STAT 501 – Mid-Term Exam 1 – Solutions – Fall 2015

1. **(8x2 = 28 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 null hypothesis for testing significance of the slope regression coefficient is written as H_0: $\beta_1 = 0$.**
   c) In an ANOVA table for regression SSTO = SSR – SSE. **False: In an ANOVA table for regression 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. It should be 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). **False. It should be H_0: $\beta_1 = \beta_2 = 0$ vs. H_a: Not both β_k are 0 (k = 1, 2).**
   f) Two variables *Y* and *X* with a zero correlation coefficient can be related. **True.**
   g) In a multiple linear regression model, all four LINE assumptions must fail for the model to be invalid. **False. Only one LINE assumption needs to fail for the model to be invalid.**
   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. Rejection of H_0 can lead to two other possibilities: a Type 1 error, which means that β_1 may indeed be zero, or the Y and X relationship could be nonlinear, for example, there could be a quadratic effect in addition to the linear effect.**

2. **(3x2 = 6 points)** State with a brief reason whether the following are valid multiple linear regression equations:
   a) $Y = \beta_0 + \beta_1 X_1 + \beta_2 (X_2/X_3) + \varepsilon$
      Yes, this is a valid multiple linear regression equation because it is linear in the parameters.
   b) $Y = \beta_0 \exp(\beta_1 X_1 + \beta_2 X_2) + \varepsilon$
      No, this is not a valid multiple linear regression equation because it is not linear in the parameters.
   c) $Y = \beta_0 + \frac{\beta_1 X}{\beta_2 + X} + \varepsilon$
      No, this is not a valid multiple linear regression equation because it is not linear in the parameters.

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 correct form of the equation is either $E(Y_i) = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2}$ or $Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \varepsilon_i$. The error term ε_i indicates the departure of individual observed Y_i from its expected value $E(Y_i)$. $E(Y_i)$ is completely determined by the regression equation and no error term can be attached with it.

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	**4–1=3**	**3000–1200=1800**	**1800/3=600**	**600/30=20**	**~0**
Error	**44–4=40**	**1200**	**1200/40=30**	**xxxx**	**xxxx**
Total	**44–1=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!)

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

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.

We are testing H_0: $\beta_4 = 0$ vs. H_a: $\beta_4 \neq 0$.

Since the t-statistic is 0.86 with a p-value of 0.392, we do not reject the null hypothesis and thus conclude that *Services* is not a statistically significant predictor of *InfctRisk* when controlling for the other predictors in the model.

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.

We are testing H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ vs. H_a: at least one of β_1, β_2, β_3, or $\beta_4 \neq 0$.

Since the F-statistic is 24.15 with a p-value of 0.000, we reject the null hypothesis and thus conclude that at least one of *Stay*, *Xrays*, *Nurses*, and *Services* is a statistically significant predictor of *InfctRisk*.

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.

The Pearson correlation coefficient is 0.994, which indicates a strong positive linear relationship between the femur and humerus measurements.

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.

The Minitab results are:

```
Humerus = - 3.66 + 1.20 Femur

Predictor      Coef  SE Coef      T      P
Constant    -3.660    4.459  -0.82  0.472
Femur      1.19690  0.07509  15.94  0.001
```

The p-value to test H_0: $\beta_0 = 0$ is 0.472, so we cannot reject H_0 at a 5% significance level, and a "regression through the origin" model may be appropriate for this data.

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.

The Minitab results are:

```
Humerus = 1.14 Femur

Predictor      Coef  SE Coef      T      P
No constant
Femur       1.13651  0.01430  79.45  0.000
```

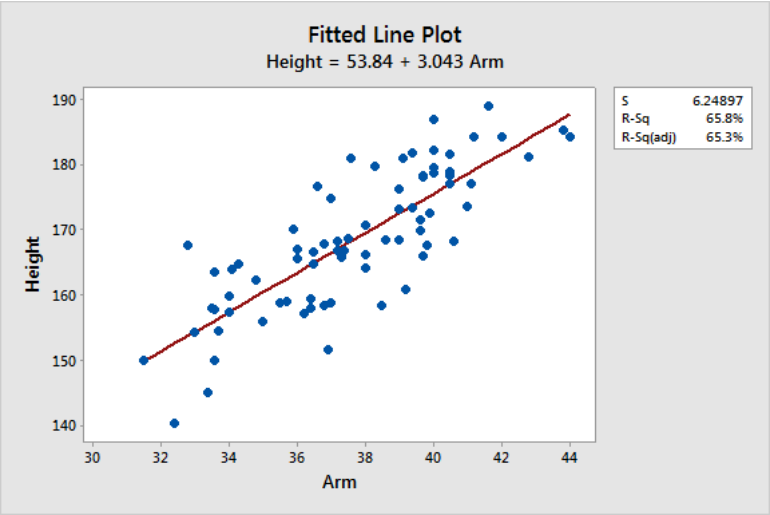
The slope estimate is similar in both cases, but the slope standard error is much smaller for the "regression through the origin" model.

