Assignment 4

MACSS Ken Chen

Question 1:

(b) How many numbers did you call? How many people responded according to your Response variable? How many people did not respond according to your Response variable? What is your response rate?

I made all 200 calls, of which 3 people responded, while the other 197 people did not respond or refused to answer at all. The response rate is 1.5%.

(c) What fraction of those for whom Response = 1 answered the voting question? What fraction of those for whom Response = 1 answered the age question?

66.7% of the respondents answered the voting question and 66.7% of them answered the age question.

(d) What time of day was it in the area codes you called when you called them? What role did the time of day play in your response rate?

I did the survey in two rounds. The first round consists of 50 calls, which were made during *9:00pm - 9:30pm* eastern time, the second round depleted the remaining numbers during *12:00 - 12:45pm* eastern time. By calculation the response rate is higher during the night (1/50 = 2%) versus during daytime (2/150 = 13.3%). But since the volume of valid responses is very limited, we can't draw a solid conclusion whether time plays a role here.

(e) What is the median age of your respondents? How does that compare to the average age in the state of the phone numbers you called? What are some reason's why your sample median does or does not match the State data?

The median age of the respondents is 38, while the median age of Washington D.C. is 37.6 according to 2012-2016 American Community Survey 5 year estimates[^1]. The surveyed median age is a tiny bit higher, but we can be sure this estimate almost has no value:

- The sample size is too small (only two respondents answered the age question)
- Our sample representativeness is jeopardized by selection bias, since only those over 18 are able to vote, and we can't be sure people's willingness to answer survey questions are independent of age

(f) What percent of your respondents voted Republican (Trump) in the 2016 U.S. Presidential election? What percent of your respondents voted Democrat (Clinton)? How do those percentages compare to the actual voting percentages from the 2016 election? How might you test if the order in which you say the candidates or categories in the survey question influences the results?

Of the two respondents who answered the voting question, none voted for Republican(Trump) in 2016. One voted for Democrat(Clinton). In reality, of those who did vote, 54.4% voted for Democrat, 38.2% voted for Republican, 7.4% voted for others[^2]. To test whether the order plays a role, we can redo the survey on other random sizable samples, but state the names of candidates in different orders, and then check if the results are robust to change of orders, by comparing them pairwisely and with the real voting result.

---

[^1] U.S. Census Bureau, (2016). American Community Survey [2012-2016 American Community Survey 5-Year Estimates]. Retrieved from **https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=CF**

[^2] POLITICO (2016). 2016 Presidential Election Results. Retrieved from **https://www.politico.com/mapdata-2016/2016-election/results/map/president/**

Question 2:

Traditional polling on the phone has long been the dominant approach to predict election results. But due to general public's diminishing willingness to respond on phone, these practices are becoming less cost-effective and statistically less reliable. This paper enlightens on how we can make the best of the non-representative polls, via innovative statistical tools, namely multilevel regression and poststratification, to generate high-quality predictions.

The primary concern of non-representative polls is the sample selection bias. One example is taken from the Xbox online survey: eight dimensions of individuals' demographic characteristics are collected, of which **sex, age and education are least representative**, since they deviate most from the 2012 electorate data. With the same measure, **race, state and the 2008 vote are the most representative three.** For race and age, we expect Xbox platform population is dominated by young men, which aligns with the data: "18 to 29-year olds comprise 65% of the Xbox dataset, compared to 19% in the exit poll; and men make up 93% of the Xbox sample but only 47% of the electorate."(Wang et al., 2015, p.2) As for education, there are significantly fewer college graduates on Xbox. The discrepancy indicates that Xbox users are generally young and less educated, and this is possibly because collage graduates are more self-disciplined from games, or many game addicts are less likely to attend college.

To tackle the non-representativeness, the authors leveraged the Multi-regression and Poststratification (MRP) method to re-weight the respondents. The two data sources used are **Xbox online survey data** and **exit poll data from the 2008 presidential election**. The general methodology is to categorize the population into various cells, conditional on all possible combinations of their demographic and political characteristics. Solid estimation of response variable is computed in each cell by using the multi-regression method, and the aggregate level is derived from the population-weighted sum of estimations in all the cells. As Fig.5 demonstrates, the adjusted prediction matches well with the 2012 exit poll.

After the MRP adjustment, prediction quality has improved significantly. As we can see from Fig.2, only using the raw(unadjusted) data from Xbox will have predicted Romney to be the winner, since the two-party Obama support is constantly below 50%. On the other hand, Pollster.com would have said the result is uncertain, because the support for Obama and Romney are almost on par during the last three weeks. However, as Fig.3 presented, the MRP-adjusted data would have predicted Obama as the winner, since his support now is steadily above 50% over the last three weeks. It even outperforms the Pollster.com and almost mirrors the real outcome.

Reference

Wang, Wei, David Rothschild, Sharad Goel, and Andrew Gelman, "Forecasting Elections with Non-Representative Polls," *International Journal of Forecasting*, 2015, 31 (3), 980-991.