POLS201 Spring 2019

Introduction to Linear Regression: Part II

March 20

Recap

- Friday: I will be here to answer questions, but regular class is cancelled.
- Spring? Consider it sprung.
- Come if you want or work on your assignment.
- Plan on meeting with me after the break for at least 15-30 minutes. I will set up a schedule.

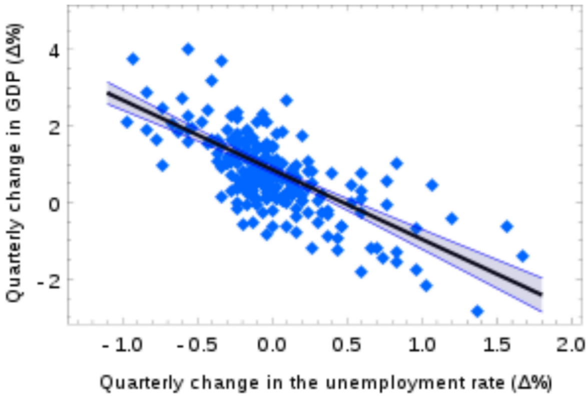
Thoughts about Descriptive Stats for April 5

- Does the number of observations seem right?
- Does the range of your values seem right?
- Are missing values treated as missing?
- Does the measurement of your data match your theoretical needs?
- Do you have categorical variables that need to be further "dummied out?"

Notice there are 4 ways to tell if β is significant

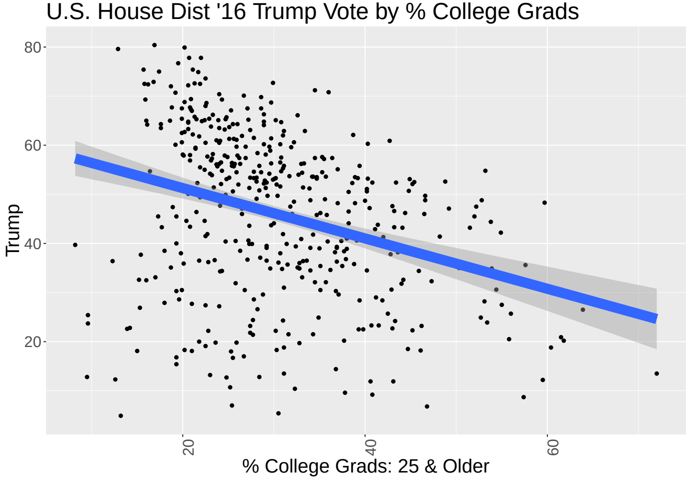
1 P-Value Less than 0.05
2 T-Statistic higher than $|1.96|$ (in large sample)
3 Zero not in the 95% confidence interval for the coefficient
4 The coefficient plus or minus 1.96 times the standard error does not cross zero (large sample)

You can generate graphs with coefficient standard errors



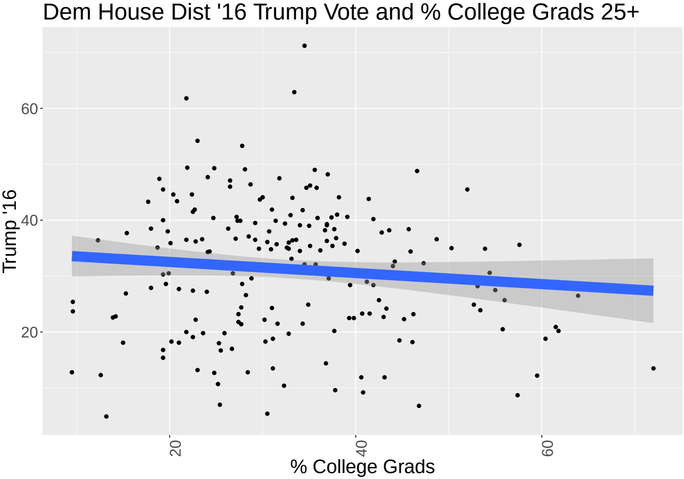
And this familiar graph with standard errors

-. . . of the slope estimate, that is



And contrast with the graph of Dem Districts only

- The overlap across the line suggests we can't reject the null.



