**Introduction to Statistical Learning**

**Assignment 2**

**Aishwarya Naidu Kotla – 16305800**

**GitHubLink -**

**Pranay Reddy Singireddy – 18222557**

**GitHubLink -**

**1. ISLR 2.4 Applied Problem 8 (page 54)**

**Synopsis:**

The scatterplots in the problem show  the aspects that may be related to student graduation rates. One of the plots there exists a weak positive relationship between the number of PhD-holding faculty and graduation rates. One more graph apparently shows that there is not proper relationship in between the number of faculty who has terminal degrees and the graduation rates.  The last graph shows that the graduation rates are decreased as the student-to-faculty ratio increases.  Lastly, an other statistic shows a significant correlation between instructional expenses per student and the graduation rates. With increase in expenditures there increases to higher rates of graduation.

**2. ISLR 2.4 Applied Problem 9 (page 56)**

**a) Which of the predictors are quantitative, and which are qualitative?**

Based on the information MPG, displacement, horsepower, weight, and acceleration comes under quantitative variables.  Based on the circumstances, we can use cylinders and the year to make both qualitative and quantitative predictions. Origin and Name are important qualitative factors. Origin represents cars nations for Example 1 representing America and 2 representing Europe. In this way it is used in numerical encoding.

**f) Suppose we wish to predict gas mileage (mpg) based on the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.**

Considering the scatter plots generated in part 5 which ties miles per gallon to the predictor's engine displacement, horsepower, car weight, and model year. Three parameters displacement, horsepower, car weight will be the most helpful in determining mpg, with the model year remaining less useful.

We can observe that there is an unsteady relation between fuel economy and inflation between the years 1970 and 1982. There is a good determined relation between low fuel efficiency and engine properties like displacement, horsepower, car weight.

Looking at the box plot, we can conclude that there is a correlation between a car's country of origin and its fuel efficiency. We can say the Japanese vehicles are the most efficient on average, followed by European vehicles and  American vehicles.

**3. ISLR 2.4 Applied Problem 10 (page 57)**

**Synopsis:**

One thing that clearly shows out from the numerical summary is that tracts with at least eight rooms per dwelling have lower crime rates, lower nitric oxides concentrations, lower proportions of Black residents and lower proportions of lower status residents when compared to the overall data set.

**4. ISLR 3.7 Applied Problem 8 (page 123)**

**Key points:**

In the Residuals vs. Fitted plot, the residuals have a noticeable U-shape, which is a shows of a data non-linearity. When combined with the plot from Part 2, we can draw the conclusion that the simple linear regression model does not fit. Moreover, when viewing the Residuals vs. Leverage plot, there are some high leverage points. The data set shows that after dropping the rows with missing value, the data set has 392 observations, giving an average leverage value of 2/3920.0051. The leverage points with high streamlined residual values which are greater than 2 is an issue for the simple linear regression model. There are some of the observations with a standardized residual value of 3 or higher, and they suggests that there is a chance of outliers if the data had not been non-linear.

**5. ISLR 3.7 Applied Problem 9 (page 123)**

**Key points:**

To begin with, none of the terms above order 2 (cubic, quartic, or quintic terms) have statistically significant p-values. The modified R2 value in the quadratic model has significantly decreased from 0.6872 to 0.6861 in the linear model. The anova() algorithm returns a p-value of 0.65, indicating that there isn't enough evidence to prove null hypothesis that the quintic model better fits than the quadratic model. Including these three pieces of data, incorporating terms more than order 2 does not improve the efficiency of model.

**6. ISLR 3.7 Applied Problem 10 (page 124)**

**b) Provide an interpretation of each coefficient in the model. Be careful -- some of the variables in the model are qualitative!**

The Price coefficient of -0.054459 indicates that raising the price of a car seat by $1 results in a loss of approximately 54.46 units in the model for the given area. Coming to the UrbanYes coefficient of -0.021916 indicates that the model predicts that for a given car seat price point and US value, urban areas will sell 22 less car seats on average than the non-urban areas. The coefficient of 1.200573 for USYes indicates that the model predicts that for a given car seat price level and Urban value, stores in the United States will purchase 1201 more car seats than stores outside the United States on average.

**c) Write out the model in equation form, being careful to handle the qualitative variables properly.**

The model's equation is as follows.

**Y^=13.043−0.054X1−0.022X2+1.200X3**

Here, y^ represents the estimated car seat sales in thousands of seats;

x1j is the price of the carseat in dollars in the jth store;

x2j and x3j are dummy variables to represent the jth store which is located in urban area and in the United States itself.

More specifically, x2j and x3j employ the following coding scheme.

x2j=1 , if the jth store is in an urban location

x2j=0 , if the jth store is not in an urban location

x3j=1 ,if the jth store is in the United States

x3j=0 , if the jth store is not in the United States

d) **For which of the predictors can you reject the null hypothesis**H0:βj=0H0:βj=0**?**

The p-values for the intercept, Price, and USYes are all close to zero which infers that the null hypothesis H0:j=0 is strongly rejected for those predictors. However, because UrbanYes has a p-value of 0.936, there is no reason to assume that it has non-zero co - efficient in the link among predictors and Sales.

**Is there evidence of outliers or high leverage observations in the model from Part (e) ?**

If we see the residuals vs. leverage plot for the model from part 5 that is developed in the part 6, we can conclude that there are lot many observations with standardized residuals which are near to 3 in the absolute value. These observations could be considered outliers. In the same plot, we can see a number of high high leverage sites with leverage values far greater than the average leverage of 3/400=0.0075, even though these high leverage observations are highly improbable to be outliers since their studentized residual values are less than 2.