**P1:** 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.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Coefficient | Std. error | t-statistic | p-value |
| Intercept | 2.939 | 0.3119 | 9.42 | < 0.0001 |
| TV | 0.046 | 0.0014 | 32.81 | < 0.0001 |
| radio | 0.189 | 0.0086 | 21.89 | < 0.0001 |
| newspaper | -0.001 | 0.0059 | -0.18 | 0.8599 |

**Ans:** In terms of sales, TV, radio, and newspaper, the associated null hypotheses are that none of these predictors for advertising budget have any effect on sales. Here, the linear model is constructed as follows:

where is the error term approximately following a normal distribution () with zero mean () and constant variance ().

The corresponding p-values for TV and radio show that both of these are highly significant whereas, newspaper has no significance for predicting sales. Therefore, we can drop newspaper from the advertisement budget analysis and reconstruct the linear model with updated coefficients.

**P2:** Carefully explain the differences between the KNN classifier and KNN regression methods.

**Ans:** The KNN classifier solves a classification problem (qualitative response) by assigning the most frequent class among the -nearest neighbors present in the locality of the object in consideration. In short, the output of a KNN classifier belongs to a discrete domain, e.g., {0, 1}.

The KNN regression method is used for solving regression problems where the output belongs to a continuous domain, e.g., . The process is similar to KNN classification, except that the prediction is performed by taking the average of all the values of the -nearest neighbors.

**P3:** Suppose we have a data set with five predictors, =GPA, = IQ, = Gender (1 for Female and 0 for Male), = Interaction between GPA and IQ, and = 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 = 50, = 20, = 0.07, = 35, = 0.01, = −10.

1. Which answer is correct, and why?
   1. For a fixed value of IQ and GPA, males earn more on average than females.
   2. For a fixed value of IQ and GPA, females earn more on average than males.
   3. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.
   4. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

(b) Predict the salary of a female 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.

**Ans:**

The fitted linear model is given by

For males, the fitted model becomes:

For females, the fitted model becomes:

So, for a fixed value of IQ and GPA, , iff, . Therefore, **iii** is the correct answer, i.e., males earn more on average than females provided that the GPA is high enough.

1. $137.1k = **$137100**
2. This is a **false** statement. We have to perform hypothesis testing and observe the p-value to claim whether GPA/IQ interaction is insignificant.

**P7:** It is claimed in the text that in the case of simple linear regression of onto , the 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 .

**Ans:** Accordingly, , and are the actual and predicted values. From OLS theory and general statistics, we have the following formulas:

Correlation coefficient =

Coefficient of determination =

Total Sum of Squares =

Explained Sum of Squares =

Residual Sum of Squares =

Now,

|  |  |
| --- | --- |
| ∴ | **[proved]** |

**P13:** Plots for 13 (d) and (f)

|  |  |
| --- | --- |
|  |  |

Variance = 0.25

|  |  |
| --- | --- |
|  |  |

Variance = 0.015625

|  |  |
| --- | --- |
|  |  |

Variance = 0.7