher education.
- H_o :There is no relationship between education and unemployment.

- Dataset: states
- Dependent variable: unemploy
- Independent variable: hs_or_more


Unemployment Rate
4
5
6
7
80
82
84
86
88
90
92
Pct. Population with HS degree or more

WHY STATISTICAL ANALYSIS PROGRAMS ARE WONDERFUL!

```
> mod.1 = lm(unemploy ~ hs_or_more, data=states)
> summary(mod.1)

Call:
lm(formula = unemploy ~ hs_or_more, data = states)

Residuals:
     Min       1Q   Median       3Q      Max
-1.45395 -0.69418  0.03689  0.38075  2.78066

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.18365    3.47714   4.079  0.00017 ***
hs_or_more  -0.10355    0.03999  -2.589  0.01270 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9552 on 48 degrees of freedom
Multiple R-squared:  0.1225,    Adjusted R-squared:  0.1043
F-statistic: 6.703 on 1 and 48 DF,  p-value: 0.0127
```

α

p-value

β

Adjusted R^2

R^2

- Estimated $\beta = -0.10355 = \mathit{-0.10}$
  - “For every 1 unit increase in education, unemployment will decrease by 0.10.”
- Estimated $\alpha = \ 14.18365 = 14.18$
  - “When education = 0, we expect that unemployment would be 14.18.”
- P-value (Pr(>|t|)) = 0.01270
  - Given the data, there is a 1.27% probability that this estimated β would occur by chance.
  - All we care about is whether this is less than or equal to 0.05.
- Adjusted R^2 = 0.1043 = 0.10
  - “The independent variable explains 10% of the variation in the dependent variable

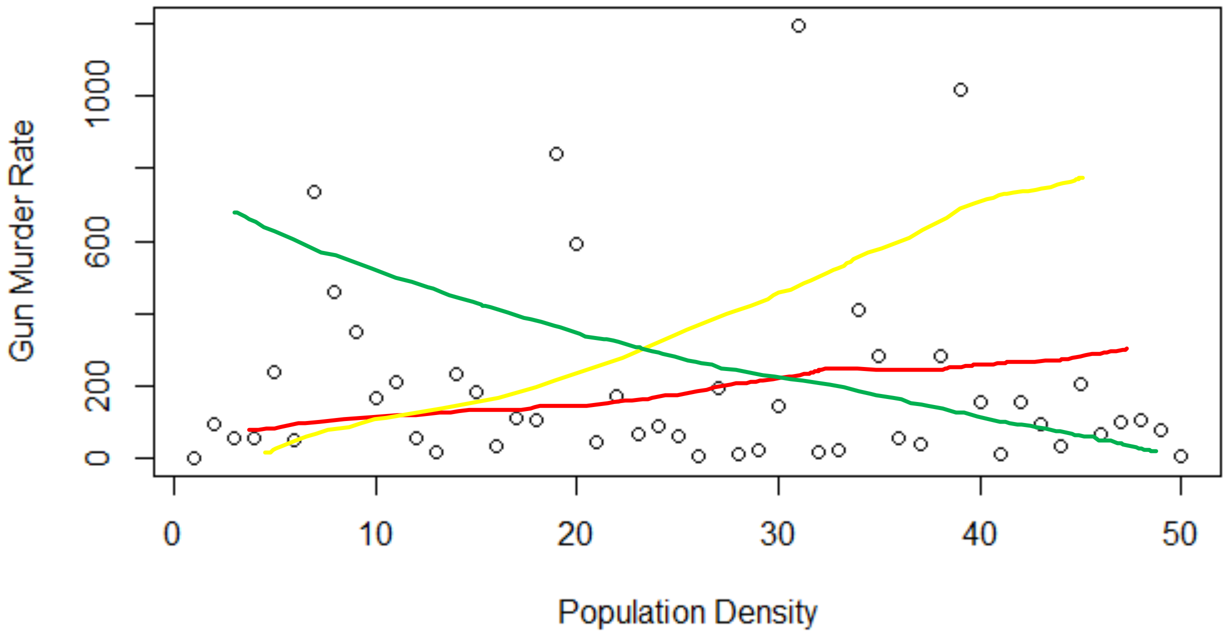
```
> mod.1 = lm(unemploy ~ hs_or_more, data=states)
> summary(mod.1)

Call:
lm(formula = unemploy ~ hs_or_more, data = states)

Residuals:
     Min       1Q   Median       3Q      Max
-1.45395 -0.69418  0.03689  0.38075  2.78066

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.18365    3.47714   4.079  0.00017 ***
hs_or_more  -0.10355    0.03999  -2.589  0.01270 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9552 on 48 degrees of freedom
Multiple R-squared:  0.1225,	Adjusted R-squared:  0.1043
F-statistic: 6.703 on 1 and 48 DF,  p-value: 0.0127
```

REJECT THE NULL HYPOTHSIS?

- P>t = 0.01270. Given the data, there is a 1.27% probability that this estimated β would occur by chance. If we use the 95% confidence level, P>t would need to be less than .05.
- WE CAN REJECT THE NULL HYPOTHESIS.

EXAMPLE 2

- H_A: In a comparison of states, those with higher population density will have higher murder rates than those with lower population density.
- H_o :There is no relationship between population density and the murder rate.
- Dataset: states
- Dependent variable: gun_murder10
- Independent variable: density


Gun Murder Rate
0
200
600
1000
0
10
20
30
40
50
Population Density

WHY STATISTICAL ANALYSIS PROGRAMS ARE WONDERFUL!

```
> summary(mod.2)

Call:
lm(formula = gun_murder10 ~ density, data = states)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.0905 -1.3197  0.0245  0.8514  5.2776 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.3319697  0.2663018   8.757 1.62e-11 ***
density     0.0008619  0.0008225   1.048      0.3    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.503 on 48 degrees of freedom
Multiple R-squared:  0.02236,	Adjusted R-squared:  0.001993 
F-statistic: 1.098 on 1 and 48 DF,  p-value: 0.3
```

α

β

p-value

Adjusted R^2

R^2

- Estimated $\beta = 0.0008619 = 0.00$
  - "For every 1 unit increase in population density, murder rate increase by 0."
- Estimated $\alpha = 2.3319697 = 2.33$
  - "When population density = 0, we expect that murder would be 2.32"
- P>t = 0.33
  - Given the data, there is a 33% probability that this estimated β would occur by chance.
- Adjusted $R^2 = 0.001993 = 0.00$
  - "The independent variable explains 0% of the variation in the dependent variable"

```
> summary(mod.2)

Call:
lm(formula = gun_murder10 ~ density, data = states)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.0905 -1.3197  0.0245  0.8514  5.2776 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.3319697  0.2663018   8.757 1.62e-11 ***
density     0.0008619  0.0008225   1.048      0.3    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.503 on 48 degrees of freedom
Multiple R-squared:  0.02236,	Adjusted R-squared:  0.001993 
F-statistic: 1.098 on 1 and 48 DF,  p-value: 0.3
```

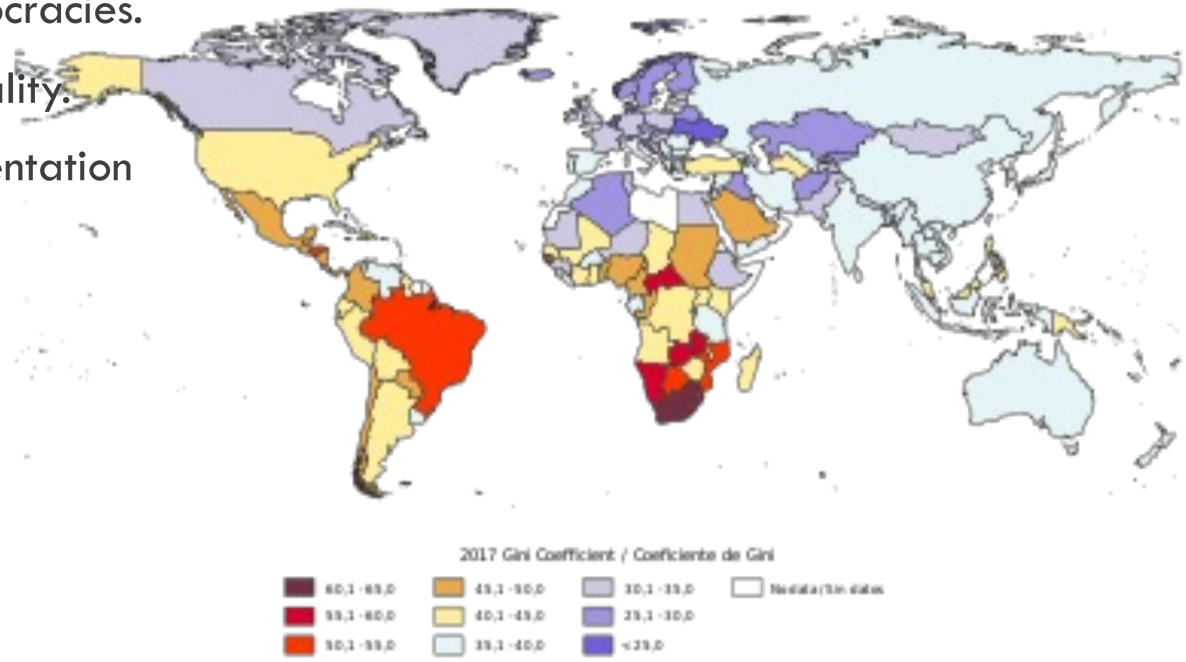
REJECT THE NULL HYPOTHSIS?

