Mini-project 2: Obama to Trump

S470/670

Upload a draft submission for questions 1 and 2 through the Assignments tab on Canvas by 11:59 pm, **Thursday 7th November**. Upload your final submission by 11:59 pm, **Tuesday 19th November**.

Work in groups of up to four. Email me your group name and the names and user IDs of your group members (one email per group) and I'll set them up on Canvas. If you're in a group of one to three people and you'd like to be paired with additional group members, tell me and I'll do some matchmaking.

Research question

To what extent do attitudes toward immigration explain the switching of votes of 2012 Obama supporters who became 2016 Trump supporters?

Polls have shown that people from certain demographic groups were more likely to switch their votes than others. But what might explain why some people within a group switched, while others didn't? One theory is that attitudes toward immigration became especially salient during the 2016 campaign. Most attempts I've seen to assess this have involved fitting big, complicated logistic regression models. This is useful but not really sufficient to study this problem, so YOU will explore the data and then fit a big and complicated logistic regression model. We won't attempt to assess cause-and-effect, but we can get a sense of whether attitudes toward immigration had explanatory power over and above demographic shifts.

Data

We'll use the 2016 Cooperative Congressional Election Study, a very large survey of a nationally representative sample of 64,600 adults. The investigators asked questions to the sample both before and after the election, although not all the pre-election respondents replied to the post-election survey. The data is available in various formats at `http://cces.gov.harvard.edu/data`. I've uploaded two files I pulled from there:

- `CCES16_Common_OUTPUT_Feb2018_VV.RData`: An R workspace containing a data frame called x, containing 64,600 observations on 563 variables.
- `CCES Guide 2016.pdf` : the codebook.

Here are the variables we'll focus on.

Technical variables:

- commonweight_vv_post: The survey weights for people who took the post-election survey.
- tookpost: Whether the respondent took the post-election survey. Limit your study to those for whom this is ”Yes.”

Demographic variables:

- gender: Male or Female.
- educ: Education (an ordered factor with six levels.)
- race: A factor with eight levels.
- pid7: Party identification (an ordered factor with seven levels from “Strong Democrat” to “Strong Republican.”) (One notable variable we omit is income, because the way it's coded in the CCES is hard to deal with.)

Voting variables:

- CC16_326: The respondent's vote in the 2012 Presidential election. Limit your study to those who voted for Barack Obama.
- CC16_410a: The respondent's vote in the 2016 Presidential election. “NA” could mean they didn't vote or that they didn't take the post-election survey. Do not limit your study to those who voted for Donald Trump; otherwise you won't be able to give probabilities.

Immigration variables:

Respondents were asked: “What do you think the U.S. government should do about immigration? Select all that apply.”

- CC16_331_1: Grant legal status to all illegal immigrants who have held jobs and paid taxes for at least 3 years, and not been convicted of any felony crimes. Here, “Yes” is a pro-immigration response.
- CC16_331_2: Increase the number of border patrols on the U.S.-Mexican border. Here, “No” is a pro-immigration response.
- CC16_331_3: Grant legal status to people who were brought to the US illegal as children, but who have graduated from a U.S. high school. Here, “Yes” is a pro-immigration response.
- CC16_331_7: Identify and deport illegal immigrants. Here, “No” is a pro-immigration response.

(Some respondents were given additional options, but we'll omit these.)

The full documentation of the variables and question wording is in the PDF.

Questions

1. Load the data in R. You'll need to pre-process your data before you can do anything. Create a data frame called `obama` that satisfies the following:

   - Only keep respondents who responded to the post-election survey.
   - Only keep respondents who voted for Obama in 2012.
   - Create a binary variable that indicates whether the respondent voted for Trump or not.
   - Create a *quantitative* variable that measures the respondent's attitude toward immigration using the four immigration variables described above. This variable should count the number of pro-immigration responses across the four items. The variable should then range from 0 to 4, with 4 being the most pro-immigration and 0 the least. (Make sure you add things the right way around.)
   - The sample sizes for some of the racial categories is small, so recode this factor to have four levels: "White", "Black", "Hispanic", and "Other." The `recode()` function in `dplyr` will be useful for this.
   - You might want to create numerical versions of the ordered categorical variables (party, education), though you can also do this later.

   After doing this, I got a data frame consisting of data for 23,395 individuals who voted for Obama in 2012, of whom 2,121 said they voted for Trump in 2016.

   Note: R code alone is sufficient for this question : it does not have to be part of your write-up.

2. All the demographic variables, as well immigration attitude, are meaningful predictors of vote switching from Obama to Trump. However, for each of the demographic categories, it could be that immigration attitude affects all groups in the same way, or it could affect different groups in different ways. (For example, if you compare white and black voters with the same attitudes toward immigration, it might be that these attitudes sway one group more than the other.) Using *weighted* logistic regression with Obama-to-Trump switching as a response or otherwise, fit models using immigration attitudes as a predictor along with each demographic variable in turn (e.g. immigration and race, immigration and party, etc.) In each case, consider whether you need an interaction. With which of the demographic variables does immigration attitude interact with? Carefully explain the substantive meaning of any large interaction effects you find.

   Note: For this question, I'd advise you to use immigration attitude as a quantitative predictor. You might want to recode some of the demographic variables as quantitative variables as well.

3. Fit TWO weighted logistic regression models to give the probability of an 2012 Obama voter switching to Trump in 2016: one without immigration attitude as a predictor, and one with immigration attitude as a predictor. Include interactions as necessary. State or display the coefficients of your models. Display the model probabilities for selected demographic groups. Compare the results of your models. Does including immigration attitudes make a substantive difference? Does it matter more for some demographic groups than others?

Instructions

Each should submit a set of answers (making sure that everybody's name is on the submission.) The submission should consist of TWO files: (i) a PDF with your write-up, and (ii) your code. (If you use any additional data, you should upload those two.) Zipped files are discouraged unless files are very large.

The initial submission should be for questions 1 and 2. If you want to get an attempt for question 2 done by then you can, or you can leave that until later.

The page limits for the final submission is 8 pages, including graphs. Any extraneous material should be put into an appendix that no one will look at.

Write for a broad audience. You should make a non-technical argument for people with limited knowledge of statistics, but also include enough technical detail to convince nerds that your results are features of the real world and not just of your method. Your argument should include graphs, numbers, and words.

Try not to start political fights with other members of your group, unless you're going to be working with different people on your final project.

Grading

Your initial submission won't be fully graded—we'll just check you're on the right track.

Grading for the final submission will follow the following:

- Question 1: zero points
- Question 2: 10 points
- Question 3: 10 points
- Communication: 10 points

Full credit for communication requires a readable, informative, comprehensive, clearly labeled set of graphs, and a comprehensible write-up with few glaring spelling and grammatical errors that makes the main points of the analysis clear.