POLS201 Spring 2019

Introduction to Linear Regression

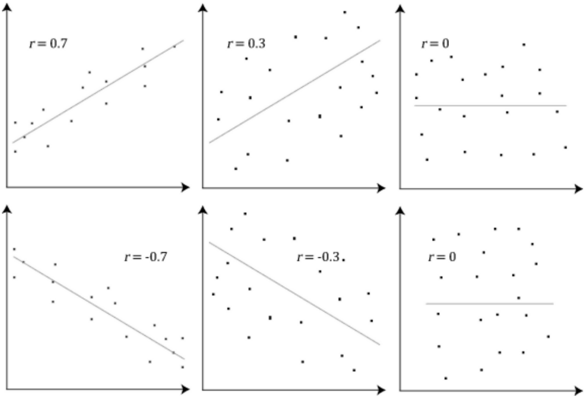
March 18

Agenda (On Moodle if you want to call this up)

- We're going to dive into Linear Regression quickly
- Bear in mind:
  - t distribution and t-tests (for diffs in sample means)
  - the idea of statistical significance: very unlikely that we're seeing a value of zero.
  - we can guage statistical significance in a couple of ways: a big absolute t-value or a small p-value.
  - Statistical significance doesn't show causal effect nor strength of the relationship.

Recall linear correlation

- as measured by the R score?



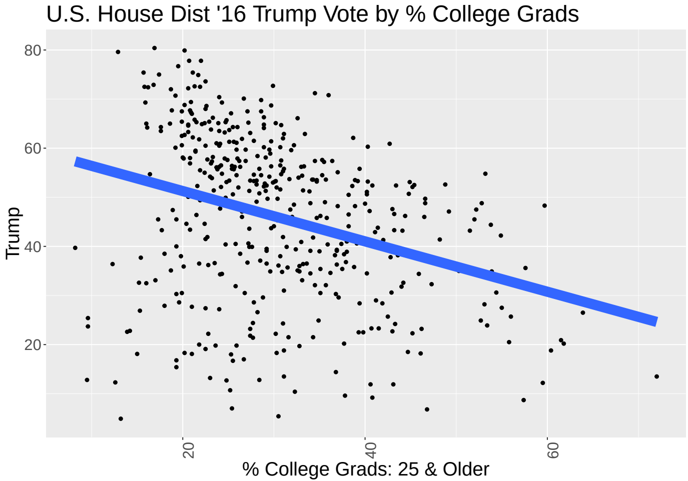
Linear Regression

- Regression is a natural extension of correlational analyses.
- A two-variable linear regression fits the best line on a scatterplot between two variables. - Y axis: DV - X axis: IV

A Line of Best Fit

POLS201
Spring 2019

- Knowing the equation for this line will be more valuable than just knowing the correlation. Why?



You might recall from algebra: $y = mx + b$ where:

- $y =$ the vertical axis
- $m =$ the slope of the line $(\Delta y \div \Delta x)$
- $x =$ the horizontal axis
- $b =$ the intercept: the value of y when $x = 0$

The Equation for a Line

- Our usage is identical with slightly different syntax

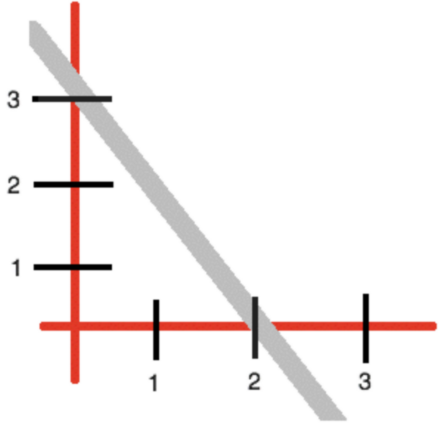
$$Y = mx + b \rightarrow \boxed{y = \propto + \beta x}$$

$\propto$: "alpha" - constant term

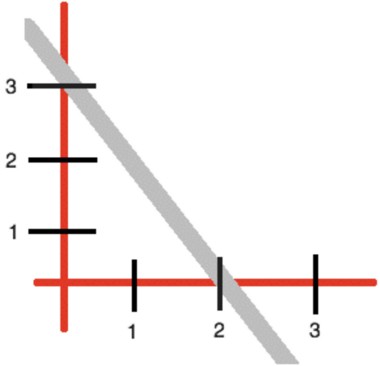
β: "beta" - slope term

y: your dependent variable

What is the equation of this line?



What is the equation of this line?



