**STAT 501 - Homework 2 – Fall 2015 - Due Sunday September 6th**

**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.

-------------------------------------------------------------------------------------------------------------------

1. (**2x5 = 10 points**) Suppose that the estimated slope for a straight-line relationship between *y* and *x* has the value 0.
2. Describe how the straight line would look in a plot of *y* versus *x*.

If we find that there is no linear relationship between X and Y (which is what estimated slope of 0 suggests) we will find that the straight line will be horizontal (or very close to being horizontal)

Since the value of the y-intercept indicates the average y score for x=0 and we know further that as x changes the value of y doesn’t change systematically (there may be random variation) we can thus say that the horizontal line will stay at (or close to) the mean of the sampled y values yBar.

1. Explain why a slope of 0 would indicate that *y* and *x* are not linearly related and why a slope not equal to 0 would indicate that *y* and *x* are linearly related.

Published:

**A slope of 0 would mean that, on average, *y* does not change when *x* is changed. Thus specific values of *x* have no linear effect on values of *y*. A slope not equal to 0 would mean that, on average, *y* does change when *x* is changed. Thus specific values of *x* have a linear effect on values of *y*.**

Before getting into the regression aspect, lets see what the slope really means:

slope = rise/run = dy/dx = [delta]y / [delta]x = [change in y over change in x]

**Stated differently 0 slope means no change in y when there is change in x.**

Now further the components of the population regression model is y = mean + deviation

This conceptual equation states that for any individual, the value of the response variable (*y*) can be constructed by combining two components:

1. The *mean,* which in the population is the line *E*(*Y* ) = β0 + β1 x, however when the slope (β1) is 0, the mean will stay at β0. Whereas in the case where slope is not 0, it will systematically change (increase / decrease) with x.

2. The individual’s deviation = y - mean, which is what is left unexplained after accounting for the mean y value at that individual’s x value.

From (1) above we can see that when slope is 0, the population mean will stay constant – showing no linear relation between y and x

On the other hand when slope ≠ 0, we will observe changes in y over changes in x. This will lead to different values of E(Y) for different values of X – linear relationship between x and y.

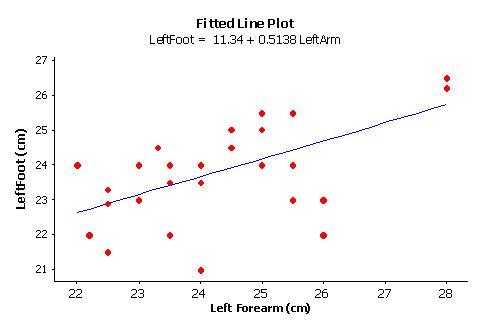
**R2 based explanation**: Another way to understand this is through the expressions for R2. When the slope of the equation is 0 we will have SSTO = SSE

Therefore R2 = 1-SSE/SSTO = 0

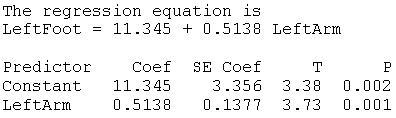
R2 = 0 means that there is no linear relation between x and y

For all the other values of the slope R2 > 0 and so indicate a different level of linear relationship.

2. (**4x5 = 20 points**) Data from *n* = 30 female college students are used to analyze a straight-line relationship between *Y* = left foot length (cm) and *X* = left forearm length (cm). A scatterplot of the data with the regression line superimposed is given below.



Minitab output with information for hypothesis tests about the intercept and slope is also given below.



1. What is the *p*-value for testing H0: β1 = 0? On the basis of this *p*-value, what can we conclude about the population slope?

The p-value for testing H0: β1 = 0 is 0.001. The value is less than the significance level of 0.01 (assumed) and therefore leads us to reject the null hypothesis in favor of the alternate i.e. β1 ≠ 0

We conclude "there is sufficient evidence at the α=0.01 level to conclude that there is a linear relationship in the population between the predictor LeftArm and response LeftFoot"

1. The value of the *t*-statistic for the test in part (a) is 3.73. Show how this *t*-statistic is calculated using other values given in the output.

The *t*-statistic for testing H0: β1 = 0 is 3.73

It is calculated by the following expression from the output: Coef / SE Coef = 0.5138 / 0.1377 = 3.73

1. Calculate a 95% confidence interval that estimates the unknown value of the population slope. (Recall that n = 30.)

t(0.025, 28) = 2.0484

Then, the 95% confidence interval for β1 is 0.5138 ± 2.0484 \* (0.1377) or (0.2317, 0.7959)

1. For this data, SSTO=55.34 and SSE=36.95. Calculate the value of *R*2. Then, write a sentence that interprets that value in the context of these data.