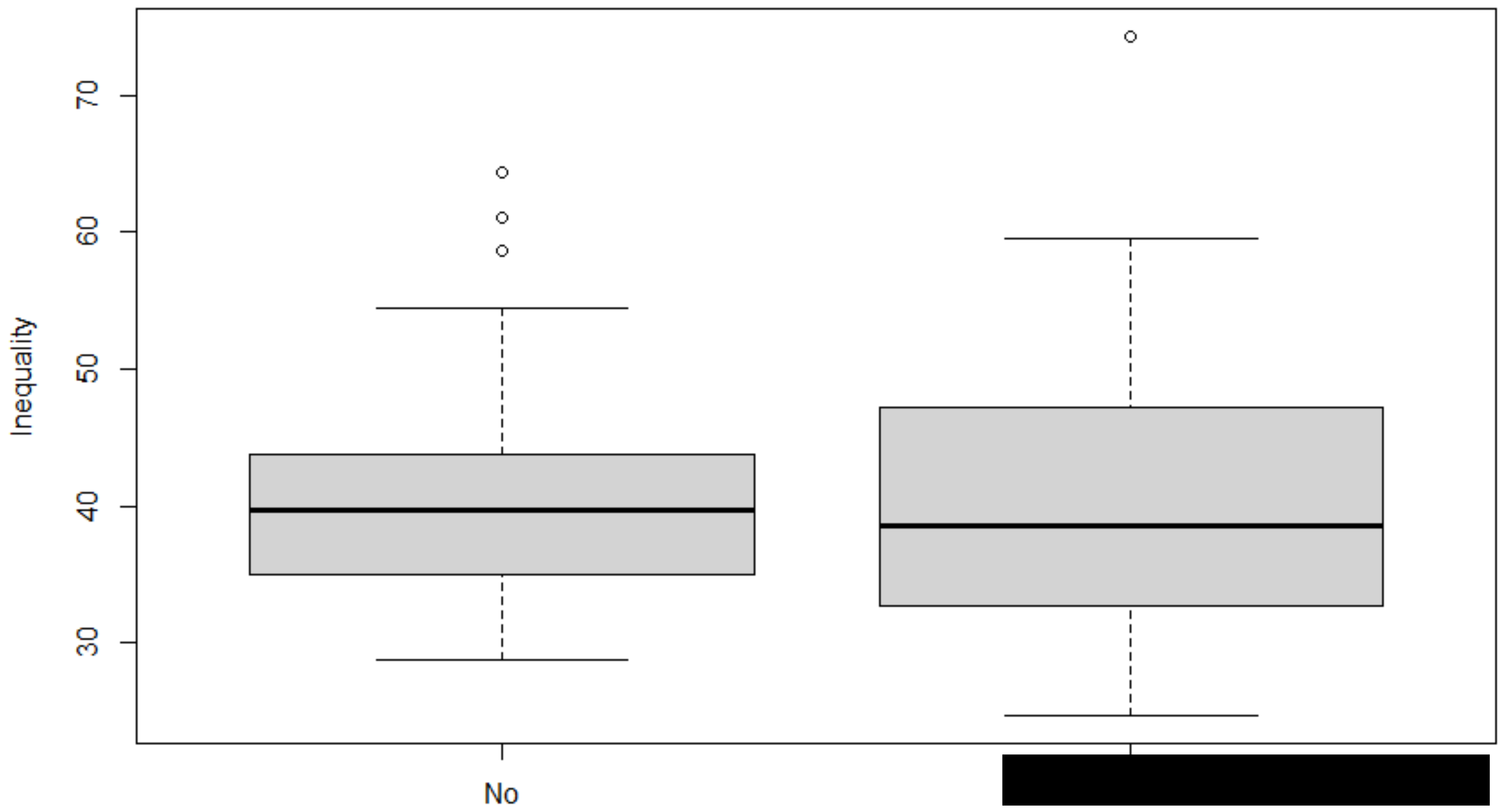
- P>t = 0.33
  - Given the data, there is a 33% probability that this estimated β would occur by chance
- If we use the 95% confidence level, P>t would need to be less than 0.05
- WE CANNOT REJECT THE NULL HYPOTHESIS.

EXAMPLE 3

- *H_A*: In a comparison of countries, those that are democracies will have lower rates of inequality than those that are not democracies.
- *H_o* :There is no relationship between deomcracy and inequality.
- In economics, the Gini index is the most widely used representation of the income or wealth distribution of a nation's residents.

- Dataset: world
- Dependent variable: gini10
- Independent variable: democ




Inequality
70
60
50
40
30
No
[redacted]

WHY STATISTICAL ANALYSIS PROGRAMS ARE WONDERFUL!

```
> mod.3 = lm(gini10 ~ democ_regime, data=world)
> summary(mod.3)

Call:
lm(formula = gini10 ~ democ_regime, data = world)

Residuals:
    Min      1Q  Median      3Q     Max
-15.124  -6.824  -1.224   5.841  34.476

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      40.6942     1.3005  31.291   <2e-16 ***
democ_regimeYes  -0.8701     1.6303  -0.534    0.594
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.378 on 141 degrees of freedom
  (24 observations deleted due to missingness)
Multiple R-squared:  0.002016,  Adjusted R-squared:  -0.005062
F-statistic: 0.2848 on 1 and 141 DF,  p-value: 0.5944
```

α

p-value

β

Adjusted R^2

R^2

- Estimated $\beta = -0.8701$ = -0.87
- "For every 1 unit increase in democracy, inequality will decrease by -0.87."
  - Because "democ_regime" is a "dummy" variable indicating whether a country is a democracy or not, this would indicate that democracies have a lower inequality rate (by -0.87) than non-domocracies.
- Estimated $\alpha = 40.6942$
  - "When democ_regime = 0, we expect that inequality (the gini coefficient) would be 40.69."
- P>t = 0.594
  - Given the data, there is a 59.4% probability that this estimated β would occur by chance.
- Adjusted R2 = -0.005062 = -0.01
  - "The independent variable explains negative 1% of the variation in the dependent variable"

```
> mod.3 = lm(gini10 ~ democ_regime, data=world)
> summary(mod.3)

Call:
lm(formula = gini10 ~ democ_regime, data = world)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.124  -6.824  -1.224   5.841  34.476 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      40.6942     1.3005  31.291   <2e-16 ***
democ_regimeYes  -0.8701     1.6303  -0.534    0.594    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.378 on 141 degrees of freedom
  (24 observations deleted due to missingness)
Multiple R-squared:  0.002016,	Adjusted R-squared:  -0.005062 
F-statistic: 0.2848 on 1 and 141 DF,  p-value: 0.5944
```

REJECT THE NULL HYPOTHSIS?

- P>t = 0.59
  - Given the data, there is a 59.4 % probability that this estimated β would occur by chance.
- If we use the 95% confidence level, P>t would need to be less than .05
- WE CANNOT REJECT THE NULL HYPOTHESIS!

DOES CRIME CAUSE INEQUALITY?

- Still too soon to impute causality to a relationship we observe…. We would need to control for other possible variables that explain our independent variable.
- THIS is when all that theorizing becomes VERY important!