STAT 501 – Homework 1 Solutions – Fall 2015

1. (**4x4 = 16 points**) Data were gathered for y = hours of sleep the previous day and x = hours of studying the previous day for n = 116 college students. The estimated regression equation is found to be $\hat{y} = 7.56 - 0.269x$.

(a) What is the estimated slope of the regression line? Write a sentence that interprets the slope in the context of the variables for this problem. That is, explain exactly what the slope indicates about the relationship between hours of study and hours of sleep.

The estimated slope is –0.269. For each one-hour increase in studying, average (or predicted) amount of sleep decreases by 0.269 hours.

(b) What is the estimated intercept of the regression line? In the context of these variables, what is the interpretation of the intercept?

The estimated intercept is 7.56 hours. It does have useful meaning for this problem and it means the average hours of sleep for people who studied 0 hours.

(c) For students who studied 2 hours the previous day, what is the estimated value of average hours of sleep the previous day?

The estimated value of average hours of sleep the previous day is $\hat{y} = 7.022$, calculated by substituting 2 hours into the regression equation.

(d) Suppose that a student studied 2 hours the previous day and slept 6 hours that day. What is the value of the residual for that person? (A residual is defined as the difference between the observed and predicted values of y for an individual.)

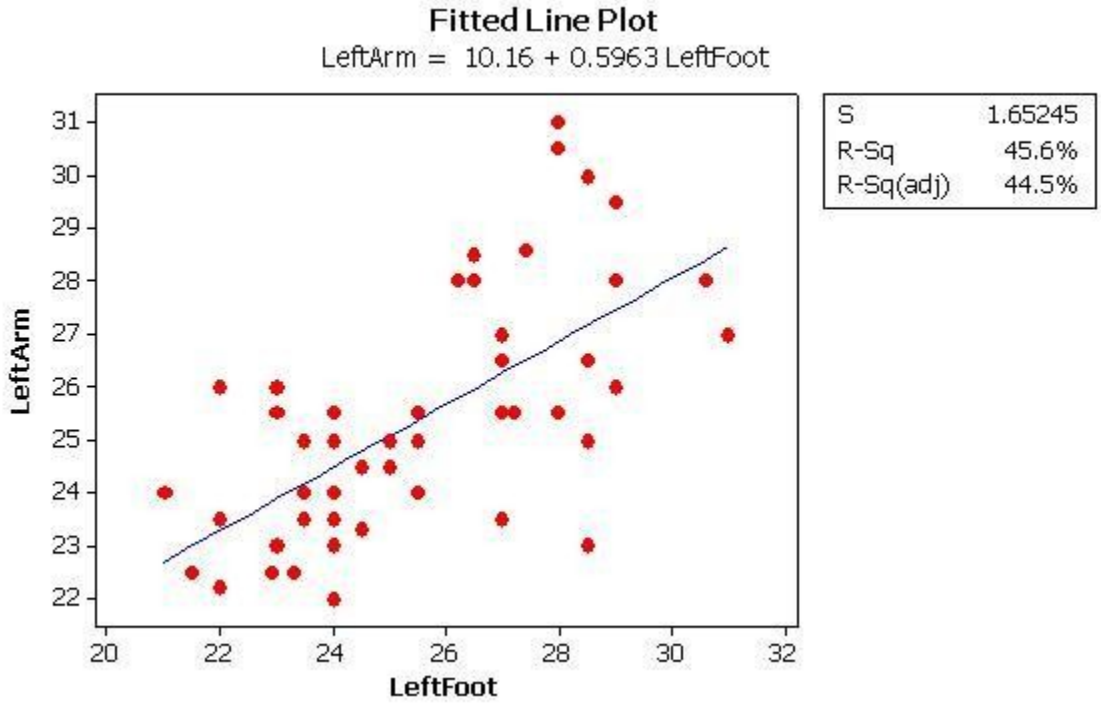
The value of the residual for that person is –1.022, calculated by finding the difference between the observed value of 6 hours and the predicted value found in the previous part.

2. (**2x4 = 8 points**) Two of the following statements about a population model for a simple regression are correct and two are incorrect. Which two statements are correct? Explain your answer.

(a) $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
(a) $Y_i = \beta_0 + \beta_1 X_i$
(b) $E(Y_i) = \beta_0 + \beta_1 X_i + \varepsilon_i$
(c) $E(Y_i) = \beta_0 + \beta_1 X_i$

(a) and (d) are correct. With E(Y_i) on the left side, we should not have the error term on the right side since the expectation would be taken for that quantity as well. E(Y_i) is given by the equation of the straight-line $\beta_0 + \beta_1 X_i$. Or, with the error term on the right side, we should have Y_i on the left, not E(Y_i).

3. (**4x4 = 16 points**) The fitted line plot below gives results for a straight-line regression between y = left forearm length (cm) and x = left foot length (cm). The data are from n = 55 college students.



(a) Write a sentence that gives the value of the slope and interprets it in the context of this situation.

