W271 Assignment 1

Due 4:00pm Pacific Time, Monday June 1, 2020

Instructions (Please Read Carefully):

- **Late submissions will not be accepted**
- No page limit, but be reasonable
- Do not modify fontsize, margin or line_spacing settings
- This assignment needs to be completed individually; this is not a group project
- Submission is by pushing to your student fork of the course repository
- Submit two files:
  1. A pdf file that details your answers (knit to pdf, do not knit to html then save as pdf). Include all R code used to produce the answers. Do not suppress the code in your pdf file
  2. The R markdown (Rmd) file used to produce the pdf file

  The assignment will not be graded unless **both** files are submitted
- Use the following file-naming convensation:
  - StudentFirstNameLastName_HWNumber.fileExtension
  - For example, if the student's name is Kyle Cartman for assignment 1, name your files follows:
    - KyleCartman_assignment1.Rmd
    - KyleCartman_assignment1.pdf
- Although it sounds obvious, please write your name on page 1 of your pdf and Rmd files
- Answers should clearly explain your reasoning; do not simply 'output dump' the results of code without explanation
- For statistical methods that we cover in this course, use 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 must provide an explanation of why such libraries and functions are used and reference the library documentation. For data wrangling and data visualization, you are free to use other libraries, such as dplyr, ggplot2, etc.
- For mathematical formulae, type them in your R markdown file. Do not e.g. write them on a piece of paper, snap a photo, and use the image file.
- Incorrectly following submission instructions results in deduction of grades
- Students are expected to act with regard to UC Berkeley Academic Integrity

1. Confidence Intervals (2 points)

A Wald confidence interval for a binary response probability does not always have the stated confidence level, $1 - \alpha$, where α (the probability of rejecting the null hypothesis when it is true) is often set to 0.05%. This was demonstrated with code in the week 1 live session file.

Question 1.1: Use the code from the week 1 live session file and: (1) redo the exercise for `n=50, n=100, n=500`, (2) plot the graphs, and (3) describe what you have observed from the results. Use the same `pi.seq` as in the live session code.

Question 1.2: (1) Modify the code for the Wilson Interval. (2) Do the exercise for `n=10, n=50, n=100, n=500`. (3) Plot the graphs. (4) Describe what you have observed from the results and compare the Wald and Wilson intervals based on your results. Use the same `pi.seq` as in the live session code.

2: Binary Logistic Regression (2 points)

Do Exercise 8 a, b, c, and d on page 131 of Bilder and Loughin's textbook. Please write down each of the questions. The dataset for this question is stored in the file *"placekick.BW.csv"* which is provided to you.

In general, all the R codes and datasets used in Bilder and Loughin's book are provided on the book's website: chrisbilder.com

For **question 8b**, in addition to answering the question, re-estimate the model in part (a) using "Sun" as the base level category for $Weather$.

3: Binary Logistic Regression (2 points)

The dataset *"admissions.csv"* contains a small sample of graduate school admission data from a university. The variables are specificed below:

1. admit - the depenent variable that takes two values: $0, 1$ where 1 denotes *admitted* and 0 denotes *not admitted*
2. gre - GRE score
3. gpa - College GPA
4. rank - rank in college major

Suppose you are hired by the University's Admission Committee and are charged to analyze this data to quantify the effect of GRE, GPA, and college rank on admission probability. We will conduct this analysis by answering the follwing questions:

Question 3.1: Examine the data and conduct EDA

Question 3.2: Estimate a binary logistic regression using the following set of explanatory variables: gre, gpa, $rank$, gre^2, gpa^2, and $gre \times gpa$, where $gre \times gpa$ denotes the interaction between gre and gpa variables

Question 3.3: Test the hypothesis that GRE has no effect on admission using the likelihood ratio test

Question 3.4: What is the estimated effect of college GPA on admission?

Question 3.5: Construct the confidence interval for the admission probability for the students with $GPA = 3.3$, $GRE = 720$, and $rank = 1$

4. Binary Logistic Regression (2 points)

Load the `Mroz` data set that comes with the *car* library (this data set is used in the week 2 live session file).

Question 4.1: Estimate a linear probability model using the same specification as in the binary logistic regression model estimated in the week 2 live session. Interpret the model results. Conduct model diagnostics. Test the CLM model assumptions.

Question 4.2: Estimate a binary logistic regression with `lfp`, which is a binary variable recoding the participation of the females in the sample, as the dependent variable. The set of explanatory variables includes `age`, `inc`, `wc`, `hc`, `lwg`, `totalKids`, and a quadratic term of `age`, called `age_squared`, where `totalKids` is the total number of children up to age 18 and is equal to the sum of `k5` and `k618`.

Question 4.3: Is the age effect statistically significant?

Question 4.4: What is the effect of a decrease in age by 5 years on the odds of labor force participation for a female who was 45 years of age.

Question 4.5: Estimate the profile likelihood confidence interval of the probability of labor force participation for females who were 40 years old, had income equal to 20, did not attend college, had log wage equal to 1, and did not have children.

5: Maximum Likelihood (2 points)

Question 18 a and b of Chapter 3 (page 192,193)

For the wheat kernel data (*wheat.csv*), consider a model to estimate the kernel condition using the density explanatory variable as a linear term.

Question 5.1 Write an R function that computes the log-likelihood function for the multinomial regression model. Evaluate the function at the parameter estimates produced by multinom(), and verify that your computed value is the same as that produced by logLik() (use the object saved from multinom() within this function).

Question 5.2 Maximize the log-likelihood function using optim() to obtain the MLEs and the estimated covariance matrix. Compare your answers to what is obtained by multinom(). Note that to obtain starting values for optim(), one approach is to estimate separate logistic regression models for $log\left(\frac{\pi_2}{\pi_1}\right)$ and $log\left(\frac{\pi_3}{\pi_1}\right)$. These models are estimated only for those observations that have the corresponding responses (e.g., a $Y = 1$ or $Y = 2$ for $log\left(\frac{\pi_2}{\pi_1}\right)$).