CS 422 - Data Mining - Homework 1

Imaduddin Sheikh - isheikh@hawk.iit.edu

Due Date: **2/16/2022 11:59:59 PM**

---

1 Exercises

1.1 ISLR 2e (Gareth James, et al.), Chapter 3

Q.6)

The simple linear regression model uses the form of $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. The error function is defined as the sum of square error of each data point. The formula is given below,

$$E(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

Using the error function $E(\beta_0, \beta_1)$ we can get the $\hat{\beta}_0$ *and* $\hat{\beta}_1$ are known to be as follows,

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

If we want to predict an estimated value of $\hat{y}$ if we provide the model with a predictor value x, our equation is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Upon plugging $\hat{\beta}_0$ in $\hat{y}$, we end up with

$$\hat{y} = (\bar{y} - \beta_1 \bar{x}) + \hat{\beta}_1 x = \bar{y} - \hat{\beta}_1 (x - \bar{x})$$

When we plug $x = \bar{x}$ into this regression equation, terms in the expression cancel each other out, and leave us with $\hat{y} = \bar{y}$.

$$\hat{y} = \bar{y} - \hat{\beta}_1 (\bar{x} - \bar{x}) = \bar{x}$$

This means that the least squares line will always pass through the point $\boldsymbol{(\bar{x}, \bar{y})}$.

1.1 ISLR 2e (Gareth James, et al.), Chapter 3

Q.1)

The null hypotheses deduced from Table 3.4 are that the advertising budgets allocated to 'TV', 'radio', or 'newspaper' have no impact on sales. Mathematically, this can be expressed as $H_0^{(1)} : \beta_1 = 0$, $H_0^{(2)} : \beta_2 = 0$ and $H_0^{(3)} : \beta_3 = 0$ for the respective predictors. The p-values that correspond to 'TV', and 'radio' show high significance, compared to the p-value of 'newspaper'. Hence, we reject $H_0^{(1)}$, $H_0^{(2)}$ that correspond to the null hypotheses of 'TV', and 'radio' but not $H_0^{(3)}$(null hypothesis of 'newspaper') with the conclusion that changing advertising budget for newspapers do not affect sales.

Q.3)

a) (C) is correct. The least square line is given by

$$\hat{y} = 50 + 20GPA + 0.07IQ + 35Gender + 0.01GPA \times IQ - 10GPA \times Gender$$

Since $Gender = 0$ for males, the equation for males is given by

$$\hat{y} = 50 + 20GPA + 0.07IQ + 0.01GPA \times IQ,$$

Since $Gender = 1$ for females, the equation for females is given by

$$\hat{y} = 85 + 10GPA + 0.07IQ + 0.01GPA \times IQ.$$

When we take the constant terms

$$0.07IQ + 0.01GPA \times IQ$$

out of both the equations, we observe that for $GPA \geq 3.5$, we get the following inequality

$$50 + 20GPA \geq 85 + 10GPA,$$

explaining that the starting salary for males is higher than that for females on aver-

age.

b) Plugging the given values in the least square line gives

$$\hat{y} = 50 + 20(4.0) + 0.07(110) + 35(1) + 0.01(4.0 \times 110) - 10(4.0 \times 1) = 137.1$$

giving off a starting salary of \$137,100.

c) False. We need to test the hypothesis $H_0^{(4)} : \hat{\beta}_4 = 0$ and draw a conclusion by looking at the p-value associated with F statistic.

Q.4)

a) Since X and Y has a true linear relationship. We can anticipate that the least square line is closer to the true regression line, and so the RSS for the linear regression may be lower than that for the cubic regression. In the end, this is a vague assumption because we don't have much knowledge about our data.