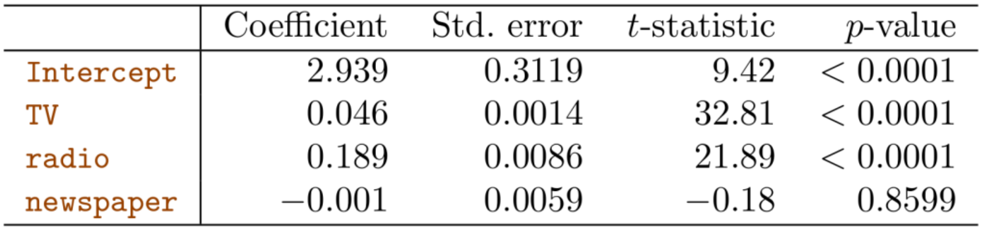
HW 2 – STAT 627
Homayoon Fotros

Question 1



	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

a. Interpreting the TV line in regression results

The null hypothesis states (*H_0*) that 'There is *no* relationship between TV advertising budget and the product's sale.' The alternative hypothesis (H_1) states that 'There is *some* relationship between TV advertising budget and the product's sale.' Mathematically, the null hypothesis corresponds to $\beta_{TV}=0$ and the alternative hypothesis corresponds to $\beta_{TV}\neq0$ in the regression model. The table shows the estimated coefficient for TV is 0.046 with a standard error of 0.0014. The p-value for this estimate is very small (<0.0001). Therefore, we can confidently reject the null hypothesis and accept the alternative hypothesis that there is a statistically significant relationship between the TV advertising budget and units of product sales. Concretely, for one thousand unit increase in TV advertising budget, we estimate that the product sales will increase by 46 units when holding other factors (i.e., radio and newspaper spending) constant.

b. Interpreting the Newspaper line in regression results

The null hypothesis states (*H_0*) that 'There is *no* relationship between newspaper advertising budget and the product's sale.' The alternative hypothesis (H_1) states that 'There is *some* relationship between newspaper advertising budget and the product's sale.' Mathematically, the null hypothesis corresponds to $\beta_{Newspaper}=0$ and the alternative hypothesis corresponds to $\beta_{Newspaper}\neq0$ in the regression model. The table shows the estimate coefficient for newspaper is close to zero (-0.001) with a standard error of 0.0059. The p-value for this estimate is large (0.8599) and greater than conventional confidence levels (e.g., 0.05). Therefore, we *can not* reject the null hypothesis that there is no relationship between newspaper advertising budget and units of product sales. The regression result indicates if we hold spending on TV and radio advertising constant, one thousand unit increase in newspaper advertising budget will not yield any significant change on the units of product sales.

Question 2

a. Comparing the residual sum squares (RSS) between the linear and cubic models

The cubic model may capture more variations in the training data set and thus generates a wiggly outcome compared to the linear model. Therefore, I expect that the sum of $\mathrm{RSS} = (y_i - \hat{y_i})^2$ for the cubic model will be smaller that the linear model. However, if the n=100 observations are perfectly fitting on a line, the linear model's RSS will be close to zero, which is likely smaller than the cubic model's RSS. In sum, the answer depends on the distribution of the data while in a typical set of 100 observations we might expect that a cubic model's overfitting effect results in a smaller RSS.

a. Previous question using RSE

For the test data, I expect that the cubic model generates a larger RSE than a linear model because the true relationship is linear and the cubic model is likely overfitting on the training set. We also know that increasing the number of predictors would increase RSE because $RSE = \sqrt{\frac{RSS}{n-p-1}}$. Therefore, chances are higher that the RSE for cubic model will be larger than the linear model's RSE, particularly if the difference between the training RSS values is very small.

c. Part (a) if the true relationship is non-linear

The cubic model is more flexible and better follows the non-linear variation in the training data set. Hence, I expect that the cubic model yields a smaller RSS compared to a linear model. However, if the relationship between X and Y is non-linear but close to a linear relationship, it might be the case that the training set is distributed in a way that a cubic model generates a RSS not too different than a linear model's RSS.

d. Part (c) but RSE on the test set

The answer depends on how far the true relationship is from a linear one. If it is close to a linear relationship, it is likely that a linear model generates a smaller RSE than a cubic model. Conversely, if the true relationship is far from a linear one, a cubic model can better capture the variance and thus generate a smaller RSE than a linear model. In sum, based on the test data set, the variance-bias trade off between the cubic and linear model will determine which one yields a better fit (i.e., smaller RSE).

Question 3

An LS model of X1 = GPA, X2 = IQ, X3 = Level, and interaction terms X4 = GPA*IQ and X5 = GPA*Level, with Starting Salary as response.

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

a. For a fixed value of IQ and GPA…

The coefficient for Level (X3) is positive and equals 35, which means that the main effect of being graduated from college is positive. However, the interaction effect of Level and GPA is negative and equals -10. This means if GPA is high enough, the interaction effect will cancel out the Level's main effect. For example, if GPA is high and equals to 4, the interaction effect of a college graduate would be -40 = -10 * 4 * 1 and the main effect of Level would be 35 = 35 * 1, which in total is -5. Instead, if it was a high school graduate, these effects would be zero. Hence, if everything else is remained unchanged, high school graduates earn more on average given a high level of GPA.

b. Predict Salary given college graduate, IQ = 100, and GPA = 4.0

Salary = 50 + 20 * GPA + 0.07 * IQ + 35 * Level + 0.01 * GPA * IQ + (-10) * GPA * Level
Salary = 50 + 20 * 4 + 0.07 * 100 + 35 * 1 + 0.01 * 4 * 100 + (-10) * 4 * 1 -->
Salary = 136 (thousand dollars)

c. Interpreting the interaction coefficient of GPA and IQ

The coefficient is small ($\beta 4 = 0.01$) but it does not mean that the interaction effect is insignificant. To test the statistical significance of the effect, we have to assess the p-value of the coefficient. To assess the extent to which this effect is substantive we should compare it to the range and average of starting salary values. For example, if the average salary is 60 thousand dollars, and the average IQ to be about 100 and the average GPA to be 3, the interaction effect between these two variables would be 3 thousand dollars, which is 5 percent of the average. But if the average salaries is 150 thousand dollars, the interaction effect will be 2 percent, which might be negligible in the overall context. In sum, the small value of the coefficient does not imply that it is not significant and substantive.