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

Final Project

Due April 08 11:59 PM (Friday)

*Note: Choose one of the two questions in each section: Section 1.R programming Questions and Section 2.Data Analysis Questions. Each question has a different level of difficulty, which will give you upper bound of your final marks for the project. For example, if you choose the easiest combination (Q1 from Section 1 and Q2 from Section 2), the upper bound is 85 points. The submission should include i) a write-up for all your answers (including figures and written responses) 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. 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.

Section 1: R Programming Questions:

1. R programming competition [Moderate]: A clinic has three doctors. Patients come into the clinic at random, starting at 9 a.m. The time after opening at which the first patient appears follows an exponential distribution with expectation of 10 minutes and then, after each patient arrives, the waiting time until the next patient is independently exponentially distributed, also with expectation of 10 minutes. When a patient arrives, he or she waits until a doctor is available. The amount of time spent by each doctor with each patient is a random variable, uniformly distributed between 5 and 20 minutes. The office stops admitting new patients at 4 p.m. and closes when the last patient is through with the doctor. [35 points]

   (a) Simulate this process once. (i) How many patients came to the office? (ii) How many had to wait for a doctor? (iii) What was their average wait? (iv) When did the office close?

   (b) Simulate the process 1000 times and estimate the median for each of the summaries in (a).

2. R programming competition [Difficult]: A lake is full of bluefish and goldfish. A researcher wants to estimate the number of each sort of fish in the lake. He catches 100 at random and tags them. Of these 100 fish, 60 are blue and 40 are gold. He asks his assistant to return a few days later and catch another 100 fish. (Assume this is another simple random sample of all the fish in the lake.) The assistant reports that 20 of these are tagged, but he forgets to record the color of the fish. Let N_b and N_g be the number of bluefish and goldfish in the lake. [40 points]

   (a) You can use any models for this problem. One possibility is a multinomial distribution. The multinomial distribution is a simple extension of a binomial distribution with multiple mutually exclusive outcomes. Based on your model, define all the relevant coefficients, missing data, and observed data.

(b) Define your likelihood function.

(c) Find the set of value of N_b and N_g that maximizes your likelihood function by simulation.

(d) Suppose you know a priori that the set of value for N_b and N_g is either ($N_b = 280$, $N_g = 220$) or ($N_b = 290$, $N_g = 210$). Which hypothesis is more likely?

Section 2: Data Analysis Questions

1. R data analysis competition [Difficult]: Load the attached two data files. The first data is Pew Research Center polls taken during the 2008 election campaign (Q1Data1.csv) and the second data (Q1Data2.csv) is about 2008 election result in the US.[60 points]

(a) Take the first data file (Q1Data1.csv). 1) Subset the data so that you have all states but Hawaii, Alaska, and Washington D.C and have only four columns "state," "marital," "heat2," and "heat4." 2) If no data is available in "heat2," replace **na** for the corresponding value in "heat4." If neither of "heat2" and "heat4" has data, erase the corresponding row. 3) Subset the data so that you only have "dem/lean dem" and "rep/lean rep" in the "heat2" column. 4) Change the label of all the variables but 'married' (married people) in the "marital" column to 'other' (which indicates non-married people).

(b) For each state, calculate 1) the proportion of the democratic supporters, 2) the proportion of the married people, 3) the ratio of the married people among the democratic supporters to the total married people, 4) the ratio of non-married among the democratic to the total non-married people, 5) the difference of 3) and 4). Multiply all values by 100 to convert to percentage. Show the first 5 observations of these new variables.

(c) Take the second data file (Q1Data2.csv). Subset the data so that 1) you have all but three states, Hawaii, Alaska, and Washington D.C, and 2) only two columns "state," and "vote_Obama_pct" (Obama's actual vote share). Show the first 5 lines of the data set.

(d) Use a logistic regression predicting vote intention given state, using the indicator for being married as a predictor. Set up a proper link function. Try three different assumptions as to the state-level heterogeneity

- Assumption 1: No state-level heterogeneity. All states have the same intercept and slope.
- Assumption 1: Complete state-level heterogeneity. All states have completely independent intercepts and slopes. No outlying coefficient is penalized.
- Assumption 3: State-level heterogeneity is unknown a priori. States have partially pooled intercepts and slopes. Outlying coefficients are penalized.

(e) Using the estimation result from the model with Assumption 3, plot your inference for the predicted vote share by state, along with the actual vote intention, plotting them vs. Obama's actual vote share. Annotate each dot with the corresponding state name.

(f) The marriage gap is defined as the difference of Obama's vote share among married and non-married people ("other"). Figure out how to infer this marriage gap from your model. Using the estimation result from the model with Assumption 3, plot your inference for the marriage gap, along with the raw marriage gaps from the data, plotting them vs. Obama's vote share.

(g) Repeat (e)-(f) for the model with Assumption 1, Discuss your result.

2. R data analysis competition [Moderate]: Load the data (Q2Data.csv). This data includes information about home radon levels ("activity") in the US. [50 points]

   (a) Get only 4 columns "state", "activity", "floor" (a floor level in the house), "county" for the state "MN" (Minnesota). Replace the zero values in "activity" column with 0.05 and then log-transform the entire values in "activity." Assign a numerical ID variable to each unique county; for example, you assign 1 to county "AITKIN," which is the first county in the MN subdata.

   (b) Fit a linear regression model with a normal error. Your outcome variable is "activity" (home radon levels in picocuries per liter) and the predictor variable is "floor" (0 for basement and 1 for first floor). Try three different assumptions as to the county-level heterogeneity:

   - Assumption 1: No county-level heterogeneity. All counties have the same intercept and slope.
   - Assumption 1: Complete county-level heterogeneity. All counties have completely independent intercepts and slopes. No outlying coefficient is penalized.
   - Assumption 3: County-level heterogeneity is unknown a priori. Counties have partially pooled intercepts and slopes. Outlying coefficients are penalized.

   (c) Visualize the difference in the estimated coefficients from three models. Discuss the difference.

   (d) Using the model with Assumption 3, plot the predicted values of radon activity against floor along with raw data. Repeat this with the model with Assumption 2. What difference can you tell? Assess the predictive power of each the two models.