**STAT 501 – Mid-Term Exam 2 Fall 2015**

**Instructions**: Use Word to type your answers within this document. Then, submit your answers in the appropriate dropbox in ANGEL by the due date and **within 3 hours of downloading the exam**. The point distribution is located next to each question.

1.(**6x2 = 12 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. Removing an outlier in a regression analysis will result in narrower confidence intervals.

True

1. The sum of leverages adds to p, the # of regression coefficients in the model (including the intercept).

True

1. Since leverages depend only on the predictors, removal of an outlier (unusual Y value) has no impact on the leverages.

False. While it is true than an outlier observation has its dependent-variable value being unusual given its predictor variables. An outlier can have unusual predictor values too. Since the computation of leverage involves the predictor values



When we remove an outlier, its predictors are removed too and therefore have an impact on the leverages.

1. In a simple linear regression (SLR) model, if a log transformation is performed on X to remedy some non-linearity, the mean value of Y is bound to change.

False. Log transformations on X will only change the X values and the Y values will stay the same. Therefore the mean value of Y will not change.

1. In model selection, the highest adjusted R2-value and the smallest S-value criteria always yield the same "best" models.

True. The two criteria are equivalent!

1. Regression models with different responses, but the same predictor X matrix, will have the same leverage values.

True

2. (**3+3+4+4+3+3 = 20 points**) Open the “Math Scores” dataset. The dataset consists of Math scores (Math) for 14 students with information about their undergraduate major (Major) and weekly hours of study (Hours). Your goal is to fit a regression model to express the dependence of Y (Math) on X (Hours) and Major.

1. Clearly define a set of indicator variables that could be used in a regression model to represent the qualitative variable Major. *[Hint: Think carefully about the number of indicator variables needed given the number of levels of Major and use “Engineering” as the reference level.]*

Since the qualitative variable (Major) defines 3 groups, we need 2 indicator variables (MajorScience and MajorHistory). Since engineering is the reference level we keep the value zero for the indicator variables

Major = Engineering: MajorScience = 0 and MajorHistory = 0

Major = Science: MajorScience = 1 and MajorHistory = 0

Major = History: MajorScience = 0 and MajorHistory = 1

1. Write a population multiple linear regression equation for predicting the Math scores in terms of Hours and Major. Since one’s major could impact the dependence of Y on X, the model should contain an interaction effect between Hoursand Major, together with their main effects. *[Hint: Your equation should include Y, X, the indicator variables you defined in part (a), interaction terms, and population regression coefficients (β’s).]*

The multiple regression equation

E(Math) = β0 + β1 \* Hours + β2 \* MajorScience + β3 \* MajorHistory + β4 \* Hours \* MajorScience + β5 \* Hours \* MajorHistory

yi = β0 + β1 \* Hours + β2 \* MajorScience + β3 \* MajorHistory + β4 \* Hours \* MajorScience + β5 \* Hours \* MajorHistory + ϵi

1. Conduct a hypothesis test at significance level 0.05 to determine if the average math score due to one hour of extra study per week differs by major (i.e., test if the slopes for two or more Major categories differ). Write out the null and alternative hypotheses, the test statistic, the p-value, and the conclusion. *[Minitab v17: Select Math as the Response, Hours as the Continuous predictor, Major as the categorical predictor, click “Model,” select both Hours and Major together in the Predictors box and click the Add button next to “Interactions through order 2.” Minitab v16: Create interaction terms using Calc > Calculator before fitting the regression model.]*

To perform the test of whether or not there is a significant interaction between hours and major 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(Math) = β0 + β1 \* Hours + β2 \* MajorScience + β3 \* MajorHistory

Full model is simply E(Math) = β0 + β1 \* Hours + β2 \* MajorScience + β3 \* MajorHistory + β4 \* Hours \* MajorScience + β5 \* Hours \* MajorHistory

Full Model Minitab:

**Regression Analysis: Math versus Hours, Major**

Method

Categorical predictor coding (1, 0)

Analysis of Variance

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

Regression 5 8419.22 1683.84 72.57 0.000

Hours 1 2310.40 2310.40 99.57 0.000

Major 2 238.43 119.22 5.14 0.037

Hours\*Major 2 0.22 0.11 0.00 0.995

Error 8 185.63 23.20

Lack-of-Fit 7 185.13 26.45 52.90 0.105

Pure Error 1 0.50 0.50

Total 13 8604.86

Reduce Model Minitab:

**Regression Analysis: Math versus Hours, Major**

Method

Categorical predictor coding (1, 0)

Analysis of Variance

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

Regression 3 8419.00 2806.33 151.00 0.000

Hours 1 5940.35 5940.35 319.62 0.000

Major 2 1334.69 667.34 35.91 0.000

Error 10 185.85 18.59

Lack-of-Fit 9 185.35 20.59 41.19 0.120

Pure Error 1 0.50 0.50

Total 13 8604.86

Reduced model from above: SSE = 185.85 (DF = 10)

Full model from above: SSE = 185.63 (DF = 8)

F = ((185.85-185.63)/2)/(185.63/8) = 0.0047 with degrees of freedom 2 and 8

We have from minitab:

F distribution with 2 DF in numerator and 8 DF in denominator

x P( X ≤ x )

0.0047 0.0046862

p-value = 1-0.0046862 = 0.9953138 > alpha (assumed=0.05), so we fail to reject the null hypothesis. Therefore we conclude that there isn’t a significant interaction between hours and major

1. Write a new population regression equation based on your conclusion to part (c). Fit this model and conduct two separate hypothesis tests to determine if the mean math score for a fixed number of weekly study hours differs by major. For each test, write out the null and alternative hypotheses, the test statistic, the p-value, and the conclusion.

E(Math) = β0 + β1 \* Hours + β2 \* MajorScience + β3 \* MajorHistory

On fitting this model we get from minitab:

**Regression Analysis: Math versus Hours, Major**

Method

Categorical predictor coding (1, 0)

Analysis of Variance

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

Regression 3 8419.00 2806.33 151.00 0.000

Hours 1 7084.32 7084.32 381.18 0.000

Major 2 1334.69 667.34 35.91 0.000

Error 10 185.85 18.59

Lack-of-Fit 9 185.35 20.59 41.19 0.120

Pure Error 1 0.50 0.50

Total 13 8604.86

Model Summary

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

4.31108 97.84% 97.19% 96.22%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 25.05 3.19 7.86 0.000

Hours 15.115 0.845 17.88 0.000 1.11

Major

History -16.78 3.01 -5.57 0.000 1.40

Science -22.40 2.73 -8.22 0.000 1.29

Regression Equation

Major

Engineering Math = 25.05 + 15.115 Hours

History Math = 8.27 + 15.115 Hours

Science Math = 2.65 + 15.115 Hours

Upon fitting MathHat = 25.05 + 15.11\* Hours – 16.78 \* MajorHistory – 22.4 \* MajorScience

Further, the mean math score for different majors will be:

Engineering: µ\_engg = β0 + β1 \* Hours ; MathHat = 25.05 + 15.115 Hours

Science: µ\_engg = (β0 + β2) + β1 \* Hours ; MathHat = 25.05 + 15.115 Hours - 22.4 = 2.65 + 15.115 Hours

History: µ\_engg = (β0 + β3) + β1 \* Hours ; MathHat = 25.05 + 15.115 Hours – 16.78 = 8.27 + 15.115 Hours

**Mean difference between science and engineering:**

µ\_science - µ\_engg = β2

In order for the mean difference to be 0, we will have to do the following null hypothesis:

Null : β2 = 0 and Alternate : β2 ≠ 0

T\* = coef / se coef = -22.40 / 2.73 = -8.22 ; p-value < 0.001

Since p-value < 0.001 < alpha (assume = 0.05) and therefore reject the null hypothesis. There is sufficient evidence at the α = 0.05 level to conclude that there is a significant difference in the mean math score between the Science and Engineering major.

**Mean difference between history and engineering:**

µ\_history - µ\_engg = β3

In order for the mean difference to be 0, we will have to do the following null hypothesis:

Null : β3 = 0 and Alternate : β3 ≠ 0

T\* = coef / se coef = -16.78 / 3.01 = -5.57 ; p-value < 0.001

Since p-value < 0.001 < alpha (assume = 0.05) and therefore reject the null hypothesis. There is sufficient evidence at the α = 0.05 level to conclude that there is a significant difference in the mean math score between the History and Engineering major.

1. Based on your conclusion to part (d), write three fitted sample regression equations that can be used to predict the Math scores for each major. *[Hint: Your equations should include number values, not β’s.]*

The mean math score for different majors will be:

Engineering: MathHat = 25.05 + 15.115 Hours

Science: MathHat = 25.05 + 15.115 Hours - 22.4 = 2.65 + 15.115 Hours

History: MathHat = 25.05 + 15.115 Hours – 16.78 = 8.27 + 15.115 Hours

1. Based on one of the equations from part (e), predict the Math score of a History major who studies 4 hours per week. *[Hint: A point estimate is sufficient.]*

History: MathHat = 25.05 + 15.115 Hours – 16.78 = 8.27 + 15.115 Hours

MathHat = 8.27 + (15.115 \* 4) = 68.73

1. (**3+2+3+2 = 10 points**) The following Minitab output resulted from a multiple linear regression model fit to response variable, *Y*, and predictor terms, *X1, X2, and X1X2*:

Coefficients

Term Coef SE Coef T-Value P-Value

Constant 4.49 1.89 2.37 0.022

X1 0.759 0.374 2.03 0.048

X2 0.965 0.426 2.26 0.028

X1\*X2 0.1742 0.0821 2.12 0.039

1. Conduct a hypothesis test for whether the interaction term, *X1X2*, can be dropped from the model. Write out the population model, null and alternative hypotheses, the test statistic, the p-value, and the conclusion.

Full model: E(Y) = β0 + β1 \* X1 + β2 \* X2 + β3 \* X1\*X2

Reduced model: E(Y) = β0 + β1 \* X1 + β2 \* X2

Null hypothesis: β3 = 0; Alternate: β3 ≠ 0

Test statistis: T\* = Coef / SE Coef = 0.1742 / 0.0821 = 2.12

p-value = 0.039 < alpha (assumed = 0.05) and therefore reject the null hypothesis. There is sufficient evidence at the α = 0.05 level to conclude that there is a significant interaction. We can’t drop β3 from the model.

1. Based on your conclusion to part (b), write the fitted sample regression equation.

yHat = 4.49 + 0.759 \* X1 + 0.965 \* X2 + 0.1742 \* X1 \* X2

1. State whether the following statements are supported by the Minitab output. (simply write “yes” or “no” for each statement).
   1. *X1* and *X2* are positively associated. NO
   2. *Y* and *X1* are positively associated for fixed values of *X2* between 0 and 10. YES
   3. The linear association between *Y* and *X1* increases as *X2* increases. YES
2. Use the fitted equation in part (b) to predict *Y* for an observation with *X1* = 6 and *X2* = 5. *[Hint: A point estimate is sufficient.]*

yHat = 4.49 + 0.759 \* X1 + 0.965 \* X2 + 0.1742 \* X1 \* X2

yHat = 4.49 + 0.759 \* 6 + 0.965 \* 5 + 0.1742 \* 6 \* 5 = 19.095

1. (**6+2+6+3+3 = 20 points**) The dataset “Savings” contains savings of 33 individuals along with their age. It is apparent that Y = Savings (in $) has a positive association with X = Age (in years). An appropriate regression model relating Savings to Age could be useful for predicting savings based on age. The most straightforward approach would be to fit a simple linear regression (SLR) model for Y vs X, provided that the LINE assumptions are satisfied. *[Consult “Worked Examples Using Minitab” in the Online Notes for help with any Minitab procedures.]*
   1. Fit an SLR model for Y vs X and perform a residual plot analysis to determine if the LINE assumptions are satisfied. Include a numerical test when checking for normality (use the Ryan Joiner test in Minitab). Discuss your findings and include any relevant graphs.

**Scatter plot:**



The scatter plots supports the L condition but some transformation may help with improving linearity. Equal variance condition is suspect.

**Residual vs fits:**



The above plots supports the L and I conditions. However equal variance condition is suspect as also indicated by the scatter plot.

**Ryan Joiner test:**



p-value < alpha (assumed 0.05), so we reject the null hypothesis and conclude that we have enough evidence to **suspect normality of data**.

* 1. Based on your conclusion in part (a), determine if any transformations are suggested for X and/or Y. *[Hint: You should find that both X and Y need to be transformed.]*

Y transformation will be helpful since there are some issues with the error terms – they don’t have equal variance and are not normal. Going by the heuristic that transforming the y values corrects problems with the error terms (unequal variances and/or non-normality) and may also help with non-linearity. Therefore it will be prudent to try transforming the y-variable to start with.

X transformation will be helpful since it will correct some of the non linearity indicated in the scatter and residual plot.

* 1. Fit an SLR model for the transformed variable(s) and comment on this model’s validity with supporting statements, numerical tests and/or plots.

We try savings vs ln(age)



We try ln(savings) vs age



We try ln(savings) vs ln(age)



The last plot looks most satisfactory for L

**Ryan Joiner test:**



p-value > alpha (assumed 0.05), so we fail to reject the null hypothesis and conclude that we have enough evidence to assume **normality of data**.

Residaul plot:



This shows that equal variance condition has been met

* 1. Use Minitab to compute a 95% confidence interval for the mean amount of savings (in $) expected for 40 year-olds based on the fitted model in part (c)*. [Hint: Remember to take into account the transformations to X and Y.]*

Regression Equation

lnsavings = 7.092 + 0.6910 lnage

We use lnage = ln(40) = 3.89

**Prediction for lnsavings**

Regression Equation

lnsavings = 7.092 + 0.6910 lnage

Variable Setting

lnage 3.69

Fit SE Fit 95% CI 95% PI

9.64120 0.163084 (9.30858, 9.97381) (8.46372, 10.8187)

So the confidence interval 11025.87431 and 21461.15852

* 1. Use Minitab to compute a 95% prediction interval for the amount of savings (in $) predicted for a randomly selected 40 year-old based on the fitted model in part (c). *[Hint: Remember to take into account the transformations to X and Y.]*

**Prediction for lnsavings**

Regression Equation

lnsavings = 7.092 + 0.6910 lnage

Variable Setting

lnage 3.69

Fit SE Fit 95% CI 95% PI

9.64120 0.163084 (9.30858, 9.97381) (8.46372, 10.8187)

So the prediction interval 4739.656767 and 49946.11484

1. (**7x2 =14 points**) The table below was obtained from the Best Subsets regression procedure for the “Infection Risk” dataset.

Response is InfctRsk

C

u

l C N

t X e u

S u r n r

t r a s s

R-Sq R-Sq Mallows a e y u e

Vars R-Sq (adj) (pred) Cp S y s s s s

1 35.5 34.8 30.2 1.1351 X

1 34.7 34.0 30.7 53.2 1.1428 X

2 53.0 52.0 48.3 14.0 0.97380 X X

2 46.3 45.1 40.9 29.2 1.0415 X X

3 57.0 55.6 51.4 7.1 0.93657 X X X

3 56.0 54.6 49.5 9.4 0.94740 X X X

4 59.3 57.5 53.1 4.0 0.91622 X X X X

4 58.7 56.9 52.0 5.4 0.92323 X X X X

5 59.3 57.1 51.0 6.0 0.92120 X X X X X

* 1. Calculate SSTO using information in the table.

MSE = 1.1351^2 = 1.288

SSE = MSE \* df = 1.288 \* 95 = 122.36

R2 = 35.5 = 1 – (122.36 / SSTO)

SSTO = SSE / (1-R) = 122.36 / (1-.355) = 189.71

* 1. Calculate the Mallows Cp for model with only *Stay* as the predictor (Show your work).



MSEall = 0.8486

n =97, p = 2, MSE=1.288: Cp = 2+((1.288-0.8486)\*(97-2)/1.288) = 10.6

Cp = 34.41

* 1. Based on the criteria listed in the table above select what you believe to be the “Best” model and write down its population regression equation. Support your answer.

E(InfctRsk) = β0 + β1 \* Stay + β2 \* Cultures + β3 \* Xrays+ β5 \* Nurses

It has lowest Cp, S. It has highest R2 adj

* 1. Would you consider this model to yield an unbiased predicted response? Support your answer.

Yes. The Cp = P = 4 for this model indicating that the bias is low or not present.

* 1. Name a model in the table that may yield a biased predicted response. Support your answer.

Model using only cultures. It has highest value of Cp

* 1. Use Minitab’s Backward Elimination procedure on this dataset and write down the fitted sample regression equation for the resulting “best” model. Use αr = 0.15 and the Minitab v17 command sequence: Stat > Regression > Regression > Fit regression model > Stepwise (select Backward Elimination for Method). For Minitab v16 use Stat > Regression > Stepwise.

Backward Elimination of Terms

Candidate terms: Stay, Cultures, Xrays, Census, Nurses

------Step 1----- ------Step 2------

Coef P Coef P

Constant -0.318 -0.331

Stay 0.2435 0.000 0.2464 0.000

Cultures 0.0477 0.000 0.04751 0.000

Xrays 0.01323 0.027 0.01315 0.025

Census 0.00016 0.922

Nurses 0.00202 0.239 0.002170 0.004

The equation is

InfctRskHat = -0.331 + 0.2464 \* Stay + 0.04751 \* Cultures + 0.01315 \* Xrays+ 0.002170 \* Nurses

* 1. State any extra useful information provided by the Backward Elimination output that is not available in the Best Subsets table above.