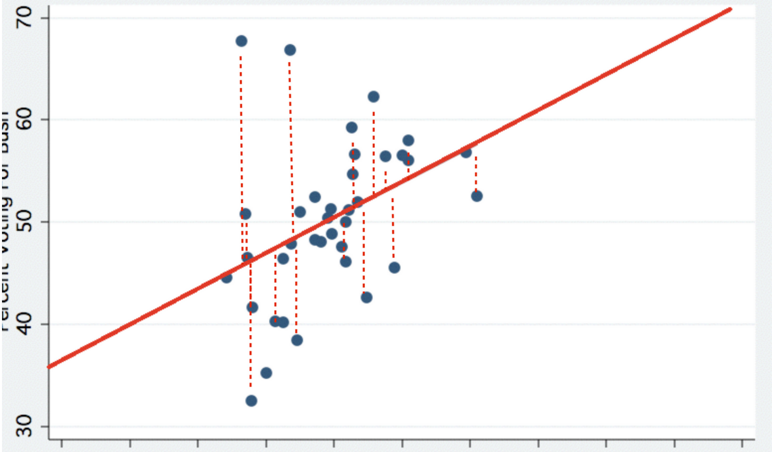
y=3-1.5x

Linear Regression

- A line through scatter plot has an imperfect fit:
- $y_i = \alpha + \beta x_i + \epsilon_i$
- ϵ_i is an error term
- Linear regression fits a line that *minimizes* those errors
- Specifically, a line that *minimizes* the total sum of the *squared* errors.

We call this type of regression Ordinary Least Squares

- Figure out the line that minimizes the sum of squared *errors*
- An error is the difference between the estimate and observed value.

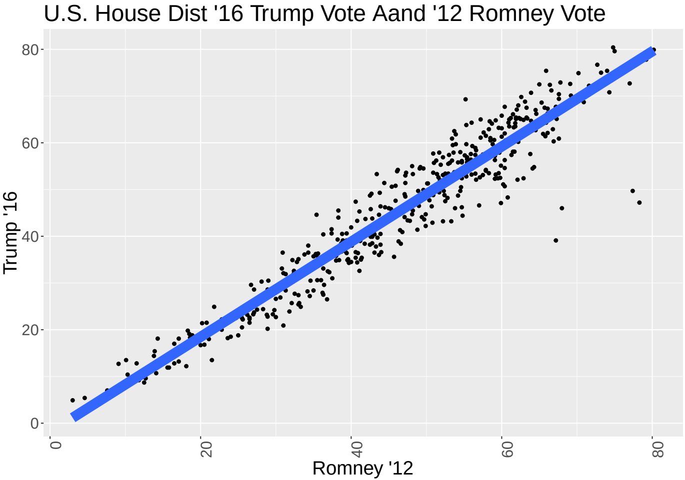


Meaning of the Slope and the Intercept

- The Intercept is the estimated value of y when $x = 0$
  - Sometime this value has meaning, but often it does not
- We will call this our "constant" term
- Meaning of slope: One unit increase in x is associated with β unit increase in the predicted value of y.
- We will refer to β terms as "coefficients"

Guess the equation for this blue line?

POLS201
Spring 2019



Actually, R will tell us


```
> simple_regression <- lm(Trump ~ Romney, data=meas_ex)
> summary(simple_regression)

Call:
lm(formula = Trump ~ Romney, data = meas_ex)

Residuals:
     Min       1Q   Median       3Q      Max
-30.6331  -2.9301   0.1322   3.1147  14.9876

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.89314    0.77123  -2.455   0.0145 *
Romney       1.01822    0.01559  65.311   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.105 on 433 degrees of freedom
Multiple R-squared:  0.9078,    Adjusted R-squared:  0.9076
F-statistic:  4265 on 1 and 433 DF,  p-value: < 2.2e-16
```

Say what? What did R tell us?

- We "ran a regression" using R's lm function (lm stands for "linear model")
- We created an object with all kinds of information
- Let's focus the most important pieces: the slope estimate, t-test of our estimate, and R-squared

What line did we estimate?

