Statistical Methods for Discrete Response, Time Series, and Panel Data (W271): Lab 1

W271 Instructional Team

September 16, 2017

Instructions:

- **Due Date: 10/01/2017 11:59 PM PT**
- Submission:
  - Submit your own assignment via ISVC
  - Submit 2 files:
    1. A pdf file including the summary, the details of your analysis, and all the R codes used to produce the analysis. Please do not suppress the codes in your pdf file.
    2. R markdown file used to produce the pdf file
  - Each group only needs to submit one set of files
  - Use the following file naming convensation; fail to do so will receive 10% reduction in the grade:
    - SectionNumber_hw01_FirstNameLastNameFirstInitial.fileExtension
    - For example, if you are in Section 1 and have two students named John Smith and Jane Doe, you should name your file the following
      - Section1_hw01_JohnS_JaneD.Rmd
      - Section1_hw01_JohnS_JaneD.pdf
  - Although it sounds obvious, please write the name of each members of your group on page 1 of your report.
  - This lab can be completed in a group of up to 3 people. Each group only needs to make one submission. Although you can work by yourself, we encourage you to work in a group.
  - When working in a group, do not use the "division-of-labor" approach to complete the lab. That is, do not divide the lab by having Student 1 completed questions 1 - 3, Student 2 completed questions 4 - 6, etc. Asking your teammates to do the questions for you is asking them take away your own opportunity to learn.
- Other general guidelines:
  - If you use R libraries and/or functions to conduct hypothesis tests not covered in this course, you will have to explain why the functions you use are appropriate for the hypothesis you are asked to test. Lacking explanations will result in a score of zero for the corresponding question.
  - Thoroughly analyze the given dataset. Detect any anomalies, including missing values, potential of top and/or bottom code, etc, in each of the variables.
  - Your report needs to include a comprehensive Exploratory Data Analysis (EDA) analysis, which includes both graphical and tabular analysis, as taught in this course. Output-dump (that is, graphs and tables that don't come with explanations) will result in a very low, if not zero, score.
  - Your analysis needs to be accompanied by detailed narrative. Remember, make sure your that when your audience (in this case, the professors and your classmates) can easily understand your your main conclusion and follow your the logic of your analysis. Note that just printing a bunch of graphs and model results, which we call "output dump", will likely receive a very low score.
  - Your rationale of any decisions made in your modeling needs to be explained and supported with empirical evidence. Remember to use the insights generated from your EDA step to guide your modeling step, as we discussed in live sessions.

  - All the steps to arrive at your final model need to be shown and explained very clearly.
- Students are expected to act with regards to UC Berkeley Academic Integrity.

Description of the Business Problem and the Data

Imagine you work in a data science consulting company. Your client is interested in selling T-shirts to voters who are likely to support politically liberal candidates (such as Bernie Sanders). Your client has data from a political survey conducted in January of 2016 and is able to identify voters who preferred Bernie Sanders over Hillary Clinton (1 = Likes Bernie more than Clinton; 0 = Likes Clinton more than Bernie). In addition, this (extremely simple) dataset contains information on respondents':

- Party affiliation (1 if Democrat , 2 if Independent or Other, and 3 if Republican);
- Race (1 if white, 0 otherwise);
- Gender (2 if female, 1 if male);
- and Birthyear.

Your client conducted a t-test and found that younger voters were more likely to support Sanders and is willing to target younger voters/shoppers based on this analysis. He thinks that you can do better.

For reference, the United States is considered a two party system. The Democratic Party tends to be associated with politically liberal polices while the Republican Party tends to be associated with politically conservative ideas. Voters are not required to be associated with these two parties and, as you will see later, a high proportion of voters are not associated with these two parties.

Note: This dataset is modified from the 2016 American National Election Survey.

1. Model the relationship between age and voters' preference for Bernie Sanders over Hillary Clinton. Select the model that you prefer the most and describe why you chose these variables and functional form.
    a. Describe your chosen model in words, along with a brief description of the variables and the model's functional form (*Note: You do not have to justify your choices at this step*).
    b. Describe the variables your have included in your model and justify why you chose these variables and the model's functional form. (*Hint: you will have to conduct a very careful EDA and use insights generated from the EDA to support your modeling decision. DO NOT USE OUTPUT-DUMP, meaning do not just print a bunch of graphs and let us interpret the graphs for you. Choose your graphs/tables very selectively and present them with narratives to suppor your modeling decisions.*)
    c. Based on your EDA, describe other models that you might have considered and why you ended up choosing your final model. Be sure to print each of the model results and any statistical tests you used to choose which model to use.
    d. Print the model results of your chosen model, even if you did so earlier.
    e. Conduct all of the relevant statistical tests on your chosen model.
    f. Interpret the impact of age on the dependent variable using odds ratios and be sure to include confidence intervals.
2. For your chosen model, graph the relationship between age and the predicted probabilty of supporting Sanders. Be sure to include any graphs that helps you understand how your model can help you answer the question at hand.
3. Comment on the importance of age and evaluate your client's decision to target younger voters.