We know the coeff, their p-values

1. **(4X3 = 12 points)** Open the “Profits” dataset . The data indicate a positive linear association between interest rates and broker profits. The data are to be primarily used to obtain a regression model and compute confidence/prediction intervals.
   1. Detect any outliers, extreme X values and/or influential data points after fitting an SLR model for Y = profits and X = interest rate. Support your answer by means of scatter plots and quantitative diagnostic measures.

Regression:

**Regression Analysis: profit versus Interest**

Analysis of Variance

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

Regression 1 3491 3491.3 12.00 0.003

Interest 1 3491 3491.3 12.00 0.003

Error 19 5528 290.9

Total 20 9019

Model Summary

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

17.0569 38.71% 35.48% 25.41%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 12.63 6.88 1.84 0.082

Interest 3.73 1.08 3.46 0.003 1.00

Regression Equation

profit = 12.63 + 3.73 Interest

Fits and Diagnostics for Unusual Observations

Std

Obs profit Fit Resid Resid

20 90.00 20.09 69.91 4.30 R

21 71.00 68.62 2.38 0.18 X

R Large residual

X Unusual X

Levergae threshold 3(*p*/*n*) = 3 \* 2 / 21 = 0.28

Row 21 has leverage > 0.28, so removed.

DFITS threshold: = 2 \* (sqrt(3/(21-2-1))) = 0.816496581

Row 20 has DFIT > threshold so removed.

* 1. Repeat your regression analysis after deleting any outliers (unusual Y values).

Removed rows 20 and 21. Regression output

**Regression Analysis: profit versus Interest**

Analysis of Variance

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

Regression 1 3664.2 3664.17 544.60 0.000

Interest 1 3664.2 3664.17 544.60 0.000

Error 17 114.4 6.73

Total 18 3778.5

Model Summary

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

2.59386 96.97% 96.79% 96.15%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 2.25 1.24 1.82 0.087

Interest 5.038 0.216 23.34 0.000 1.00

Regression Equation

profit = 2.25 + 5.038 Interest

Fits and Diagnostics for Unusual Observations

Std

Obs profit Fit Resid Resid

3 13.815 8.698 5.117 2.14 R

R Large residual

**Correction: explicitly state which is influential, outlier and extreme etc**

* 1. Compare the results of your regression analyses and plots obtained from parts a and b.

R2 increased substantially

S / MSE decreased

T-stat value increase showing higher significance

* 1. In the context of this problem, comment on any detrimental effects if outliers were not removed.

The R2 value is very low . The se-coef is high so the confidence intervals will be much wider

1. (**2+4+3+3 = 12 points**) There is evidence in general that broker profits increase as interest rates increase while # of sales decline as interest rates increase. The tables below give results from two SLR regression analyses carried out using n = 5 observations.
   1. Fill in the two blanks in the tables below. *[There are different ways to approach this. One is to use formulas and properties to calculate the various measures – this will get you the most points. The other is to use Minitab to find the missing numbers – this can be used to check your calculations but will get you less points if you don’t also show how to calculate the numbers using formulas and properties.]*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Observation # | 1 | 2 | 3 | 4 | 5 |
| Interest rate (X) | 8.7 | 9.2 | 4.0 | 15.1 | 18.0 |
| Profit in thousands (Y) | 46.5 | 45.8 | 75.0 | 10.1 | 71.0 |
| Leverage | 0.243 | 0.226 | 0.597 | 0.336 | 0.597 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Observation # | 1 | 2 | 3 | 4 | 5 |
| Interest rate (X) | 8.7 | 9.2 | 4.0 | 15.1 | 18.0 |
| # of sales (Y) | 20 | 15 | 30 | 10 | 11 |
| Fitted values | 20.277 | 19.608 | 26.564 | 11.715 | 7.836 |
| Cook’s Distances | 0.001 | 0.261 | 1.415 | 0.073 | 1.20065 |

* 1. Which observation(s), if any, are extreme X values in predicting the broker profits? Justify your answer.

Case 1 – obs 3 and 5 have the highest leverage

Case 2 – Obs 3 has the highest cook’s distance. We know if Di is greater than 1, then the data point is quite likely to be influential.

**Correction: No extremes**

* 1. Which observation(s), if any, would you consider to be influential data point(s) in predicting the # of sales? Justify your answer.

Case 2 – Obs 3 and 5

Case 1 – none