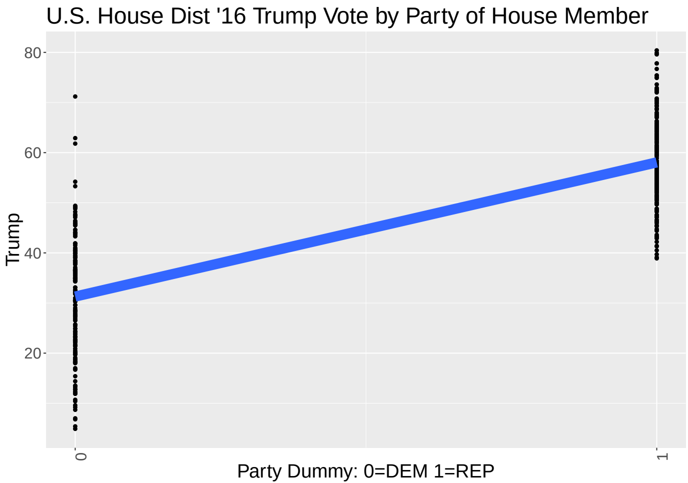
The Basic Interpretation of a Regression

- The β coefficient" or "coefficient on your IV" represents the slope of a line
- Which means we predict:
- A one unit increase in x leads to an β unit increase in y
- The formula $\hat{y}_i = \alpha + \beta x_i$ gives our *predicted* value
- The hat or caret over the y_i means it's estimated.
- The actual value of $y_i - \hat{y}_i = \epsilon_i$, aka the residual

Regression with Dichotomous Independent Variables

- "Dichotomous" or "dummy" variables are very common in OLS ("Ordinary Least Squares")
- We customarily code a dummy variable as zeros and ones
- This allows means we interpret a coefficient as the marginal effect of moving from a value of "0" to "1"
- The constant term will be the predicted value when $x_i = 0$

The scatterplot of an independent dummy variable is worthless



Can you find? WRITE ON THE WORKSHEET

- The marginal effect of a Republican district?
- Prediction for Democratic districts?
- Prediction for Republican districts
- On the worksheet!

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

Residuals:
    Min      1Q  Median      3Q     Max 
-26.392  -6.873  -0.454   6.996  39.908 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  31.2919     0.7260   43.10   <2e-16 ***
Party_dum    26.7617     0.9835   27.21   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.22 on 433 degrees of freedom
Multiple R-squared:  0.631,	Adjusted R-squared:  0.6301 
F-statistic: 740.4 on 1 and 433 DF,  p-value: < 2.2e-16
```

That was fun. Now let's go multivariate

- A regression with one independent variable is rare and uninteresting
- We can add more variables and see more effects

The Joy of Going Multivariate

- With more variables can see the effect of confounds and other issues of endogeneity.
- Multivariate regression is a helpful though imperfect solution.

What is a control variable?

- A variable that is held constant, i.e. whose effect is isolated in the equation.
- We can isolate them and see if the result of the IV persists
- In a sense: Control variables allow you to make your "treatment" and "control" groups (or comparisons between high and low values of IV)
- We can improve the quality of our β estimate

What do we choose to control?

- Confounds
- Other variables known to affect the DV
- But try to avoid:
  - Intervening Variables
  - Multiple measures of the same variable that are highly correlated with each other

Example: Regression Estimates for Voter Turnout

- Interpretation of coefficients remains the same... but now add "ALL ELSE EQUAL"

Table 2. Determinants of Voter Turnout in Legislative Elections in 51 selected Latin American and European countries, 2004-2008

Independent Variables	Model 5	Model 6	Model 7	Model 8
Proportional Representation	6.907 (10.343)	6.907 (10.343)	13.937 (7.903)	10.583 (10.197)
Freedom House Score	* 7.174 (3.423)	* 7.174 (3.423)	** 7.476 (2.520)	** 10.125 (3.045)
Compulsory Voting	10.588 (5.640)	10.588 (5.640)	* 9.540 (4.693)	8.568 (4.994)
Latin America	-1.641 (7.296)	–	–	–
Europe	–	1.641 (7.296)	–	–
New Democracy	–	–	7.949 (4.867)	–
Spoiled Votes	–	–	–	0.585 (0.352)
Constant	9.923	8.282	0.171	-15.210
Adjusted R-Squared	0.173	0.173	0.206	0.205
Number of Observations	43	43	45	40

Note: Coefficients reflect percentage change in voter turnout; $*p < 0.05$, $**p < 0.01$

Notice something powerful for your own work

- You can mix up your model with variations.
- Add, subtract, and change your variables.
- Papers typically focus on one or two models but with variations as the questions demand.
- Doing that will give you more to report.

Trump Vote and % of College Grads by Cong District

- Forget about ecological inference problems. What else is wrong with this research design?
- What are the possible confounds?
- Think of a confound this way: if this were an experiment, what might interfere with random assignment to the "treatment" group?
- One is certainly partisanship
- Another reliable, persistent confounder in U.S. politics is ____________?
    - Really, everything in U.S. politics starts here.

We started with this model

