Assignment #4

MACS 30000, Dr. Evans

Due Wednesday, Oct. 31 at 11:30am

Ruixi Li

Problem 1

(a) See the attached PhoneSurvey.xlsx.

(b) I called 200 numbers. Only one people responded. 199 people did not respond. The response rate is 0. 5%.

(c) The people who responded answered the whole three questions, so the fraction of those for whom Response = 1 answered the voting question is 100% and the fraction of those for whom Response = 1 answered the age question is also 100%.

(d) I called in the evening, around 8:30 p.m. to 9:30 p.m. One of the people who picked up the phone said it was too late for him to do a survey, so he refused. Therefore, I believe if I called a little bit early I could have a higher response rate.

(e) Since only 1 people answered my age question, the median age of my respondent is 23. Since the average age in IL is 37.4[1], 23 is smaller compared

[1] 2016 Illinois Presidential Election Results. Retrieved from https://www.politico.com/2016-election/results/map/president/illinois/

to the average age in the state of the phone numbers I called. There are some reasons for the mismatch. First is that younger people tended to use cellphone more. When we called, it is more likely for younger people to pick up the phone than the older people. Secondly, the sample size is too small that can't represent the population.

(e) Since only 1 people answered my vote question, the percentage of respondent who voted Republican (Trump) in the 2016 U.S. Presidential election is 0. Since the percentage in IL is 39.4[2], 0 is smaller compared to the actual voting percentages from the 2016 election in the state of the phone numbers I called. A proposed way to check if the order matters is that we conduct another round of 200 numbers and doing exactly what we do in this round except for the different order when asking the voting. If we want to further explore the question, we should include a larger sample and randomly divided them into two groups. Then, we will repeat what we do in this research except for slightly difference in asking order in each group.

Problem 2

The paper demonstrated a new approach to forecast the election result with non-representative dataset. In this paper, the authors using the daily polls in 2012 presidential election on the Xbox gaming platform and adjusted it with multilevel regression and post stratification (MRP). They found that this adjusted non-representative data, Xbox post-stratified data, even forecasting more accurate than the representative polls which is Pollster.com forecast data.

---

[2] U.S. Census Bureau. Retrieved from https://factfinder.census.gov/faces/nav/jsf/pages/community_facts.xhtml

Firstly, let's look at the raw data. By comparing the demographic, partisan and 2008 vote distributions between the Xbox dataset and the adjusted exit poll in the 2012 electorate, from Fig. 1 we can see that the least three representative variables of the eight variable collected are sex, age and education, and the most three representative variables of the eight variable collected are race, state and 2008 vote. For the three least representative variables, the reason why the Xbox sample would be so different from the broader voting population is that Xbox users are mostly young males. "As one might expect, young men dominate the Xbox population: 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., 2005). As for the third variable education, younger people normally have lower average education than the voting population. First is that they are relatively young so they might not yet graduate from high school. Second is that those who vote tend to have a higher education level because the lower class might not care about the politics and turn out to vote.

In order to adjusted the non-representative data into a well-representative data, the authors used the MRP approach and two datasets, one is the Xbox dataset and the other is the exit poll data. Firstly, the authors divided the population into groups with different characteristics by all possible combinations of the variables in the Xbox dataset. Secondly, the authors calculated the weight of each cell by the proportion of the electorate in the corresponding subgroups. Thirdly, they obtained the weights from the exit poll data. Lastly, they applied the obtained weights to the Xbox dataset.

To verify the validity of the adjusted Xbox dataset with MRP approach, the authors compared the forecasting results of the three datasets, Xbox raw data, Pollster.com forecast data, and Xbox post-stratified data. For Xbox raw data, from Fig. 2 we can see that it predicted that Romney would have a landslide victory over Obama during the last three weeks before 2012 U.S. Presidential election. For the Pollster.com forecast data, since the poll kept moving up and down around the 50%, so it would predict that the outcome is undecided. For the Xbox post-stratified data, from Fig. 3 we see that it would predicts that Obama would win during the last three weeks before 2012 U.S. Presidential election because the support rate was always above 50%.

References

2016 Illinois Presidential Election Results. Retrieved from https://www.politico.com/2016-election/results/map/president/illinois/

U.S. Census Bureau. Retrieved from https://factfinder.census.gov/faces/nav/jsf/pages/community_facts.xhtml