**STAT 501 – Mid-Term Exam 1 – Fall 2015 – Due Oct 11**

**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. **(8x2 = 16 points)** State which of the following statements is TRUE and which is FALSE. For the statements that are false, (briefly) explain why they are false.
2. MSE provides an estimate for σ2.

TRUE

The MSE  estimates σ2, the common variance of the many subpopulations.

1. The null hypothesis for testing significance of the regression coefficient is written as H0: b1 = 0.

TRUE

1. In an ANOVA table for regression SSTO = SSR – SSE.

FALSE It is SSTO = SSR + SSE

1. In a simple linear regression model, the F-test for the ANOVA is equivalent to performing the following test: H0: β1 = 0 vs. Ha: β1 > 0.

TRUE

1. In a multiple linear regression model, the F-test for the ANOVA is equivalent to performing the following test: H0: β0 = β1 = β2 = 0 vs. Ha: Not all βk are 0 (k = 0, 1, 2).

FALSE. It is the test where null hypothesis is H0: β1 = β2 = 0 and Ha: Not all βk are 0 (k = 1, 2)

1. Two variables *Y* and *X* with a zero correlation coefficient can be related.

FALSE. Zero correlation coefficient means that there is no relationship.

1. In a multiple linear regression model, all four LINE assumptions must fail for the model to be invalid.

FALSE. All the assumptions should be valid for the model to be valid.

1. In a simple linear regression model with *Y*= β0 + β1 *X*, if the null hypothesis H0: β1 = 0 is rejected in favor of Ha: β1 ≠ 0, then the only possibility is that *Y* and *X* are linearly related.

FALSE. The hypothesis only indicates that there is a relationship. To find whether it is linear or curvilinear etc, we must perform other tests like the Lack of Linear Fit test.

1. **(3x2 = 6 points)** State with a brief justification whether the following are valid multiple linear regression equations:
2. Y = β0 + β1 X1 + β2 (X2/X3) + ε

Valid. The linear regression has the parameters and the response in a linear relationship. Transformation of the predictor variables is valid.

1. Y = β0 exp (β1 X1 + β2 X2) + ε

Invalid. The linear regression has the parameters and the response in a linear relationship. The above equation is having an exponential relation with the parameters.

1. Y = β0 + + ε

Invalid. The linear regression has the parameters and the response in a linear relationship. The linear regression model provides one parameter per one predictor or its transformation.

1. **(5 points)** Suppose that someone who has not taken STAT 501 writes a multiple linear regression model with two predictors as E(Yi) = β0 + β1 X i,1 + β2 X i,2 + εi. Explain what is wrong here and rectify the equation.

Incorrect. The linearity condition is that the mean of the error, E(𝜀𝑖), at each value of the predictor, xi, is zero.

Therefore E(Yi ) = 𝛽0 + β1 X i,1 + β2 X i,2

The population regression line μY=E(Y)=β0 + β1 X i,1 + β2 X i,2

1. **(10 points)** Suppose 3 predictors (plus an intercept term) are candidates to be in a model for predicting *Y*. The SSE = 1200, SSTO = 3000 and n = 44. Complete the following ANOVA table with numerical values.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source | df | SS | MS | F | *p*-value |
| Regression | 3 | 1800 | 600 | 20 | 0.000 |
| Error | 40 | 1200 | 30 | xxxx | xxxx |
| Total | 43 | 3000 | xxxx | xxxx | xxxx |

p-value: 1 – F(3,40, 20) = 1- 1.00000 = 0.000

1. **(3x5 = 15 points)** Consider the following sample data:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *xi,1* | 12 | 15 | 19 | 22 | 15 | 14 |
| *xi,2* | 33 | 37 | 41 | 31 | 30 | 39 |
| *xi,3* | 4 | 1 | 0 | 0 | 7 | 9 |
| *yi* | 100 | 82 | 96 | 110 | 90 | 91 |

Suppose we wish to fit a multiple linear regression model (with the three predictors plus the intercept). Write the ***X***-matrix, the ***Y***-vector and the ***β***-vector for this problem. (Notice that I only request the ***β***-vector and not the ***b***-vector!)

