Homework 2

Roly Vicaría
STAT501 Fall 2015

September 6, 2015

0.0.1 Question 1

a. If the estimate slope is 0, then the straight line would be a horizontal line with $y = b_0$

b. A slope of 0 means that y is always the same value regardless of the x value. That means that x has no effect on y, so they're not linearly related. A non-zero slope, on the other hand, says that a change in x will give a different value for y, i.e. they're linearly related.

0.0.2 Question 2

a. The p-value for testing $H_0 : \beta_1 = 0$ against the alternative hypothesis $H_A : \beta_1 \neq 0$ is 0.001. Since this p-value is double the p-value of testing against $H_A : \beta_1 > 0$, then the p-value for that would be 0.0005. There is then sufficient evidence, at the 0.05 level, to conclude that $\beta_1 > 0$.

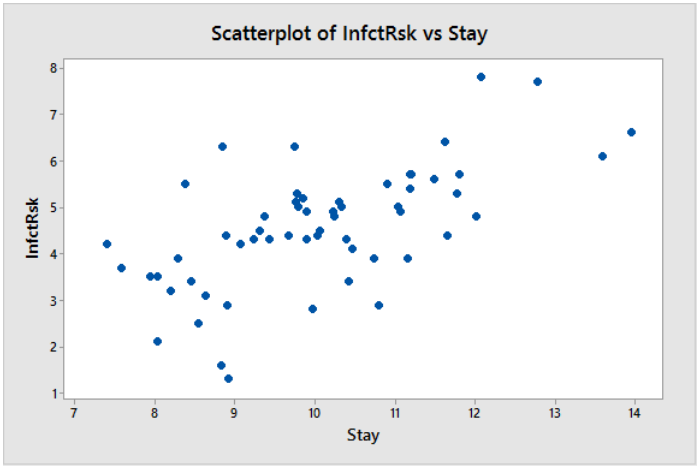
b. The t-statistic is calculated using the "Coef" value divided by the "SE Coef" value. In other words, $t = \frac{b_1}{s\{b_1\}} = \frac{0.5138}{0.1377} \approx 3.73$

c. 95% confidence interval: $0.5138 \pm t_{0.025,28}(0.1377)$ or (0.2318, 0.7958)

d. $R^2 = 1 - \frac{SSE}{SSTO} = 1 - \frac{36.95}{55.34} \approx 0.3323$. This can be interpreted as saying that roughly 33% of the variation in the left foot lengths can be 'explained by' the variation in the left forearm lengths.

0.0.3 Question 3

a. The scatterplot does suggest that there may be a positive association between length of stay and risk of infection. There are a couple x values at the far right, but not far enough that I would label them as outliers.

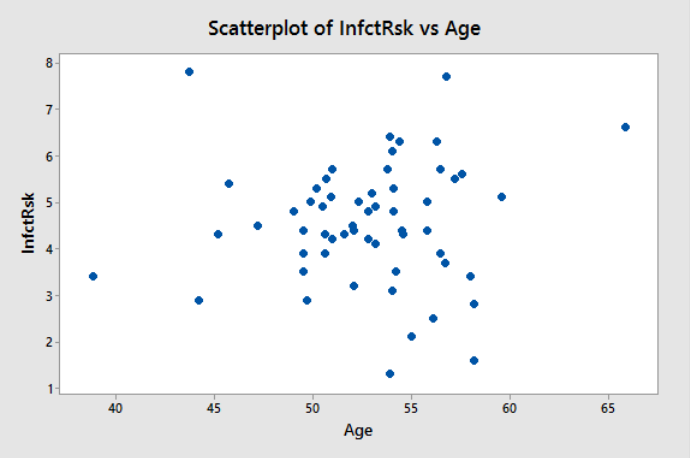


b. i) The estimated regression equation is: $InfctRsk = -1.160 + \ 0.5689 \ Stay$
ii) The slope can be interpreted as saying that for every additional day of hospital stay, the infection risk rises by 0.5689%

c. i) The t-statistic given by Minitab is 6.04 with a p-value of 0.000.
ii) We can conclude that at the 0.05 level, there is enough evidence to say that the slope is not 0 and that there is a positive association between InfctRsk and Stay

d. R^2 for the regression is 39.46%. We can interpret this as saying that about 39.46% of the variation of InfctRsk is 'explained by' the variation in the length of stay.

e. The scatterplot does not show any obvious relationship between Age and InfctRsk. There is a lot of variability for InfctRsk at all ages. There are two potential outliers for Ages less than 40 and above 65.



f. The estimated regression line is: $InfctRsk = 3.663 + 0.01694Age$. The evidence in the output that there might not be a linear relationship is: 1) the low R-sq value, 0.3%, 2) the high p-value, 0.675, which means that at the 0.05 level, we cannot conclude that slope is non-zero.

0.0.4 Question 4

a. The estimated slope can be interpreted as saying that the 28-day strength will increase by 0.96 units for every additional unit of 7-day strength.

b. The correlation between X and Y in this example seems strong since the slope is very close to 1. Therefore the line would be almost a 45 degree angle. It suggests a stronger correlation than a smaller slope. We can't really comment on the strength of the correlation without more information about the regression, e.g. R^2, p-value of β_1, etc

0.0.5 Question 5

a. Population regression model equation: $E(Income) = \beta_0 - \beta_1 Age$

b. Estimated regression equation: $Income = 652.1 - 3.328 Age$

c. 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			

d. $H_A : \beta_1 < 0$ t-statistic given is -4.45. The P(t < -4.45) = 0.000, from which we can reject the null hypothesis of zero slope.

e.
1. We committed a Type I error. The slope really is zero, but our data is unusual suggesting that it was less than 0.
2. The relationship between x and y is indeed linear.
3. A linear function fits the data okay, but a curved function would fit the data even better.

f. From the table above, we calculated an F-value for lack of fit to be 55.373 with a p-value of 0.000.

g. Based on the conclusion from the lack of fit test, outcome 3 seems most likely: a linear function fits the data okay, but a curved function would fit the data even better.