Solutions of the exercises from Chapter 3

Conceptual

Q1. Describe the null hypotheses to which the p-values given in Table 3.4 correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of sales, TV, radio, and newspaper, rather than in terms of the coefficients of the linear model.

The null hypotheses associated with table 3.4 are that advertising budgets of "TV", "radio" or "newspaper" do not have an effect on sales. More precisely $H_0^{(1)} : \beta_1 = 0$*,* $H_0^{(2)} : \beta_2 = 0$ *and* $H_0^{(3)} : \beta_3 = 0$*. The corresponding p-values are highly significant for "TV" and "radio" and not significant for "newspaper"; so we reject* $H_0^{(1)}$ *and* $H_0^{(2)}$ *and we do not reject* $H_0^{(3)}$*. We may conclude that newspaper advertising budget do not affect sales.*

Q2. Carefully explain the differences between the KNN classifier and KNN regression methods.

The KNN classifier is typically used to solve classification problems (those with a qualitative response) by identifying the neighborhood of x_0 *and then estimating the conditional probability* $P(Y = j|X = x_0)$ *for class* j *as the fraction of points in the neighborhood whose response values equal* j*. The KNN regression method is used to solve regression problems (those with a quantitative response) by again identifying the neighborhood of* x_0 *and then estimating* $f(x_0)$ *as the average of all the training responses in the neighborhood.*

Q3. Suppose we have a data set with five predictors, $X1 =$ GPA, $X2 =$ IQ, $X3 =$ Gender (1 for Female and 0 for Male), $X4 =$ Interaction between GPA and IQ, and $X5 =$ Interaction between GPA and Gender. 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, males earn more on average than females.
ii. For a fixed value of IQ and GPA, females earn more on average than males.
iii. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.
iv. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

The least square line is given by

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

which becomes for the males

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

and for the females

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

So the starting salary for males is higher than for females on average iff $50 + 20GPA \geq 85 + 10GPA$ *which is equivalent to* $GPA \geq 3.5$*. Therefore iii. is the right answer.*

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

It suffices to plug in the given values in the least square line for females given above and we obtain

$$\hat{y} = 85 + 40 + 7.7 + 4.4 = 137.1,$$

which gives us a starting salary of 137100\$.

(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.

False. To verify if the GPA/IQ has an impact on the quality of the model we need to test the hypothesis $H_0 : \hat{\beta}_4 = 0$ *and look at the p-value associated with the* t *or the* F *statistic to draw a conclusion.*

Q4. I collect a set of data (n = 100 observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$.

(a) Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \varepsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

Without knowing more details about the training data, we are not able to know which training RSS is lower between linear or cubic. Essentialy we are using a wrong model, so our results may show some inconsistence.

(b) Answer (a) using test rather than training RSS.

In this case the test RSS depends upon the test data, so we have not enough information to conclude.

(c) Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

Here also, there is not enough information to tell. If the actual data is pretty linear, then a linear regression may yield better results than a cubic regression. However, if the actual data is more cubic, then cubic regression may be the better option. Whichever one is the better option will yield lower RSS than the other.

(d) Answer (c) using test rather than training RSS.

See answer for (c). above.

Q5. Consider the fitted values that result from performing linear regression without an intercept. In this setting, the i-th fitted value takes the form $\hat{y_i} = x_i \hat{\beta}$, where

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{k=1}^{n} x_k^2}.$$

Show that we can write

$$\hat{y}_i = \sum_{j=1}^{n} a_j y_j.$$

What is a_j ?

We have immediately that

$$\hat{y}_i = x_i \frac{\sum_{j=1}^n x_j y_j}{\sum_{k=1}^n x_k^2} = \sum_{j=1}^n \frac{x_i x_j}{\sum_{k=1}^n x_k^2} y_j = \sum_{j=1}^n a_j y_j.$$

Q6. Using (3.4), argue that in the case of simple linear regression, the least squares line always passes through the point $(\overline{x}, \overline{y})$.

The least square line equation is $y = \hat{\beta}_0 + \hat{\beta}_1 x$, so if we substitute $\overline{x}$ for x we obtain

$$y = \hat{\beta}_0 + \hat{\beta}_1 \overline{x} = \overline{y} - \hat{\beta}_1 \overline{x} + \hat{\beta}_1 \overline{x} = \overline{y}.$$

We may conclude that the least square line passes through the point $(\overline{x}, \overline{y})$.

Q7. It is claimed in the text that in the case of simple linear regression of Y onto X, the R^2 statistic (3.17) is equal to the square of the correlation between X and Y (3.18). Prove that this is the case. For simplicity, you may assume that $\overline{x} = \overline{y} = 0$.

We have the following equalities

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_j y_j^2};$$

with $\hat{y}_i = \hat{\beta}_1 x_i$ we may write

$$R^2 = 1 - \frac{\sum_i (y_i - \sum_j x_j y_j / \sum_j x_j^2 x_i)^2}{\sum_j y_j^2} = \frac{\sum_j y_j^2 - (\sum_i y_i^2 - 2 \sum_i y_i (\sum_j x_j y_j / \sum_j x_j^2) x_i + \sum_i (\sum_j x_j y_j / \sum_j x_j^2)^2 x_i^2)}{\sum_j y_j^2}$$

and finally

$$R^2 = \frac{2(\sum_i x_i y_i)^2 / \sum_j x_j^2 - (\sum_i x_i y_i)^2 / \sum_j x_j^2}{\sum_j y_j^2} = \frac{(\sum_i x_i y_i)^2}{\sum_j x_j^2 \sum_j y_j^2} = Cor(X, Y)^2.$$