```
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
```

Let's add Percent White as a second IV (or a control)

- Notice the increase in the R-Square

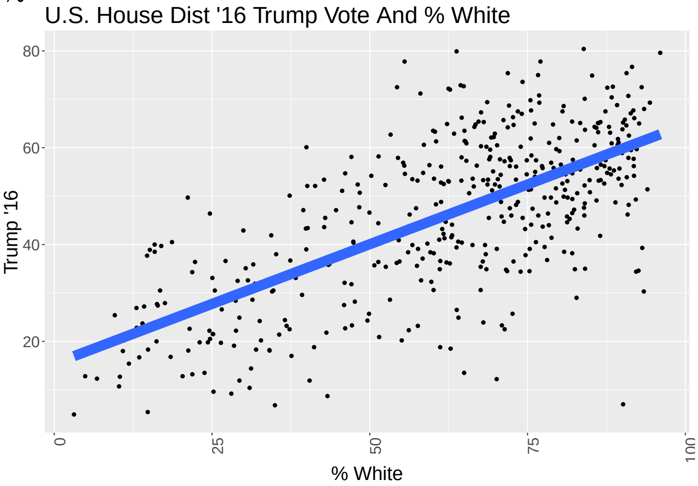
```
Call:
lm(formula = Trump ~ coll_grad + Pct_White, data = meas_ex)

Residuals:
    Min      1Q  Median      3Q     Max 
-56.512  -7.706  -0.028   7.245  29.615 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 32.35781    2.07046   15.63   <2e-16 ***
coll_grad   -0.59645    0.05012  -11.90   <2e-16 ***
Pct_White    0.51369    0.02288   22.45   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.85 on 432 degrees of freedom
Multiple R-squared:  0.585,	Adjusted R-squared:  0.5831 
F-statistic: 304.5 on 2 and 432 DF,  p-value: < 2.2e-16
```

The R-Squared is much improved: from .09 to .58

- But importantly: the independent effect of College Grad Percent persists too

Unsurprisingly the effect of race matters

- But (perhaps surprisingly) it doesn't confound college grad %



If run a simple regression with net worth, we see significance


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

Residuals:
    Min      1Q  Median      3Q     Max 
-39.043 -10.991   1.155  11.559  41.050 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 29.02499    2.00888   14.45   <2e-16 ***
Net_Worth    0.11497    0.01275    9.02   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.43 on 433 degrees of freedom
Multiple R-squared:  0.1582,	Adjusted R-squared:  0.1562 
F-statistic: 81.37 on 1 and 433 DF,  p-value: < 2.2e-16
```

But we can also see that race confounds the effect of net worth ON THE WORKSHEET

- In a sense, net worth "drops out" of the model when we add race
- Which causal story seems persuasive? What can we explain best?

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

Residuals:
    Min      1Q  Median      3Q     Max 
-55.717  -7.845   0.041   7.465  29.896 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 31.809642   2.176620  14.614   <2e-16 ***
coll_grad   -0.595630   0.050148 -11.878   <2e-16 ***
Pct_White    0.501244   0.027471  18.246   <2e-16 ***
Net_Worth    0.008815   0.010758   0.819    0.413    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.85 on 431 degrees of freedom
Multiple R-squared:  0.5857,	Adjusted R-squared:  0.5828 
F-statistic: 203.1 on 3 and 431 DF,  p-value: < 2.2e-16
```

We could keep adding one IV after another

- But at some point, if many variables all correlate, their power to explain the DV is lost
- The formal name for this problem is multi-collinearity
- What if we add party to the model?

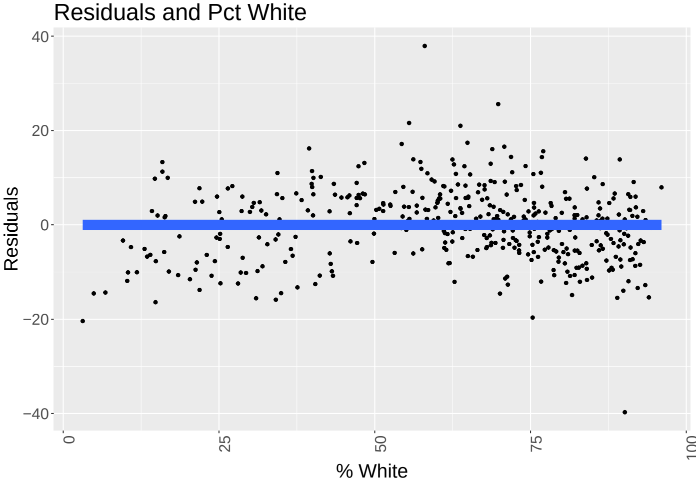
This is starting to look pretty good

- Except. . . I have a tiny reverse causality problem

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

Residuals:
    Min      1Q  Median      3Q     Max 
-39.722  -5.118   0.478   4.907  37.920 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 29.71810    1.50782   19.71   <2e-16 ***
coll_grad   -0.40675    0.03760  -10.82   <2e-16 ***
Pct_White    0.30325    0.01973   15.37   <2e-16 ***
Party_dum   18.06829    0.91492   19.75   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.867 on 431 degrees of freedom
Multiple R-squared:  0.7822,    Adjusted R-squared:  0.7806 
F-statistic: 515.8 on 3 and 431 DF,  p-value: < 2.2e-16
```

