Intro to Statistical Learning

Chapter 3 Solutions

Conceptual

1.

The null hypothesis is that there is no significant relationship on product sales with radio, TV, and newspaper advertising budgets. From Table 3.4, newspaper advertising budgets do not have a significant relationship with product sales, but radio and TV do have a significant relationship based on the p-values.

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

2. KNN classifiers are used to solve classification problems while KNN regression is used to solve regression problems. KNN classifiers identify the neighborhood x_0 and estimates the conditional probability $P(Y = j|X = x_0)$ for class j as a proportion of points in the neighborhood where the response values equal j. KNN regression identifies the neighborhood x_0 and then estimates $f(x_0)$ as the average of all training responses in the neighborhood.

3.

Least Squares Best Fit Line

$\hat{y}$ = 50 + 20GPA + .07IQ + 35Gender + .01GPA*IQ -10GPA*Gender

Males

$\hat{y}$ = 50 + 20GPA + .07IQ + .01GPA*IQ

Females

$\hat{y}$ = 85 + 10GPA + .07IQ + .01GPA*IQ

A) The solution is iii). For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough because of the following:

50 + 20 GPA ≥ 85 + 10GPA if and only if GPA ≥ 3.5.

B) For a female with an IQ of 110 and GPA of 4.0, predicted salary is $137,100.

$\hat{y}$ = 85 + 10(4.0) + .07(110) + .01(4.0*110) = 137.1

C) This is false. We want to test the null hypothesis that $\beta_4 = 0$, and find out the p-value associated with the t-statistic or F-statistic to determine if the interaction of GPA/IQ is significant with salary.

4.

A) $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

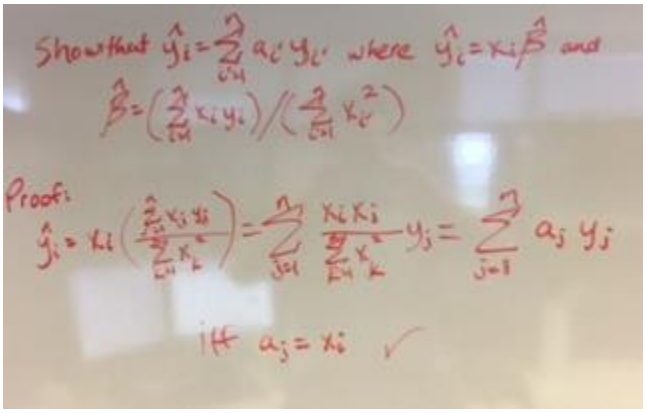
Since the relationship between X and Y is linear, RSS would be less in the linear regression model than the cubic regression model because there is less variability between the difference of the actual and the predicted values determined from the best fit line.

B) Test data is not available so there is no way to know for sure. If test data was available, RSS would be higher with the cubic regression model than the linear regression model.

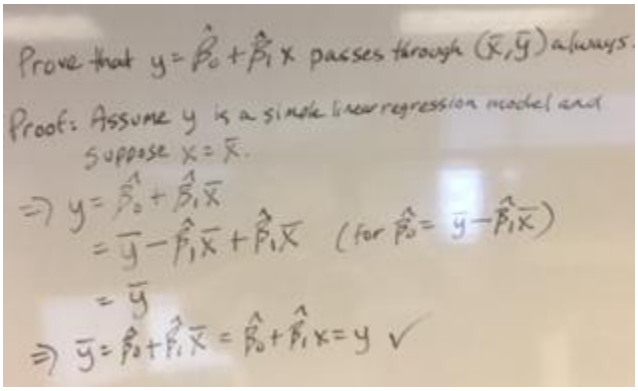
C) Figure 2.9 on p. 31 of Chapter 2 shows higher flexibility for the cubic regression model which results in the training RSS being less than the linear regression model.

D) Since test data is not available there is not enough information to conclude the test RSS being lower for either regression model.

