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, explain why they are false.
   a) MSE provides an estimate for σ^2. TRUE
   b) The null hypothesis for testing significance of the slope regression coefficient is written as H_0: $b_1 = 0$. FALSE. The term b_1 is the estimator of the population slope β_1.
   c) In an ANOVA table for regression SSTO = SSR – SSE. FALSE. SSTO = SSR + SSE.
   d) In a simple linear regression model, the F-test for the one-way ANOVA is equivalent to performing the following test: H_0: $\beta_1 = 0$ vs. H_a: $\beta_1 > 0$. FALSE. The F-test is equivalent to testing H_0: $\beta_1 = 0$ vs. H_a: $\beta_1 \neq 0$
   e) In a multiple linear regression model, $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, the F-test for the one-way ANOVA is equivalent to performing the following test: H_0: $\beta_0 = \beta_1 = \beta_2 = 0$ vs. H_a: Not all β_k are 0 (k = 0, 1, 2). TRUE
   f) Two variables *Y* and *X* with a zero correlation coefficient can be related. TRUE. There could be a non-linear relationship.
   g) In a multiple linear regression model, all four LINE assumptions must fail for the model to be invalid. TRUE
   h) In a simple linear regression model with $Y = \beta_0 + \beta_1 X$, if the null hypothesis H_0: $\beta_1 = 0$ is rejected in favor of H_a: $\beta_1 \neq 0$, then the only possibility is that *Y* and *X* are linearly related. FALSE. Not necessarily LINEARLY related. But related somehow such that a change in X causes a change in Y.

2. **(3x2 = 6 points)** State with a brief reason whether the following are valid multiple linear regression equations:
   a) $Y = \beta_0 + \beta_1 X_1 + \beta_2 (X_2/X_3) + \varepsilon$ Valid because we can label $X'_2 = (X_2 / X_3)$
   b) $Y = \beta_0 \exp(\beta_1 X_1 + \beta_2 X_2) + \varepsilon$ Valid because we can take the log of the whole thing
   c) $Y = \beta_0 + \frac{\beta_1 X}{\beta_2 + X} + \varepsilon$ Invalid because we can't separate the β_1 and β_2 terms

3. **(5 points)** Suppose that someone who has not taken STAT 501 writes a multiple linear regression model with two predictors as $E(Y_i) = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \varepsilon_i$. Explain what is wrong here and rectify the equation.

   The equation should be either $Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \varepsilon_i$ or $E(Y_i) = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2}$. If you have the expectation on the left, you shouldn't have the error term on the right since the expectation would be taken for that quantity as well which is 0.

4. **(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

5. **(3x5 = 15 points)** Consider the following sample data:

$x_{i,1}$	12	15	19	22	15	14
$x_{i,2}$	33	37	41	31	30	39
$x_{i,3}$	4	1	0	0	7	9
y_i	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!)

$$\mathbf{Y} = \begin{bmatrix} 100 \\ 82 \\ 96 \\ 110 \\ 90 \\ 91 \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & 12 & 33 & 4 \\ 1 & 15 & 37 & 1 \\ 1 & 19 & 41 & 0 \\ 1 & 22 & 31 & 0 \\ 1 & 15 & 30 & 7 \\ 1 & 14 & 39 & 9 \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

6. **(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 X_1 = *Stay* the average length of patient stay (days), X_2 = *Xrays* the percentage of patients who get x-rays at the hospital, X_3 = *Nurses*, the number of nurses employed at the hospital, and X_4 = *Services*, a measure of how many medical services are available at the hospital. We fit a multiple regression model,

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon,$$

to the data with results as follows:

```
The regression equation is
InfctRsk = - 2.14 + 0.439 Stay + 0.0184 Xrays + 0.00151 Nurses + 0.0095 Services

Predictor       Coef    SE Coef       T       P
Constant     -2.1360     0.7574   -2.82   0.006
Stay         0.43851    0.08483    5.17   0.000
Xrays       0.018437   0.006047    3.05   0.003
Nurses      0.001513   0.001183    1.28   0.204
Services     0.00947    0.01101    0.86   0.392

Analysis of Variance
Source           DF       SS       MS       F       P
Regression        4   94.560   23.640   24.15   0.000
Residual Error   90   88.117    0.979
Total            94  182.677
```

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

a) 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.

