STAT 8330: **Homework 1**

Due on September 10, 2020; 5pm

Instructions: To minimize the grading effort, ONLY report the results that are asked for. Submit as a pdf document preferably using R Markdown.

1. JWHT (2013), Chap 2, Prob. 5
2. JWHT (2013), Chap 2, Prob. 6
3. SEEDS ANALYSIS: The dataset `seeds.csv` consists of measurements of geometrical properties of 210 wheat kernels belonging to three different varieties (Kama, Rosa and Canadian). High quality visualization of the internal kernel structure was detected using a soft X-ray technique to construct the following seven, real-valued attributes:

   - Area
   - Perimeter
   - Compactness $C = 4\pi \frac{A}{P^2}$
   - Length of kernel
   - Width of kernel
   - Asymmetry coefficient
   - Length of kernel groove

   (a) Use the read.csv() function to read the data into R. Redefine the "Type" variable from (1,2,3) to (Kama, Rosa, Canadian). Ensure that R knows that this variable is a factor with the line

   ```
   > seeds$Type = factor(seeds$Type)
   ```

   (b) Produce a summary table of the dataset.

   (c) What proportion of observations have perimeter values greater than 15?

   (d) Which observation has the largest Asymmetry coefficient? To which class does this observation belong?

   (e) Produce a scatterplot matrix of all variables and note some relationships between them. Which attributes are highly related? Which attributes do a good job of distinguishing type?

   (f) Select two or three attributes and produce boxplots of these attributes vs. Type (don't forget axis labels!). What two or three variables might you would want to use as predictor variables in a model for kernel type based on these plots. Why?

   (g) Augment your pairs() function from part (e) to give a different color to each type.
   Hint: col="green" would make all the points green, so col = [variable] will color them according to a variable. What additional information does this figure give you in building a model for seed type?

4. JWHT (2013), Chap 3, Prob. 5
5. JWHT (2013), Chap 3, Prob. 15

   This problem involves the `Boston` data set, which we saw in the lab for this chapter. We will now try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

(a) For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

(b) Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$?

(c) How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis.

(d) Is there evidence of non-linear association between any of the predictors and the response? To answer this question, for each predictor X, fit a model of the form

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon.$$

6. JWHT (2013), Chap 4, Prob 3

7. TUMOR ANALYSIS: The file tumor.csv was created from data compiled in the mid 1990s. Each record was generated from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. It is of interest to classify the mass as benign or malignant based on a number of features which describe the mass. The columns of the dataset are as follows:

- Radius (mean of distances from center to points on the perimeter)
- Texture (standard deviation of gray-scale values)
- Perimeter
- Area
- smoothness (local variation in radius lengths)
- Compactness ($perimeter^2/area - 1.0$)
- Concavity (severity of concave portions of the contour)
- Concave points (number of concave portions of the contour)
- Symmetry
- Fractal dimension ("coastline approximation" - 1)

(a) Explore the data graphically in order to investigate the association between Diagnosis and the other features. Which features seem useful in predicting Diagnosis? Are any features highly related to each other? Describe your findings.

(b) Split the data into a 90% training set and a 10% test set, being sure to set a seed of 1 for consistency. How many rows in the test set?

(c) Using the training data, fit a logistic regression model predicting the probability of a malignant tumor using Radius, Symmetry, and Concave.Points as predictor variables. Using 0.5 as the decision threshold, compute the confusion matrix and the overall fraction of correct predictions for the held out data. Report the misclassification rate of the model.

(d) Repeat part (e) but decrease the threshold to 0.25 for classification and discuss the differences. What is the misclassification rate for this threshold?

(e) Perform LDA on the training data to predict Diagnosis based on the same variables used in part (d). What is the test error?

(f) Perform QDA on the training data to predict Diagnosis based on the same variables used in part (d). What is the test error?

(g) Perform KNN on the training data using several values of K to predict Diagnosis based on the same variables used in (d). Report test errors and which value of K works best for these data.

8. JWHT (2013), Chap 4, Prob 13

Additional Useful Exercises: (go through these but don't submit answers)

- JWHT (2013), Chap 2., Prob. 8, 9, 10
- JWHT (2013), Chap 3, Prob 8, 9, 10
- JWHT (2013), Chap 4, Prob 11