Tiffany Wong A20442087
CS422 - HW 2 09/19/21

1. Exercises

1.1 ISLR 2e (Gareth James, et al.): Section 3.7 (Exercises), page 120: Exercises 1, 3, 4-a.

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	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 TV, radio, and newspaper advertising budgets.*

H_0^1: The amount of budget dedicated to TV advertisements has no effect on sales.
H_0^2: The amount of money in the radio advertisements budget does affect sales.
H_0^3: There is no relationship between sales and newspaper advertisement.
Based on these p-values, I can conclude that the p-values for TV and radio are highly significant, while the p-value for newspaper advertising is not significant in comparison. This means I can reject H_0^1 and H_0^2 , but not H_0^3. Therefore, I can say that the budget for newspaper advertising does not affect sales.

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

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.*
   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.*

The least squares line is given by:
$\hat{y} = 50 + 20*GPA + 0.07*IQ + 35*Level + 0.01*(GPA \times IQ) - 10*(GPA \times Level)$

For college students, (plug in Level=1) it becomes:
$\hat{y}$ = 85 + 10*GPA + 0.07*IQ + 0.01*(GPA×IQ)

and for high school students, (plug in Level=0) it becomes:
$\hat{y}$ = 50 + 20*GPA + 0.07*IQ + 0.01*(GPA×IQ)

With high school students and college students, their starting salary differ with the following equation: 50 + 20*GPA = 85 + 10*GPA.
10*GPA = 35, GPA=3.5

So the starting salary for high school students is higher than for college students on average if 50+20*GPA ≥ 85+10*GPA which is equivalent to GPA ≥ 3.5.

Therefore iii. is the right answer.

b. *Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.*

$\hat{y}$ = 85 + 10*GPA + 0.07*IQ + 0.01*(GPA×IQ)
= 85 + 10*4.0 + 0.07*110 + 0.01*(4.0*110) = 137.1
A college graduate with IQ of 110 and a GPA of 4.0 is predicted to get a salary of $137,100.

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.*

This statement is false because this isn't the proper way to test the impact of the interaction effect. In order to verify the impact of GPA and IQ on the quality of the model, I would have to test the null hypothesis of setting $\hat{\beta}4=0$. Then I will look at the p-value associated with the t of the F statistic to draw a conclusion.

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. $Y = \beta0 + \beta1*X1 + \beta2*X2 + \beta3*X3 + \varepsilon$.

a. *Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta0 + \beta1X + \varepsilon$. 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.*

Without knowing more details about the training data, it is difficult to know which training RSS is lower between linear or cubic. However, as the true relationship between X and Y is linear, we may expect the least-squares line to be close to the true regression line, and consequently, the RSS for the linear regression may be lower than for the cubic regression.