We know that: 

So R2 = 1 – (36.95/55.34) = 0.3323

We can say that 33.23% of the variation in the LeftFoot is reduced by taking into account LeftArm. Or, we can say that 33.23% of the variation in LeftFoot is 'explained by' LeftArm.

3. (**6x5 = 30 points**) The “Hospital Infection Risk” dataset gives characteristics of *n* = 58 hospitals in the eastern and north central areas of the United States. The overall purpose for the data set is to analyze factors that predict *InfctRsk*, the infection risk for patients staying in the hospital. The infection risk value is the percentage of patients who get an infection while they are hospitalized. The variable *Stay* is the average length of stay (days) for patients at the hospital.

1. Use statistical software to graph *y* = I*nfctRsk* versus *x* = *Stay*. In Minitab use Graph > Scatterplot and then select Simple. Discuss noteworthy features of the plot. Specifically, does the relationship look to be described by a straight line? Is there a positive or a negative association? Are there any outliers?

The graph is:



Salient features:

* The relationship seems to be linear
* There appears to be a positive relationship between the variables
* There are data points that seem to be far from the trend line but not outliers. However we don’t have all the tools as yet to perform a detailed analysis of presence of outliers.



1. Using your software, estimate a simple linear regression model with *y* = *InfctRsk* and *x* = *Stay*. In Minitab, use Stat > Regression > Regression > Fit Regression Model. Enter *InfctRsk* as the Response variable and enter *Stay* in the Continuous Predictors box.
   1. Write the estimated regression equation provided by the software.
   2. Write a sentence that interprets the numerical value of the slope in the context of this situation.
2. Estimated regression equation is: InfctRsk = -1.160 + 0.5689 Stay
3. Here, it tells us that we predict that the mean Infection Risk will increase by 0.5689 percent (i.e. the mean percentage of patients who get an infection while they are hospitalized will increase by 0.5689) for every additional day of stay.
4. Refer to the output created for the previous part.
5. Give the values of the t-statistic and the p-value for testing the null hypothesis that the population slope equals 0.
6. On the basis of the *p*-value for testing that the slope is 0, what conclusion can we draw about the linear relationship between the variables *InfctRsk* and *Stay*?
7. For the testing of the null hypothesis that the population slope equals 0, the *t*-statistic = 6.04 and the *p*-value = 0.000 (< 0.001)
8. Since the p-value < alpha = 0.01 and therefore we reject the null hypothesis that the population slope equals 0. Therefore, the conclusion we draw is that there is a linear relationship between the variables InfctRsk and Stay
9. What is the value of *R*2 for the regression? (You'll find it somewhere in the output.) Write a sentence that interprets this value in the context of this situation.

So R2 from the output is: 39.46%

We can say that 39.46% of the variation in the InfectionRisk (as measured by percentage of patients who get an infection while they are hospitalized) is reduced by taking into account Stay (as measured by number of days at the hospital). Or, we can say that 39.46% of the variation in InfectionRisk is 'explained by' Stay.

1. Graph *y* = *InfctRsk* versus *x* = *Age*. The variable *Age* is the average age for patients at the hospital. Discuss noteworthy features of the plot. Specifically, does the relationship look to be described by a straight line? Is there a positive of a negative association (or perhaps, no association)? Are there any outliers?

The graph is:



Salient features:

* The relationship doesn’t seems to be linear
* There appears to be a no association between the variables
* There are data points that seem to be outliers since they are at the peripheries of the plot. However we don’t have all the tools as yet to perform a detailed analysis of presence of outliers.



1. Using your software, estimate a straight-line model with *y* = *InfctRsk* and *x* = *Age*. What is the evidence in the output that there might not be a linear relationship between *InfctRsk* and *Age* (Hint: What is the *p*-value for testing that the population slope is 0?)

The evidence is that the testing of the null hypothesis that the population slope equals 0 has a p-value = 0.675

Since the p-value > alpha = 0.05 and therefore we fail to reject the null hypothesis that the population slope equals 0.

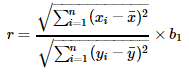
**4. (2x5 = 10 points**) Concrete road pavement gains strength over time as it cures. Highway builders use regression lines to predict the strength after 28 days (when curing is complete) from measurements made after 7 days. Let X be the strength after 7 days (in pounds per square inch) and Y the strength after 28 days. The estimated regression equation for the observed data is = 1389 + 0.96 x.

