W271 Group Lab 1

Due 11:59pm Pacific Time Sunday February 9 2020

Instructions (Please Read Carefully):

- 20 page limit (strict)

- Do not modify fontsize, margin or line_spacing settings

- One student from eah group should submit the lab to their student github repo by the deadline; submission and revisions made after the deadline will not be graded

- Answers should clearly explain your reasoning; do not simply 'output dump' the results of code without explanation

- Submit two files:

    1. A pdf file that details your answers. Include all R code used to produce the answers. Do not suppress the codes 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

- Name your files to include all group members names. For example the students' names are Stan Cartman and Kenny Kyle, name your files as follows:

    - `StanCartman_KennyKyle_Lab1.Rmd`
    - `StanCartman_KennyKyle_Lab1.pdf`

- Although it sounds obvious, please write your name on page 1 of your pdf and Rmd files

- All answers should include a detailed narrative; make sure that your audience can easily follow the logic of your analysis. All steps used in modelling must be clearly shown and explained

- 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.

Investigation of the 1989 Space Shuttle Challenger Accident

Carefullly read the Dalal et al (1989) paper (Skip Section 5).

Part 1 (25 points)

Conduct a thorough EDA of the data set. This should include both graphical and tabular analysis as taught in this course. Output-dump (that is, graphs and tables that don't come with explanations) will result in a very low, if not zero, score. Since the report has a page-limit, you will have to be selective when choosing visuals to illustrate your key points, associated with a concise explanation of the visuals. This EDA should begin with an inspection of the given dataset; examination of anomalies, missing values, potential of top and/or bottom code etc.

Part 2 (20 points)

Answer the following from Question 4 of Bilder and Loughin Section 2.4 Exercises (page 129):

(a) The authors use logistic regression to estimate the probability an O-ring will fail. In order to use this model, the authors needed to assume that each O-ring is independent for each launch. Discuss why this assumption is necessary and the potential problems with it. Note that a subsequent analysis helped to alleviate the authors' concerns about independence.

(b) Estimate the logistic regression model using the explanatory variables in a linear form.

(c) Perform LRTs to judge the importance of the explanatory variables in the model.

(d) The authors chose to remove Pressure from the model based on the LRTs. Based on your results, discuss why you think this was done. Are there any potential problems with removing this variable?

Part 3 (35 points)

Answer the following from Question 5 of Bilder and Loughin Section 2.4 Exercises (page 129-130):

Continuing Exercise 4, consider the simplified model $logit(\pi) = \beta_0 + \beta_1 Temp$, where π is the probability of an O-ring failure. Complete the following:

(a) Estimate the model.

(b) Construct two plots: (1) π vs. Temp and (2) Expected number of failures vs. Temp. Use a temperature range of 31° to 81° on the x-axis even though the minimum temperature in the data set was 53°.

(c) Include the 95% Wald confidence interval bands for π on the plot. Why are the bands much wider for lower temperatures than for higher temperatures?

(d) The temperature was 31° at launch for the Challenger in 1986. Estimate the probability of an O-ring failure using this temperature, and compute a corresponding confidence interval. Discuss what assumptions need to be made in order to apply the inference procedures.

(e) Rather than using Wald or profile LR intervals for the probability of failure, Dalal et al. (1989) use a parametric bootstrap to compute intervals. Their process was to (1) simulate a large number of data sets (n = 23 for each) from the estimated model of Temp; (2) estimate new models for each data set, say and (3) compute at a specific temperature of interest. The authors used the 0.05 and 0.95 observed quantiles from the simulated distribution as their 90%

confidence interval limits. Using the parametric bootstrap, compute 90% confidence intervals separately at temperatures of 31° and 72°.27

(f) Determine if a quadratic term is needed in the model for the temperature.

Part 4 (10 points)

With the same set of explanatory variables in your final model, estimate a linear regression model. Explain the model results; conduct model diagnostic; and assess the validity of the model assumptions. Would you use the linear regression model or binary logistic regression in this case? Explain why.

Part 5 (10 points)

Interpret the main result of your final model in terms of both odds and probability of failure. Summarize the final result with respect to the question(s) being asked and key takeaways from the analysis.