Homework 3

1 Exercises

1.1 Using (3.4), argue that in the case of simple linear regression, the least squares line always passes through the point $(\bar{x}, \bar{y})$.

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

$$\beta_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Substituting, we see that:

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

$$\hat{y} = \left(\bar{y} - \hat{\beta}_1 \bar{x}\right) + \hat{\beta}_1 x$$

$$0 = \left(\bar{y} - \beta_1 \bar{x}\right) + \hat{\beta}_1 x - \hat{y}$$

The right hand side will only equal 0, if x = $\bar{x}$ and $\hat{y} = \bar{y}$, this means that $(\bar{x}, \bar{y})$ is part of the line.

1.2.1 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.

The null hypotheses in Table 3.4 correspond to whether the advertising mediums—TV, radio, and newspaper—have any effect on sales when the other variables are accounted for.

For TV ads: The null hypothesis states that, in the presence of radio and newspaper ads, TV ads do not affect sales.

For radio ads: The null hypothesis states that, in the presence of TV and newspaper ads, radio ads do not affect sales.

For newspaper ads: The null hypothesis states that, in the presence of TV and radio ads, newspaper ads do not affect sales.

The low p-values for TV and radio ads suggest that their null hypotheses are likely false, meaning that both TV and radio ads significantly impact sales when the other variables are included. The high p-value for newspaper ads suggests that its null hypothesis is likely true, indicating that newspaper ads do not have a significant effect on sales when TV and radio ads are accounted for.

1.2.3 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.

This is incorrect because females earn more when GPA is less than 3.5

ii. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates.

This is incorrect because males earn more when GPA is greater than 3.5.

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.

This is correct because high school graduates earn more, on average, than college graduates if the GPA is high enough. The positive impact of GPA on salary is stronger for high school graduates (20 per GPA point) than for college graduates (10 per GPA point), and at a high enough GPA, the higher GPA effect for high school graduates will outweigh the dollar advantage that college graduates initially have.

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.

This is incorrect because males earn more when GPA is greater than 3.5.

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

$\text{Salary} = \beta_0 + \beta_1(\text{GPA}) + \beta_2(\text{IQ}) + \beta_3(\text{Level}) + \beta_4(\text{GPA}\times\text{IQ}) + \beta_5(\text{GPA}\times\text{Level})$

$\text{Salary} = 50 + 20(4.0) + 0.07(110) + 35(1) + 0.01(4.0\times110) - 10(4.0\times1)$

Salary = $ 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.

False. The magnitude of the coefficient alone does not indicate the presence or absence of an interaction effect. Without knowing the standard error or p-value for the GPA/IQ interaction term, we cannot determine whether the interaction is statistically significant. A small coefficient might still represent a significant effect if the associated p-value is low, and conversely, a larger coefficient might not be significant if its p-value is high.

1.2.4 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. $Y = \beta0 + \beta1X + \beta2X2 + \beta3X3 + \epsilon$.

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

I expect the training RSS for the cubic regression to be lower than the linear regression. This is because the cubic regression has more parameters and can fit the training data more flexibly, even if the true relationship is linear. However, this does not necessarily mean the cubic regression model is better, as it may overfit the training data and not generalize well to new data.

2 Programming Problems

Q2(a)(ii) Why is using name as a predictor not a reasonable thing to do?

Using name as a predictor is not reasonable because it is a categorical variable with many unique values. Including name would add noise rather than informative variance to the model.

Q2(a)(vi) Does the histogram follow a Gaussian distribution? What can you say about the distribution of the residuals?

The histogram does not follow a perfect Gaussian distribution, indicating that the residuals are not normally distributed. This suggests potential nonlinearity in the model's fit.

Q2(b)(iv) Please provide a justification in 1-2 sentences for your choice in the above question.

The model created in 2(a)(i) is better because it has a higher R^2 value, indicating that it explains a larger proportion of the variance in the response variable. Additionally, the adjusted R^2 is higher, suggesting that it provides a better fit, improving predictive performance on unseen data.