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

Professor Jeffrey Yau

Instructions:

- **Due Date: Monday of Week 6 4p.m. Pacific Time**
- **Page limit of the pdf report: 20 (not include title and the table of content page**
- Use the margin, linespace, and font size specification below:
  - fontsize=11pt
  - margin=1in
  - line_spacing=single
- Submission:
  - Each group makes one submission to Github; please have one of your team members made the submission
  - Submit 2 files:
    1. A pdf file including 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
  - Use the following file-naming convensation; fail to do so will receive 10% reduction in the grade:
    - FirstNameLastName1_FirstNameLastName2_FirstNameLastName3_LabNumber.fileExtension
    - For example, if you have three students in the group for Lab Z, and their names are Gerard Kelley, Steve Yang, and Jeffrey Yau, then you should name your file the following
      - GerardKelley_SteveYang_JeffreyYau_LabZ.Rmd
      - GerardKelley_SteveYang_JeffreyYau_LabZ.pdf
  - Although it sounds obvious, please write the name of each members of your group on page 1 of your pdf and Rmd files.
  - This lab can be completed in a group of up to 3 students in your session. Students are encouraged to work in a group for the lab.
- Other general guidelines:
  - For statistical methods that we cover in this course, use only the R libraries and functions that are covered in this course. If you use libraries and functions for statistical modeling that we have not covered, you have to provide (1) explanation of why such libraries and functions are used instead and (2) reference to the library documentation. Lacking the explanation and reference to the documentation will result in a score of zero for the corresponding question.
  - In this particular lab, simply answer the following questions stated in Question 12 of chapter 3 (on page 189 and 190) of Bilder and Loughin's *"Analysis of Categorical Data with R"*
    - **No need to include introduction, data examination, EDA, and conclusion sections.**
    - Since this question has **part a to h**, please write down each of the questions in your

report so that we can easily follow your answers.

- Students are expected to act with regards to UC Berkeley Academic Integrity.

Strategic Placement of Products in Grocery Stores

Answer **Question 12 of chapter 3 (on page 189 and 190)** of Bilder and Loughin's *"Analysis of Categorical Data with R"*. Here is the background of this analysis, taken as an excerpt from this question:

In order to maximize sales, items within grocery stores are strategically placed to draw customer attention. This exercise examines one type of item—breakfast cereal. Typically, in large grocery stores, boxes of cereal are placed on sets of shelves located on one side of the aisle. By placing particular boxes of cereals on specific shelves, grocery stores may better attract customers to them. To investigate this further, a random sample of size 10 was taken from each of four shelves at a Dillons grocery store in Manhattan, KS. These data are given in the **cereal_dillons.csv** file. The response variable is the shelf number, which is numbered from bottom (1) to top (4), and the explanatory variables are the sugar, fat, and sodium content of the cereals.

a. The explanatory variables need to be reformatted before proceeding further.
   - First, divide each explanatory variable by its serving size to account for the different serving sizes among the cereals.
   - Second, rescale each variable to be within 0 and 1.
   - Some sample code is provided

b. Construct side-by-side box plots with dot plots overlaid for each of the explanatory variables.
   - Some sample code is provided
   - Also, construct a **parallel coordinates plot** for the explanatory variables and the shelf number. Discuss if possible content differences exist among the shelves.

c. The response has values of $1, 2, 3$, and 4. Under what setting would it be desirable to take into account ordinality. Do you think that this setting occurs here?

d. Estimate a **multinomial regression model with linear forms of the sugar, fat, and sodium variables**. Perform **LRTs** to examine the importance of each explanatory variable.

e. Show that there are no significant interactions among the explanatory variables (including an interaction among all three variables).

f. Kellogg's Apple Jacks (http://www.applejacks.com) is a cereal marketed toward children. For a serving size of 28 grams, its sugar content is 12 grams, fat content is 0.5 grams, and sodium content is 130 milligrams. Estimate the shelf probabilities for Apple Jacks.

g. Construct a plot similar to **Figure 3.3** where the estimated probability for a shelf is on the *y-axis* and the sugar content is on the *x-axis*. Use the mean overall fat and sodium content as the corresponding variable values in the model. Interpret the plot with respect to sugar content.

h. Estimate odds ratios and calculate corresponding confidence intervals for each explanatory variable. Relate your interpretations back to the plots constructed for this exercise.