MACS 30000 Assignment 4

Shanglun Li

October 29, 2018

Problem 1.

(a) See `PhoneSurvey_filled.xlsx`

(b) I called all the 200 numbers. Around 10 people picked up the phone, but only 2 people answered the research questions. Thus, the number of people with $Response = 1$ is 2. Therefore, there are 198 people did not respond. The corresponding response rate is $2/200 = 1\%$.

(c) All of the two people whom $Response = 1$ answered the voting question (100%). And, all the two people answered the age question (100%).

(d) My area code is 248, where the time zone is Eastern. I called some of the numbers around 4pm E.D.T. on Wednesday, and the rest of them around 7pm E.D.T on Wednesday. I think the time to call the number will affect the response rate. For example, in my case, I called some of the numbers around 4pm, which is the working hour for most people, which means many people will not pick up the phone. In addition, for the rest of the numbers I called around 7pm, some people are hanging out with friends or having dinner. Thus, wisely choosing a time to call people can improve the response rate.

(e) The ages of the two respondents are 22 and 50, respectively. Thus, the median of age is 36. The area code 248 is from Oakland county in Michigan. According to American FactFinder website,

the median age there is 40.9 years old (https://factfin-der.census.gov/faces/nav/jsf/pages/index.xhtml). Since the sample size is 2, which is way too small to be representative to the population, the bias of age is expected.

(f) For the two respondents, one person voted Trump and the other one voted other. Thus, 50% of my respondent voted Trump, and 0% of my respondent voted Clinton. According to Politico website, 47.6% of the population voted Trump, and 47.3% of the population voted Clinton (https://www.politico.com/mapdata-2016/2016-election/results/map/president/). The actual result is expected to be different from the result I collected, since the number of respondents is too small. To test if the order in which I say the candidates or categories in the survey question influences the results, we should have a large enough sample of respondents. Then, we can randomly group these respondents into two groups with different order in which I say the candidates or categories. Next, we can compare the result of two groups and find if there is any influence on the results.

Problem 2.

The paper aimed to demonstrate a proper adjustment technique can make the non-representative polls be predictive to the presidential election as accurately as those traditional representative polls do. The paper took the Xbox survey as an example.

As shown in Figure 1 in Wang et al., 2005, we can see that sex, age and education from the Xbox sample are the least representative of the data among the eight variables collected from the respondents. On the other hand, race, state and 2008 vote are the most representative of the data. The reason that sex and age from the Xbox sample are the least representative of the data is that it is well-known that young men contribute to most of the Xbox users, which resulted in the bias of the data: 93% of the Xbox users who took the survey are male, while only 47% of the electorates are males; people who are 18 to 29 years old make up 65% of the Xbox dataset, but only 19% in the exit poll (Wang et al., 2005). To explain the bias of education level variable, there are fewer college graduates in the Xbox population, since about 50% the population who took Xbox survey were college graduates but only about 30% in 2012 Exit Poll were college graduates in Wang et al., 2005. This comparison result suggests that Xbox users generally have a lower education level than the voting population, since the dominate population of Xbox Users are young people, who tend to have lower education level.

To solve the problem of non-representative data, the paper utilized an adjustment technique called multilevel regression and post-stratification (MRP). In order to perform a post-stratification re-weighting of the respondents, authors first divided the population into characteristics subgroups by all possible combinations of the variables in the Xbox survey. Then, the authors calculated the weights of each characteristics subgroup by the proportion of the electorate in the corresponding subgroups. They used exit poll data from the 2008 presidential election to obtain the weights, and applied the weights to the Xbox survey data.

The Xbox raw data, according to Figure 2 (Wang et al., 2005), anticipated that Romney would win during the last three weeks before 2012 U.S. Presidential election, since the two-party Obama support was below 50% most of the time in Xbox raw data. On the other hand, for Pollster.com forecast data, it would predict that the election outcome as uncertain during the last three weeks before election, since although the two-party Obama support was above 50% from Sep 24th to Oct 8th, it dropped very close to 50% later on, and the trend kept moving up and down. Thus, the support to two candidates are evenly distributed, which made the outcome uncertain. According to Fig. 3 (Wang et al., 2005), the Xbox post-stratified data would predict that Obama would win during the last three weeks before 2012 election, since the two-party Obama support was above 50% almost all the time in Xbox post-stratified data.

References

2016 Election Results: President Live Map by State, Real-Time Voting Updates. (n.d.). Retrieved from https://www.politico.com/mapdata-2016/2016-election/results/map/president/

Data Access and Dissemination Systems (DADS). (2010, October 05). Retrieved from https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml

Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. International Journal of Forecasting, 31(3), 980-991. doi:10.1016/j.ijforecast.2014.06.001