**STAT 4533/5533 - Homework 1 - Due February 2, 2024**

*Instructions: Everyone must answer Questions 1, 2, and 3. Graduate students are also required to answer Questions 4 and 5. Questions 6 - 10 are optional and will not be graded. Solutions must be created using R Markdown, and you will upload your HTML file to D2L to submit your answers. Your markdown document must be logically organized and easy to read. Do not print entire data sets in your output.*

1. The following table contains a training set of five observations with two quantitative predictor variables and one categorical response variable that will be used for classification with $k$-nearest neighbors:

| $X_1$ | $X_2$ | $Y$ |
|---|---|---|
| 4 | 4 | Yes |
| 1 | 8 | No |
| 7 | 5 | No |
| 8 | 8 | Yes |
| 5 | 9 | Yes |

The new test observation that we want to classify has $X_1 = 8, X_2 = 6$.

(a) Calculate the distance from each of the five training observations to the test point.

(b) Which three of the training observations are closest to the test point?

(c) Using your answer in part (b), what is the predicted value of $Y$ for the test point?

2. This question uses the *vgsales.csv* file found on D2L. The data set contains observations for video game sales and was derived from `https://www.kaggle.com/gregorut/videogamesales`.

(a) How many total observations are in the data set? How many of the games are in the sports genre?

(b) Find the minimum, maximum, mean, median, and variance of the North American sales variable.

(c) Create a histogram of the natural logarithm of the North American sales variable.

(d) Create pairwise scatterplots of the three sales variables and comment on the results.

3. This question uses the *BP.csv* data found on D2L. It contains measurements from a medical study evaluating factors that influence blood pressure. Variables included are sex (coded numerically with 1 = male and 0 = female), age in years, height in inches, weight in pounds, race (coded numerically with 2 = Asian, 3 = Black, 4 = Hispanic, and 5 = Caucasian), blood pressure with a cold compress, blood pressure while performing mental arithmetic, resting blood pressure, and a family history of hypertension indicator (1 = yes and 0 = no).

   (a) Fit a multiple linear regression model to predict resting blood pressure using age, height, weight, and race. What is the resulting regression equation?

   (b) What is your interpretation of the regression coefficient for height?

   (c) Which of the variables are significant at $\alpha = 0.1$?

   (d) How well does this model fit the data? Provide a relevant statistic and interpretation.

   (e) Predict the resting blood pressure for a 20-year-old Hispanic person who is 70 inches tall and weighs 160 pounds.

   (f) Fit a new multiple regression model that also includes an interaction between age and weight. Does this new model improve fit? Is the interaction term significant?

   (g) Investigate an additional multiple regression model to predict resting blood pressure that includes at least one transformation of the variables (such as $log(X)$, $\sqrt{X}, X^2$) and comment on your results. You are free to use any of the variables in the data set. Can you improve upon the results from part (a)?

4. GRADUATE STUDENTS ONLY: Chapter 2 Exercise 6

5. GRADUATE STUDENTS ONLY: Chapter 3 Exercise 13 (use your student number in the `set.seed()` function)

   *The following questions are recommended, but will not be graded:*

6. Chapter 2 Exercise 8

7. Chapter 2 Exercise 9

8. Chapter 2 Exercise 10

9. Chapter 3 Exercise 9 not part (d)

10. Chapter 3 Exercise 10 not part (h)