Estadística

U5 5C

David Aarón Ramírez Olmeda

Mayo 2023

Linear regression

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.

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

The p-values in the table suggest that TV and radio advertising expenditures are significantly related to sales, while newspaper advertising expenditure does not show evidence of being significantly related to sales in the presence of the other two variables.

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

KNN classifier predicts classes or categories, while KNN regression predicts numeric values. The classifier assigns the majority class, while the regression method calculates the average value among the nearest neighbors.

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

$\beta\hat{}0 = 50, \beta\hat{}1 = 20, \beta\hat{}2 = 0.07, \beta\hat{}3 = 35, \beta\hat{}4 = 0.01, \beta\hat{}5 = -10$

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 coefficient for the gender variable, $\beta\hat{}3$, is positive (35). A positive coefficient indicates that, on average, males earn more than females when other predictors. (iii)

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

```
beta0 <- 50
beta1 <- 20
beta2 <- 0.07
beta3 <- 35
beta4 <- 0.01
beta5 <- -10

IQ <- 110
GPA <- 4
Gender <- 1

salary <- beta0 + beta1 * GPA + beta2 * IQ + beta3 * Gender + beta4 * (GPA * IQ) + beta5 * (GPA * Gende
salary
```

```
## [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.

False. The statement does not imply that there is very little evidence of an interaction effect. The size of the coefficient alone does not determine the presence or absence of an interaction effect. A small coefficient value does not necessarily mean there is no 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. $Y = \beta 0 + \beta 1X + \beta 2X2 + \beta 3X3 + e$

(a) **Suppose that the true relationship between X and Y is linear, i.e.** $Y = \beta 0 + \beta 1X + e$ **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.**

(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.**

a. If the true relationship between X and Y is linear, the linear regression model is better suited to capture a linear relationship between X and Y, we would expect the training residual sum of squares (RSS) for the linear regression to be lower than the training RSS for the cubic regression

b. We would still expect the test RSS for the linear regression to be lower than the test RSS for the cubic regression when the true relationship between X and Y is linear.

c. The cubic regression might have a lower training RSS if the true relationship is closer to a cubic function, although we don't know how far it deviates from linearity, so it is difficult to determine whether the training RSS for the linear regression or the cubic regression would be lower.

d. There is still not enough information to determine whether the test RSS for the linear regression or the cubic regression, but we still go with maybe lower to keep it safe.

5. Consider the fitted values that result from performing linear regression without an intercept. In this setting, the ith fitted value takes the form:

$y_i^{'} = x_i\beta$,

where

$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$ **(3.38)**

Show that we can write the i-th fitted value as:

$y_i^{'} = \sum_{j=1}^{n} a_{ij} y_j$

What is a_{ij} ?

To show that:

$y_i^{'} = \sum_{j=1}^{n} a_{ij} y_j$,

holds, we need to find the expression for a_{ij}.

Given that $\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$ (3.38), we can substitute it into the equation $y_i^{'} = x_i\beta$:

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

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

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

Notice that we can rewrite $\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$ as a_{ij}.

Therefore:

$y_i^{'} = \sum_{j=1}^{n} a_{ij} y_j$,

where $a_{ij} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \cdot x_i$.

In this case, a_{ij} represents the weight or contribution of each observation y_j in the calculation of the i-th fitted value.

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

In the case of simple linear regression, the least squares line is defined by the equation: $y = \beta_0 + \beta_1 x$

So, from 3.4: $\beta_0 = y + \beta_1 x$

Notice that x and y are constants, not dependent on any specific observations. Therefore, the least squares line always passes through the points, which represents the center of the data in terms of the predictor and response variables.

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.

To prove that the coefficient of determination R^2 in simple linear regression is equal to the square of the correlation between the predictor variable X and the response variable Y, we can start with the definitions of R^2 and the correlation coefficient.

The coefficient of determination R^2 is defined as:

$$R^2 = 1 - \frac{\text{Residual Sum of Squares}}{\text{Total Sum of Squares}} \quad (1)$$

The correlation coefficient r between X and Y is defined as:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}} \quad (2)$$

Given that $\bar{X} = \bar{Y} = 0$ (as assumed), we can simplify the equations further.

First, let's calculate the Residual Sum of Squares (RSS). In simple linear regression, the predicted values are given by $\hat{Y}_i = \hat{\beta} X_i$, where $\hat{\beta}$ is the estimated slope coefficient.

The RSS is calculated as:

$$\text{RSS} = \sum(Y_i - \hat{Y}_i)^2 = \sum(Y_i - \hat{\beta} X_i)^2 \quad (3)$$

Next, let's calculate the Total Sum of Squares (TSS). The TSS is the sum of squares of the differences between the observed Y_i and the mean $\bar{Y}$.

The TSS is given by:

$$\text{TSS} = \sum(Y_i - \bar{Y})^2 \quad (4)$$

Since $\bar{Y} = 0$ (as assumed), the TSS simplifies to:

$$\text{TSS} = \sum Y_i^2 \quad (5)$$

Now, substituting equations (3), (4), and (5) into equation (1) for R^2, we have:

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\sum(Y_i - \hat{\beta} X_i)^2}{\sum Y_i^2} \quad (6)$$

Next, let's simplify the correlation coefficient r given in equation (2) when $\bar{X} = \bar{Y} = 0$.

The correlation coefficient r is simplified to:

$$r = \frac{\sum X_i Y_i}{\sqrt{\sum X_i^2 \sum Y_i^2}} \quad (7)$$

Now, let's square the correlation coefficient r to obtain r^2:

$$r^2 = \left(\frac{\sum X_i Y_i}{\sqrt{\sum X_i^2 \sum Y_i^2}} \right)^2 = \frac{(\sum X_i Y_i)^2}{\sum X_i^2 \sum Y_i^2} \quad (8)$$

Comparing equations (6) and (8), we can see that they have the same terms in the numerator and the denominator. Therefore, we can conclude that:

$$R^2 = r^2$$

Hence, the coefficient of determination R^2 in simple