SHEET - An Introduction to Statistical Learning
Chapter 2 - Statistical Learning

PLAYE Nicolas

1 December 2023

0.1 Answers of Exercises

Question 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..

1-Intercept: The null hypothesis is that the intercept is equal to zero. In other words, it suggests that if no money is spent on TV, radio, or newspaper advertising, sales would be zero. The very low p-value (¡ 0.0001) allows us to reject this null hypothesis, indicating that the intercept is significantly different from zero. This means there are some baseline sales even without advertising. 2-TV: The null hypothesis is that TV advertising has no effect on sales, meaning there is no relationship between the amount spent on TV ads and sales. The very low p-value (¡ 0.0001) strongly rejects this hypothesis, suggesting a significant and positive relationship between TV advertising and sales. 3-Radio: The null hypothesis is that radio advertising has no effect on sales, implying no relationship between the radio ad budget and sales. The p-value (¡ 0.0001) is also very small, which allows us to reject this null hypothesis. This shows there is a significant and positive relationship between radio advertising and sales. 4-Newspaper: The null hypothesis is that newspaper advertising has no effect on sales. With a p-value of 0.8599, we fail to reject this hypothesis, indicating that newspaper advertising does not have a statistically significant effect on sales in this model.

Conclusions: -TV and radio advertising have a significant and positive impact on sales. -Newspaper advertising does not have a statistically significant effect on sales. -The model predicts a level of sales that is significantly different from zero, even when no advertising money is spent.

Question 2: Carefully explain the differences between the KNN classifier and KNN regression methods

The KNN regression method is closely related to the KNN classifier. The key difference between the KNN (k-nearest neighbors) classifier and regression methods lies in the type of output they generate and the nature of the problem they are used to solve. when KNN classification predicts categories (classes), KNN regression predicts continuous numerical values. The output is computed: - In classification, the prediction is made based on a majority vote among the k neighbors, -In regression, the prediction is made by averaging the target values of the k neighbors.

Question 3: Suppose we have a data set with five predictors X1 = GPA, X2 = IQ, X3 = Level (1 for College and 0 for High School), X4 = Interaction between GPA and IQ, X5 = 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

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

(a)Which answer is correct, and why?
i. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates.
ii. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates. 128 3. Linear Regression
iii. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates provided that the GPA is high enough.
iv. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates provided that the GPA is high enough.
We have

$$Y = 50 + 20X_1 + 0.07X_2 + 35X_3 + 0.01X_4 - 10X_5$$

When IQ and GPA are fixed value we have

$$Y = Cte + 35X_3 - 10X_5$$

Reformulated we Have

$$Salary = Cte + 35Level - 10Level * GPA$$

We are searching the response by comparaison of level. For college level we have

$$Salary_{college} = Cte + 35 - 10GPA$$

For High School level we have

$$Salary_{hs} = Cte$$

Then

$$Salary_{college} \geq Salary_{hs}$$

$$Cte + 35 - 10 * GPA \geq Cte$$

$$35 - 10 * GPA \geq 0$$

$$3.5 \geq GPA$$

The result shows that knowing that college graduates earn more on average than high school graduates depends on GPA. In fact if GPA is lower than 3.5, then College graduates earn more on average than high school graduates. So if GPA is high enough (over 3.5) then high school graduates earns more one average than college graduates. Which is answer iii.

(b) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0. We have

$$Y = 50 + 20 * GPA + 0.07 * IQ + 35 * Level + 0.01 * GPA * IQ - 10 * GPA * Level$$

Then

$$Y = 50 + 20 * 4.0 + 0.07 * 110 + 35 * 1 + 0.01 * 4.0 * 110 - 10 * 4.0 * 1 = 137.1$$

(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 The size of the coefficient for an interaction term in this case,

$$\hat{\beta}_4 = 0.01$$

alone does not provide enough information to conclude whether or not there is little evidence of an interaction effect. A small coefficient means that the effect of the interaction between GPA and IQ on the response (starting salary) is small in magnitude, but it does not directly indicate whether the interaction is statistically significant.To determine if there is little evidence of an interaction effect, we would need to examine: - The p-value associated with the interaction term. A high p-value would indicate weak evidence against the null hypothesis (no interaction effect). - The confidence interval for the interaction term. If the confidence interval includes zero, it would suggest that the interaction effect might not be significant. Thus, the small size of the coefficient suggests a minor practical effect, but we would need more information (such as statistical significance) to conclude whether the interaction effect is supported by the data.

