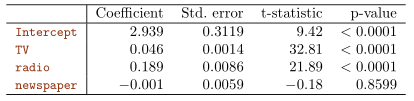
# 3.7 Exercises

## Conceptual

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.*In a multiple linear regression, when conducting a null hypothesis on an individual predictor, you are hypothesising that in the presence of the other predictors, the predictor analysed has no effect on response. E.g. in this example the null hypothesis for the first predictor would be; given the presence of radio and newspaper, TV has no effect on sales. From the resulting p-values, we can conclude that TV and radio and statistically significant predictors, while newspaper is not.
2. *Carefully explain the differences between the KNN classifier and KNN regression methods.*Both KNN models are similar in formula, however the former is used for classification (qualitative) outputs and the latter for quantitative outputs.
3. *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  
   ˆβ 0 = 50, ˆ β 1 = 20, ˆ β 2 = 0.07, ˆ β 3 = 35, ˆ β 4 = 0.01, ˆ β 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.*

Resulting formula:  
^Y = 50 + 20X1 + 0.07X2 + 35X3 + 0.01X4 + -10X5  
^Y = 50 + 20X1 + 0.07X2 + 35X3 + 0.01(X1\*X2) + -10(X1\*X3)  
When X3 = 1:  
^Y = 50 + 20X1 + 0.07X2 + 35 + 0.01(X1\*X2) + -10(X1)  
When X3 = 0:  
^Y = 50 + 20X1 + 0.07X2 + 0.01(X1\*X2)   
Therefore, statement 3 is correct; given a high enough GPA, men will earn more than women on average.

*(b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.* ^Y = 50 + 20\*4 + 0.07\*110 + 35 + 0.01(4\*110) + -10(4)  
 ^Y = 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 the interaction term determines how large the affect it has on the response, it does not determine whether it is statistically significant or not; we would need to look at the p-value to determine evidence of whether the interaction effect is there or not.

*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 = β 0 + β 1 X + β 2 X 2 + β 3 X 3 + ?.*

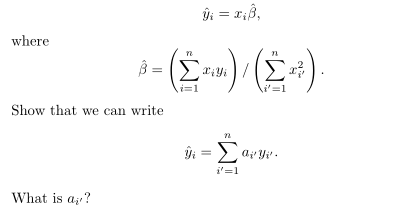
*(a) Suppose that the true relationship between X and Y is linear, i.e. Y = β 0 + β 1 X + ?. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.*It would depend on the amount of variance present in the data. A dataset with low variance would be fit by a more rigid, linear model, however one with a high level of variance would be fit better by a more flexible polynomial model.

*(b) Answer (a) using test rather than training RSS.*If the true relationship is linear, then a rigid linear regression would have a lower test RSS as the cubic regression will overfit any errors present in the training data, and so have a worse test RSS result.

*(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 is non-linear, then we would expect the training RSS for the cubic regression to be lowest as it will fit the variance in the training data better than the linear model will.

*(d) Answer (c) using test rather than training RSS.*Given “*we don’t know how far it is from linear”*, there’s not enough information present to predict this. The more linear the relationship is, the more the cubic regression will overfit the variance in the training data and the worse the test RSS will be. The more non-linear the relationship is, the more the linear regression will underfit the training data and the higher the resulting bias will be for the test data, leading to a worse test RSS.

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

**

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

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

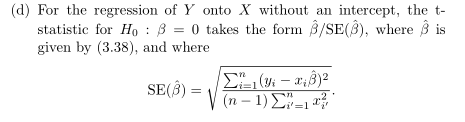
  
If is a point on our line, then .Given:

The least squares line always passes through

*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*  





11)