The null hypothesis is H_0: $\beta_4 = 0$, and the alternative hypothesis is H_a: $\beta_4 \neq 0$. The test statistic is $t = \frac{0.00947}{0.01101} = 0.860$ This has a p-value of 0.392 which is not less than the significance level of 0.05. Therefore, we cannot reject the null hypothesis and conclude that $\beta_4 = 0$.

b) 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.

The test in the ANOVA table is testing the null hypothesis H_0: All $\beta_k = 0$ vs H_a: not all β_k are equal to 0. The test statistic is $F = \frac{MSR}{MSR} = \frac{23.640}{0.979} = 24.15$ This has a p-value of 0.000 which is less than the significance level of 0.05. Therefore, we can reject the null hypothesis and conclude that at least one β_k is not equal to 0.

7. **(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).

a) Calculate the Pearson correlation coefficient and describe what this means in terms of the strength and direction of the relationship.

$r = 0.99415$

The r value indicates a strong positive linear relationship between humerus measurement and femur measurement.

b) Suppose we wish to produce a regression model where the humerus measurement is the response and the femur measurement is the predictor.

i. 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.

Humerus = -3.66 + 1.1969 Femur
A "regression through the origin" model would be where the null hypothesis, H_0: $\beta_0 = 0$ vs H_a: $\beta_0 \neq 0$. The test statistic is t = -0.82 with p-value of 0.472. Based on this, we cannot reject the null hypothesis and conclude that a "regression through the origin" model would be appropriate.

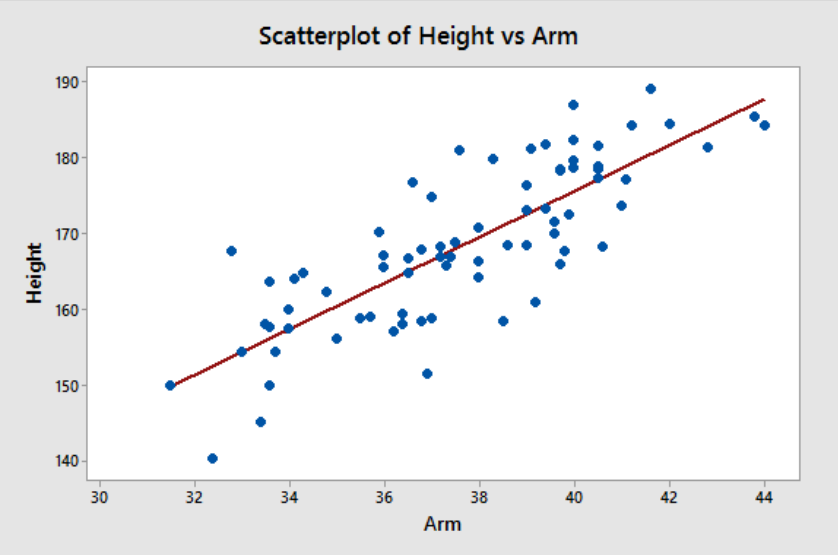
ii. Regardless of the outcome of the test in part (i), fit a "regression through the origin" model in Minitab and report the fitted regression equation. Describe any notable differences between the two sets of results.

Humerus = 1.1365 Femur
The most notable difference between the two results is that the p-value for the coefficient as well as p-value for the regression improved significantly. The t-value for femur as well as the F-value for regression went up.