Question 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.

$$\mathbf{Y} = \beta_\mathbf{0} + \beta_\mathbf{1}\mathbf{X} + \beta_\mathbf{2}\mathbf{X}^\mathbf{2} + \beta_\mathbf{3}\mathbf{X}^\mathbf{3} + \varepsilon$$

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

$$\mathbf{Y} = \beta_\mathbf{0} + \beta_\mathbf{1}\mathbf{X} + \varepsilon$$

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.

If the true relationship between X and Y is linear, i.e.,

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

then we would expect the training residual sum of squares (RSS) for the cubic regression to be **lower** than the RSS for the linear regression. - The linear model only includes terms for X and the intercept, fitting the true underlying relationship between X and Y. This would result in a good fit with the appropriate level of complexity for the data. - The **cubic model** includes additional terms for X^2 and X^3, which means it has more flexibility. Even though the true relationship is linear, the cubic regression can still fit the data more closely because it has more parameters, allowing it to "bend" and potentially overfit the noise in the training data. This additional flexibility leads to a lower training RSS because the model can adapt more to the specific points in the training data. Therefore, even though the true relationship is linear, the cubic model will likely have a lower training RSS due to its greater capacity to fit the data points closely, even if some of that extra fit is to random noise. However, this lower training RSS does not imply that the cubic model is a better representation of the true relationship—it might perform worse on new data due to overfitting.

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

When considering the test residual sum of squares (RSS) instead of the training RSS, we would expect the linear regression model to have a lower test RSS than the cubic regression model, assuming the true relationship between X and Y is linear. - The linear model reflects the true underlying relationship between X and Y (i.e., $Y = \beta_0 + \beta_1 X + \varepsilon$). Therefore, it is expected to generalize well to unseen data, leading to a lower test RSS. - The cubic model has more flexibility and includes unnecessary higher-order terms (X^2 and X^3) that do not represent the true relationship. While it can fit the training data better (lower training RSS), this extra flexibility can lead to overfitting. Overfitting occurs when the model captures not just the underlying pattern but also the random noise in the training data. This results in worse performance on new (test) data because the model is too complex for the true linear relationship. Therefore, in the test set, where the goal is to generalize to new data, the linear model should have a lower test RSS than the cubic model. The cubic model's overfitting on the training data will typically result in a higher test 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.

If the true relationship between XX and YY is not linear, but we do not know how far it is from linear, we would expect the training RSS for the cubic regression to be lower than the training RSS for the linear regression. -The linear regression model assumes a simple linear relationship between XX and YY. If the true relationship is not linear, the linear model will not be flexible enough to capture the more complex relationship, leading to a higher residual sum of squares (RSS) on the training data. - The cubic regression model is more flexible because it includes higher-order terms (X2X2 and X3X3). This extra flexibility allows the cubic model to better fit the training data, even if the true relationship is non-linear. The cubic regression will be able to capture more complex patterns in the data, including non-linear trends, resulting in a lower training RSS compared to the linear regression model. - Since the cubic regression model can fit the data more closely by including additional terms, it will almost always have a lower training RSS than the linear model, regardless of how non-linear the true relationship is. This is because adding more terms to the model increases its ability to fit the data, even if the higher-order terms are not necessary or beneficial for generalization. In summary, the cubic model will likely have a lower training RSS due to its greater flexibility, regardless of how far the true relationship is from linear. However, this lower training RSS does not necessarily mean the cubic model is better—it could be overfitting the training data.

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

If the true relationship between X and Y is not linear, and we consider the test RSS rather than the training RSS, the situation changes: - Linear Regression Model: This model assumes a linear relationship between X and Y. If the true relationship is not linear, the linear model might not fit the training data perfectly, leading to a higher training RSS. However, because it is simpler, it may generalize better to new, unseen data. - Cubic Regression Model: This model includes additional terms (X_2 and X_3) and can fit a more complex relationship. It will likely have a lower training RSS because it can adjust more closely to the specific training data, including any non-linear patterns. Lower Test RSS for Linear Model: If the true relationship is not extremely complex or non-linear, the linear model might generalize better to new data. The cubic model, while fitting the training data well, may overfit and not perform as well on the test data. This overfitting can lead to a higher test RSS for the cubic model because it captures not only the underlying pattern but also the noise in the training data. Potential for Lower Test RSS for Cubic Model: If the true relationship is significantly non-linear and the cubic model is appropriately capturing the underlying pattern, it might have a lower test RSS compared to the linear model. This would be the case if the additional complexity in the cubic model is genuinely useful and not just capturing noise. Finally, without knowing the exact nature of the non-linearity, we generally expect that if the non-linearity is moderate or if the cubic model overfits, the linear regression model will likely have a lower test RSS. This is because it avoids overfitting and maintains better generalization to unseen data. However, if the non-linearity is significant and the cubic model is correctly capturing the true relationship, it might have a lower test RSS. In practice, the cubic model tends to have a lower training RSS but may not always outperform the linear model on test data due to potential overfitting.