- The slope coefficient is very close to 1. Why is that?

```
> simple_regression <- lm(Trump ~ Romney, data=meas_ex)
> summary(simple_regression)

Call:
lm(formula = Trump ~ Romney, data = meas_ex)

Residuals:
     Min       1Q   Median       3Q      Max
-30.6331  -2.9301   0.1322   3.1147  14.9876

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.89314    0.77123  -2.455   0.0145 *
Romney       1.01822    0.01559  65.311   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.105 on 433 degrees of freedom
Multiple R-squared:  0.9078,	Adjusted R-squared:  0.9076
F-statistic:  4265 on 1 and 433 DF,  p-value: < 2.2e-16
```

DV ~ IV

We generate regression estimates in R with the lm() function

- The basic format: regression <- lm(dv ~ iv [+ more iv's], data = dataframe)
- lm() creates an object that we can retrieve and summarize (in part or in whole)
- We retrieve a summary with: summary(regression_object_name)
- E.g. : pope_regression <- lm(pope_approval ~ religion, data = Pew)
- The view results with summary(pope_regression)

The estimate of our line is:

- y_i = -1.89 + 1.02x_i + $\hat{\epsilon}_i$
- Note that our estimate of y_i, $\hat{y}_i$, differs from y_i by $\hat{\epsilon}_i$

```
> simple_regression <- lm(Trump ~ Romney, data=meas_ex)
> summary(simple_regression)

Call:
lm(formula = Trump ~ Romney, data = meas_ex)

Residuals:
     Min       1Q   Median       3Q      Max 
-30.6331  -2.9301   0.1322   3.1147  14.9876 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.89314    0.77123  -2.455   0.0145 *  
Romney       1.01822    0.01559  65.311   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.105 on 433 degrees of freedom
Multiple R-squared:  0.9078,	Adjusted R-squared:  0.9076 
F-statistic:  4265 on 1 and 433 DF,  p-value: < 2.2e-16
```

DV ~ IV

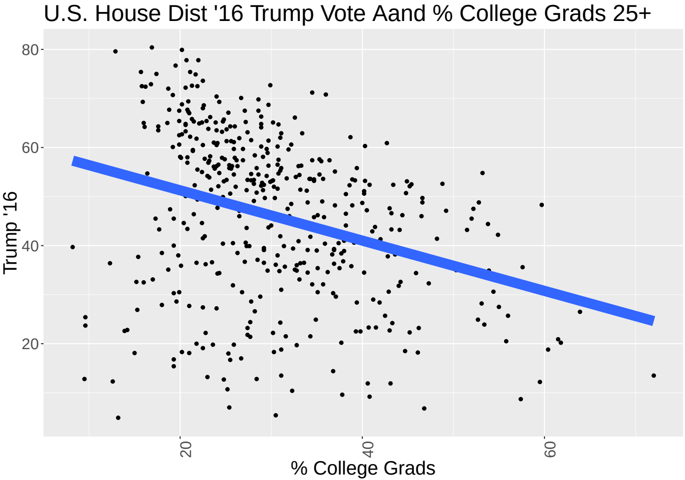
The t value and p value tell us:

- There is a virtually certain relationship between Romney vote and Trump vote

```
> simple_regression <- lm(Trump ~ Romney, data=meas_ex)
> summary(simple_regression)

Call:
lm(formula = Trump ~ Romney, data = meas_ex)

Residuals:
     Min       1Q   Median       3Q      Max
-30.6331  -2.9301   0.1322   3.1147  14.9876

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.89314    0.77123  -2.455   0.0145 *
Romney       1.01822    0.01559  65.311   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.105 on 433 degrees of freedom
Multiple R-squared:  0.9078,    Adjusted R-squared:  0.9076
F-statistic:  4265 on 1 and 433 DF,  p-value: < 2.2e-16
```

DV ~ IV

Guess the equation for this blue line?



What did we estimate?

POLS201
Spring 2019

```
Call:
lm(formula = Trump ~ coll_grad, data = meas_ex)

Residuals:
    Min      1Q  Median      3Q     Max 
-49.856  -9.799   3.921  11.795  28.732 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 61.52163    2.36999  25.959  < 2e-16 ***
coll_grad   -0.51257    0.07348  -6.976 1.14e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.94 on 433 degrees of freedom
Multiple R-squared:  0.101,	Adjusted R-squared:  0.09896 
F-statistic: 48.66 on 1 and 433 DF,  p-value: 1.14e-11
```

The estimate of our line is:

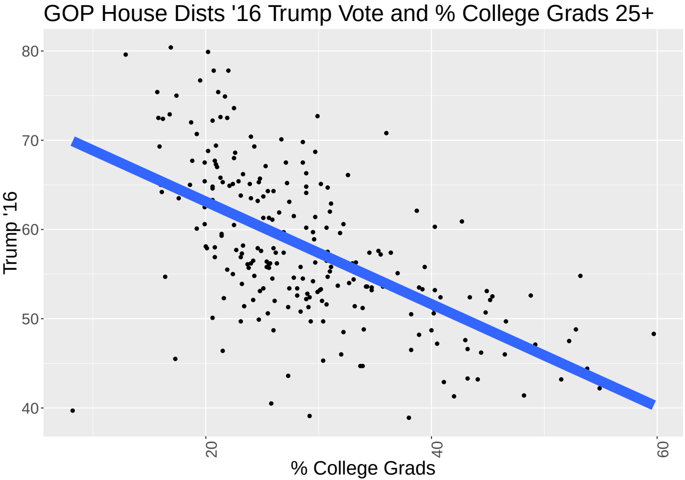
- $y_i = 61.52 - 0.51x_i + \hat{\epsilon}_i$

```
Call:
lm(formula = Trump ~ coll_grad, data = meas_ex)

Residuals:
    Min      1Q  Median      3Q     Max 
-49.856  -9.799   3.921  11.795  28.732 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 61.52163    2.36999  25.959  < 2e-16 ***
coll_grad   -0.51257    0.07348  -6.976 1.14e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.94 on 433 degrees of freedom
Multiple R-squared:  0.101,	Adjusted R-squared:  0.09896 
F-statistic: 48.66 on 1 and 433 DF,  p-value: 1.144e-11
```

The t value and p value tell us:

- There is a virtually certain relationship between % Coll grads and Trump vote
- But there is reason to wonder if this relationship is particularly strong
- The visual evidence in the scatter plot is one clue

Suppose we separately analyze districts that elected a GOP member vs. Dem districts.



Spoiler Alert: The Effect in GOP Districts is Apparent


```
Call:
lm(formula = Trump ~ coll_grad, data = rdist)

Residuals:
     Min       1Q   Median       3Q      Max
-30.2040  -4.2100  -0.2025   4.2848  16.8915

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 74.61591    1.56402   47.71   <2e-16 ***
coll_grad   -0.57463    0.05188  -11.07   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.054 on 235 degrees of freedom
Multiple R-squared:  0.343,     Adjusted R-squared:  0.3402
F-statistic: 122.7 on 1 and 235 DF,  p-value: < 2.2e-16
```

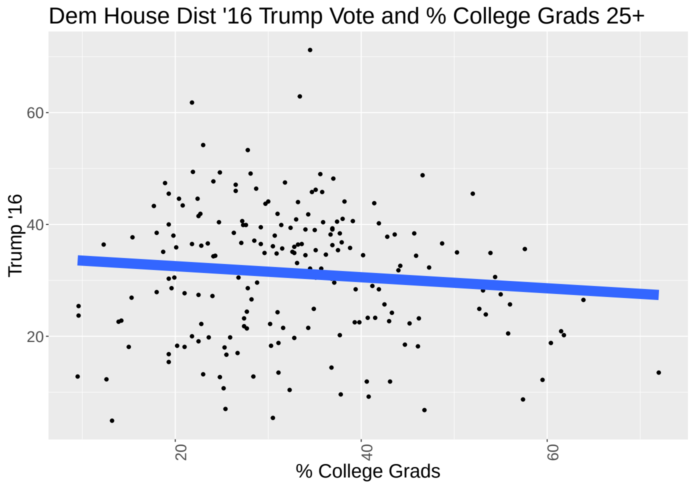
The Most Important Null Hypothesis in a Regression

- That the coefficient (slope) of a variable is zero
- Do we learn enough to reject that hypothesis?

Note the R-Squared

- R-squared, is a number that indicates how well data fit a statistical model – sometimes simply a line or a curve.
- An R-squared of 1 indicates that the regression line perfectly fits the data, while an R- squared of 0 indicates that the line does not fit the data at all
- For our GOP dists: The R-Square is 0.34
- For Dem dists: 0.01

Notice the very flat line. Can we say it isn't zero?



Here, we do not see enough to reject the null:


```
Call:
lm(formula = Trump ~ coll_grad, data = ddist)

Residuals:
    Min      1Q  Median      3Q     Max 
-28.309  -8.420   1.749   8.064  40.099 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 34.51569    2.47405  13.951   <2e-16 ***
coll_grad   -0.09896    0.07148  -1.384    0.168    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.76 on 196 degrees of freedom
Multiple R-squared:  0.009685,	Adjusted R-squared:  0.004632 
F-statistic: 1.917 on 1 and 196 DF,  p-value: 0.1678
```

If we can reject the null, we have a bit of support . . .

- . . . for a claim that the IV drives the DV
- "If there is smoke, there may be fire. . . "
- "But maybe not"
  - Regression estimates rely on certain assumptions:
  - A consistent normal distribution of residuals with a mean of zero.
  - A reasonably robust r-square
  - It's no evidence of a causal connection. Confounders lurk everywher
  - And we cannot just pile on an infinite number of controls
  - Finally: are we dealing with continuous variables? If not, can we proceed?
- Stay tuned. . . .