**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.

TRUE. When there is an interaction effect we have different slopes and intercepts as opposed to same slope and different intercepts in the absence of interaction effect. This is due to the fact that when the regression model contains interaction effects, the response function is not additive.

* 1. 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. Such coding is not advised. The ordering of the allocated numbers start to take significance even though there may not be any significance in the underlying data.

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.

SSE: 60417 DF: 48

1. 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!]***

We will test: H0 : β6 = β7 = 0

HA : at least one of {β6, β7} ≠ 0

The full model is: yi = β0 + β1xi,1 + β2xi,2 + β3xi,3 + β4xi,4 + β5xi,5 + β6xi,6 + β7xi,7 + εi

The reduced model is: yi = β0 + β1xi,1 + β2xi,2 + β3xi,3 + β4xi,4 + β5xi,5 + εi

Minitab output for the full model:

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 7 150577 21511.0 17.09 0.000

Edu 1 5637 5636.9 4.48 0.040

Nwt 1 47640 47639.8 37.85 0.000

Jant 1 13387 13386.8 10.64 0.002

Rain 1 10423 10423.4 8.28 0.006

Nox 1 14281 14281.3 11.35 0.001

Hum 1 446 445.9 0.35 0.555

Inc 1 88 88.3 0.07 0.792

Error 48 60417 1258.7

Minitab output for the reduced model:

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 5 150045 30009 24.62 0.000

Edu 1 7129 7129 5.85 0.019

Nwt 1 48905 48905 40.12 0.000

Jant 1 13653 13653 11.20 0.002

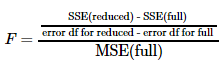
Rain 1 10550 10550 8.65 0.005

Nox 1 14754 14754 12.10 0.001

Error 50 60948 1219

Total 55 210994

To calculate the F-statistic, we see that SSE(full) = 60417 with error df = 48, MSE(full) = 1258.7, and SSE(reduced) = 60948 with error df = 50. Thus,



= ((60948-60417)/2) / 1258.7 = 0.2109

with df 2 and 48 (i.e., this F-statistic comes from an F2,48 distribution)

F distribution with 2 DF in numerator and 48 DF in denominator

x P( X ≤ x )

0.2109 0.189399

Therefore p-value = 1-0.189399 = 0.810601

Since the p-value > alpha = 0.05, we fail to reject the null hypothesis and conclude that Hum and Inc **don’t** provide significant information about the response, Mort, beyond the information provided by the other predictor variables.

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.***

Normality test

Ryan-Joiner test:



p-value > alpha (assumed 0.05), so we fail to reject the null hypothesis and conclude that we don’t have enough evidence to suspect normality of data. The data is normal is also supported by the Histogram, Normal probability plot as shown below:

, 

L, I, E conditions

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |

Overall the residual plots above support the L, I, E conditions. The residuals versus Nwt and Jant show some departure from the equal variance condition but nothing very significant.

Residuals versus each of the predictors not in the model

, 

The plots above don’t show any patterns and provide no evidence that including them will improve the model any further. This conclusion is in line with our computation of the F test in the previous part of the question.

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.

We get the following minitab output:

Fit SE Fit 95% CI 95% PI

962.609 8.13329 (**946.273, 978.945**) (890.605, 1034.61)

response variable for analysis is Mort = age adjusted mortality per 100,000

We can be 95% confident that the average mortality of city with Edu = 10, Nwt = 15, Jant = 35, Rain = 40, and Nox = 2 is between 946.273, 978.945 persons per 100,000

1. 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).

We get the following minitab output:

Fit SE Fit 95% CI 95% PI

962.609 8.13329 (946.273, 978.945) (**890.605, 1034.61**)

Interpretation: We can be 95% confident that the mortality of a randomly selected city with Edu = 10, Nwt = 15, Jant = 35, Rain = 40, and Nox = 2 will be between 890.605, 1034.61 persons per 100,000.

The predition interval is different because it depends on two components:

The variation due to estimating the mean µY with y^h and,

The variation in the responses y, which we denote as "σ2." (Note that quantity is estimated, as usual, with the mean square error MSE.)

