STAT0032: Exercise Sheet #4

The exercises in this sheet focus on linear regression. Most of the questions are from James et al., "An Introduction to Statistical Learning" (ISLR). This book is freely available as a PDF file, see link in our Moodle page. As before (see Sheet #3), questions marked with the indication "(COMPUTER IMPLEMENTATION)" require programming. In particular, ISLR poses questions which explicitly require R. Feel free to use an equivalent software package for linear regression, but all solutions provided in Moodle will be based around R.

1. (Exercise 1 of Chapter 3, ISLR.) 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

TABLE 3.4. *For the* `Advertising` *data, least squares coefficient estimates of the multiple linear regression of number of units sold on radio, TV, and newspaper advertising budgets.*

2. (Exercise 3 of Chapter 3, ISLR.) Suppose we have a data set with five predictors, $X_1 =$ GPA, $X_2 =$ IQ, $X_3 =$ Gender (1 for Female and 0 for Male), $X_4 =$ Interaction between GPA and IQ, and $X_5 =$ 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?

      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.

   (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.

3. (Exercise 4 of Chapter 3, ISLR.) 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. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$.

   (a) Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$. 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.

4. (Exercise 5 of Chapter 3, ISLR.) Consider the fitted values that result from performing linear regression without an intercept. In this setting, the ith 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 $a_{i'}$?

   Note: We interpret this result by saying that the fitted values from linear regression are linear combinations of the response values.

5. The estimated correlation coefficient (or just **sample correlation**) of two variables X and Y is defined as

$$\text{Cor}(X, Y) \equiv \frac{\sum_{i=1}^{n} (x^{(i)} - \bar{x})(y^{(i)} - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x^{(i)} - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y^{(i)} - \bar{y})^2}}$$

   for a dataset of bivariate measurements $\{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\}$. It is a measure of linear association.

   (a) What would this correlation be if each $y^{(i)}$ was a constant multiple of the respective $x^{(i)}$, that is, $y^{(i)} = a x^{(i)}$ for some constant a? What if $y^{(i)} = a x^{(i)} + b$ for some other constant b? Conclude that correlation coefficients are a linear measure of association between -1 and 1, with the extremes corresponding to deterministic linear dependencies.

(b) (Exercise 7 of Chapter 3, ISLR.) 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 $\bar{x} = \bar{y} = 0$.

6. (COMPUTER IMPLEMENTATION) Exercise 8 of Chapter 3, ISLR.

7. (COMPUTER IMPLEMENTATION) Exercise 9 of Chapter 3, ISLR.

8. (COMPUTER IMPLEMENTATION) Exercise 10 of Chapter 3, ISLR.