Question 5: 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

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

Show that we can write

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

What is $a_{i'}$?

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

$$\hat{y}_i = x_i \frac{\sum_{i=1'}^{n} x_{i'} y_{i'}}{\sum_{i'=1}^{n} x_{i'}^2}$$

$$\hat{y}_i = \sum_{i=1'}^{n} \frac{x_{i'} x_i}{\sum_{i'=1}^{n} x_{i'}^2} y_{i'}$$

Then

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

The equation $\hat{y}_i = \sum_{i'=1}^{n} a_{i'} y_{i'}$ represents the linear regression of the $y_{i'}$ variables to explain $\hat{y}_i$. Since $\hat{y}_i = x_i\hat{\beta}$, the coefficients $a_{i'}$ represent the transformation from a simple linear regression dependent on a single parameter $\hat{\beta}$ to a multiple linear regression dependent on the $y_{i'}$ coefficients. Therefore, $a_{i'}$ expresses $\hat{\beta}$ in terms of the individual coefficients of the multiple linear regression model involving the $y_{i'}$ variables.

Question 6: Using (3.4), argue that in the case of simple linear regression, the least squares line always passes through the point $(\bar{x}, \bar{y})$
Let $\hat{y} = \hat{\beta}_1 x_i + \hat{\beta}_0$
We know that

$$f(\hat{\beta}_0, \hat{\beta}_1) = \min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

We get the minimum where

$$\frac{\partial f}{\partial \hat{\beta}_k} = 0 \text{ with k} \in \{0, 1\}$$

We have

$$n\bar{y} = \sum_{i=1}^{n} y_i \textbf{ and } n\bar{x} = \sum_{i=1}^{n} x_i$$

Then

$$\frac{\partial f}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x} = 0$$

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

Finally, the least squares line always passes through the point $(\bar{x}, \bar{y})$

Question 7: It is claimed in the text that in the case of simple linear regression of Y onto X, the R^2 statistic (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 $\bar{x} = \bar{y} = 0$ We suppose having $\bar{x} = \bar{y} = 0$

The simple linear regression model is $\hat{y}_i = \beta x_i$, where β is the regression coefficient. We have $R^2 \frac{SSE}{SST}$

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2$$

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} \hat{y}_i^2$$

The Pearson correlation coefficient between X and Y is:

$$r_{XY} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

In simple linear regression, the estimated coefficient $\hat{\beta}$ is:

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

$$\hat{y}_i = \hat{\beta} x_i = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} x_i$$

The explained sum of squares SSE is the sum of squared predicted values $\hat{y}_i$:

$$SSE = \sum_{i=1}^{n} \hat{y}_i^2 = \sum_{i=1}^{n} \left(\frac{\sum_{i'=1}^{n} x_{i'} y_{i'}}{\sum_{i'=1}^{n} x_{i'}^2} x_i \right)^2$$

$$SSE = \left(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \right)^2 \sum_{i=1}^{n} x_i^2$$

$$SSE = \frac{\left(\sum_{i=1}^{n} x_i y_i\right)^2}{\sum_{i=1}^{n} x_i^2}$$

The total sum of squares SST is the sum of squared actual values of y_i:

$$SST = \sum_{i=1}^{n} y_i^2$$

Using the formula for $R^2 = \frac{SSE}{SST}$, we have:

$$R^2 = \frac{\frac{\left(\sum_{i=1}^{n} x_i y_i\right)^2}{\sum_{i=1}^{n} x_i^2}}{\sum_{i=1}^{n} y_i^2}$$

$$R^2 = \frac{\left(\sum_{i=1}^{n} x_i y_i\right)^2}{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}$$

The square of the Pearson correlation coefficient r_{XY} is:

$$r_{XY}^2 = \left(\frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}}\right)^2$$

$$r_{XY}^2 = \frac{\left(\sum_{i=1}^{n} x_i y_i\right)^2}{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}$$

$$R^2 = r_{XY}^2$$