HOMEWORK 3

STUDENT NAME HERE

2/10/2022

PROBLEM 1

Using (3.4) on p. 62, show that in the case of simple linear regression, the least squares line always passes through the point $(\bar{x}, \bar{y})$.

PROBLEM 2

Suppose we have a data set with five predictors, X_1 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = Interaction between GPA and Level. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 50$, $\hat{\beta}_1 = 20$, $\hat{\beta}_2 = 0.07$, $\hat{\beta}_3 = 35$, $\hat{\beta}_4 = 0.01$, $\hat{\beta}_5 =$ -10.

(a) Which answer is correct, and why?

i. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates.

ii. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates.

iii. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates provided that the GPA is high enough.

iv. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates provided that the GPA is high enough.

(b) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

(c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

PROBLEM 3

This question involves the use of simple linear regression on the `Auto` data set. Information can be found here: `https://rdrr.io/cran/ISLR/man/Auto.html`

(a) Read in the data and use the `lm()` function to perform a simple linear regression with `mpg` as the response and `horsepower` as the predictor. Use the `summary()` function to print the results.

i. Is there a relationship between the predictor and the response?

ii. How strong is the relationship between the predictor and the response?

iii. Is the relationship between the predictor and the response positive or negative?

iv. What is the predicted `mpg` associated with a `horsepower` of 98? What are the associated 95% confidence and prediction intervals?

(b) Plot the response and the predictor. Use the `abline()` function to display the least squares regression line.

(c) Use the `parfor()` and `plot()` functions to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit in terms of the appropriateness of a linear model, normality, heteroscedasticity, outliers, and high leverage observations.

PROBLEM 4

This question involves the use of multiple linear regression on the `Auto` data set.

(a) Create a new data frame called `Auto2` which includes all of the variables in the data set except for `name` (you learned to remove a variable in HWK2). Then produce a scatterplot matrix of `Auto2`.

(b) Compute the matrix of correlations between the variables in `Auto2` using the function `cor()`. You will need to exclude `origin`, which is qualitative.

(c) Recode `origin` as follows:

```
```{r}
library(dplyr)
Auto2$origin <- recode(Auto2$origin, "1" = "American", "2" = "European", "3" = "Japanese")
attach(Auto2)
```
```

Then use the `lm()` function to perform a multiple linear regression with `mpg` as the response and all other variables except `name` as the predictors. Use the `summary()` function to print the results. Comment on the output. For instance:

i. Is there a relationship between the predictors and the response?

ii. Which predictors appear to have a statistically significant relationship to the response?

iii. What does the coefficient for the year variable suggest?

(d) Use the `parfor()` and `plot()` functions to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit in terms of the appropriateness of a linear model, normality, heteroscedasticity, outliers, and high leverage observations.

(e) The plots seem to indicate heteroscedasticity. Try three different transformations of `mpg`, such as `log(mpg)`, `sqrt{mpg}`, `mpg^2` and determine which one seems best in eliminating this issue. Only show the results for the best of the three.

(f) Use the `*` or `:` symbols to fit a multiple linear regression model which includes the two most significant predictors from part (c) and their interaction effect. Comment on the statistical significance of the result.