STAT 501 – Homework 7 (covering Lessons 7 and 8) – due date Oct 18th

1. **(5x2 = 10 points)** State which of the following statements is TRUE and which is FALSE. For the statements that are false, explain why they are false.
   (a) A small *p*-value associated with the Ryan-Joiner test for normality indicates that data are normally distributed. **False: A large p-value indicates normality.**
   (b) A confidence interval for the mean response and a predicted interval for a new response will be valid only if all LINE conditions are satisfied. **False: The mean response confidence interval will be valid even if normality is not satisfied, provided the sample size n is large enough.**
   (c) Increasing the sample size, n, ensures that the widths of both the mean response confidence interval and the new response prediction interval will be decreased regardless of the confidence level. **True.**
   (d) A statistically significant interaction effect between a continuous predictor X1 and a qualitative predictor X2 indicates that the SLR model equations expressing the Y vs X1 relationship at different levels of X2 have different slopes and intercepts. **False: A significant interaction effect implies that the slopes are different, but not the intercepts.**
   (e) Suppose in a regression model there is a qualitative predictor variable X1 with 5 levels. You will have to code the variable X1 as -2, -1, 0, 1 and 2 corresponding to the 5 levels. **False: You need to introduce 5 – 1 =4 indicator variables**

2. **(3 + 4 + 10 + 3 + 5 = 25 points)** Use the "SMSA" dataset. Researchers at General Motors analyzed data on 56 U.S. Standard Metropolitan Statistical Areas (SMSAs) to study whether air pollution contributes to mortality. These data were obtained from the "Data and Story Library" at lib.stat.cmu.edu/DASL/ (the original data source is the U.S. Department of Labor Statistics). The response variable for analysis is *Mort* = age adjusted mortality per 100,000 population (a mortality rate statistically modified to eliminate the effect of different age distributions in different population groups). The dataset includes predictor variables measuring demographic characteristics of the cities, climate characteristics, and concentrations of the air pollutant nitrous oxide (NO_x). In particular, *Edu* is median years of education, *Nwt* is percentage nonwhite, *Jant* is mean January temperature in degrees Fahrenheit, *Rain* is annual rainfall in inches, *Nox* is the natural logarithm of nitrous oxide concentration in parts per billion, *Hum* is relative humidity, and *Inc* is median income in thousands of dollars.
   (a) Fit the multiple linear regression model, $E(Mort)=b_0 + b_1\,Edu + b_2\,Nwt + b_3\,Jant + b_4\,Rain + b_5\,Nox + b_6\,Hum + b_7\,Inc$. Report the SSE (sum of squared errors) and degrees of freedom (df) for error.

SSE = 60,417 and df = 56 – 8 = 48.

(b) Do a general linear *F*-test (using a significance level of 5%) to see whether *Hum* and *Inc* provide significant information about the response, *Mort*, beyond the information provided by the other predictor variables.
[In Minitab, you can find the information for the *F*-statistic either by selecting Sequential sums of squares in the regression options for the model in the previous part or by fitting a reduced model without *Hum* and *Inc*. You'll also need to calculate a *p*-value for the *F*-statistic: select Calc > Probability Distributions > F. . . . Select Cumulative probability and leave the Noncentrality parameter set to 0.0. Next, enter in the respective Numerator degrees of freedom and Denominator degrees of freedom. Finally, enter the *F* value of interest into the box that says Input constant. For the output, we will see the *F* value (given as *x*) and the probability $P(X \leq x)$. The p-value for this problem is $1 - P(X \leq x)$. Make sure you know why!]

