**STAT 501 – Homework #2 – Due date January 25th**

**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. If there are multiple parts, then the points are divided equally over the subparts.

1. (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.

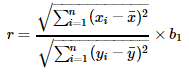
1. Interpret the estimated slope in the context of this problem.

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?

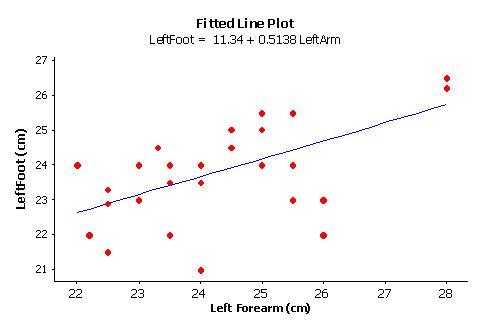
We can say the following:

* 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. The relation between r and b1 is:

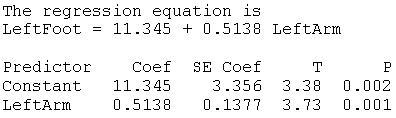


Therefore more information is required to say anything about strength of r when only b1 is given.

2. (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 tests of hypothesis about the intercept and slope is also given below.



(a) 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"

(b) 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

(c) 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)

(d) 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. (30 points) The “Hospital Infection Risk Data” 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.

(a) Use statistical software to graph *y* = InfctRsk versus *x* = Stay. The variable Stay is the average length of stay (days) for patients at the hospital. 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



(b) 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.

i. Write the estimated regression equation provided by the software.

ii. Write a sentence that interprets the numerical value of the slope in the context of this situation.

1. Estimated regression equation is: InfctRsk = -1.160 + 0.5689 Stay
2. 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.

(c) Refer to the output created for the previous part.

i. Give the values of the *t*-statistic and the *p*-value for testing the null hypothesis that the population slope equals 0.

ii. On the basis of the *p*-value for testing that the slope is 0, what conclusion can we make about the linear relationship between the variables InfctRsk and Stay?

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)

(d) 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.

(e) 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



(f) 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 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. (15 points) The “GPA Data” contains grade point average (GPA) at the end of the freshman year (Y) and corresponding ACT test score (X) for a randomly selected 120 students. We will study the relationship between these two variables.

(a) Draw a scatterplot of the data (with the variables on the appropriate axes) and briefly describe the appearance of the plot.

(b) Fit a simple linear regression model to this data and report the estimated model equation.

(c) Do a formal test of H0: β0 = 0 vs. H0: β0 ≠ 0 at the α = 0.05 significance level. Report your *p*-value and state your conclusion in the context of the problem.

(d) Do a formal test of H0: β1 = 0 vs. H0: β1 ≠ 0 at the α = 0.05 significance level. Report your *p*-value and state your conclusion in the context of the problem.

(e) Suppose we have a new student with ACT score xh = 28. Using the model from part (b), what is their predicted GPA?

5. (10 points) Fill in the missing numbers (??) in the following analysis of variance table resulting from a simple linear regression analysis:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source | DF | SS | MS | F |
| Regression | ?? | ?? | 3.588 | 9.24 |
| Residual Error | ?? | ?? | ?? |  |
| Lack of Fit | ?? | ?? | ?? |  |
| Pure Error | 99 | 39.332 | ?? |  |
| Total | 119 |  |  |  |

Retain three decimal places in your calculations.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source | DF | SS | MS | F |
| Regression | 1 | 3.588 | 3.588 | 9.24 |
| Residual Error | 118 | 118\*2.575=  303.850 | 9.24/3.588 = 2.575 |  |
| Lack of Fit | 19 | 264.518 | 13.922 |  |
| Pure Error | 99 | 39.332 | 0.397 |  |
| Total | 119 |  |  |  |

6. (15 points) The “Crop Data” contains the observed crop yield (y) for different fertilizer levels (x).

(a) Fit a simple linear regression model to this data and use the ANOVA table to test

H0: β1 = 0 against Ha: β1 ≠ 0.

In Minitab, use Stat > Regression > Regression> Fit Regression Model

(b) Based on your conclusion for (a) above, state the three possible outcomes concerning the slope and/or the relationship between crop yield and fertilizer level. (Refer to section 2.1 in Lesson 2.)

(c) Perform a lack of fit test to determine the adequacy of the linear regression model.

(d) Plot cropyld (y) versus fertlevel (x) using the Graph > Scatterplot command in Minitab.

(e) Based on parts (c) and (d) above, which of the three possible outcomes in (b) is most likely?