**ISL – Assignment 2 (Lab)**

**RajendraKumar(16320872)**

**Github Link:** [**https://github.com/Rajendraganji/ISL-LAB**](https://github.com/Rajendraganji/ISL-LAB)

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

**Summary:**

Some of the elements that may be associated to student graduation rates can be seen in the scatterplots. According to one of the plots, there is a weak positive link between the number of PhD-holding faculty and graduation rates. Another graph appears to show that the number of faculty with terminal degrees and graduation rates are unrelated. The bottom graph shows that as the student-to-faculty ratio rises, graduation rates tend to fall. Finally, the other figure appears to indicate a clear relationship between instructional expense per student and graduation rates, with higher expenditures equating to higher graduation rates.

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

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

MPG, displacement, horsepower, weight, and acceleration are the quantitative variables. Depending on the situation, we may choose to use cylinders and the year as quantitative or qualitative predictions. Finally, name and origin are qualitative determinants. origin is a numerical encoding of a car's nation of origin, with 1 representing America, 2 representing Europe, and 3 representing Japan.

**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.**

Based on the scatter plots I created in part 5 that link miles per gallon to the predictor’s engine displacement, horsepower, car weight, and model year, it appears that the first three parameters will be the most useful in predicting mpg, with model year still being useful but less so. There is a definite link between increased engine displacement/horsepower/car weight and lower fuel efficiency. There is also a shaky link between fuel economy and inflation between 1970 and 1982.

Looking at the above box plot, we can see that there is a link between a car's place of origin and its fuel efficiency, with Japanese automobiles being the most efficient on average, followed by European cars, and finally American cars.

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

**Summary:**

From the numerical summary, one thing that stands out is that the tracts which average at least eight rooms per dwelling have low crime rates, low concentrations of nitric oxides, low proportions of Black residents (high values of B), and low proportions of lower status residents compared to the overall data set.

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

**Comments:**

The residuals have an obvious U-shape in the Residuals vs. Fitted plot, which is a strong indicator of non-linearity in the data. When paired with the plot from Part 2, we can conclude that the simple linear regression model is not a good fit. Furthermore, when looking at the Residuals vs. Leverage plot, there are some high leverage points (remember that after dropping the rows with null values, the data set has 392 observations, giving an average leverage value of 2/3920.0051) that also have high standardized residual values (greater than 2), which is a concern for the simple linear regression model. There are also a handful of observations with a standardized residual value of 3 or greater, which suggests that they could be outliers if the data wasn't already non-linear.

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

**Comments:**

First, none of the terms above order 2 (cubic, quartic, and quintic terms) exhibit statistically significant p-values. Furthermore, the modified R2 value has decreased significantly from 0.6872 in the quadratic model to 0.6861 in the linear model. Finally, the anova() algorithm returns a p-value of 0.65, indicating that there is insufficient evidence to reject the null hypothesis that the quintic model fits better than the quadratic model. Adding terms beyond order 2 does not improve the model, according to these three pieces of data.

**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 increasing the price of a car seat by $1 leads in a loss in sales of around 54.46 units in the model for a given area (i.e. fixed values of Urban and US). The UrbanYes coefficient of -0.021916 indicates that, for a given carseat price point and US value, the model predicts that urban areas will sell 22 fewer carseats on average than non-urban areas. The coefficient of 1.200573 for USYes indicates that the model forecasts that for a given carseat price point and Urban value, stores in the United States will sell 1201 more carseats on average than stores outside the United States.

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

The model has the following equation.

Y^=13.043−0.054X1−0.022X2+1.200X3

Here, y^ is the estimated carseat sales, in thousands of car seats; x1j is the price of the carseat at the jth store, in dollars; and x2j and x3j are dummy variables to represent whether the jth store at is located in an urban area and in the United States, respectively. More concretely, x2j and x3j use 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 almost zero, indicating that the null hypothesis H0:βj=0 for those predictors is strongly rejected. However, because the p-value for UrbanYes is 0.936, there is no reason to believe that it has a non-zero coefficient in the genuine link between the predictors and Sales.

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

We can see that there are several observations with standardized residuals close to 3 in absolute value when we look at the residuals vs. leverage plot for the model from Part 5 that I developed in Part 6. Those observations could be regarded as outliers. In the same plot, we can observe that there are a number of high leverage sites with leverage values far higher than the average leverage of 3/400=0.0075, albeit those high leverage observations are unlikely to be outliers because their studentized residual values are less than 2.