Mini-project 2: New version

S470/670

Upload your draft project through the Assignments tab on Canvas by 11:59 pm, **Friday 3rd April**. Upload your final submission by 11:59 pm, **Monday 20th April**.

For both the initial and the final submission, work individually or in pairs. If you want to work in a pair, send me an email. If you work in a pair, it is STRONGLY RECOMMENDED that you meet virtually, not in person.

The thinktank Data for Progress collected survey data (in `DFP_WTHH_release.csv`) that represents the population of people registered to vote in the 2018 midterm elections.

We wish to study **swing voters**. Define the following groups:

- **Loyal Democrats:** People who voted for Hillary Clinton in 2016 and a Democratic House candidate in 2018.
- **Loyal Republicans:** People who voted for Donald Trump in 2016 and a Republican House candidate in 2018.
- **Swing voters:** All other people who voted in 2018.

In addition, define the following two subsets of swing voters:

- **Switch to D:** People who didn't vote for Hillary Clinton in 2016 but voted for a Democratic House candidate in 2018.
- **Switch to R:** People who didn't vote for Donald Trump in 2016 but voted for a Republican House candidate in 2018.

Note that some swing voters don't fall into either of these two groups, i.e. some people didn't vote for either a Republican or a Democratic House candidate in 2018.

Basic variables

- `presvote16post`: 2016 Presidential election vote. 1 means Clinton, 2 means Trump, 3–7 or NA means voted for someone else or didn't vote for President in 2016.
- `house3`: 2018 House of Representatives vote. 1 means Democrat, 2 means Republican, 3 means other.
- `weight_DFP`: survey weights. You should use these!

Issue variables

Respondents were asked to give their support for the following programs on a 1–5 scale, where 1 means strongly support and 5 means strongly oppose. (6 means "Not sure.")

- `M4A`: Medicare for All
- `GREENJOB`: A Green Jobs program
- `WEALTH`: A tax on wealth over $100 million
- `MARLEG`: Legalizing marijuana
- `ICE`: Defunding Immigration and Customs Enforcement
- `GUNS`: Gun control [1]

The codebook (`http://filesforprogress.org/datasets/wthh/WTHH_Core_and_DFP_modules.pdf`) contains full question wording for all issue variables.

Populism variables

Respondents were indicate their agreement with the following statements on a 1–5 scale, where 1 means strongly agree and 5 means strongly disagree. (6 means "Not sure.")

- `POP_1`: "It doesn't really matter who you vote for because the rich control both political parties."
- `POP_2`: "The system is stacked against people like me."
- `POP_3`: "I'd rather put my trust in the wisdom of ordinary people than in the opinions of experts and intellectuals."

Questions to answer

First you need to create three subsets of the data: Switch to D voters, Switch to R voters, and Swing Voters. The code in `DFP.Rmd` does this (creating data frames `switchD`, `switchR`, and `swingers`), or you might prefer to do it yourself. Then answer the following questions:

1. **How do Switch to D and Switch to R voters differ on the issue variables?**

   On which issue variables do Switch to D and Switch to R voters differ a lot? On which issue variables are they reasonably similar? Describe these differences.

2. **How do swing voters differ from loyal Democrats and loyal Republicans on the issue variables?**

   Some hypotheses might be:

[1] The response choices were slightly different for this question; see the codebook.

- Swing voters are moderates, and tend to the in the middle of the distribution when Democrats are on one side and Republicans are on the other.
- On most issues, swing voters are split, with some of them acting more like Democrats and others acting more like Republicans.
- Swing voters think more like Democrats on some issues and more like Republicans on other issues.
- Swing voters are ideologically incoherent and don't have consistent patterns in their issue positions.

Which of these hypotheses (or which mixture of them) fits the data best, and for which issues?

3. **What predicts being a swing voter?**

   Build two models to probabilistically predict whether a registered voter is a swing voter:

   - One model should use ONLY the issue variables as predictors.
   - The other model should use ONLY the populism variables as predictors.

   Clearly display both models, showing how the predicted probability changes with each predictor you include.

   How well do your models do? Which of your models does better? If you had to guess, what factors are most important in determining what makes a voter a swing voter?

   Note: You don't have to fit the best possible models, just sensible and interpretable ones. You should try to have some idea of how good your model is, although it may not be possible to improve classification error if nobody has a sufficiently high probability of being a swing voter.

Write a PDF report of no more than EIGHT pages, including graphs, addressing these questions. The body of the report should *not* include code — it should be readable to someone who has never used R (generally you should not just copy-paste output.) Additional graphs for model checking can be placed in an appendix that does not count toward the page limit and which probably no one will read.

What to submit

- A PDF or other file containing your report.
- A .Rmd or other file containing your code.
- Any other supplementary files required to reproduce your work.

Grading

- Question 1: 5 points; question 2: 5 points; question 3: 10 points.
- Communication: 10 points. Full credit for presentation 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.