Statistics 12, Homework 3

Michael Wu
UID: 404751542

February 3rd, 2019

Chapter 6, Problem 14

a) Plot 1 shows little or no association.

b) Plot 4 shows a negative association.

c) Plots 2 and 4 show a linear association.

d) Plot 3 shows a moderately strong association.

e) Plots 2 and 4 show a very strong association.

Chapter 6, Problem 20

a) This graph has a correlation of -0.977.

b) This graph has a correlation of 0.736.

c) This graph has a correlation of 0.951.

d) This graph has a correlation of -0.021.

Chapter 7, Problem 6

a) The outlier condition is violated here, since the 32TB drive is causing the regression line to have a more positive slope than it should. This causes it to underestimate the price of lower capacity drives and overestimate the price of higher capacity drives, with the exception of the 32TB drive.

b) I would recommend to remove the 32TB drive from the data set and run the regression again.

Chapter 7, Problem 16

This suggests that I would get

$$36.25 - 3.867 \times 4 = 20.782$$

miles per gallon.

Chapter 8, Problem 2

a) For the larger venue this plot shows a linear, positive, weak correlation between talent cost and total revenue. There appears to be one outlier with a talent cost of over 100,000 and a revenue of over 160,000.

For the smaller venue this plot shows a linear, positive, weak correlation between talent cost and total revenue. There appears to be one outlier with a talent cost of over 50,000 and a total revenue of over 40,000.

b) The results of the two venues are similar because they both have a linear, positive, weak correlation between talent cost and total revenue. They also both have an outlier with a lot of leverage that would move the regression line upwards. Without the outlier it would seem that there would be weaker correlation between talent cost and total revenue.

c) The results of the two venues are different because the smaller venue usually hosts talent that costs less and brings in less revenue. This means that the range of data for the smaller venue is smaller than the range of data for the larger venue.

Chapter 8, Problem 12

No, correlation does not imply causation. Most likely cell phone usage is correlated to life expectancy because people who live in places where cell phone usage is high have a better standard of living than people who live in places where cell phone usage is low. Therefore they may experience less hardship and illnesses, leading to longer lives.

Chapter 8, Problem 16

a) There is a clear pattern since the percentage of smokers appear to be decreasing over time. It decreases linearly up to the 90's, increases slightly in the 90's, then begins decreasing again in the 2000's.

b) There appears to be strong association, since the data points follow a trend line fairly closely. There is very little spread since it seems that the smoking rate does not randomly go up and down by a large amount every year.

c) I do not think a linear model is appropriate. It would not be able to account for the slight increase in the 90's, as the overall trend is negative. Furthermore, smoking is a social phenomenon whose popularity can behave erratically. It does not have any intrinsic property that would make it have a linear relationship with respect to time.

Chapter 8, Problem 20

a) A low R^2 value by itself does not necessarily mean that a linear model is not appropriate. Perhaps the variables are truly linearly correlated and not enough data was collected in ranges such that the correlation could be seen clearly. Or perhaps the correlation is very slight.

b) This model will not allow the student to make accurate predictions because there is too much variability that the regression line will not account for. So the predictions could be very different from the actual values.

Chapter 8, Problem 22

a) The trends in smoking behavior are similar in men and women because the percentage of men and women who smoke have been steadily decreasing over time.

b) The smoking rate for women is usually lower than the smoking rate for men, though in 80's the smoking rate for women was higher in some years.

c) The trend for women also violates the linearity condition, since there appears to be a different rates of change over time before the 80's, in the 80's, in the 90's, and in the 2000's. I do not believe a linear model is appropriate, since smoking is a social phenomenon whose popularity could change at any time. Smoking rates could increase if it becomes popular again. A better model would be something like the amount of money spent on smoking advertisements versus the percentage of smokers.