CONCEPTUAL QUESTIONS

CHAPTER - 3

1. 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.

Roughly speaking, p value is the probability that the given result is true, considering the null hypothesis. If the p value is very less, we say that the p value is significant and we reject the null hypothesis. In a given linear model, null hypothesis is that there is no relationship between a predictor and response, and there low or significant p means that we reject null hypothesis and that there is a relationship between predictor and response.

In the given table. Table 3.4, the p value of TV, radio and newspaper are <0.001,<0.001, and 0.8599 respectively. P-values for TV and radio and significant, and hence there is a relationship between TV and sales, and , radio and sales, However there is a very large p value for newspaper, and hence we can conclude that there is no relationship between newspaper and radio.

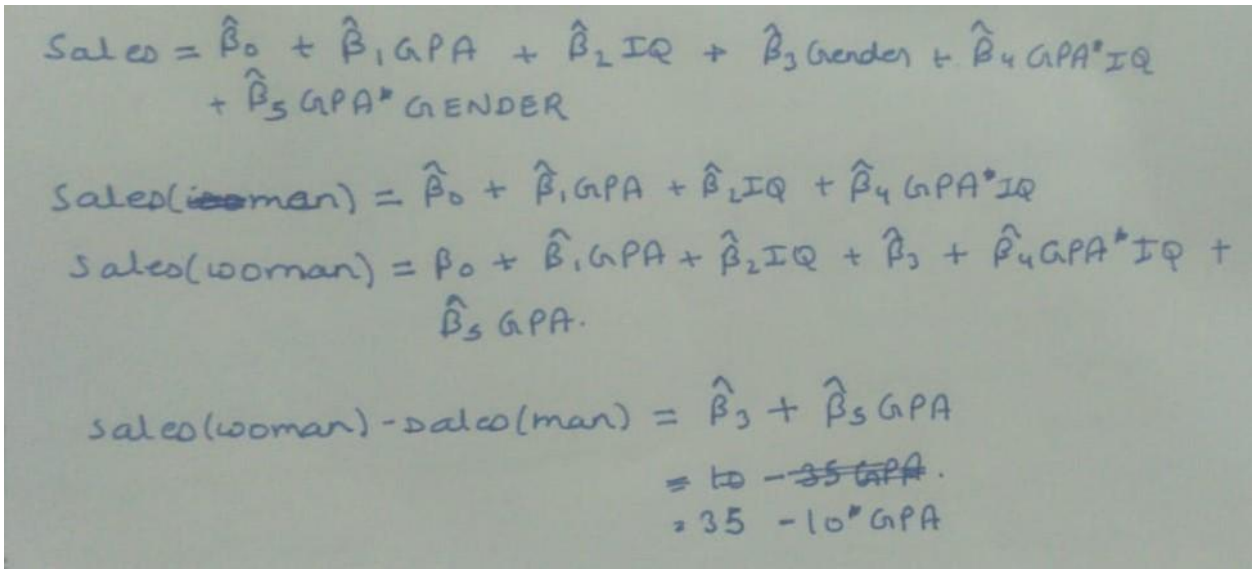
2. Carefully explain the differences between the KNN classifier and KNN regression methods.

KNN classifier is used for classification and for a given input Xo, it selects the k points which are closest Xo, and outputs the label contained by the maximum of the k points. In KNN regression, which is used to predict quantative values, it selects k points closest to the input Xo, and returns the average of all the values.

3. 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?

(iii)



Therefore, for fixed value of IQ and GPA, if the value of (10 – 35*GPA)>0, than sales of male will be greater than female, ie, for GPA > 3.5, sales of male is larger than sale of female.

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

Sales = 50 + 20*GPA + 0.007*IQ + 35*GENDER + 0.01(GPA*IQ) – 10(GPA*GENDER)

Python code

```
x = [4,110,1]

Sales = 50 + x[0]*20 + x[1]*0.07 + x[2]*35 + (x[0]*x[1])*0.01 + (x[0]*x[2])*(-10)

Print(Sales)
```

137.1

(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. For calculating the significance the value of coefficient is not the best choice. Hence we can't conclude that it is there is little evidence of an interaction effect.

4. 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 = β0 + β1X + β2X2 + β3X3 + .

(a) Suppose that the true relationship between X and Y is linear, i.e. Y = β0 + β1X + . 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.

For training data, RSS decreases as we increase model complexity. The model will overtry to find the patterns and will overfir the data. So, RSS for cubic regression will be lower than RSS of linear regression

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

In the case of test data, for cubic regression which is overfitted, will produce results far from the true values, whereas linear regression, which is close to the true function will perform better on the test data. And hence RSS will be lower on test data for linear regression

(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.