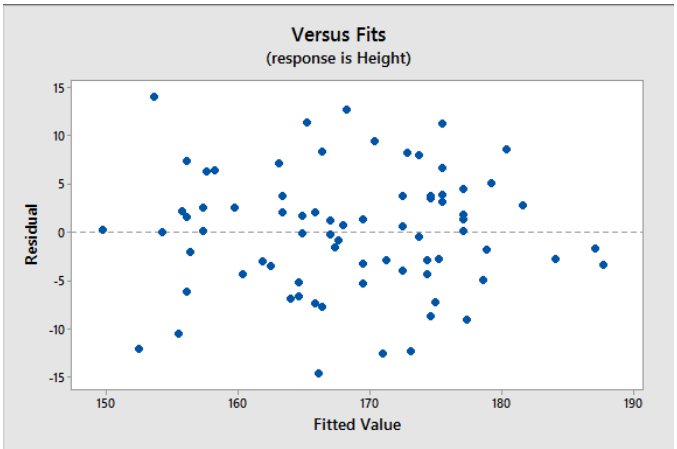
8. **(5x5 = 25 points)** Open the "HeightArm" dataset. The data are *X* = Arm (upper arm length in cm) and *Y* = Height (standing height in cm) of individuals with height over 140 cm randomly selected from the 2007-8 National Health and Nutrition Examination Survey. We would like to examine the relationship between these two variables. Include relevant output from Minitab (or any other software you may be using) for this analysis.

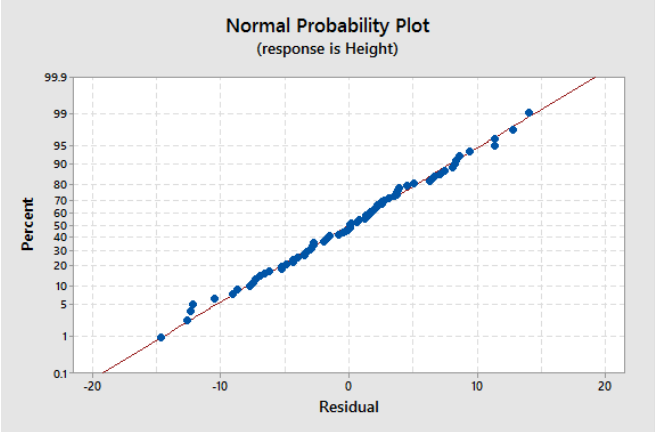
a) Report the fitted simple linear regression model for this data and provide a scatterplot of the data with the fitted regression line overlaid on it.

Height = 53.84 + 3.043 Arm



b) 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?





Based on the residual vs fit plot, it looks like the horizontal band is evenly distributed around the residual = 0 line. There are no obvious outliers and the residuals appear to be independent. The variance looks to be constant and the normality assumption seems to hold.

c) 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.)

The test here is H_0: $\beta_1 = 0$ vs H_a: β_1 not equal to 0. According to minitab, the test statistic T is 11.85 (with 73 degrees of freedom). The p-value of 0.000 is less than 0.005, so we can reject the null hypothesis and conclude that β_1 is not equal to 0. The F-test confirms this same conclusion. The test statistic F* = 140.41 with p-value of 0.000 says that we can reject the null hypothesis.

d) Construct a 95% confidence interval for the slope parameter and interpret the interval.
$3.043 - 1.993\ (\ 0.257\ ) \leq \beta_1 \leq 3.043 + 1.993\ (\ 0.257\ )$
$2.531 \leq \beta_1 \leq 3.555$

The interval indicates that we can be 95% confident that the standing height in cm of an individual will increase by somewhere between 2.531 and 3.555 cm for each additional cm in arm length.

e) Construct a 95% prediction interval for a new individual's height for an upper arm length of 38 cm and interpret the interval.
The prediction interval is (156.928, 182.004). The interval indicates that with 95% confidence, we can predict that the new individual's height will be somewhere between 156.928 and 182.004 cm tall.