Are the residuals correlated with the IV's

- If not, we may have dealt with endogeneity successfully
- At least one of them isn't: Pct_White



But this model? It seems to "gild the lily"

- Here, we have arguably added some statistical junk. But Sharkansky's mother would be proud.

Variable	**Donald Trump's GOP Share of Vote**		
STATE POLITICAL CULTURE			
Moralistic Subculture	-17.430 (3.728)***		-1
Sharkansky's Typology		2.379 (0.842)**	
STATE PARTISANSHIP			
2012 Obama Vote	0.589 (0.218)	0.868 (0.279)	
STATE DEMOGRAPHICS			
% White	0.217 (0.137)	0.267 (0.192)	
Per Capita Income	0.195 (0.314)	.350 (0.379)	
% Urban	0.103 (0.104)	0.275 (0.116)	-
% Aged 65+	1.071 (1.065)	0.743 (1.269)	
% College Graduate	-0.500 (0.521)	-0.906 (0.618)	
Constant	-18.967 (18.037)	-58.549 (28.396)*	1
R^2	.686	.569	
F	9.050***	5.460***	

A Final Note: Our Syntax changes slightly when we add multiple variables

- A multivariate model might be described as:
- $Y_i = \alpha_i + \beta 1 x_i + \beta 2 x_i ... \beta n x_i + \epsilon_i$
- Or simply
- $Y_i = \alpha_i + \beta X_i + \epsilon_i$ where big X captures all the independent variables. Sometimes we say "explanatory" variables.

Many simple research projects might boil down to this drill

- Maybe yours??
- Run a simple regression and then (carefully) add control variables
- Test to see if a few basic assumptions hold
  - Are the residuals random relative to the IV's
  - Do the controls that fit with your causal story change the significance of the coefficients
  - Report the results of the various tweaks

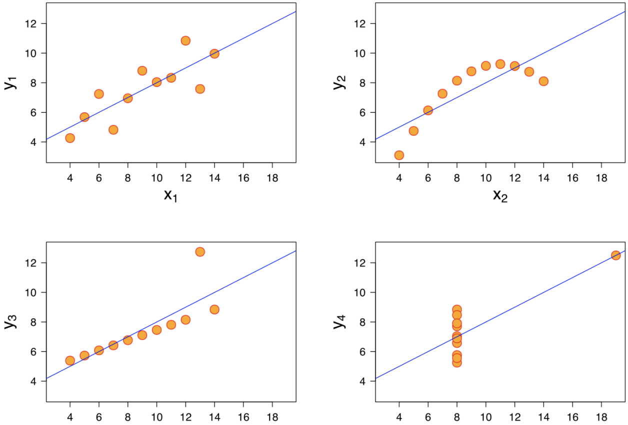
Naked Statistics c.12 lists seven big cautionary points about linear regression.

- Your data aren't linear. Remember Anscombe's Quartet?
- Correlation <> Causation. Whudda thunk?
- Multicollinearity can hide actual effects
- Reverse causality (endogeneity)
- Ommitted Variable Bias (endogeneity redux)
- Extrapolating beyond the data. Ecological inference anyone?
- Data mining: Too many variables

Anscombe's Quartet Redux

POLS201
Spring 2019

- The four correlations are identical. The northwest example might be appropriate for regression. The others? Neigh way, José.



And some final points to remember

- No one cares about predicting the past
  - An overly perfect model using past data is useless
- Wonder if your predictions miss, a lot, and not randomly
- If your DV is discrete (i.e, two or just a few values), you need to run a slightly different variation of the lm() function.
- Have you thought of all the confounders? No really, have you?
- Linear regression is most appropriate for continuous variables. What if the DV isn't continuous> We have some tricks up our sleeve. Stay tuned.