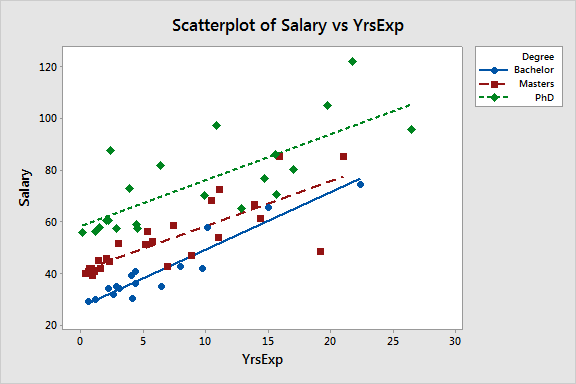
1. **(10 points)** Measurements on systolic blood pressure, body weight, gender (male or female), and age are recorded for n = 200 people in the 50-70 year old age group. The data will be used to estimate a multiple regression model for predicting systolic blood pressure from four predictors: body weight, 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?).***

# The model is

ySysPressureHat\_i = β0 + β1 \* WEIGHT\_i + β2 Gender\_i + β3 \* Age\_i + β4 \* Gender\_i \* Age\_i + εi

# The matrix X will have 200 rows with the following columns

1. Column of 1’s
2. Body weight of the 200 people
3. Gender coding (one coding is Male=0, Female=1) of the 200 people
4. Age of the 200 people
5. Multiplication of Gender Coding and age (col3 \* col4)
6. (**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.
7. 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.



There are following observations:

The slope for Bachelor line is different than the slope for Masters and PhD. This indicates that there may be an interaction affect here since a regression model contains interaction effects if the response function is not additive and cannot be written as a sum of functions of the predictor variables. There may be an interaction between YrsExp and Deg3

The slope for Masters and PhD appears to be the same. This indicates an absence of interaction affect. It seems that there is no interaction between YrsExp and Deg1 or Deg2.

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?

Minitab output for the reduced model is:

Analysis of Variance

Source DF Seq SS Seq MS F-Value P-Value

Regression 1 13104.6 13104.6 64.40 0.000

YrsExp 1 13104.6 13104.6 64.40 0.000

Error 61 12412.8 203.5

Lack-of-Fit 60 12161.9 202.7 0.81 0.730

Pure Error 1 250.9 250.9

Total 62 25517.4

Model Summary

S R-sq R-sq(adj) R-sq(pred)

14.2649 51.36% 50.56% 47.54%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 40.77 2.70 15.10 0.000

YrsExp 2.158 0.269 8.02 0.000 1.00

Regression Equation

Salary = 40.77 + 2.158 YrsExp

SSE = 12412.8

DF = 61

1. 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?

Minitab output for the full model is:

Analysis of Variance

Source DF Seq SS Seq MS F-Value P-Value

Regression 3 20569 6856.3 81.75 0.000

YrsExp 1 13105 13104.6 156.25 0.000

Deg1 1 6329 6329.4 75.46 0.000

Deg2 1 1135 1135.0 13.53 0.001

Error 59 4948 83.9

Total 62 25517

Model Summary

S R-sq R-sq(adj) R-sq(pred)

9.15815 80.61% 79.62% 77.37%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 29.27 2.55 11.48 0.000

YrsExp 1.841 0.176 10.45 0.000 1.04

Deg1 28.36 3.08 9.19 0.000 1.59

Deg2 10.71 2.91 3.68 0.001 1.54

Regression Equation

Salary = 29.27 + 1.841 YrsExp + 28.36 Deg1 + 10.71 Deg2

SSE = 4948

DF = 59

1. 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.***

The F = ((12412.8-4948)/2)/(4948/59) = 44.5052

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. The degrees of freedom for this F-statistic are 2 and 59

p-value computation

We have from minitab:

F distribution with 2 DF in numerator and 59 DF in denominator

x P( X ≤ x )

44.5052 1.00000

1. p-value = 1-1.0000 = 0.000 < 0.001

Since the p-value < 0.001 < alpha (assume=0.01) we can reject the null hypothesis and conclude that at least one and perhaps both of Deg1 and Deg2 are useful in the model.

1. Refer to the regression model you fit in part (c) (with three predictor variables).
   1. Write the estimated sample regression equation.

