Statistical Methods for Discrete Response, Time Series, and Panel Data (W271): Lab 2

W271 Instructional Team

February 10, 2017

Instructions:

- **Due Date: 3/4/2018**

- **Page Limite: 10 pages**

- 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.
        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:

        * Student1FirstNameLastName_Student2FirstNameLastName_Student3FirstNameLastName.fileExtension
        * For example, Chole Kim, Shawn White, and Red Gerard are in the group, name your files as
            · ChloeKim_ShawnWhite_RedGerard.Rmd
            · ChloeKim_ShawnWhite_RedGerard.pdf

    - 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, I encourage you to work in a group for benefits I have discussed in live sessions.

- Other general guidelines:

    - Please read the instructions carefully.

    - Please read the questions carefully.

    - For statistical libraries, use only those are covered in this course.

    - If you use R libraries and/or functions to conduct hypothesis tests and other statistical modeling not covered in this course, you will have to explain why the function you use is appropriate and why do you decide not to use those covered in the course

    - 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.

    - 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 the logic of your analysis. Note that just printing a bunch of graphs and model results, which is called "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.

Introduction

The authors give an example (pp. 220 in the text) regarding alcohol consumption, which is based on a study in which moderate-to-heavy drinkers (defined as at least 12 alcoholic drinks/week for women, 15 for men) were recruited to keep a daily record of each drink that they consumed over a 30-day study period. Participants also completed a variety of rating scales covering daily events in their lives and items related to self-esteem. Among the researchers' hypotheses was that negative events — particularly those involving romantic relationships — might be related to amount of alcohol consumed, especially among those with low self-esteem.

The Lab

This lab is a lot more open-ended than lab 1.

First, read the example "Alcohol Consumption" on pp. 220-221 in the text, but please do not copy and paste their codes, as you analysis will be different. I also would not suggest that you read the original study, as I expect you to conduct your own study using the data set and the stated hypothesis. So, please do not make use of or reference the original study. Doing so will result in points deduction.

The researchers stated the hypothesis as follow: *"We hypothesized that negative interactions with romantic partners would be associated with alcohol consumption (and an increased desire to drink). We predicted that people with low trait self-esteem would drink more on days they experienced more negative relationship interactions compared with days during which they experienced fewer negative relationship interactions. The relation between drinking and negative relationship interactions should not be evident for individuals with high trait self-esteem."*

In this lab, you will use **DeHartSimplified.csv** data set to conduct an analysis (using everything covered in week 1-5) and build a model (using techniques covered in week 4 and 5) to address this hypothesis. In this study, trait self-esteem (a long-term view of self-worth) is measured by the variable *rosn*, while the measure of negative relationship interactions is *nrel*. As mentioned in the instruction section, an examination of the data set as well as a thorough EDA, which you will use to guide your model specification, are required. You will also have to explain your model specification you have considered and substantiated your final model choice. Importantly, interpret the model result and conduct residual diagnostics.

Your report needs to include an introduction section (that includes a summary of your final models and key results), an EDA section, a modeling section, and a conclusion section that summarizes the key takeaways, lesson learned, and any caveats of your study.