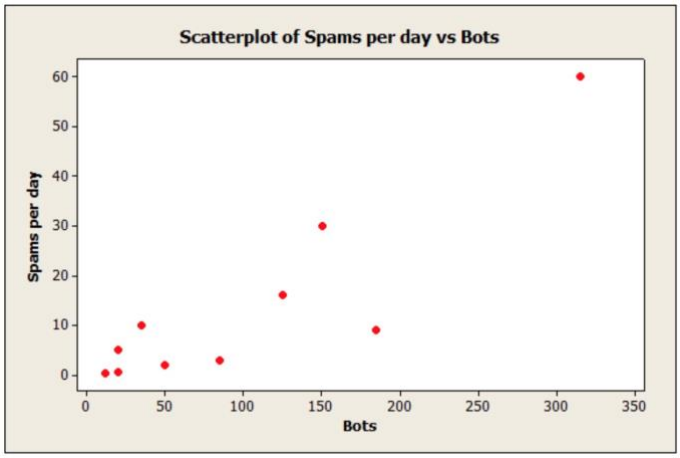
2019Spring_st305_hw2_solution

(Total points: 58)

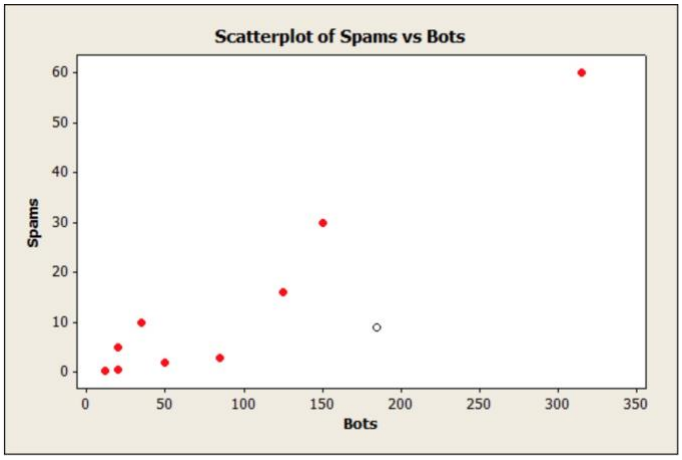
Ch 2.1: 7, 10, 14, 20, 24, 33, 36

2.7 (a) (1 point) The spreadsheet that contains the spam botnet data is as follows:

Botnet	Bots (thousands)	Spam's per day (billions)
Srizbi	315	60
Bob ax	185	9
Rustock	150	30
Cut wail	125	16
Strom	85	3
Grum	50	2
Ozdok	35	10
Nucrypt	20	5
Wopla	20	0.6
Spamthru	12	0.35



(b) (1 point) For the label Bobax the number of Bots (in thousands) is 185 and the Spam's per day (billions) is 9. This is the second from the right of the scatter plot (in open circle).



2.10 (1 point) Find the sample mean and standard deviation for 2006 and 2007 respectively:

$\bar{x}_{2006} = 154.91,\ \bar{x}_{2007} = 173.99$

$s_{2006} = 127.95,\ s_{2007} = 143.20$

Find a linear transformation that will change the debts in 2006 into new scores where the mean 200 with a standard deviation of 150. (The values of the transformed mean and standard deviation can be chosen by yourself.)

The form of a linear transformation is

$$x_{new} = a + bx.$$

We can get the value of a and b by solving the following system of equations:

$$\begin{cases} \bar{x}_{new} = a + b\bar{x} \\ s_{new} = bs \end{cases}, \text{ i.e., } \begin{cases} 200 = a + b\bar{x} \\ 150 = bs \end{cases}.$$

For that data of 2006, we can get $x_{new}^{2006} \approx 18.394 + 1.172x$; similarly, $x_{new}^{2007} \approx 17.832 + 1.047x$ is the linear transformation for 2007 debt.

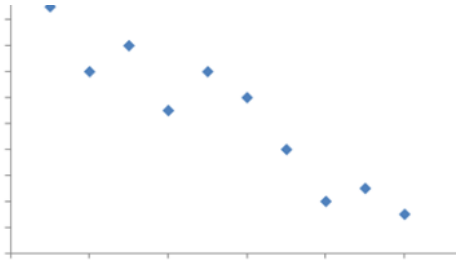
Using the transformed equations, we can find the new mean debt amounts per year:

In 2006, substitute in the mean, $\bar{x}_{new}^{2006} \approx 18.394 + 1.172(154.91) = 199.949.$

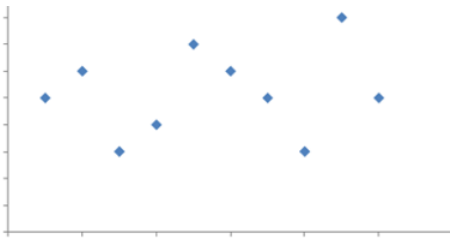
In 2007, substitute in the mean, $\bar{x}_{new}^{2007} \approx 17.832 + 1.047(173.99) = 199.999.$

The new amounts come out within 0.05 and are therefore approximately the same accounting for inflation.