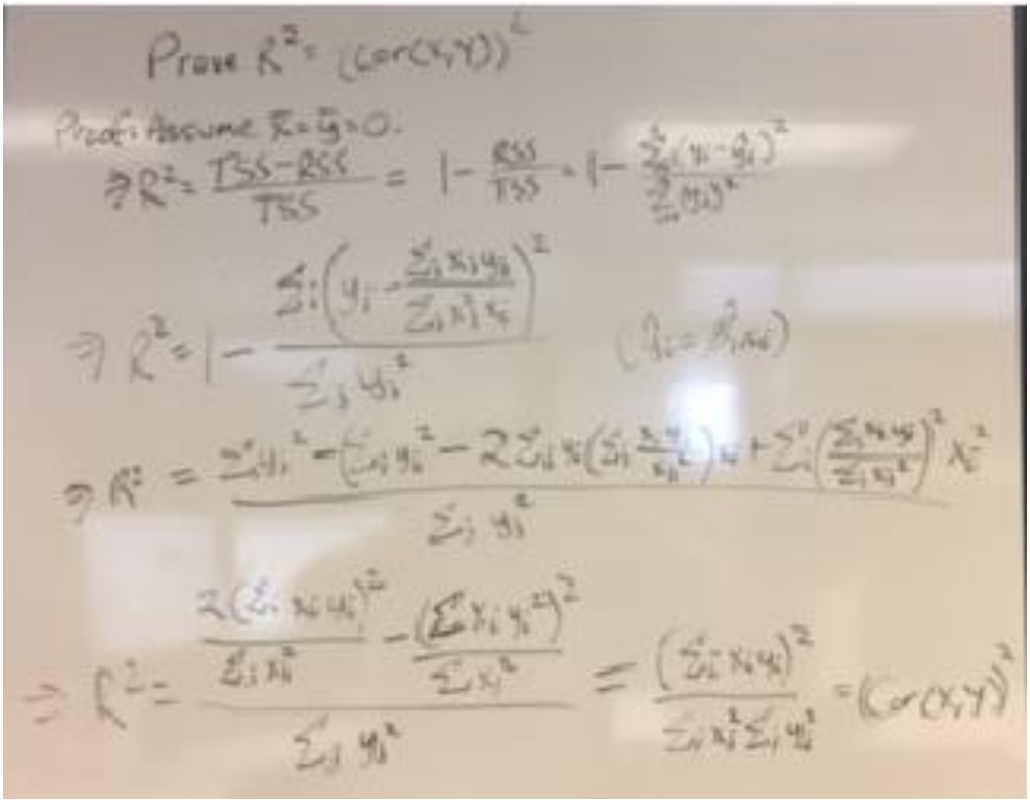
5.



6.


Prove that $y = \hat{\beta}_0 + \hat{\beta}_1 x$ passes through $(\bar{x}, \bar{y})$ always.
Proof: Assume y is a simple linear regression model and suppose $x = \bar{x}$.
$\Rightarrow y = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$
$= \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x}$ (for $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$)
$= \bar{y}$
$\Rightarrow \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \hat{\beta}_0 + \hat{\beta}_1 x = y$ ✓

7.


Prove $R^2 = (\text{Cor}(X,Y))^2$
Proof: Assume $\bar{x} = \bar{y} = 0$.
$\Rightarrow R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i)^2}$
$\Rightarrow R^2 = 1 - \frac{\sum_i \left(y_i - \frac{\sum_i x_i y_i}{\sum_i x_i^2} x_i\right)^2}{\sum_i y_i^2}$ $(\hat{y}_i = \hat{\beta}_1 x_i)$
$\Rightarrow R^2 = \frac{\sum_i y_i^2 - \sum_i y_i^2 - 2\sum_i x_i \left(\sum_i \frac{x_i y_i}{x_i^2}\right) x_i + \sum_i \left(\frac{\sum_i x_i y_i}{\sum_i x_i^2}\right)^2 x_i^2}{\sum_i y_i^2}$
$\Rightarrow R^2 = \frac{2\frac{(\sum_i x_i y_i)^2}{\sum_i x_i^2} - \frac{(\sum_i x_i y_i)^2}{\sum_i x_i^2}}{\sum_i y_i^2} = \frac{\left(\sum_i x_i y_i\right)^2}{\sum_i x_i^2 \sum_i y_i^2} = (\text{Cor}(X,Y))^2$

Applied

8.

A) i. The p-value of the F-statistic (p-value < 2.2e-16) shows strong evidence of a relationship between mpg and horsepower.

ii. Since $R^2 = 0.6059$, there is about 60.59% variability in mpg explained by horsepower.

iii. There is a negative relationship between mpg and horsepower because the coefficient for horsepower is negative.

iv. The predicted mpg with a horsepower of 98 is 24.467. The 95% confidence interval is (23.973, 24.961) and the 95% prediction interval is (14.809, 34.125).

C) The Residuals vs Fitted plot shows a nonlinear relationship between mpg and horsepower. There is also presence of some outliers from the Normal Q-Q plot for values of the Standardized Residuals that are greater than 2 and less than -2.

9.

C) i. Since the p-value corresponding to the F-statistic is significant (p-value < 2.2e-16), there is evidence of a relationship between mpg and the other predictors.

ii. From the p-values corresponding to the t-statistics, cylinders, horsepower, and acceleration are not statistically significant predictors of mpg.

iii. The coefficient for year states that for an increase in one year results in an increase in mpg by 0.75. This implies that cars are more fuel efficient by 0.75 mpg per year.

D) The Residuals vs Fitted plot shows a slight non linear relationship between mpg and the predictors. There is also presence of some outliers from the Normal Q-Q plot for values of the Standardized Residuals that are greater than 2 and less than -2. Also the Residuals vs Leverage plot shows one high leverage point (point 14).

E) From the p-values corresponding to the t-statistics, the interaction between displacement and weight is statistically significant but cylinders and displacement is not.

F) In the example of analyzing horsepower and mpg using the logarithmic, square root, and square transformation, the logarithmic transformation results in the most linear plot.