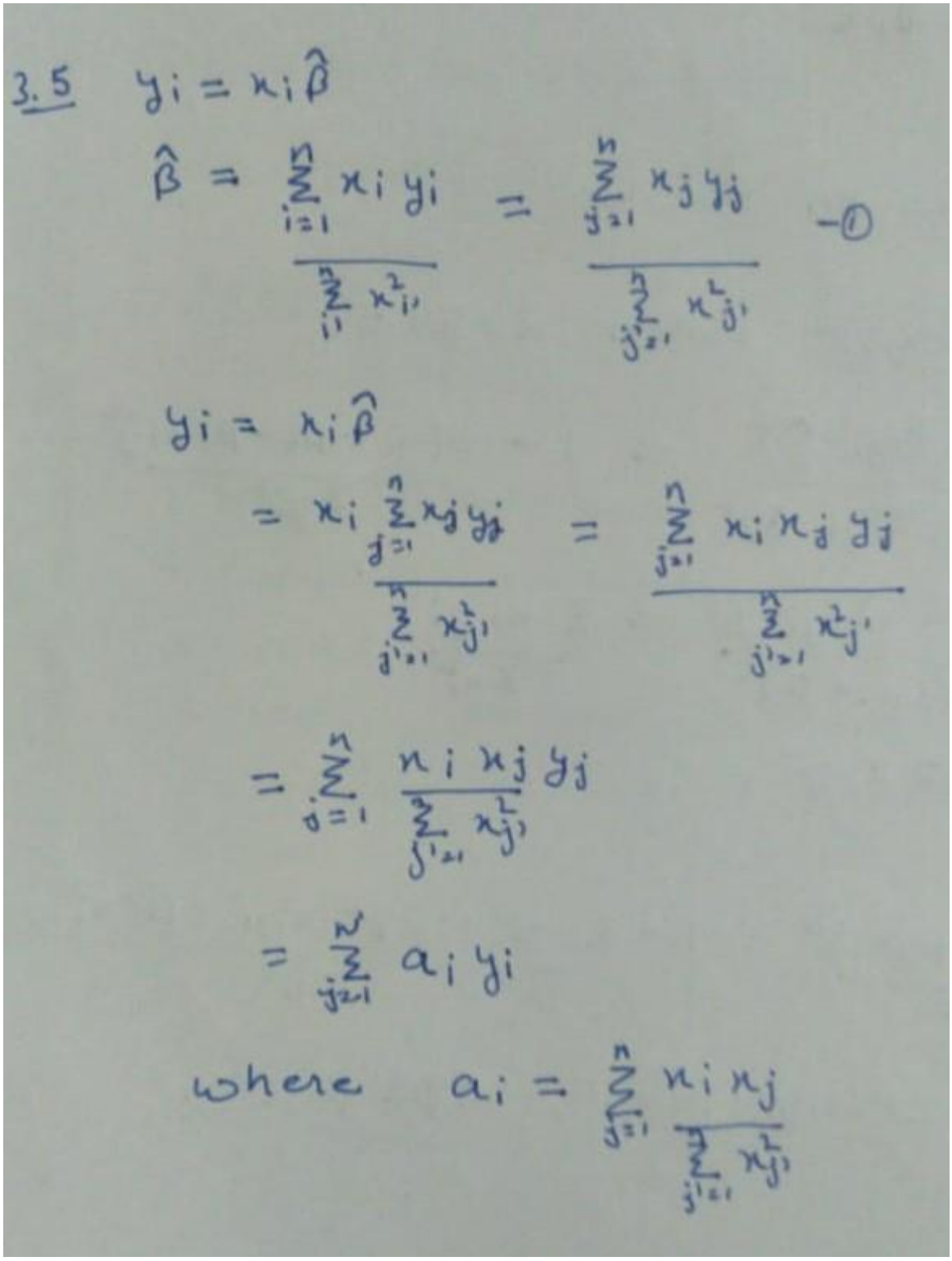
Since, training RSS decreases with increasing complexity, we will observe lower RSS for cubic regression.

