POLS201 Spring 2019

The Descriptive Statistics Assignment and Regression Topics

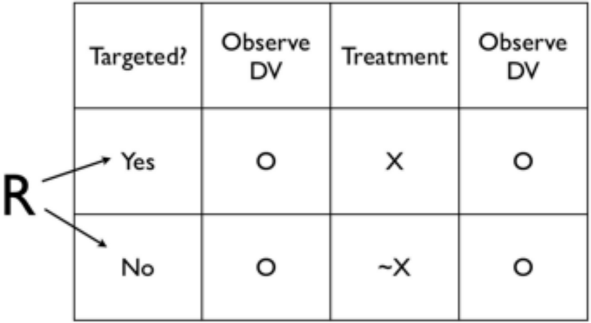
April 5

Agenda

- Correction to Question #9 on Homework
- Finishing the Descriptive Statistics Assignment
- Some risks of using regressions
    - Not typically relevant for your paper, but note for the final
- Brief Mention of the Cherie Berry paper

Refresh on Omitted Variable Bias

- An omitted variable is correlated with an IV and partly determines the DV.
- It distorts the estimate of the IV coefficient.
- It does not mean "you neglected to include every variable that you could imagine".
- Jacobson thinks he fixed problem.
  - an R^2 approaching 1 is an indication.
  - capturing candidate spending before the AFL-CIO program is another.

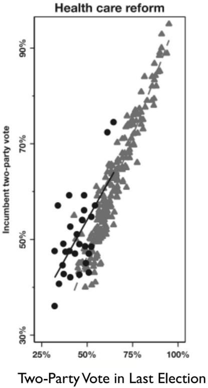


R

Targeted?	Observe DV	Treatment	Observe DV
Yes	O	X	O
No	O	~X	O

The simplicity of Jacobson's model is a virtue

- But beware:



- Lingering issues
  - An "Identification" problem: Is there enough variation on the IVs? It would be best if there were un-targeted candidates with similar district/incumbent profiles.
  - A "linearity assumption" problem: Is accounting for a "linear" effect of our control sufficient?

Basic Regression Problems:

- Regression won't run:
    - If your number of variables is greater than your number of observations
    - If your X variable perfectly predicts your Y variable
- Regression will drop variables if:
    - If your X variable doesn't vary
    - If your X variable is identical or nearly identical to another X variable

The List of Basic Problems (even if the regression does run)

- Too Little Variation
- Outliers
- Too Many or Too Few Observations
- Collinear Variables
- Non-Linear Effects
- Error-Term Issues: Residuals vary in range or are systematic

Outliers

- Outliers are data points that take on extreme values (high or low) of either your IV or DV.
- The slope of your line (the marginal effect) can be heavily influenced by outliers.
    - Especially if you don't have a lot of observations

Multicollinearity

- Beware of putting multiple variables that are highly correlated into the same regression.
- Practical effect: Null results or even false results.
  - Increases risk of false negatives
- Don't lard your model with highly correlated IV's

Too many / Too few number of observations

- Too Many Observations (~10,000 plus):
  - Easy to achieve significance
  - But all you have done is explain the data, not the process
  - Machine learning uses large observations much more effectively
- Too Few Observations (~less than 30): -Too hard to achieve significance
- But also remember: you need to work with the data you have, not the data you wish you had :)

The Problem We Will Consider Next Week:

- Binary or Categorical **Dependent** Variables
  - We considered binary independent variables buy binary dependent variables pose a different and more serious problem.
- We solve that problem by transforming the depending variable into an expression of its probability that it is one or zero.
- We use a technique called logit regression (or its close cousin probit regression)

Finishing the Descriptive Statistics Assignment

- There are a few certainties in this life.

Finishing the Descriptive Statistics Assignment

- There are a few certainties in this life.
- Death.

Finishing the Descriptive Statistics Assignment

- There are a few certainties in this life.
- Death.
- Taxes.

Finishing the Descriptive Statistics Assignment

- There are a few certainties in this life.
- Death.
- Taxes.
- The sun rises in the. . . thinking. . .

Finishing the Descriptive Statistics Assignment

- There are a few certainties in this life.
- Death.
- Taxes.
- The sun rises in the. . . thinking. . .
- . . . east.

Finishing the Descriptive Statistics Assignment

- There are a few certainties in this life.
- Death.
- Taxes.
- The sun rises in the. . . thinking. . .
- . . . east.
- You will encounter frustration when you wrangle with data in a statistics package.

Finishing the Descriptive Statistics Assignment

- There are a few certainties in this life.
- Death.
- Taxes.
- The sun rises in the. . . thinking. . .
- . . . east.
- You will encounter frustration when you wrangle with data in a statistics package.
- Don't be alarmed. Every problem you are encountering at this stage has been met by students before you, and all of them are solvable.

The assignment for today:

- Write a short report about your data that can eventually form the "data" section of your final paper.
- Any political science research paper describes its data apart from the methdology or analysis results. That's your template.
- Key considerations: where did you get it? What does it represent? What do the values mean? How much does it vary? How is it expressed in terms of you unit analysis? What choices did you make? What choices would you prefer to have made?

“My Data don’t load”

- The word “data”" is presumably plural, by the way.
- To load data into R (or any place else), you need some kind of raw data file.
- A pdf generally does not take this form. Usually you will see file with an extension of .csv, .txt, .dta, .spss, .sav, .xls, or similar.
- Sometimes, R Studio is smart enough to solve it for you. Click on a raw data file in the R Studio file list and it will ask if you want to import it.
- But it (or you) will generate commands that you should save in a script file.

The basic template for loading a data file

- Two libraries should do the trick: tidyverse and haven
- Use a read function. For instance, if you have .dta file generated in Stata, use read_dta. Example:

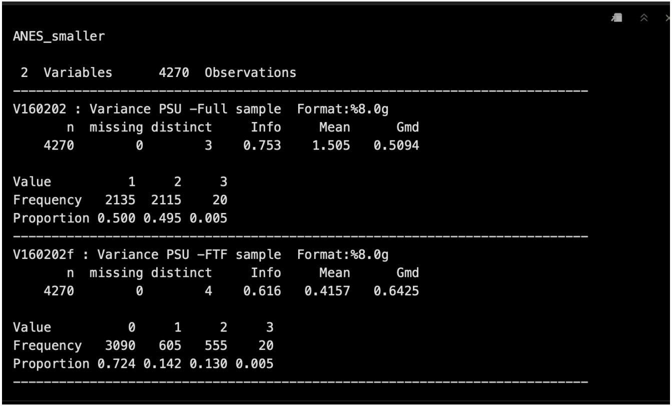
```
suppressMessages(library(tidyverse))
suppressMessages(library(haven))
ANES <- read_dta("anes_timeseries_2016.dta")
```

- ..and now I have an R data table called ANES
- I can use all the descriptive statistics functions on the handout
- Notice that I have 1836 variables (yikes!) I need to boil that down

```
ANES_smaller <- ANES %>% select(c("V160202", "V160202
```

Now I can run some functions that will give me descriptive information

```r
Hmisc::describe(ANES_smaller)
```



```
ANES_smaller

 2  Variables      4270  Observations
---------------------------------------------------------------------------
V160202 : Variance PSU -Full sample  Format:%8.0g
       n  missing distinct     Info     Mean      Gmd
    4270        0        3    0.753    1.505   0.5094

Value          1     2     3
Frequency   2135  2115    20
Proportion 0.500 0.495 0.005
---------------------------------------------------------------------------
V160202f : Variance PSU -FTF sample  Format:%8.0g
       n  missing distinct     Info     Mean      Gmd
    4270        0        4    0.616   0.4157   0.6425

Value          0     1     2     3
Frequency   3090   605   555    20
Proportion 0.724 0.142 0.130 0.005
---------------------------------------------------------------------------
```

A fail safe plan

- Notice that RStudio saves your history (look in the upper right corner)
- Once you have run a function successfully, you can save it forever
- Load your work into a project in RStudio Cloud and share it
- “Work” means: your script file and any data files you’re trying to load into R.

Quick note on the Cherie Berry paper

- Moral: You can find variation and natural experiments everywhere
- Some people saw Cherie Berry's face every day, twice a day, for years. I was one of them.
- Implication: Without campaigning, places with elevators gave her greater electoral success than places without them.
- The theory: subtle awareness of a government official is a kind of advertising.