Management Sciences, University of Waterloo
Statistical Methods for Data Analytics
Winter 2022
Instructor: Jangho Yang
TA: Yekta Amirkhalili & Muhammad Saadi Azi

Assignment 2

Due Mar 21 11:59 PM (Monday)

*Note: The assignment submission should include i) a write-up for all your answers in a pdf form and ii) a separate R script. If you use Rmarkdown and combine i) and ii), you can submit the Rmd file only. The questions with *R coding only* do not require a written response. We will check your R coding only. Your R script needs to be fully commented, meaning that you need to explain your code line by line. Further, your R script needs to be self-sufficient so that graders can execute your code without any bug.

1. Hypothesis testing with R: Simulate data x from **rnorm**$(\text{n} = 50, \text{mean} = 0, \text{sd} = 2)$ and y from **rnorm**$(\text{n} = 50, \text{mean} = 2 + 1.5 \times x, \text{sd} = 10)$ using **set.seed(4)**. Fit a linear regression of y on x using $y = \beta_0 + \beta_1 x + \varepsilon$, where $\varepsilon \overset{\text{iid}}{\sim} \text{N}(0,\ \sigma^2)$. [40 points]
   (a) Plot the data and add the fitted line. [4 point]
   (b) Extract the p-value attached to the estimator $\hat{\beta}_1$ from **summary()**. This is the p-value according to the null hypothesis that $\beta_1 = 0$. Given the significance level $\alpha = 0.05$, are you going to reject the null hypothesis? [4 point]
   (c) Repeat this exercise by generating another set of data. Keep the same structure of the data generating process but use **set.seed(200)**. What is your p-value and what's your decision on the null hypothesis? [2 point]
   (d) Read this article and discuss the limitations of the classical hypothesis testing. [15 point]
   (e) Read this article and discuss how prevalent p-hacking is in science. [15 point]
2. Binomial Logit Regression with R: Load the data file **voting_data.csv**. This data is Pew Research Center Polls taken during the 2008 election campaign. The data consists of a respondent's vote intention, marital status and state information. "dem_vote" is a vote intention indicator between a republican candidate and a democrat candidate (0 = republican leaning, 1 = democrat leaning), "marital_id" is a marital status indicator (0 = non-married, 1 = married), and "state" is the state where the respondent lives. Use a binomial logistic regression to predict vote intention (y) using the indicator for being married (x_1) and the state indicator (x_2) as predictors. [30 points]
   (a) Explain potential problems if we use a multiple linear regression model with a normal error term for this exercise. [5 point]
   (b) Write down and explain the outcome distribution and the link function of the binomial logistic regression model. [5 point]

(c) Show the summary of the regression results and interpret the the coefficient related to x_1. What is your conclusion as to the vote intention for married and non-married voters? [20 points]

3. R programming: Suppose we have two separate bags of balls. Bag A has N number of balls in it, some white some black but we don't know how many of each color. Bag B has K_w white ball(s) and K_b black ball(s). You draw R number of balls from Bag A and move them to Bag B. Then, you draw T balls sequentially from Bag B. Suppose the sequence of the these balls turns out to be $[S_1, S_2, \ldots, S_T]$. We are interested in the composition of balls in Bag A. List your hypotheses and count up all the ways the observed data can happen. Which hypothesis would you believe more? [30 points]

   (a) Write a code to to calculate the number of ways the observed sequence $[S_1, S_2, \ldots, S_T]$ can happen for each hypothesis. Your function should work for any value of N, K_w, K_b, R, T and for any sequence of balls $[S_1, S_2, \ldots, S_T]$. [R coding only] [20 points]

   (b) Which hypothesis is most likely when $N = 20, K_w = 10, K_b = 10, R = 5$ and the observed sequence is $[W, B, B, W, B, W]$. [10 points]