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

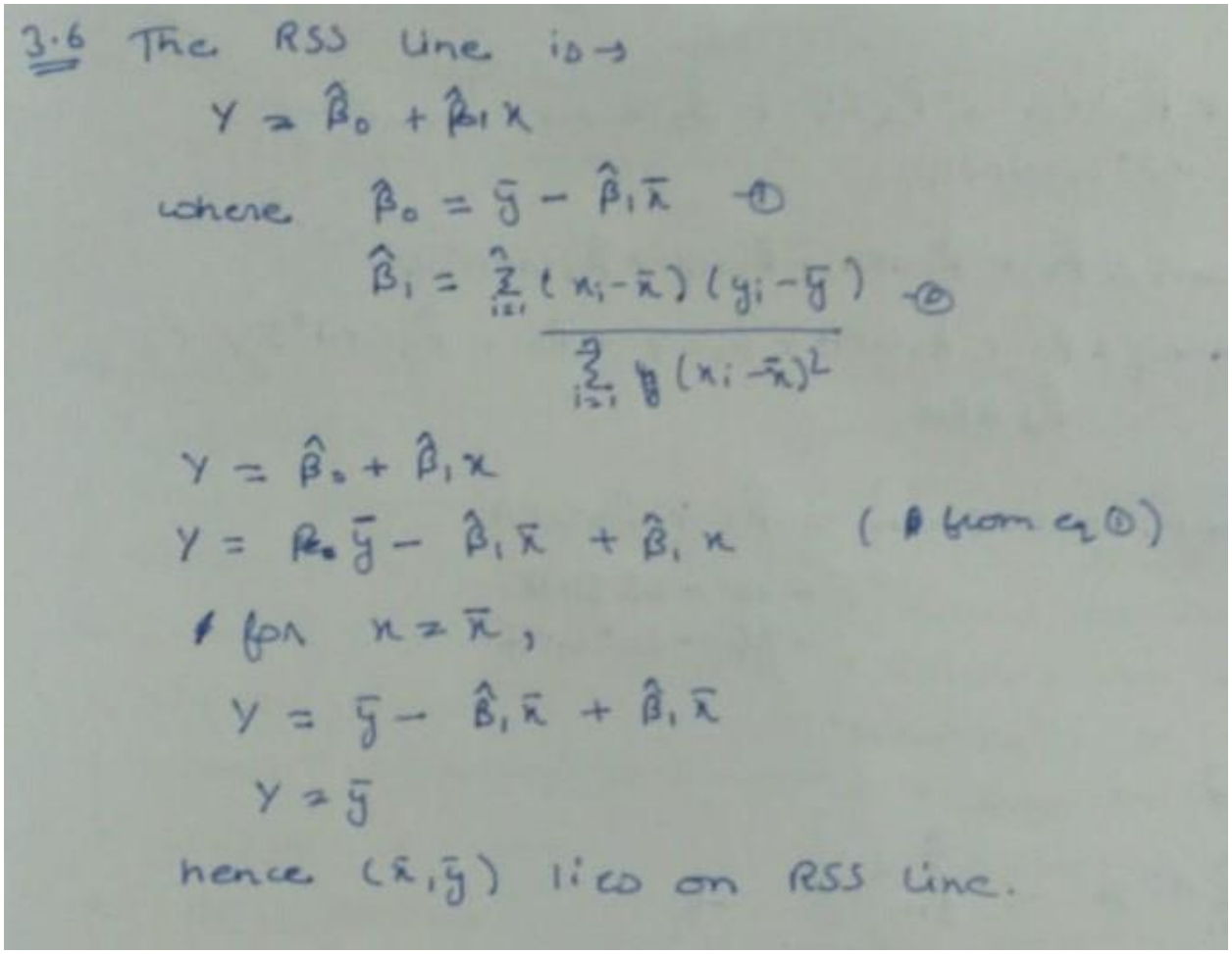
We can't predict whether RSS for test data will be lower for cubic regression or linear regression, since we have no clue what the true function is like.

We can find out the value of RSS for test data for cubic regression and linear regression, If test RSS of cubic is less than test RSS of linear regression, than we can say that true relationship is more non-linear.

5. Consider the fitted values that result from performing linear regression without an intercept. In this setting, the ith fitted value takes the form yˆi = xiβ, ˆ where βˆ = n i=1 xiyi / n i=1 x2 i . (3.38) Show that we can write yˆi = n i=1 aiyi . What is ai?



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



7. It is claimed in the text that in the case of simple linear regression of Y onto X, the R2 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 ¯x = ¯y = 0.

3.7 To prove: $R^2 = \text{cor}(x,y)^2$ when $\bar{x} = \bar{y} = 0$.

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\Sigma(y_i - \hat{y})^2}{\Sigma(y_i - \bar{y})^2} = 1 - \frac{\Sigma(y_i - \beta_0 - \beta_1 x_i)}{\Sigma y_i^2}$$

where, $\beta_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{\Sigma x_i y_i}{\Sigma x_i^2}$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} = 0,$$

hence,

$$R^2 = 1 - \frac{\Sigma(y_i - \beta_1 x_i)^2}{\Sigma y_i^2} = 1 - \frac{\Sigma(y_i^2 - 2\beta_1 x_i y_i + \beta_1^2 x}{\Sigma y_i^2}$$

$$R^2 = \frac{\Sigma y_i^2 - \Sigma y_i^2 + 2\beta\Sigma x_i y_i - \beta_1^2 \Sigma x_i^2}{\Sigma y_i^2} = \frac{2\Sigma x_i y_i \Sigma x_i y_i}{}$$

$$R^2 = \frac{2\Sigma x_i y_i}{\Sigma x_i^2} \cdot \Sigma x_i y_i - \frac{(\Sigma x_i y_i)^2}{(\Sigma x_i)^2} (\Sigma x_i)^2$$

$$\Sigma y_i^2$$

$$R^2 = \frac{\frac{2(\sum x_i y_i)^2}{\sum x_i^2} - \frac{\sum (x_i y_i)^2}{\sum x_i^2}}{\sum y_i^2}$$

$$R^2 = \frac{\sum (x_i y_i)^2}{\sum x_i^2 \sum y_i^2} = \frac{\sum (x_i y_i)^2}{\sum x_i^2 \times \sum y_i^2} \quad -①$$

$$corr(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 . \sum (y_i - \bar{y})^2}}$$

$$= \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 . \sum y_i^2}}$$

$$CORR(x, y)^2 = \frac{\sum (x_i y_i)^2}{\sum x_i^2 . \sum y_i^2} \quad -②$$

from ① & ②

$$R^2 = CORR(x, y)^2$$