

## CS 601C Fall 2021 - Final Project Due: Dec 22, 2021

Complete the exercises and answer the questions outlined below. **Submit your solutions in an R Notebook or Markdown file.** Your code must run successfully and you must answer the questions to get credit.

1. Download the “`stack.df.csv`” file, and load the data into R. This file contains stack loss data from the operation of a plant for the oxidation of ammonia to nitric acid, measured on 21 consecutive days. The data include

| Variable                | Description   |
|-------------------------|---|
| <code>stack_loss</code> | percent of ammonia lost (times 10)  |
| <code>Air_Flow</code>   | air flow to the plant   |
| <code>Water_Temp</code> | cooling water inlet temperature   |
| <code>Acid_Conc</code>  | acid concentration as a percentage (coded by subtracting 50 and then multiplying by 10) |

- (a) Display a summary of the data. Interpret the output and describe the data.
- (b) Plot a pairwise scatterplot. What are your observations regarding the relationships.
- (c) Use the `lm` function to fit a regression model with `stack_loss` as the dependent variable and the other three as explanatory variables. Display a summary of the modeling results and interpret the findings.
- (d) Plot the diagnostic plots and interpret.
- (e) Revise your model to remove any terms that are not significant and display the summary of your model.
- (f) Compare the original and revised models and discuss what you find.

[50 points]

2. Download the “`kyphosis.csv`” file, and load the data into R. The file contains 81 observations representing data on 81 children who have had corrective spinal surgery. The outcome Kyphosis is a binary variable, the other three variables are numeric.

| Variable              | Description  |
|-----------------------|--|
| <code>Kyphosis</code> | a factor telling whether a postoperative deformity (kyphosis) is “present” or “absent” |
| <code>Age</code>      | the age of the child in months   |
| <code>Number</code>   | the number of vertebrae involved in the operation                                      |
| <code>Start</code>    | the beginning of the range of vertebrae involved in the operation                      |

- (a) Make sure `Kyphosis` is a `factor`. If not, convert it from `character` to `factor`. Display the summary statistics for the data.

- (b) Set up your plot for one row of three plots (`par(mfrow = c(1, 3))`). Then make three boxplots side by side one for each variable against `Kyphosis`.
- (c) Interpret your plots. What can you say about the distributions when kyphosis is present or absent?
- (d) Fit a logistic regression model that relates the probability of developing Kyphosis to the three predictor variables, Age, Number, and Start. Fit the model using `glm`.
- (e) Display the summary statistics for your model. Interpret the coefficients to describe how the variables influence the probability of Kyphosis.
- (f) Call `anova(<model name>, test = "Chi")` and identify any insignificant variables. Revise your model based on this.

[50 points]