Y Vector is: (dimension: 6X1)

100

82

96

110

90

91

Y’ (Y vector transpose) = [100, 82, 96, 110, 90, 91] (dimension: 1X6)

X Matrix is: (dimension: 6X4)

1 12 33 4

1 15 37 1

1 19 41 0

1 22 31 0

1 15 30 7

1 14 39 9

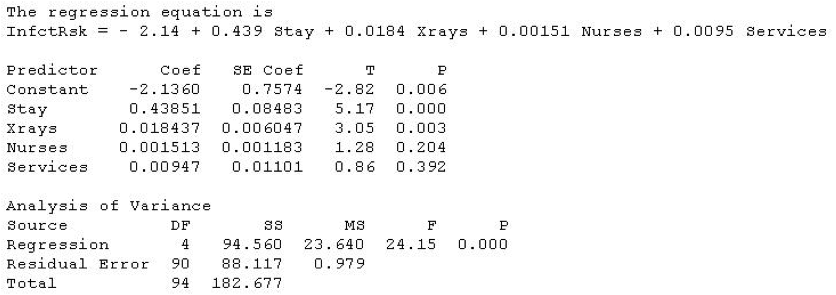
β vector is: (dimension: 4X1)

|  |
| --- |
| β0 |
| β1 |
| β2 |
| β3 |

1. **(2x5 = 10 points)** With data from n = 95 hospitals, a regression is done to analyze the relationship between *Y* = *InfctRsk*, the infection risk at a hospital (as a percent) and *X1* = *Stay* the average length of patient stay (days), *X2* = *Xrays* the percentage of patients who get x-rays at the hospital, *X3* = *Nurses*, the number of nurses employed at the hospital, and *X4* = *Services*, a measure of how many medical services are available at the hospital. We fit a multiple regression model,

*Yi = β0 + β1X1 + β2X2 + β3X3 + β4X4 +ε*,

to the data with results as follows:



Scatterplots and residual plots (not shown here) suggest no difficulties with the data or the model.

1. Interpret the result of the test of the regression coefficient for the variable *Services* using a significance level of 0.05 by indicating the null and alternative hypotheses, the test statistic, and the p-value, and stating a conclusion in this context.
2. Interpret the result of the test given in the ANOVA table using a significance level of 0.05 by indicating the null and alternative hypotheses, the test statistic, and the p-value, and stating a conclusion in this context.
3. The test and its interpretation is:

Null hypothesis H0 : β4 = 0

Alternative hypothesis Ha: β4 ≠ 0

Test Statistic t\* = Coef / SECoef = 0.00947/0.01101 = 0.86

Since the reported p-value is 0.392 > alpha (0.05).

Because the P-value is greater than alpha, we fail to reject the null hypothesis and conclude that β4= 0

We don’t have sufficient evidence to reject the null hypothesis that slope β4= 0. We can’t say that there is a statistically significant linear relationship between InfectionRisk and Services

1. The test and its interpretation is:

Null hypothesis H0 : β1 = β2 = β3 = β4 = 0

Alternative hypothesis Ha: Not all βk are 0 (k = 1, 2, 3, 4)

F Statistic = MSR / MSE = 23.640/0.979 = 24.15

The P-value for the analysis of variance F-test (P < 0.001) suggests that the model containing Stay, Xray, Nurses and Services is more useful in predicting InfctRsk than not taking into account the predictors. **It suggests that at least one of the four predictors is significant.**

1. **(3+5+5 = 13 points)** Open the “Archaeopteryx” dataset. Archaeopteryx is an extinct beast having feathers like a bird, but teeth and a long boney tail like a reptile. Only six fossil specimens are known, but because these specimens differ greatly in size, some scientists believe they are different species rather than individuals from the same species. The dataset consists of femur and humerus bone measurements for five of the six specimens (the sixth specimen did not have these bones preserved).
2. Calculate the Pearson correlation coefficient and describe what this means in terms of the strength and direction of the relationship.
3. Suppose we wish to produce a regression model where the humerus measurement is the response and the femur measurement is the predictor.
   * 1. Fit a simple linear regression model to the data using Minitab. Test whether a “regression through the origin” model would be appropriate for this data. Report the p-value of the test statistic and your conclusion from this test.
     2. Regardless of the outcome of the test in part (i), fit a “regression through the origin” model in Minitab and report the fitted regression equation. Describe any notable differences between the two sets of results.

a. Pearson correlation of Humerus and Femur = 0.994

The relationship is very strong and +ve. It means that if one variable increases than the other one will increase too.

* The relationship is positive. As the humerus increases, the femur increases (linearly).
* The relationship is quite strong (since the value is pretty close to 1)

b. i. The simple linear regression model is (not RTO):

**Regression Analysis: Humerus versus Femur**

Analysis of Variance

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

Regression 1 998.21 998.215 254.10 0.001

Femur 1 998.21 998.215 254.10 0.001

Error 3 11.79 3.928

Total 4 1010.00

Model Summary

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

1.98203 98.83% 98.44% 96.51%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant -3.66 4.46 -0.82 0.472

Femur 1.1969 0.0751 15.94 0.001 1.00

Regression Equation

Humerus = -3.66 + 1.1969 Femur

In order to test whether “regression through the origin” model would be appropriate for this data we perform the following test:

Null hypothesis H0 : β0 = 0

Alternative hypothesis Ha: β0 ≠ 0

Test Statistic t\* = Coef / SECoef = -3.66 / 4.46 = -0.82

Since the reported p-value is 0.472 > alpha (0.05).

