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

**Instructions**: Use Word to type your answers within this document. Then, submit your answers in the appropriate dropbox in ANGEL by the due date. The point distribution is located next to each question. If there are multiple parts, then the points are divided equally over the subparts.

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.
   1. A small p-value associated with the Ryan-Joiner test for normality indicates that data are normally distributed.

FALSE. The null hypothesis is that the errors follow a normal distribution. Small p-values reject the null hypothesis.

* 1. 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.

TRUE. The confidence interval for the mean response works even if the error terms are only approximately normal. And, if you have a large sample, the error terms can even deviate substantially from normality. However the prediction interval depends strongly on the condition that the error terms are normally distributed.

* 1. 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. The formulas for the two are similar and show that as we increase the sample size n, the width of the interval decreases.

* 1. 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.
  2. 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.

1. **(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 (NOx). 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.
2. Fit the multiple linear regression model, E*(Mort)*=b0 + b1 *Edu* + b2 *Nwt* + b3 *Jant* + b4 *Rain* + b5 *Nox* + b6 *Hum* + b7 *Inc*. Report the SSE (sum of squared errors) and degrees of freedom (df) for error.
3. 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 ≤ x). The p-value for this problem is 1 – P(X ≤ x). Make sure you know why!]***

1. Fit the multiple linear regression model, E(Mort)=b0 + b1 Edu + b2 Nwt + b3 Jant + b4 Rain + b5 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.” 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:***

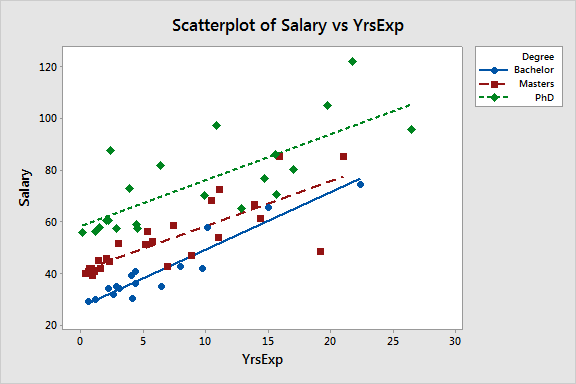
***Stat*🡪*Basic Statistics* 🡪 *Normality Test, enter RESI1 for “variable” and click “Ryan-Joiner” for Tests for Normality.]***

***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.***

1. 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.
2. 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).
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?).***
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.
5. 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.



1. 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).
   1. What is the value of the SSE (sum of squared errors) for this regression?
   2. What is the value of the error df for this regression?
2. 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).
   1. What is the value of the SSE (sum of squared errors) for this regression?
   2. What is the value of the error df for this regression?
3. For the model in part (c), calculate the value of an *F*-statistic for testing H0: β2 = β3 = 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:***

***.***

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

1. Refer to the *F*-statistic calculated in part (d).
   1. What are the df values of this *F*-statistic? [Hint: The numerator degrees of freedom is given by dfE(reduced) – dfE(full) and the denominator degrees of freedom is given by dfE(full).]
   2. What is the *p*-value for the test in part (d) and what is the appropriate conclusion for the test?
2. Refer to the regression model you fit in part (c) (with three predictor variables).
   1. Write the estimated sample regression equation.
   2. 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?)*
   3. On the basis of this model, what is the estimated sample regression equation for those with a Master's degree?
   4. On the basis of this model, what is the estimated sample regression equation for those with a Ph.D. degree?
3. Refer to the estimated sample regression equation from part (f).
   1. 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.]*
   2. Write a sentence that interprets the numerical value of the sample regression coefficient that multiplies the variable *Deg2*.
4. 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****.*]

1. 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.]***

1. **(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. Which of the following is a true statement? Justify briefly.
   1. For a fixed value of IQ and GPA, males earn more on average than females.
   2. For a fixed value of IQ and GPA, females earn more on average than males.
   3. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is higher than 3.5.
   4. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is higher than 3.5.
2. 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.
3. 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***

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

137

False