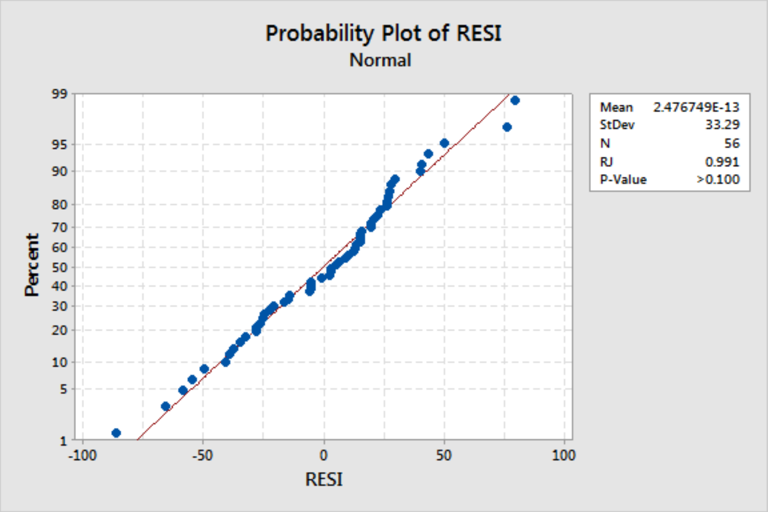
H_0: $\beta_6 = \beta_7 = 0$ versus H_a: β_6 or β_7 or both not equal to 0.
F = ((60,948−60,417)/(7−5)) / (60,417/48) = 0.21.
p-value from an F distribution with 2 numerator and 48 denominator degrees of freedom is 1 – 0.19 = 0.81.
Since 0.81 > 0.05 we fail to reject the null hypothesis, which means we can drop *Hum* and *Inc* from the model.

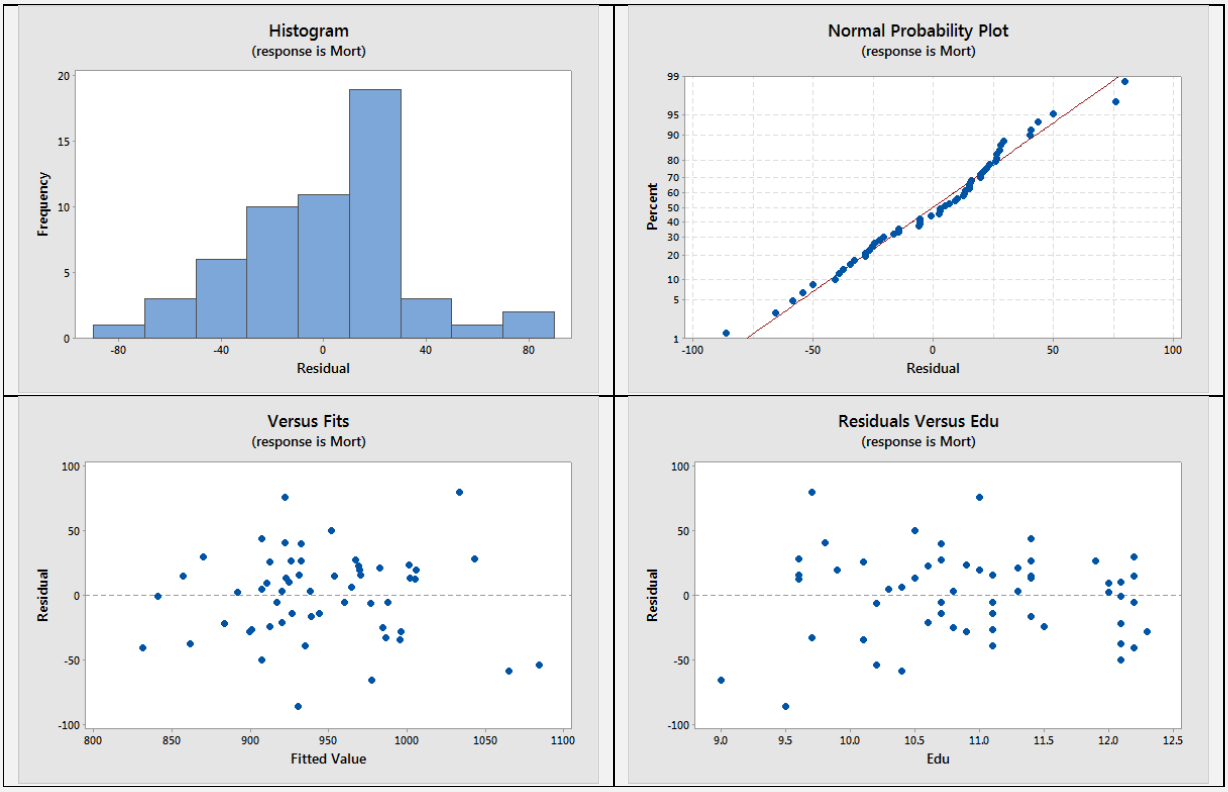
(c) Fit the multiple linear regression model, $E(Mort)=b_0 + b_1\,Edu + b_2\,Nwt + b_3\,Jant + b_4\,Rain + b_5\,Nox$. Check the LINE model assumptions for this model. To do this click Graphs in the Minitab regression dialog box. Then select Histogram of residuals, Normal probability plot of residuals, and Residuals versus fits. Also click in the Residuals versus the variables box and type "Edu-Inc." Also click Storage in the regression dialog box and select "Residuals." The resulting 10 plots can be used as follows:

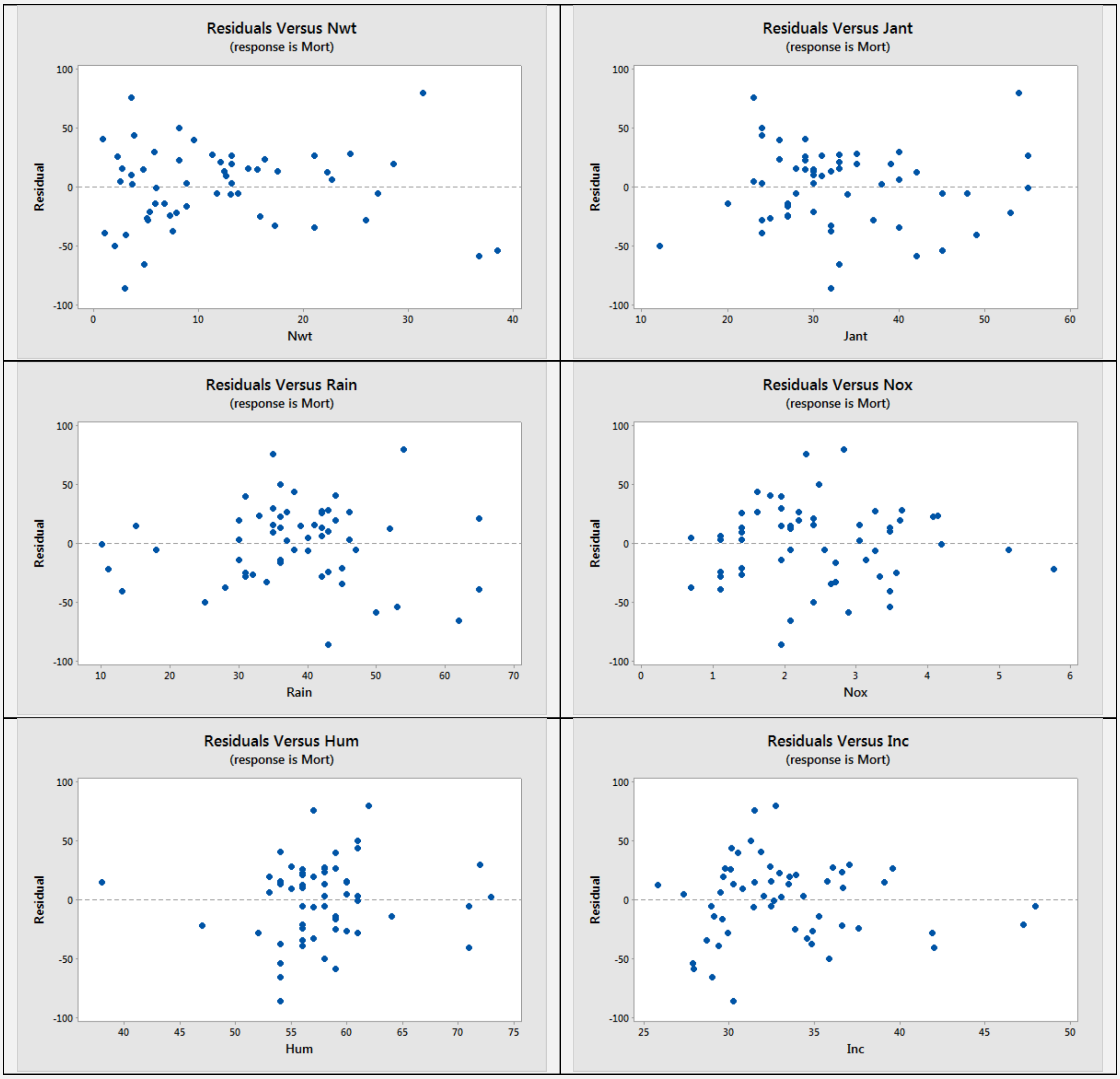
- Use the Histogram, Normal probability plot, and the Ryan-Joiner test to assess the N condition. [To do the Ryan-Joiner test, do the following: Stat→Basic Statistics → Normality Test, enter RESI1 for "variable" and click "Ryan-Joiner" for *Tests for Normality*.]
- Use the 6 scatterplots of Residuals versus fits and Residuals versus each of the predictors in the model to assess the L, I, and E conditions.
- Use the 2 scatterplots of Residuals versus each of the predictors *not* in the model to assess whether there are systematic patterns to suggest these predictors ought to be in the model.

Include all the plots in your write-up and briefly describe the dominant patterns in each plot and your conclusions.

The following plots show that the four assumptions seem reasonable. The large p-value of the Ryan-Joiner test indicates the normality of the residuals. Also, the histogram is reasonably symmetric and bell-shaped, and the normal probability plot points lie close to the diagonal line, supporting the N condition. Moving across each of the first 6 scatterplots from left to right, the residuals appear to average close to zero and remain equally variable, providing support for the L and E conditions. The lack of clear non-random patterns supports the I condition. The lack of clear non-random patterns in the residual plots versus the predictors not in the model indicates neither of these two predictors ought to be included in the model.







Residuals Versus Nwt
(response is Mort)

Residuals Versus Jant
(response is Mort)

Residuals Versus Rain
(response is Mort)

Residuals Versus Nox
(response is Mort)

Residuals Versus Hum
(response is Mort)

Residuals Versus Inc
(response is Mort)

(d) Based on the model from part (c), calculate a 95% confidence interval for E*(Mort)* for cities with the following characteristics: *Edu* = 10, *Nwt* = 15, *Jant* = 35, *Rain* = 40, and *Nox* = 2. Interpret your interval.

Using Minitab software, the 95% confidence interval is (946.273, 978.945). We are 95% confident that the average mortality rate of all cities with *Edu* = 10, *Nwt* = 15, *Jant* = 35, *Rain* = 40, and *Nox* = 2 is between 946 and 979 per 100,000.

(e) Based on the model from part (c), calculate a 95% prediction interval for *Mort* for a city with the following characteristics: *Edu* = 10, *Nwt* = 15, *Jant* = 35, *Rain* = 40, and *Nox* = 2. Interpret your interval (and say how and why it differs from the interval in the previous part).

Using Minitab software, the 95% prediction interval is (890.605, 1034.61). We are 95% confident that the mortality rate of a randomly selected city with *Edu* = 10, *Nwt* = 15, *Jant* = 35, *Rain* = 40, and *Nox* = 2 is between 891 and 1035 per 100,000. It is wider than the confidence interval for the mean because it has to take into account the random error in the model in addition to estimation error; there is more uncertainty involved when predicting an individual value versus estimating a mean.

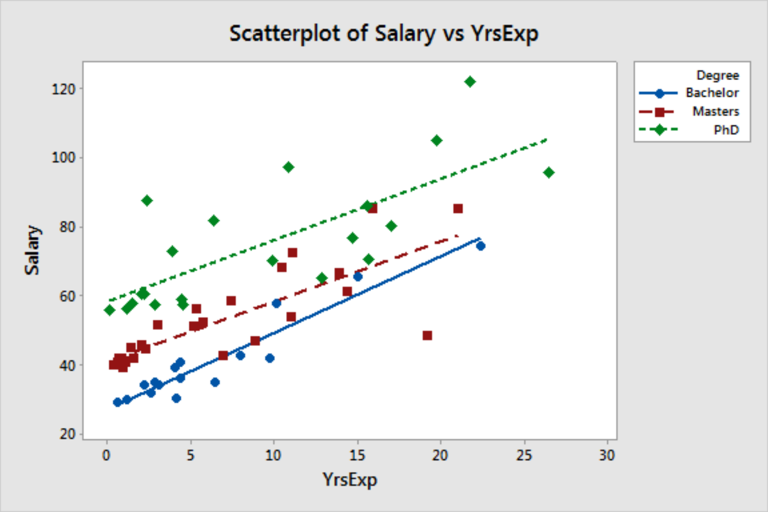
3. **(10 points)** Measurements on systolic blood pressure, body weight, gender (male or female), and age are recorded for n = 200 people in the 50-70 year old age group. The data will be used to estimate a multiple regression model for predicting systolic blood pressure from four predictors: body weight, gender, age, and an interaction between gender and age. Consider matrix notation for the model. Describe the columns of the X matrix. (How many columns will there be and what will be the numerical values in each column?).

There are p = 5 columns and n = 200 rows in the X matrix with the following entries:

- **All values in the first column are 1.**
- **The second column gives body weights for the 200 people.**
- **The third column gives values for an indicator variable (say 1 for male and 0 for female, or the reverse).**
- **The fourth column gives ages for the 200 people.**
- **The fifth column gives values that are the product of the third and fourth columns.**

4. (**4 + 6 + 4 + 4 + 4 + 7 + 4 + 4 + 8 = 45 points**) Use the "Salary" dataset. Three variables in the dataset are *Salary* = annual salary (thousands of U.S. dollars), *YrsExp* = years of work experience, and *Degree* = highest education degree for managers in software companies. The dataset also includes three indicator variables defined as *Deg1* = 1 if highest degree is Ph.D. and 0 otherwise, *Deg2* = 1 if Master's degree and 0 otherwise, and *Deg3* = 1 if highest degree is Bachelor's degree and 0 otherwise. The sample size is *n* = 63.

a) Below is a graph of salary versus experience with separate regression lines and symbols for the three different degrees. Discuss the important features of this graph, including whether you think that there may (or may not) be an interaction between degree and years of experience.



Average salary increases (linearly) with years of experience. For the same years of experience, the average salaries of Ph.D. people are greater than salaries for the Masters people, which in turn are larger than the average salaries of the Bachelors group. The slope between salary and years of experience is more or less the same for the three degree groups so the interaction may not be significant.

b) Fit a simple linear regression model with y = *Salary* and x = *YrsExp*. We will refer to this as the “reduced model” for parts (d) and (e).

a. What is the value of the SSE (sum of squared errors) for this regression?
b. What is the value of the error df for this regression?

c. SSE = 12412.8 and error df is 61.

c) Fit a multiple linear regression model with y = *Salary* and predictor variables *YrsExp*, *Deg1* and *Deg2*. We will refer to this as the “full model” for parts (d) and (e).

i. What is the value of the SSE (sum of squared errors) for this regression?
ii. What is the value of the error df for this regression?

SSE = 4948 and error df is 59.

d) For the model in part (c), calculate the value of an *F*-statistic for testing H_0: $\beta_2 = \beta_3 = 0$. [Assume that the variables were entered in the order described in part (c).] We are testing whether there is any degree effect on mean salary. The null hypothesis is that the coefficients multiplying the degree indicators are both 0. In other words, that there is no degree effect. The regression models that you fit in parts (b) and (c) will come into play here. Specifically, the *F*-statistic for this question is calculated as:

$$F = \frac{\frac{SSE(reduced) - SSE(full)}{df_E(reduced) - df_E(full)}}{\frac{SSE(full)}{df_E(full)}}.$$

This is what is referred to as a general linear *F*-statistic.

$$\boldsymbol{F = \frac{\frac{12412.8 - 4948}{61 - 59}}{\frac{4948}{59}} = 44.5.}$$

e) Refer to the *F*-statistic calculated in part (d).

i. What are the df values of this *F*-statistic? [Hint: The numerator degrees of freedom is given by df_E(reduced) – df_E(full) and the denominator degrees of freedom is given by df_E(full).]

The df are 2 and 59.

ii. What is the *p*-value for the test in part (d) and what is the appropriate conclusion for the test?

The *p*-value is essentially 0. We conclude that degree variables affect average salary. (By the way, we could tell this also by looking at the *t*-tests for the coefficients multiplying *Deg1* and *Deg2* in the model, both of which show statistical significance).

f) Refer to the regression model you fit in part (c) (with three predictor variables).

i. Write the estimated sample regression equation.

The regression equation is Predicted Salary = 29.27 + 1.841 YrsExp + 28.36 Deg1 + 10.71 Deg2.

ii. On the basis of this model, what is the estimated sample regression equation for those with a Bachelor's degree? (Hint: For a Bachelor's degree person, what are the values of *Deg1* and *Deg2*?)

For a Bachelor's Degree, Deg1 = 0 and Deg2 = 0, so Predicted Salary = 29.27 + 1.841 YrsExp + 28.36 (0) + 10.71 (1) = 29.27 + 1.841 YrsExp.

iii. On the basis of this model, what is the estimated sample regression equation for those with a Master's degree?

For a Master's Degree, Deg1 = 0 and Deg2 = 1, so Predicted Salary = 29.27 + 1.841 YrsExp + 28.36 (0) + 10.71 (1) = 39.98 + 1.841 YrsExp.

iv. On the basis of this model, what is the estimated sample regression equation for those with a Ph.D. degree?

For a Ph.D. Degree, Deg1 = 1 and Deg2 = 0, so Predicted Salary = 29.27 + 1.841 YrsExp + 28.36 (1) + 10.71 (0) = 57.63 + 1.841 YrsExp.

g) Refer to the estimated sample regression equation from part (f).

i. Write a sentence that interprets the numerical value of the sample regression coefficient that multiplies the variable *Deg1*. [Hint: Be careful. This coefficient describes the *difference* between two degree groups.]

The coefficient that multiplies the variable *Deg1* estimates the difference in average salaries for Ph.D.'s versus Bachelor's managers with the same experience. We estimate that the difference in average salaries is 28.36 thousand dollars per year.

ii. Write a sentence that interprets the numerical value of the sample regression coefficient that multiplies the variable *Deg2*.

The coefficient that multiplies the variable *Deg2* estimates the difference in average salaries for Master's versus Bachelor's managers with the same experience. We estimate that the difference in average salaries is 10.71 thousand dollars per year.

h) The model you fit in part (c) allows us to determine mean salary differences between Ph.D and Bachelor's managers and between Master's and Bachelor's managers. Which indicator variables would you need to include in a regression model that could be used to determine whether there is a significant difference between the mean salaries of Master's and Ph.D managers? Which estimated regression coefficient would estimate the difference in mean salaries? You don't actually have to estimate your model for this question. [Hint: Think about how the indicator variable that is left out affects the interpretation of the coefficients for the indicator variables that are included.]

There are two possible answers. We could do a multiple regression using *YrsExp*, *Deg1* and *Deg3* as *x*-variables. The coefficient multiplying *Deg1* would then give the difference between the Ph.D. and Master's groups. Alternatively, we could do a multiple regression using *YrsExp*, *Deg2* and *Deg3* as *x*-variables. The negative of the coefficient multiplying *Deg2* would estimate the desired difference.

i) Calculate two interaction variables that are the products, *Deg1*YrsExp* and *Deg2*YrsExp*. These are the variables needed for an interaction model. Carry out a test of whether or not there is a significant interaction between years of experience and highest degree achieved. Describe all details and state a conclusion. [Hint: To start, fit a multiple regression model that includes the interaction terms as well as the variables from part (c). This will be the full model. Think about what is the reduced model.]

The reduced model is the model from part (c). The full model adds the two interaction terms to this model. We find that the *F*-statistic is ((4948–4867.1)/(59–57)) / (4867.1/57) = 0.47 with 2 and 57 df, giving a *p*-value of 0.627. We cannot reject the null hypothesis so there is not a statistically significant interaction.

5. **(3 + 4 + 3 = 10 points)** Suppose we have a data set with five predictors, X1= GPA, X2= IQ, X3 = Gender (1 for Female and 0 for Male), X4= Interaction between GPA and IQ, and X5 = 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 b0 = 50, b1= 20, b2= 0.07, b3 = 35, b4 = 0.01 and b5= –10.

a) Which of the following is a true statement? Justify briefly.

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. TRUE Statement: For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is higher than 3.5.

iv. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is higher than 3.5.

(b) Predict the salary (in thousands of dollars) of a female with IQ of 110 and a GPA of 4.0 for the situation described above.

Predicted salary = 50 + 20(4) + 0.07(110) + 35(1) + 0.01(4)(110) –10(4)(1) = 50 + 80 + 7.7 + 35 + 4.4 – 40 = 137

(c) Do you agree with the following statement? Briefly justify your answer.
Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect

FALSE: The magnitude of the t-statistic (which depends on the coefficient estimate's standard error) determines the significance of a regression coefficient. The coefficient estimate tells us nothing by itself since it's scale depends on the measurement scale of the predictor.