1. Interpret the estimated slope in the context of this problem.

Here, it tells us that we predict that the mean strength after 28 days will increase by 0.96 pounds per square inch for every additional one-pound per square inch increase in strength after 7 days. In general, we can expect the mean response to increase by 0.96 units for every one unit increase in x.

1. Based on this slope, what can be said about the correlation between X and Y? Can you say anything about the strength of the correlation?

The relation between r and b1 is:



Observing the expression, we can say the following:

* There is +ve correlation between strength after 7 days and strength after 28 days. This is because the estimated slope and the correlation coefficient r always share the same sign.
* No, we can’t say anything about the strength of the correlation.

Therefore more information is required to say anything about strength of r when only b1 is given.

**5.** **( 3X2 + 12 + 3X4 = 30 points)** The Minitab outputs displayed below were obtained from a regression analysis conducted to determine the relationship between income and age.

Retain three decimal places in your calculations.

Analysis of Variance

Source DF SS MS F-Value P-Value

Regression ?? ?? 57575.5 19.80 0.000

Error ?? ?? ??

Lack-of-Fit ?? ?? ?? ?? ??

Pure Error 5 348 ??

Total 21 ??

Coefficients

Term Coef SE Coef T-Value P-Value

Constant 652.1 48.1 13.55 0.000

age -3.328 0.748 -4.45 0.000

1. Write down the population regression model equation assumed in this analysis.

The population regression model equation is:

population regression line and

Individual: 

This is the simple linear regression model with the subscripts modified to recognize the existence of replications.

1. Write down the estimated regression equation.

We have the estimated regression equation: y^i=b0+b1xi

Therefore the estimated regression equation for this particular case is:

y^i=652.1 – 3.328 \* xi

1. Fill in the missing numbers (??) in the table above.

Analysis of Variance

Source DF SS MS F-Value P-Value

Regression 1 57575.5 57575.5 19.80 0.000

Error 20 58157.08 2907.854

Lack-of-Fit 15 57809.08 3853.939 55.373 0.000\*

Pure Error 5 348 69.6

Total 21 115732.58

\* p-value = 1- P(F15,5 < 55.373) = 1-0.999843 = 0.000157

1. Test the hypothesis that income decreases linearly with age and state if the null hypothesis of zero slope is to be rejected.

Null hypothesis H0 : β1 = 0

Alternative hypothesis HA : β1 ≠ 0

Since the reported p-value is 0.000 (< 0.001) < alpha (assumed of 0.01).

Because the P-value is so small (less than 0.001), we can reject the null hypothesis and conclude that β1 ≠ 0

Therefore income changes linearly with age and we can reject the null hypothesis of zero slope.

**OR (perform 1-tailed test to indicate that slope is –ve)**

Null hypothesis H0 : β1 >= 0

Alternative hypothesis HA : β1 < 0 ( since we have been asked to show that income decreases requires the slope to be –ve)

Since our alternative hypothesis is the one-tailed β1 < 0, we have to divide the P-value in the summary table of predictors by 2. However the reported value is 0.000 (< 0.001) << alpha (assumed of 0.01).

Because the P-value is so small (less than 0.001), we can reject the null hypothesis and conclude that β1 < 0

Therefore income decreases linearly with age and we can reject the null hypothesis of zero slope.

1. Based on your conclusion for (d) above, state the three possible outcomes concerning the slope and/or the relationship between income and age. (Refer to section 2.1 in Lesson 2.)

The three possible outcomes are:

* We committed a Type I error. That is, in reality β1 = 0, but we have an unusual sample that suggests that β1 ≠ 0.
* The relationship between x and y is indeed linear.
* A linear function fits the data okay, but a curved ("curvilinear") function would fit the data even better.

1. Perform a lack of fit test to determine the adequacy of the linear regression model.

The null and alternative hypotheses:

Ho: There is no lack of linear fit.

Ha: There is lack of linear fit.

F-statistic: F∗=MSLF/MSPE = 55.373

P-Value: = 1- P(F15,5 < 55.373) = 1-0.999843 = 0.000157

Since the P-value is smaller than the significance level α of 0.01 (assumed), we reject the null hypothesis in favor of the alternative. We conclude "there is sufficient evidence at the α=0.01 level to conclude that there is lack of linear fit."

1. Based on part (f) above, which of the three possible outcomes in (e) is most likely?

Based on the above results it seems that the most likely outcome is: “A linear function fits the data okay, but a curved ("curvilinear") function would fit the data even better.”