**ISQA 8080 Assignment 1 Due: By Tuesday, Sep. 24 2019, 5:30 PM (see Canvas for potential changes of the due date)**

**Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**NOTES:**

1. Use R for the calculations and implementation.
2. You need to submit this answer sheet and your R code.
3. Submit all documents in a zip file and upload it to Canvas. Name your Zip Folder with your name, A1, and the course # (Example: LastName-A1-ISQA 8080).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. **Regression Analysis (60 points)**

This question should be answered using the Carseats data set that you have available in Canvas. You can find a detailed description of the data variables by looking up the help function about the data set in the package ISLR: (library(ISLR) ?Carseats)

1. Fit a multiple regression model to predict Sales, using Price, Urban, and US as predictor variables. Show the R output for the lm model here.
2. Write out the model in equation form (note that you have qualitative predictors in the model).
3. Provide an interpretation of each coefficient in the model. What does it tell you about the relationship between the target and each of the predictor variables?
4. For which of the predictors can you reject the null hypothesis H0: βj = 0?
5. Based on your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome (significance).
6. How well do the models in (a) and (e) fit the data? Use both R2 and MSE/RMSE for this.
7. Test the assumptions of the linear regression model, i.e., use the plot() function (or autoplot()) to test the constant variance and normality assumptions. Do they seem to be reasonable?
8. Using the model from (e), obtain 95% confidence intervals for the coefficient(s).
9. Using the model from (e), predict the Sales for following observation:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| CompPrice | Income | Advertising | Population | Price | ShelveLoc | Age | Education | Urban | US |
| 138 | 73 | 11 | 276 | 120 | Bad | 42 | 17 | Yes | Yes |

1. **Regression Analysis – Collinearity (40 points)**

This problem focuses on the *collinearity* problem and its potential effects on linear regression models.

1. Perform the following commands in R:

> set .seed (1)

> x1 <- runif (100)

> x2 <- 0.5\* x1 + rnorm (100) /10

> y <- 2 + 2\* x1 +0.3\* x2+rnorm (100)

The last line corresponds to creating a linear model in which y is

a function of x1 and x2. Write out the parametrized form of the linear model.

What are the regression (beta) coefficients?

1. What is the correlation coefficient between x1 and x2? Create a scatterplot displaying the relationship between the variables.
2. Using this data, fit a least squares regression to predict y using x1 and x2. Describe the results obtained. What are , , and ? How do these relate to the true *, ,* and ? Can you reject the null hypothesis *H*0 : = 0? How about the null hypothesis *H*0: = 0?
3. Now fit a least squares regression to predict y using only x1. Comment on your results. Can you reject the null hypothesis *H*0: = 0?
4. Now fit a least squares regression to predict y using only x2. Comment on your results. Can you reject the null hypothesis *H*0: = 0?
5. Do the results obtained in (c)–(e) contradict each other? If yes, explain why this happens in this example.
6. Now suppose we obtain one additional observation, which was unfortunately incorrectly measured.

> x1 <- c(x1 , 0.1)

> x2 <- c(x2 , 0.8)

> y <- c(y,6)

Re-fit the linear models from (c) to (e) using this new data. What effect does this new observation have on the each of the models?

In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers.