STAT 501 - Homework 2 Solutions – 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.

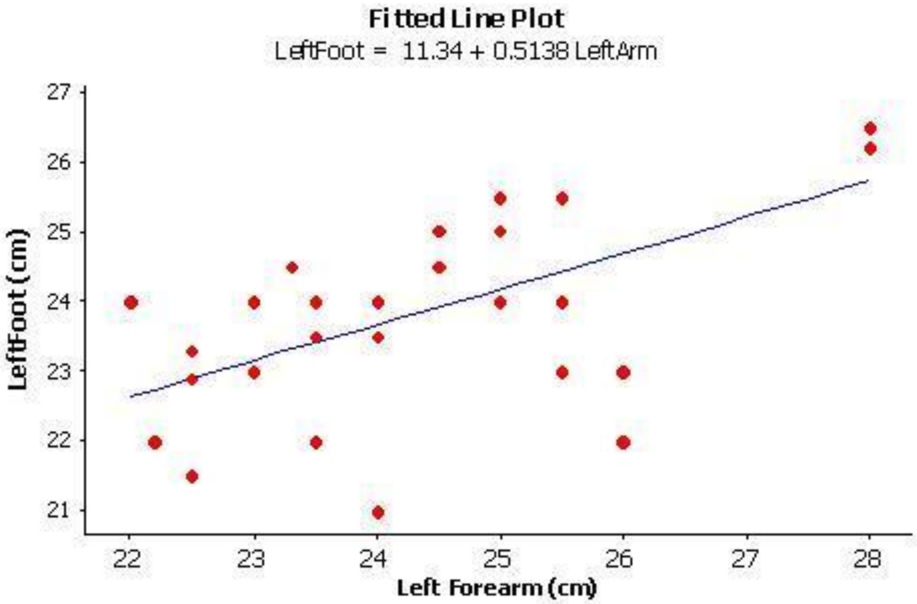
(a) Describe how the straight line would look in a plot of *y* versus *x*.

With a slope of 0, the line would be horizontal (with same value of mean *y*) as you moved across the *x*-values.

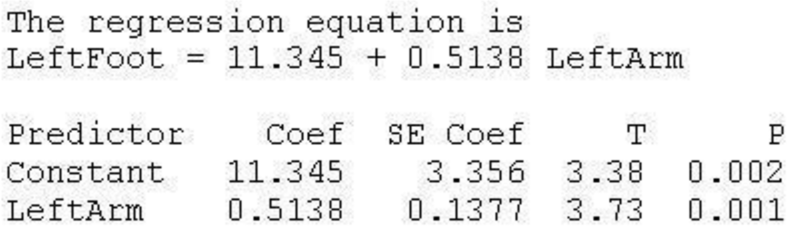
(b) Explain why a slope of 0 would indicate that *y* and *x* are not linearly related and why a slope not equal to 0 would indicate that *y* and *x* are linearly related.

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*.

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.



```
The regression equation is
LeftFoot = 11.345 + 0.5138 LeftArm

Predictor    Coef  SE Coef     T      P
Constant   11.345    3.356  3.38  0.002
LeftArm    0.5138   0.1377  3.73  0.001
```

(a) What is the *p*-value for testing H_0: $\beta_1 = 0$? On the basis of this *p*-value, what can we conclude about the population slope?

The *p*-value for testing if the slope = 0 is 0.001. We therefore reject the null hypothesis and conclude that the population slope is likely not 0.

(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.

***t* = 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.)

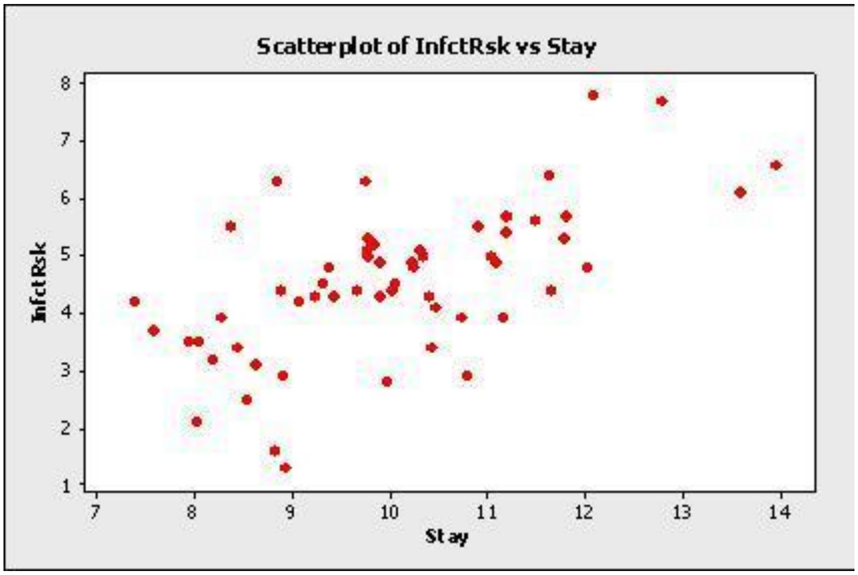
Confidence interval estimate of unknown value of the population slope is 0.5138 ± (2.05 × 0.1377), which is 0.5138 ± 0.2823 or (0.2315, 0.7961). For the multiplier, df = 30 – 2 = 28; *t* = 2.05.

(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.

R^2 = SSR / SSTO = (55.34–36.95) / 55.34 = 18.39 / 55.34 = 0.332. The interpretation is that the left forearm length variable "explains" (or accounts for) 33.2% of the observed variation in left foot lengths.

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.

