Applied Regression and Time Series Analysis (DATASCI W271) 2016 Fall

Professors Devesh Tiwari and Jeffrey Yau

Lab 1: 08/20/2016

Due: Friday, 10/21/2016 (11:59 p.m. Pacific Standard Time) of Week 8

Instructions:

- You can work on this lab by yourself or with a group of up to 3 students.
- A report (in pdf format) detailing your answers and all the steps to arrive at your answers
- A well-documented R-script, Rmd file, or R jupyter notebook detailing all of the codes used to arrive at your answers.
- Please use the following naming convention of your file:
    - SectionName_LastNameFirstLetterofFirstNameOfStudent1_ LastNameFirstLetterofFirstNameOfStudent2_LastNameFirstLetterofFirstNameOfStudent3.fileExtention
    - For example, Jeffrey Yau and Devesh Tiwari are in Section 1. Their pdf report will have the file name "**Section1_YauJ_TiwariD.pdf**" and their Rmd file will have the file name "**Section1_YauJ_TiwariD.Rmd**"
    - Fail to comply with this naming convention will receive a 10% reduction in the grade
- Fail to submit either the pdf report or a well-written set of codes used to generate the results will receive a 50% reduction in the grade.
- Late submission will not receive any credit.
- All the steps used to arrive at your final answers need to be shown clearly. These steps are as important as the final answer.
- In the instructions below, when we say "examine a variable", use both graphical techniques and non-graphical summary, such as description of various statistical moments, description of quantiles, etc., we study in this course.
- The final answer of each question needs to be very easy identified; the use of bold fonts, highlights, or circling will help.
- If working as a group project, DO NOT use the "division-of-labor" approach. Each of the students in a group is expected to make sufficient contribution to the lab. If any of your teammate does not make sufficient contribution, please contact your instructor.

Question: United States Public Opinion

The dataset, "**us_public_opinion.csv**," contains a subset of variables from the American National Election Survey 2016 Pilot Study. The pilot study conducted a nationally representative survey between January 22 and January 28, 2016. This surveys uses weights for methodological purposes, but we will not be using these weights. For that reason, among others, the results from your analysis might not be nationally representative.

This dataset contains the following variables:

Name	**Description**
ftsanders	Rating of Bernie Sanders from 0-100
fthrc	Rating of Hillary Clinton from 0-100
ideo5	Respondent's evaluation of their own ideology, 5 points scale (1 = Very liberal, 5 = Very conservative, 6 = Not sure).
pid3	Respondent's party affiliation (1 = Democrat, 2 = Republican, 3 = Independent, 4 = Other, 5 = Not sure)
race_white	Indicator variable taking the value of 1 if respondent is white, 0 otherwise.
gender	Indicator variable taking the value of 1 if respondent is male, 2 if respondent is female.
birthyr	Birthyear of the respondent.

You are hired as a political consultant to conduct a data science project. Your empirical results will be used to formulate a strategy to ask voters for donations for causes championed by Bernie Sanders.

Background: Bernie Sanders challenged Hillary Clinton in the Presidential primary elections within the Democratic Party in 2016. These two candidates competed to represent this party in the general Presidential election. The conventional wisdom was that voters who supported Bernie Sanders were more ideologically liberal than those who supported Hillary Clinton.

As part of this endeavor, you decide to explore how Sanders's supporters differ from Clinton's supporters, especially on personal ideology. Specifically, you ask "**Are Sanders's supporters more liberal than Clinton's?**" Or, phrase it differently, "**Do liberal voters rate Sanders more highly than they do Clinton?**"

To turn this question into a data science problem, this lab will ask you (or your group) to conduct data analysis and regression model building in order to test certain hypotheses that can address this question.

You will be given enough guidance to conduce this analysis and arrive at your conclusion that will be used to devise the strategy to ask voters for donations for causes championed by Bernie Sanders.

Part 1: Ideology and Candidate Support

1. As in any data science project, check your data, such as examining the structure and the integrity of the data (including the number of observations, number of variables, types of the variables, number of missing values (or oddly coded values) in each of the variables, descriptive statistics of each of the variables, etc). Do not simply print tables and summary statistics without providing context and discussing what conclusions you drew from the initial exploration.

2. As emphasized throughout the course, we do not start from building statistical models right away. Instead, we first thoroughly examine the data: Conduct Exploratory Data Analysis (EDA) on the Dependent Variable
    a. Examine the distribution of the variables *fthrc* and *ftsanders*. Comment on their distributions.
    b. Examine the relationship between *fthrc* and *ftsanders* and comment on their relationship.
    c. Create a variable called *diff,* which is the difference between *ftsanders* and *fthrc*. Examine this new variable. What does this variable mean? Is there anything noteworthy about its distribution?
    d. We could actually answer this question without creating this variable. Please describe either an alternative coding or alternative modeling strategy.

