Math 4392 HW1

Washington University, SP 2024

Instruction: Please type up your assignment neatly including R code, relevant console output, graphs and written portion in a single document. You are encouraged to use Rmarkdown, but other format is acceptable as long as the report demonstrates your work clearly and coherently.

If possible, submit the resulting file in `pdf` format due to limitation of the grading comment feature on Canvas. Raw `.Rmd` or `.R` files are not accepted in case of trouble compiling.

Please note - there is always some freedom in deciding which methods to use, in what order to apply them, and how to interpret the results. So there may not be one clear right answer and good analysts may come up with different models.

This assignment covers approximately the first three lectures.

Due Date: Sunday, Feb 4, 11:59 pm on Canvas

1. Consider the dataset `attitude` and use `rating` as the response variable
   a) Create an initial data analysis that explores the numerical and graphical characteristics of the data. Be sure to look up the dataset to get a sense of what the variables mean.
   b) Use variable selection to choose the best model (hint: look up the `step()` function)
   c) Explore any transformations to improve the fit of the model.
   d) Perform diagnostics to check the assumptions of your model.
   e) Interpret the meaning of the model and comment on potential application
2. The National Institute of Diabetes and Digestive and Kidney Diseases conducted a study on 768 adult female Pima Indians living near Phoenix. The purpose of the study was to investigate factors related to diabetes. The data may be found in the the dataset `pima` (available in the Faraway R package).
   a) Create a factor version of the test results and use this to produce an interleaved histogram to show how the distribution of insulin differs between those testing positive and negative. Do you notice anything unbelievable about the plot?
   b) Replace the zero values of insulin with the missing value code NA. Recreate the interleaved histogram plot and comment on the distribution.
   c) Replace the incredible zeroes in other variables with the missing value code. Fit a model with the result of the diabetes test as the response and all the other variables as predictors. How many observations were used in the model fitting? Why is this less than the number of observations in the data frame?
   d) Refit the model but now without the insulin and triceps predictors. How many observations were used in fitting this model? Devise a test to compare this model with that in the previous question.
   e) Use AIC to select a model. You will need to take account of the missing values. Which predictors are selected? How many cases are used in your selected model?