(a) 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?



There's a positive association between infection risk and average length of stay. It appears that a straight line can describe the average pattern of association. There are no particularly obvious 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.

The Minitab equation is InfctRsk = -1.160 + 0.5689 Stay.

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

For each increase of 1 day in *Stay*, the average (or, predicted) *InfctRisk* increases 0.5689%.

(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.

***t* = 6.04 and *p*-value is 0.000.**

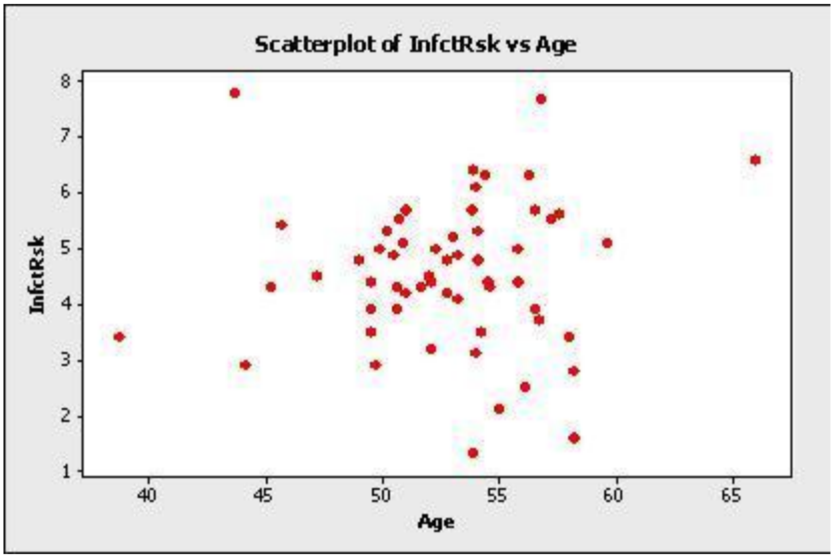
ii. On the basis of the *p*-value for testing that the slope is 0, what conclusion can we draw about the linear relationship between the variables *InfctRsk* and *Stay*?

Since the *p*-value is clearly less than 0.05, we reject the null hypothesis that the population slope is 0 and conclude that the variables are linearly related.

(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.

R^2 = 39.46%. The variable Stay "explains" (or accounts for) 39.46% of the observed variation in infection risks.

(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?



There is a fairly random-looking appearance in the scatterplot. It's not clear whether the variables are linearly associated. There are one or two possible outliers in the infection risk variable - in particular, the observations at Ages of about 44 and 57 have unusually high infection risk. Some may think that without the outliers there is a positive association. That turns out not to be true. I tried deleting the possible outliers and it's still the case that there is not a statistically significant linear relationship.

(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 linear relationship between *InfctRsk* and *Age* (Hint: What is the p-value for testing that the population slope is 0?)

For testing the slope, t = 0.42 and the p-value is 0.675. The p-value is not smaller than 0.05 and so we cannot reject the null hypothesis that the population slope is 0. Thus, there is not evidence of a linear relationship between these two variables.

7. (**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 $\hat{y}$ = 1389 + 0.96 x.

(a) Interpret the estimated slope in the context of this problem.

An estimated slope of 0.96 would mean that, on average, *Y* increases by 0.96 units per one unit increase in X, over the range of sampled X-values.

(b) 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 correlation is positive since the slope is positive. However, we do not have enough information to state whether or not this is a strong or weak correlation. We would need further summary statistics or a scatterplot of the data to make a comment about the strength.

5. (2 X3 + 12 + 4X3 = 30 points) The Minitab outputs displayed below were obtained from a regression analysis conducted to determine the relationship between income and age.

```
 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

Source          DF    SS          MS          F-Value     P-Value
Regression      1     57575.50   57575.50     19.80       0.000
Error           20    58157.07   2907.85
  Lack-of-Fit   15    57809.07   3853.938      55.37       0.0002*
  Pure Error     5    348          69.6
Total           21    115732.57
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
```

*p-value = 1-P($F_{15,5} \leq$ 55.37) = 1-.9998 = 0.0002
Note: F distribution with 15 DF in numerator and 5 DF in denominator

```
    x  P( X ≤ x )
55.37    0.999843
```

a. Write down the population regression model equation assumed in this analysis.
E(income) = β_0 + β_1 age, where β_0 and β_1 are population intercept and the slope respectively.

b. Write down the estimated regression equation.

$\widehat{income}$ = 652.1 – 3.328 age.

c. Fill in the missing numbers (??) in the table above.
See table above.

d. Test the hypothesis that income decreases linearly with age and state if the null hypothesis of zero slope is to be rejected.
Ho : ***β_1 = o vs β_1 < 0***
Test statistic: t = -4.45 with p-value ≈ 0.000 (Note: this is a one-sided p-value = two-sided p-value displayed/2). This small p-value allows us to reject the null hypothesis that the slope is zero.

e. 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.)
1. Possibility of a Type I error, which means that in reality the slope is zero even though the sample led to rejecting that the slope is zero.
2. There is indeed a decreasing linear trend between age and income.
3. A non-linear function may fit the data better than a linear function.

f. Perform a lack of fit test to determine the adequacy of the linear regression model.
H_0: There is no lack of linear fit versus H_a: There is lack of linear fit.
Test statistic: F = 55.37 with p ≈ 0.0002.
We reject H_0, indicating that there is lack of fit in the linear model.

g. Based on part (f) above, which of the three possible outcomes in (e) is most likely?
Outcome 3 is most likely. The lack of fit test attests to the inadequacy of the linear model.