SalaryHat = y^ = 29.27 + 1.841 YrsExp + 28.36 Deg1 + 10.71 Deg2

* 1. 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?)*

Degree Bachelor, Deg1 = 0, Deg2 = 0, so

SalaryHat = y^ = 29.27 + 1.841 YrsExp

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

Degree Master, Deg1 = 0, Deg2 = 1, so

SalaryHat = y^ = 29.27 + 1.841 YrsExp + 10.71 = 39.98 + 1.841 YrsExp

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

Degree Master, Deg1 = 1, Deg2 = 0, so

SalaryHat = y^ = 29.27 + 1.841 YrsExp + 28.36 = 57.63 + 1.841 YrsExp

1. 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.]*

Difference in mean salary of managers for PhD holders versus Bachelor degree holders, assuming the same values for YrsExp

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

Difference in mean salary for Masters degree holders versus Bachelor degree holders, assuming the same values for YrsExp

1. 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****.*]

If we include Deg 1 and Deg 3 in the model, the coefficient that multiplies the variable Deg1 gives the mean salary differences between Ph.D and Master’s degree holders

Similarly, if we include Deg 2 and Deg 3 in the model, the coefficient that multiplies the variable Deg2 gives the mean salary differences between Master’s and PhD holders

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.]***

Minitab output for full model

Analysis of Variance

Source DF Seq SS Seq MS F-Value P-Value

Regression 5 20650.3 4130.1 48.37 0.000

YrsExp 1 13104.6 13104.6 153.47 0.000

Deg1 1 6329.4 6329.4 74.13 0.000

Deg2 1 1135.0 1135.0 13.29 0.001

Deg1YrsExp 1 8.8 8.8 0.10 0.749

Deg2YrsExp 1 72.5 72.5 0.85 0.361

Error 57 4867.1 85.4

Total 62 25517.4

Model Summary

S R-sq R-sq(adj) R-sq(pred)

9.24052 80.93% 79.25% 75.29%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 26.94 3.51 7.67 0.000

YrsExp 2.206 0.416 5.30 0.000 5.71

Deg1 31.26 4.74 6.59 0.000 3.69

Deg2 13.73 4.42 3.11 0.003 3.49

Deg1YrsExp -0.427 0.492 -0.87 0.389 6.99

Deg2YrsExp -0.471 0.511 -0.92 0.361 5.05

Regression Equation

Salary = 26.94 + 2.206 YrsExp + 31.26 Deg1 + 13.73 Deg2 - 0.427 Deg1YrsExp - 0.471 Deg2YrsExp

To perform the test of whether or not there is a significant interaction between years of experience and highest degree achieved we use the following:

The null hypothesis that makes this happen is H0 : β4 = β5 = 0

HA : at least one of the interaction parameters is not 0

The reduced model is simply E(Y) = β0 + β1YrsExp + β2Deg1 + β3Deg2

Full model is simply E(Y) = β0 + β1YrsExp + β2Deg1 + β3Deg2 + β4Deg1YrsExp + β5Deg2YrsExp

Reduced model is computed in part c with SSE = 4948 (DF = 59)

Full model from above: SSE = 4867.1 (DF = 57)

F = ((4948-4867.1)/2)/(4867.1/57) = 0.4737 with degrees of freedom 2 and 57

We have from minitab:

F distribution with 2 DF in numerator and 57 DF in denominator

x P( X ≤ x )

0.4737 0.374877

p-value = 1-0.374877 = 0.625123 > alpha (assumed=0.05), so we fail to reject the null hypothesis. Therefore we conclude that there isn’t a significant interaction between years of experience and highest degree achieved

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. (TRUE)
   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.

3rd is TRUE. The estimated regression function is

yHat = 50 + 20 \* x1\_gpa + 0.07 \* x2\_iq + 35 \* x3\_gender + 0.01 \* x1x2 - 10 \* x1x3

Further for females:

yHat = 50 + 20 \* x1\_gpa + 0.07 \* x2\_iq + 35 + 0.01 \* x1x2 - 10 \* x1

For males:

yHat = 50 + 20 \* x1\_gpa + 0.07 \* x2\_iq + 0.01 \* x1x2

The difference between the two equations above is: 35 – 10\*x1. The inflection point of this equation is 3.5 where it reverses direction.

1. 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.

We have: yHat = 50 + 20 \* x1\_gpa + 0.07 \* x2\_iq + 35 \* x3\_gender + 0.01 \* x4 - 10 \* x1x3

Plugging the values: yHat = 50 + (20 \* 4) + (0.07 \* 110) + (35 \* 1) + (0.01 \* 4 \* 110) – (10 \* 1 \* 4) = 137.1

1. 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 coefficient size doesn’t speak to the evidence of the interaction term. We will need the SE and that will help us compute the t-value which can lead us conclusion about the evidence.