Because the P-value is greater than alpha, we fail to reject the null hypothesis and conclude that β0 = 0

The conclusion is “regression through the origin” model will be appropriate for this data

1. **(5x5 = 25 points)** Open the “HeightArm” dataset. The data are *X* = Arm (upper arm length in cm) and *Y* = Height (standing height in cm) of individuals with height over 140 cm randomly selected from the 2007-8 National Health and Nutrition Examination Survey. We would like to examine the relationship between these two variables. Include relevant output from Minitab (or any other software you may be using) for this analysis.
2. Report the fitted simple linear regression model for this data and provide a scatterplot of the data with the fitted regression line overlaid on it.



Model: Height = 53.84 + 3.043 Arm

1. Produce a plot of the residuals versus the fitted values and a normal probability plot of the residuals. What are your impressions based on these plots? What do the plots tell us about our fitted model?





The residuals indicate the following:

1. It appears that the residuals bounce randomly around the residual = 0 line. There is no obvious pattern in the residuals indicating that the model is valid
2. The variance of the residuals appears to be almost constant across the different fitted values indicating that the equal variance assumption is valid.

The residual plot above generally supports the assumptions of the model. There is no obvious curvature and the variance looks to be constant. There is no evidence of outliers and the residuals appear to be independent.

1. Test whether or not there is a statistically significant linear relationship at a 5% significance level by using the results from the ANOVA table. (For full credit you must report the F-statistic, the degrees of freedom, the p-value and a conclusion in the context of the problem. Do not just cite the ANOVA table, but clearly identify all the relevant values.)

Analysis of Variance

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

Regression 1 5482.8 5482.85 140.41 0.000

Arm 1 5482.8 5482.85 140.41 0.000

Error 73 2850.6 39.05

Lack-of-Fit 50 2329.8 46.60 2.06 0.031

Pure Error 23 520.9 22.65

Total 74 8333.5

The ANOVA table is above.

Null hypothesis H0 : β0 = 0

Alternative hypothesis Ha: β0 ≠ 0

The test statistic F\* = MSR / MSE = 5482.85 / 39.05 = 140.41

The F-Statistic is distributed as F distribution with 1 numerator degree of freedom and 73 denominator degree of freedom. The P-value is determined by comparing F\* to an F distribution with 1 numerator degree of freedom and 73 denominator degrees of freedom.

The p-value for the F-Test < 0.001

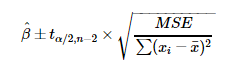
The P-value is very small. It is unlikely that we would have obtained such a large F\* statistic if the null hypothesis were true.

Therefore, we reject the null hypothesis H0: β1 = 0 in favor of the alternative hypothesis HA: β1 ≠ 0. There is sufficient evidence at the α = 0.05 level to conclude that there is a linear relationship between Height and Arm.

1. Construct a 95% confidence interval for the slope parameter and interpret the interval.

(1-α)100% t-interval for the slope parameter β1 is: Sample estimate ± (t-multiplier × standard error)

The confidence interval is given by:



**Regression Analysis: Height versus Arm**

Analysis of Variance

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

Regression 1 5482.8 5482.85 140.41 0.000

Arm 1 5482.8 5482.85 140.41 0.000

Error 73 2850.6 39.05

Lack-of-Fit 50 2329.8 46.60 2.06 0.031

Pure Error 23 520.9 22.65

Total 74 8333.5

Model Summary

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

6.24897 65.79% 65.32% 63.93%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 53.84 9.70 5.55 0.000

Arm 3.043 0.257 11.85 0.000 1.00

t(0.025, 73) = 1.993

Then, the 95% confidence interval for β1 is 3.043 ± (1.993\*0.257) = Interval is 2.5308, 3.5552

Interpretation: We can be 95% confident that the population slope is between 2.5308, 3.5552. That is, we can be 95% confident that for every additional one cm increase in arm length, the mean height increases between 2.5308, 3.5552 cms

1. Construct a 95% prediction interval for a new individual’s height for an upper arm length of 38 cm and interpret the interval.

The general formula in words is as always: Sample estimate ± (t-multiplier × standard error)



The unbiased estimator is: 

t(0.025, 73) = 1.993

The prediction output is

**Prediction for Height**

Regression Equation

Height = 53.84 + 3.043 Arm

Variable Setting

Arm 38

Fit SE Fit 95% CI 95% PI

169.466 0.726508 (168.018, 170.914) (156.928, 182.004)

The interval is between 156.928, 182.004

Interpretation: The interval represents the height of person for a randomly selected individual whose arm length is 38 cms.