Exercises Linear Regression

Exercise 1

Question 1

- Describe the null hypotheses to which the p-values given in Table . 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

Answer :

The **Null hypothesis** for each variable is

$$H_0 : \text{TV=Radio=Newspaper=0}$$

There is no association between the mediums and **Sales**

- The TV and $Radio\ p-values$ are both extremely low, means a **strong evidence to reject the null hypothesis** which indicate TV and $Radio$ advertising are **Significantly associated with sales**
- For $Newspaper$ the $p-value$ is very high , means we **fail to reject the null hypothesis** which indicate that the $Newspaper$ doesn't affect **sales**

Question 2

- Carefully explain the differences between the KNN classifier and KNN regression methods

Answer

- **KNN** classifier is a method to predict the class of a **qualitative** response, **approximates** the Bayes classifier, it finds K nearest Training Data points and assigns the most common class label among them(majority vote)
- **KNN** regression its a method to predict **quantitative** responses its takes the average of the responses of the K nearest neighbors to predict the value of the observation

Question 3

- Suppose we have a data set with five predictors,$X_1 = \text{GPA}$, $X_2 = \text{IQ}$
, $X_3 =$ **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

$$\hat{\beta}_0 = 50, \hat{\beta}_1 = 20, \hat{\beta}_2 = 0.07, \hat{\beta}_3 = 35, \hat{\beta}_4 = 0.01, \hat{\beta}_5 = -10$$

Answer

- The **College graduates always earn more** than high school grads,provided that the **GPA** is high enough

$$\text{Salary}_{college} = 50 + 20 \times \text{GPA} + 0.07 \times \text{IQ} + 35 + 0.01 \times (\text{GPA} \times \text{IQ}) - 10 \times (\text{GPA})$$

$$\text{Salary}_{college} = 50 + 10 \times \text{GPA} + 0.07 \times \text{IQ} + 35 + 0.01 \times (\text{GPA} \times \text{IQ})$$

$$\text{Salary}_{highschool} = 50 + 20 \times \text{GPA} + 20 \times \text{IQ} + 0.01 \times (\text{GPA} \times \text{IQ})$$

$$\text{Salary}_{college-highschool} = 35 - 10 \times \text{GPA}$$

- If **GPA** is high enough **College grads** have higher salary
- Predicted **Salary** for a college grad with **IQ** of 110 and **GPA** of 4.0

$$\text{Salary} = 50 + 10.4 + 0.07.110 + 35 + 0.01 \times (4.110)$$

$$\text{Salary} = 137.1$$

- Since the **coefficient** for **GPA/IQ** interaction term us very small, there is very little evidence of an interaction effect
  - **False** Having small coefficient for the interaction term doesn't automatically imply very little interaction effect
  - $p-value$ should be taken into consideration

- A small coefficient can still be **statistically significant** if the variance is low
- Also the interaction term might have a practical impact

Question 3

- I collect a set of data (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 and 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?
- b. Answer(a) using test rather than training RSS
- 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
- d. Answer (c) using test rather than training RSS.
  Answer
- We expect the **cubic regression** to have lower RSS cause its more **flexible** than the simple linear regression model(It fits better than the more "Linear" model)
- The test for the **cubic regression** model would likely be higher due to it overfitting the Training Data
- We expect the **Cubic Regression** model to have lower RSS cause its more **flexible** than the simple linear regression model
- There is not enough information to tell , it depends on how non-linear the true relationship is

Question 5

Consider the fitted linear model without an intercept

$$\hat{y} = x_i \hat{\beta}$$

Where :

$$\hat{\beta} = \frac{\left(\sum_{i=1}^{n} x_i y_i \right)}{\left(\sum_{i'}^{n} x_{i'}^2 \right)}$$

Show that :

$$\hat{y}_i = \sum_{i'=1}^{n} a_{i'} y_{i'}$$

Answer :

$$\begin{aligned} \hat{y}_i &= x_i \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i'}^{n} x_{i'}^2} \\ &= x_i \frac{\sum_{i'=1}^{n} x_{i'} y_{i'}}{\sum_{i''=1}^{n} x_{i''}^2} \\ &= \frac{\sum_{i'=1}^{n} x_i x_{i'} y_{i'}}{\sum_{i''=1}^{n} x_{i''}^2} \\ &= \frac{\sum_{i'=1}^{n} x_i x_{i'}}{\sum_{i''=1}^{n} x_{i''}^2} y_{i'} \end{aligned}$$

Therefore :

$$a_{i'} = \frac{\sum_{i'=1}^{n} x_i x_{i'}}{\sum_{i''=1}^{n} x_{i''}^2}$$

Question 6

Knowing

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

- Argue that in the case of simple linear regression the least squares line always passes through the point ($\hat{x}$,$\hat{y}$)

Answer :

When $x = \hat{x}$

$$\hat{y} = \bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1\bar{x}$$

$$\hat{y} = \bar{y}$$

Question 7

- It is claimed in the text that in the case of simple linear regression of Y onto X, the R^2 statistic is equal to the square of the correlation between X and Y . Prove that this is the case. For simplicity, you may assume that $\bar{x} = \bar{y} = 0$
  Answer :
  We know

$$R^2 = \frac{TSS - RSS}{TSS}$$

$$Cor(x, y) = \frac{\sum_i(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i(x_i - \bar{x})^2}\sqrt{\sum_i(y_i - \bar{y})^2}}$$

when $\bar{x} = \bar{y} = 0$

$$Cor(x, y)^2 = \frac{(\sum_i x_i y_i)^2}{\sum_i x_i^2 \sum_i y_i^2}$$

$$\begin{aligned}
R^2 &= \frac{\sum_i y_i^2 - \sum_i(y_i - x_i\frac{\sum_j x_j y_j}{\sum_j x_j^2})^2}{\sum_i y_i^2} \\
&= \frac{\sum_i y_i^2 - \sum_i(y_i^2 - 2y_i x_i\frac{\sum_j x_j y_j}{\sum_j x_j^2} + x_i^2(\frac{\sum_j x_j y_j}{\sum_j x_j^2})^2)}{\sum_i y_i^2} \\
&= \frac{2\sum_i(y_i x_i\frac{\sum_j x_j y_j}{\sum_j x_j^2}) - \sum_i(x_i^2(\frac{\sum_j x_j y_j}{\sum_j x_j^2})^2)}{\sum_i y_i^2} \\
&= \frac{2\sum_i(y_i x_i)\frac{\sum_j x_j y_j}{\sum_j x_j^2} - \sum_i(x_i^2)\frac{(\sum_j x_j y_j)^2}{(\sum_j x_j^2)^2}}{\sum_i y_i^2} \\
&= \frac{2\frac{(\sum_i x_i y_i)^2}{\sum_j x_j^2} - \frac{(\sum_i x_i y_i)^2}{\sum_j x_j^2}}{\sum_i y_i^2} \\
&= \frac{(\sum_i x_i y_i)^2}{\sum_i x_i^2 \sum_i y_i^2}
\end{aligned}$$