10.

B) The coefficient for Price means that for an increase in $1 in price, the unit sales decreases by 54,459 units. Coefficient for Urban means that unit sales in an urban location are 21,916 less than a rural location. The US variable coefficient means that unit sales for a US store are 1200.573 more than a non-US location.

C)

Sales = 13.043469 + (-0.0544588)*Price + (-0.0219162)*Urban + (1.2005727)*US + ε

For Urban = 1 if store is in an urban location (Urban = 0 for rural) and US = 1 for stores in the United States (US = 0 for non US stores).

D) Based off of the p-values corresponding to the t-statistics, we can reject the null hypothesis for Price and US.

F) Model (e) fits the data slightly better than model (a) from the Adjusted R^2.

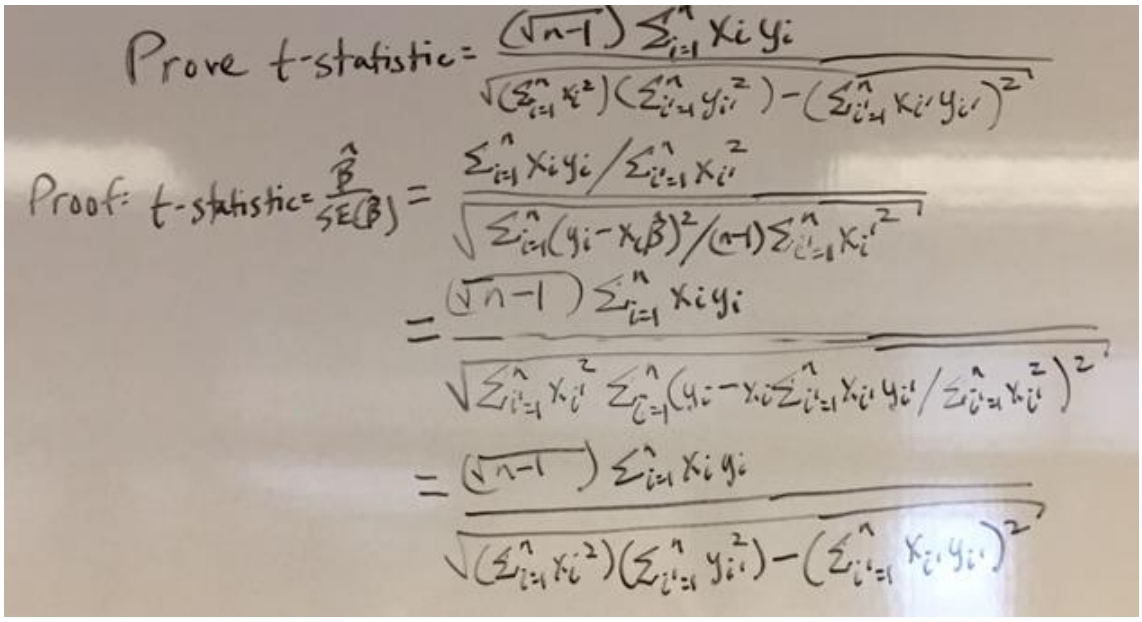
H) Based off of the Normal Q-Q Plot and Residuals vs Leverage Plot, there is evidence of outliers (Normal Q-Q Plot with standardized residuals less than -2 and greater than 2) and high leverage observations (standardized residuals greater than 0.01).

11.

A) The coefficient for x is 1.9939, standard error is 0.1065, t-statistic is 18.73 and the p-value is <2e-16. The null hypothesis can be rejected based off of the low p-value.

B) The coefficient for y is 0.39111, standard error is 0.02089, t-statistic is 18.73, and the p-value is <2e-16. The null hypothesis can be rejected based off of the low p-value.

C) Since the t-statistic and p-values are the same for A) and B), the relationship is a is the same line.

D) 

E) For D), if we replace x_i with y_i for the formula for the t-statistic, the result will be the same.

12.

A) For the regression of Y onto X, the coefficient estimate is:

$$\hat{\beta} = \frac{(\sum_{i=1}^{n} x_i y_i)}{\sum_{j=1}^{n} x_j^2}$$

The regression of X onto Y has a coefficient estimate:

$$\hat{\beta}' = \frac{(\sum_{i=1}^{n} x_i y_i)}{\sum_{j=1}^{n} y_j^2}$$

If and only if $\sum_{j=1}^{n} x_j^2 = \sum_{j=1}^{n} y_j^2$.

13.

D) The relationship between x and y are linear.

E) The coefficients from both models are close to one another. The null hypothesis can be rejected in the model because of the large F-statistic with a very low p-value.

G) Since the p-value is higher than 0.05, there is not sufficient evidence that the quadratic term improves the model fit.

14.

C) Since the p-value < 0.05 for β_1, we may reject H_0. We cannot reject H_0 for β_2 because the p-value > 0.05.

D) We may reject the null hypothesis since the p-value < 0.05.

E) We may reject the null hypothesis since the p-value < 0.05.