Average left forearm length increases 0.5963 cm for each 1 cm increase in left foot length.

(b) Write a sentence that gives the value of R^2 and interprets it in the context of this situation.

Left foot length explains 45.6% of the observed variation in left forearm lengths.

(c) Use the equation to estimate the average left forearm length of college students with a left foot length of 25 cm.

$\hat{y}$ = 10.16 + 0.5963(25) = 25.0675. (By the way, an interesting feature of the human body is that forearm length and foot length are about the same.)

(d) The Minitab output includes the information that "**S**=1.65245". Explain what is measured by this statistic.

The value S = 1.65245 measures the standard deviation of the residuals (differences between actual and predicted forearm lengths). A more conceptual explanation is that this is about the average absolute size of differences between actual and predicted forearm lengths.

4. (**10x4 = 40 points**) Infant weights in pounds have an upward linear trend with age in months. Data from a sample of 5 babies in a local community, including one newborn and four others who are 1 month, 2 months, 3 months and 5 months old, were used to obtain an estimated regression equation based on the least squares criterion with a slope of 0.2838 pounds per month. Some of the information is given in the following table.

Age (x_i)	0	1	2	3	5
Weight (y_i)	**8.0**	8.1	8.4	9.3	**9.2003**
Predicted weight ($\hat{y}_i$)	**7.9757**	**8.2595**	**8.5433**	8.8271	**9.3947**
Residual error (e_i)	0.0243	**-0.1595**	**-0.1433**	**0.4729**	**-0.1944**

(a) What is the equation of the population regression line in this setting? *[Hint: There should be no numbers in this equation, just β 's.]*

E(*weight*) = β_0 + β_1 *age*, where β_0 and β_1 are the population intercept and slope respectively.

(b) What is the estimated regression equation? *[Hint: There should be numbers in this equation. Use the information in the question and in the table, particularly the 3-month old baby.]*

$\widehat{weight}$ **= 7.9757 + 0.2838 *age*.**

(c) Based on the estimated regression equation, what is the predicted birth weight of a newborn in this community?

7.9757 lbs.

(d) What is the observed birth weight of the newborn in the sample?

8.0 lbs.

(e) Complete the remaining entries in the table above.

See above.

(f) Comment on the validity of using the estimated regression equation to predict the weight for a one year old.

The predicted weight is not valid because one year is beyond the scope of this model.

(g) Calculate SSE, the sum of residual error squares.

SSE = $(0.0243)^2 + (-0.1595)^2 + (-0.1433)^2 + (0.4729)^2 + (-0.1944)^2 = 0.3079914$.

(h) Calculate the sample estimate of the variance, σ^2 , for the regression model.

Estimate of σ^2 = MSE = SSE/(n-2) = 0.1026638.

(i) Calculate the value that would be given in Minitab for "**S=**". Write a sentence that interprets this value.

S in Minitab = sqrt(MSE) = 0.320411922. The average absolute size of differences between actual and predicted weights is about 0.3204 lbs.

(j) Calculate the value of R^2. To start, you will have to calculate the value of SSTO. Write a sentence that interprets the value of R^2.

$\bar{y}$ **=8.60006 so SSTO =** $\sum (y_i - \bar{y})^2$ **= 1.500360072.**

***R*² = (SSTO−SSE)/SSTO = (1.500360072−0.3079914)/1.500360072 = 0.7864 (or 78.64%). The interpretation is that 78.64% of the variation in y = weight is "explained by" the variation in x = age.**

5. (**4x5 = 20 points**)

(a) Briefly describe the four assumptions (or conditions) that underlie the simple linear regression model.

The four assumptions that underlie the simple linear regression model are:

1. **The mean of the response at each value of the predictor is a linear function of the predictors. (Equivalently, the mean of the error at each value of the predictor is zero.)**
2. **The errors are independent.**
3. **The errors at each value of the predictor are normally distributed.**
4. **The errors at each value of the predictor have equal variances.**

(b) The scatter plot below shows sample data for y = selling price of a house and x = square foot area of the house.

(i) Name one condition that may be satisfied by the selling price vs square foot area data and justify your choice.

Condition 1, which is linearity of selling price with respect to square foot area, appears to be satisfied.

(ii) Name one condition that may not be satisfied by the data and justify your choice.

Condition 4 may not be satisfied. It appears that that the variability of selling prices increases with square foot area rather than remaining constant.

(iii) Would you expect the magnitude of the sample correlation coefficient to be near 0, closer to +1, or closer to –1? Justify your choice.

The sample correlation coefficient (r) will be closer to 1, as the plot indicates a positive linear association between selling price and square foot area.

(iv) Based on the estimated regression equation, $\widehat{Price}$ = 5049 + 64.87 SQFT, what is the y-intercept estimate? Is this value meaningful? Why or why not?

The y-intercept estimate is $5049. This value is not meaningful since it is not possible to have a house with zero area.