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.

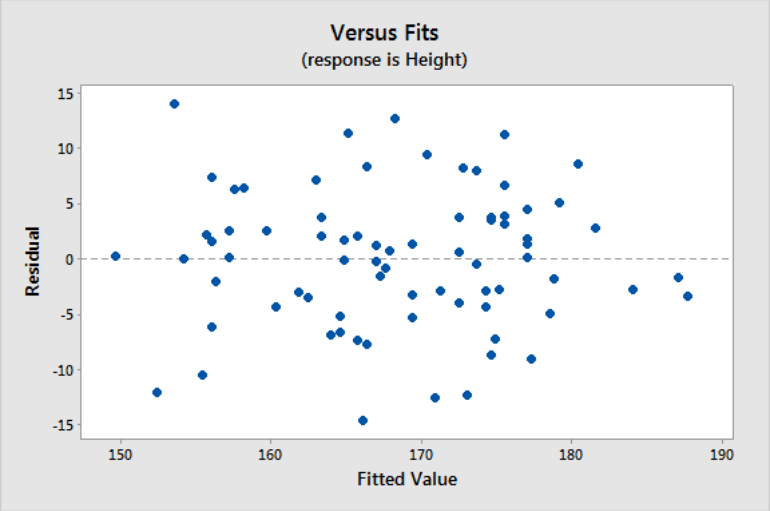
a) Report the fitted simple linear regression model for this data and provide a scatterplot of the data with the fitted regression line overlaid on it.

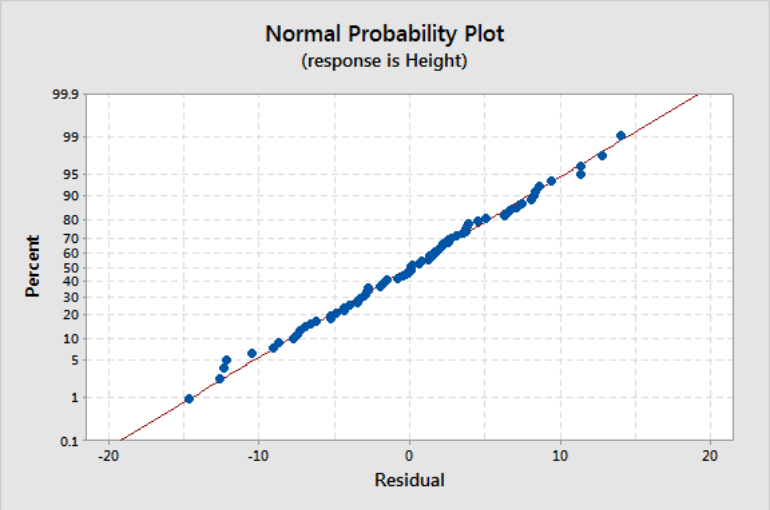
```
Model Summary
       S    R-sq  R-sq(adj)  R-sq(pred)
 6.24897  65.79%     65.32%      63.93%

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
```



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?





The appearance of the plots supports the LINE assumptions of simple linear regression since the points in the residual plot appear in a horizontal band (equal variance) with

no strong trends (linearity) or unusual patterns (independence) and the points in the normal probability plot lie close to the diagonal line (normality).

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.)

```
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
```

Test H_0: Height is not linearly associated with Arm using the ANOVA F-statistic of 140.41, which under H_0 has an F-distribution with 1 numerator and 73 denominator degrees of freedom. The p-value is 0.000, so we reject H_0 at a 5% significance level and conclude that Height is linearly associated with Arm.

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

A 95% confidence interval for the slope parameter is given by $3.043 \pm t(73, 0.975) \times 0.257 = 3.043 \pm 1.993 \times 0.257 = 3.043 \pm 0.512 = (2.531, 3.555)$. We're 95% confident that when upper arm length increases by 1 cm average standing height increases by between 2.531 and 3.555 cm.

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

A 95% prediction interval for a new individual's height for an upper arm length of 38 cm is given in the following output:

```
Variable  Setting
Arm            38

    Fit     SE Fit         95% CI               95% PI
169.466  0.726508  (168.018, 170.914)  (156.928, 182.004)
```

We're 95% confident that the standing height of a new individual with an upper arm length of 38 cm will be between 156.928 and 182.004 cm.