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.

P-values that are very small indicate that the model for that predictor is likely to account for a significant amount of the association between the predictor and the response. If that is true, then, we reject the null hypothesis, and conclude that a relationship exists between the predictor and the response. The p-values computed from the response of sales to marketing budget for each marketing paradigm indicate will give us insight into which of these predictors have a strong relationship with sales of this product.

TV marketing and radio marketing both have a strong relationship to sales, according to their linear regression p-values, but newspaper advertising does not appear to be effective, given that the linear model does not account for much of the variation in sales across that domain. We can conclude that cutting back on newspaper advertising will likely have little effect on the sales of the product, and that increasing TV and radio advertising budgets likely will have an effect. Furthermore, we can see that radio advertising spending has a stronger relationship with sales, as the best-fit slope is significantly more positive than the best fit for TV advertising spending, so increasing the radio advertising budget will likely be more effective.

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 $\beta_0 = 50$, $\beta_1 = 20$, $\beta_2 = 0.07$, $\beta_3 = 35$, $\beta_4 = 0.01$, $\beta_5 = -10$.

This is the model: $\hat{y} = 50 + 20\,X_1 + 0.07\,X_2 + 35\,X_3 + 0.01\,X_4 + -10\,X_5$

For fixed IQ and GPA, we can infer that the starting salary for a female sharing an IQ and GPA with her male counterpart will make (35*1 - 10*(GPA*1)) more starting salary units than her male counterpart. This means that at very low GPAs(maybe this includes people who didn't attend school?), males have a lower starting wage, and as GPA grows, males make a larger starting salary from that point, overtaking females at GPA=3.5. Therefore,

(a) Which answer is correct, and why? → iii. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

This one is correct.

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

$\hat{y} = 50 + 20*4.0 + 0.07*110 + 35*1 + 0.01*(4.0*110) - 10*(4.0*1)$

→ $\hat{y} = 137.1$ salary units

(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. There is still a noticeable effect because the coefficient for IQ's effect alone is only 7 times greater than the coefficient of the interaction term. So, this term holds significant weight compared to the overall response of the model to IQ.

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. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$.

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

For the training data, the cubic regression might return a better RSS than the linear regression, but this would only be because the cubic is fitting points that are varied according to the ε random error. It also may not, depending on how that random error expressed itself in this case.

(b) Answer (a) using test rather than training RSS.

For the test error, the RSS will almost certainly be greater for the cubic model than the linear model, because the random error ε will likely express itself in a way that is inconsistent with the noise that the cubic model adopted during its training. The linear model will be more likely to have a lower RSS the more test data is used against the models.

(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.

The cubic model will pick up more information because of its additional degrees of freedom. If the true relationship is more complex than linear, then the cubic model will likely have a lower RSS over the linear model. If the model is less complex than linear (E.G. perhaps it is just a constant scalar relationship) then the linear model will still be more likely to have a smaller RSS, because the cubic will again pick up information from the ε noise that is not inherent in the real relationship.

8. This question involves the use of simple linear regression on the Auto data set.

(a) Use the lm() function to perform a simple linear regression with mpg as the response and horsepower as the predictor. Use the summary() function to print the results. Comment on the output. For example:

There is definitely a correlation between horsepower and mpg. The RSE is ~4.9, which is not insignificant and does indicate that the response may not be truly linear, but it is small enough relative to the mpg magnitude that it's clear a relationship exists. The R^2 statistics corroborates this by indicating (it has a small value at ~0.6) that a large proportion of the mpg variability is explained by the model. mpg has a negative correlation with horsepower, indicated by the negative coefficient on the horsepower factor.

For example, for a vehicle with 98 horsepower, one can expect with 95% confidence that the mpg will be within 23.97 and 24.96, if the vehicles follow our model. However, after incorporating the irreducible error, the prediction turns out to be much less precise, with a 95% prediction interval spanning 14.8 to 34.1. Some of this variability may also be reduced by using a quadratic model, from visual inspection of the plot.

(b) Plot the response and the predictor. Use the abline() function to display the least squares regression line.

Attached.

(c) Use the plot() function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

Attached. From these four plots it's clear there is a lot of variability that remains unexplained by the linear model. The standardized residuals plotted against the fitted values shows clearly that the variability is strong, with values consistenly lying outside 1 standardized residual unit, but still within a tight range that doesn't extend past 3, which is often considered an approximate threshold to indicate values that aren't explained well by the model. There are many points with high leverage, and these values have less residual by default, of course, and in both of these graphs we see a few points (323, 330) that are rearing their ugly heads. These seems to be the bit of "uptick" toward the higher end of the horsepower scale that would probably be picked up by a quadratic fit.