MACS 3000: Perspectives on Computational Analysis
Problem Set 4

Ma Adelaida Martinez Cabrera

October 30, 2018

1 Non-probability sampling phone survey

(a) My file is in the folder

(b) I called 200 numbers. Only 3 people respond and 197 did not. From the non-responses, most numbers were invalid or send to voice-mails.12 numbers answered the call, but they hung-up. Most of them told me they were not interested in surveys. 2 were under 18.

(c) Of those for whom Response = 1, $\frac{2}{3}$ answered the voting and the age questions. And $\frac{1}{3}$ neither.

(d) I called at 10 am, 1 pm and 2 pm in my area code that was Salt Lake City. More calls where answer at 10 am, and most of the 1pm-2pm calls where directed to voice mail. We might think people were having lunch at this hour and they were not able to answer at that time.

(e) The median age of my respondents is 51.5 (I only have 2 data points so is equal to the mean). This age is much higher than the average age in the state of Utah that is 30.7 years[1]. First, it is a tiny sample and second younger people don't answer their phones if they don't know the caller.

(f) None of my respondents vote Republican or Democrat. 1 of them did not answer the question, 1 did not vote and 1 vote for other. Is not possible to compare the results with two data points but the fact that the option outside Democrat and Republican was more than 20 percent [2] in Utah is interesting because at random we got one of those voters. I don't think the candidates or categories in the survey question influences the results, in this case, is obvious which category matches which candidate. However, given that the Republican candidate is now the president of the United States, people can condition their response on the policies that have occurred since the election day, many of which have been controversial.

[1] https://datausa.io/profile/geo/utah/
[2] https://www.politico.com/mapdata-2016/2016-election/results/map/president/

2 Predicting elections survey, Wang, Rothschild, Goel, and Gelman (2015)

This paper shows how, after the correct statistical adjustments, non-representative polls can perform well as election forecasts. The authors used surveys that where conducted daily on the Xbox gaming platform, three weeks before the 2012 presidential election. They find that after adjusting the results by multilevel regression and poststratification, a highly non-representative survey dataset like this one can produce accurate forecasts. Taking into account that this forecasts are cheaper and can be updated faster than standard polls, this is a relevant and applicable finding.

In Figure 1 the authors show graphically the non-representativeness of the Xbox sample comparing eight variables with their actual values in the 2012 polls. We can see that the least representative variables are gender, age, and education. The population that plays Xbox or other video games is highly selected. First, younger people and specifically males tend to play more video games; therefore, young men are overrepresented in this type of samples. Second, the first Xbox appear in 2001, most of the adults were not exposed to this technology at younger ages, is not probable that they play video games later in their adult lives. The non-representativeness of the education variables is due to the age selection, here is clear how most of the people that play Xbox is in college and stop playing when they graduate. We can think that when people start working, they stop having the flexibility and the time to play these games. Also, preferences might change with age. Nonetheless, the Xbox sample is similar to the exit pool of 2008 regarding race, state, and 2008 vote.

To solve the issue of non-representativeness, the authors use other data sources that are representative of their target population. They use the exit polls from the 2008 presidential election. They argue not using the Current Population Survey (CPS) that is commonly used because they need political identity variables, such as party, as key poststratification variables. They also argue using 2008 polls instead of 2012 because they wanted to show that this forecasts will work in real time, before the 2012 election the only information available for the poststratification and with all the relevant characteristics was the 2008 exit polls. They think about combining CPS data and 2008 polls, but they limit their analysis to the 2008 polls for simplicity and transparency. As is shown in Figure 5, compared to Figure 1, the reweighted data is now representative of all eight variables.

As we can see in Figure 2, the Xbox raw data in the last three weeks of the election suggests that Romney wins the 2012 U.S. Presidential election. In this same figure, Pollster.com that is an average of traditional polls shows an uncertain result close to 50 percent. However, in Figure 3 we can see how the Xbox post-stratified data predict the correct outcome, Obama wins. Even more, the prediction on the vote share in the last few days is closer to the actual vote share compared to the Pollster.com prediction. In conclusion, in this paper the authors show how using the proper statistical tools, one can generate accurate predictions from non-representative samples.