3. Conduct EDA on our explanatory variable of interest: Ideology
    a. Visually inspect and comment on the distribution of the variable, *ideo5*. Would you include this variable "as is" in a model or does it require any sort of transformation?
    b. How would you describe the ideological distribution of American voters?

4. Conduct EDA on other explanatory variables
    a. Create a variable for *age*.
    b. Create a dummy variable for gender that takes a value of one if the respondent is female and is zero otherwise.
    c. Examine each of these explanatory variables. Do you think they require transformation, including binning the variable or creating an additional indicator variable to capture a mass of values, if applicable?
    d. What is the average age of respondents?
    e. What proportion of respondents are white?
    f. What proportion of respondents are female?

5. Examine Bivariate and Multivariate Relationships, including, but not limited to, the following:
    a. Examine the bivariate relationship between the dependent variable and ideology. Based on this initial exploration, do liberal voters have a higher level of support for Sanders over Clinton?
    b. Examine (and comment) on the bivariate relationships between the dependent variable and each of the explanatory variables as well as among the explanatory variables themselves. Comment on each of these relationships. Are there any transformations and/or creation of additional variables that you think maybe useful? Might multicolleniarity be a problem? If so, how would it impact your model's results? (Note that the classical linear regression model does not specify that each and everyone of these relationships has to be linear. The most obvious case is binary explanatory variable; it clearly is not related to the dependent variable linearly. The

CLM just assumes that the conditional expectation function is a linear, conditional on the set of explanatory variables.)

   c. In some cases, multivariate analyses could be useful before the modeling stage. Do they apply in this case? If so, conduct some multivariate analyses.

6. Model Building Part 1: Treat *ideo5* as a continuous variable. Consider the following when building your model.
   a. How does your EDA influence the inclusion of the variables in your model?
   b. Start from a parsimonious model (with just the independent variable of interest) and gradually build it up. (Note that this is not the only way to build a regression model.) Does the inclusion of additional explanatory variables improve your model? Does the direction of the estimated coefficients of explanatory variables make sense? What, after all, does "improve" your model mean (in the context of answer the data science question at hand)?
   c. Based on the best model thus far, do liberal voters have a higher level of support for Sanders over Clinton?
   d. To what extent does this relationship have any practical significance?
   e. Note: Do not just use some sort of automatic selection methods, such as forward- or backward selection. If you use them, please provide your rationale.
   f. Remember to conduct regression diagnostics.
   g. Remember to conduct formal statistical tests to test all of the underlying assumptions of the CLM.
   h. When assumptions are not satisfied, describe its potential impacts on the estimates and statistical inference.

7. Model Building Part 2: Treat *ideo5* as a categorical variable
   a. All the suggestions in Step 6 apply to this step.
   b. Based on this model, do liberal voters have a higher level of support for Sanders over Clinton?
   c. Compare and contrast your final model in step 6 and your final model in step 7. Do your answers differ? Is one model more appropriate than the other? Is one model easier to interpret than the other?

8. The Final Model
   a. Choose your final model.
   b. Does this model satisfy all of the CLM assumptions?
   c. What changes would you make to your existing models based on these results?
   d. What changes are possible to make with the existing data?
   e. What changes would require more data?
   f. What changes cannot be plausibly made with additional data only?
   g. Most importantly, do not forget to answer the original question posted by the company that hires you. Based on your model, how would you formulate the strategy to ask voters for donations for causes championed by Bernie Sanders.

Part 2: Including partisanship

The above analysis does not take into account the partisanship of the survey respondents. The conventional wisdom in American Politics is that liberal voters are members of the Democratic Party while conservative voters are members of the Republican Party. The exclusion of partisanship could lead to biased estimates.

1. By excluding this variable, which of the 5 CLM assumptions did you violate? Explain why.
2. Based on your residuals analysis in Part 1, were you able to diagnose this problem? Why or why not?
3. How closely are partisanship and ideology related? Should you be worried about multicollinearity?
4. Use your final model from Part 1 and include partisanship in your model.
5. Does the addition of this variable improve the model? (If so, in what sense?) How do your answers change?
6. Suppose you think that the relationship between ideology and the DV has a different *slope* for each party. How would you account for this in your model?
7. Make this adjustment and comment on the relationship between ideology, partisanship, and the DV.
8. Do you think that taking account of respondents' party affiliation is necessary for your purposes? As a political consultant, how would you use this information when deciding to whom solicitations should be made?