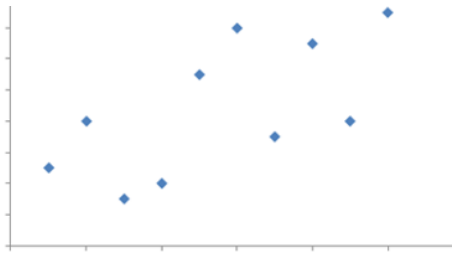
2.14 (a) (1 point)



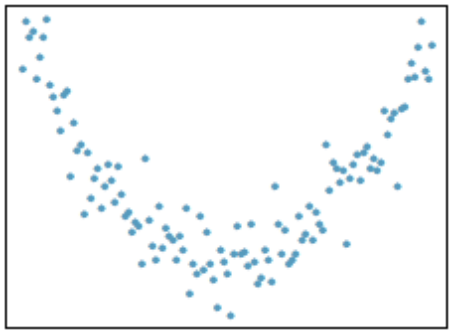
(b) (1 point)



(c) (1 point)



(d) (1 point)



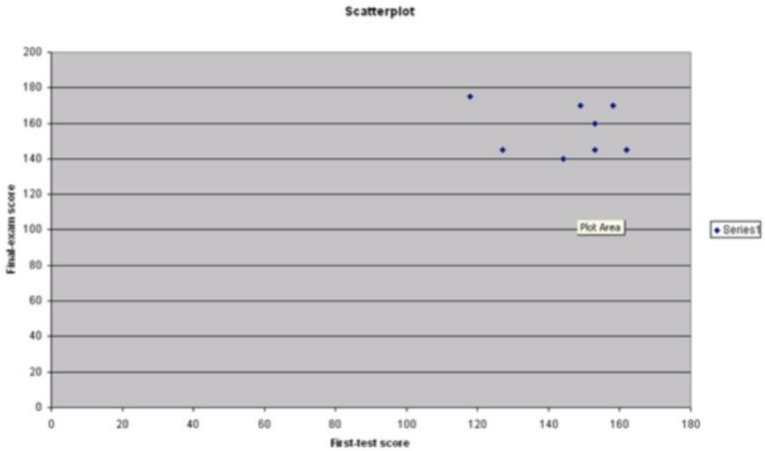
It is a non-linear relationship.

20. (a) (1 point) Looking at the scatterplot, if you ignore data points left of 10 on the x-axis, there doesn't appear to be a relationship between the internet users per 100 people in a country. For those countries that have few internet users, the results vary greatly.

(b) (1 point) The scatterplot does not give a reason to conclude that using the internet will increase life expectancy. There is great variance between countries with few internet users, but low variance between counties with a large number of internet users. The differences in life expectancy for countries with few internet users is probably due to income or quality of life rather than internet use (countries with low income or quality of life are unlikely to have a lot of internet users).

24. (a) (1 point) First-test score should play the role of explanatory variable in describing the relationship.

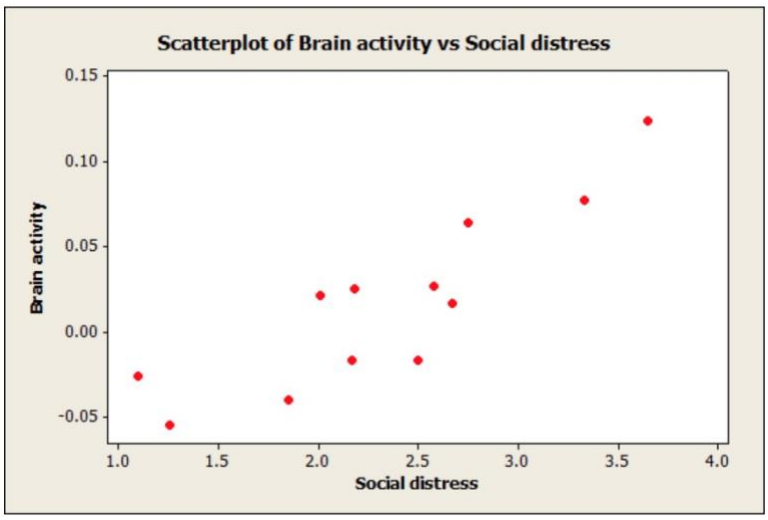
(b) (1 point)



From the scatter plot, it can be observed that the observations are not close to any particular form such as line. Therefore, we can say that the association between the variables is very weak.

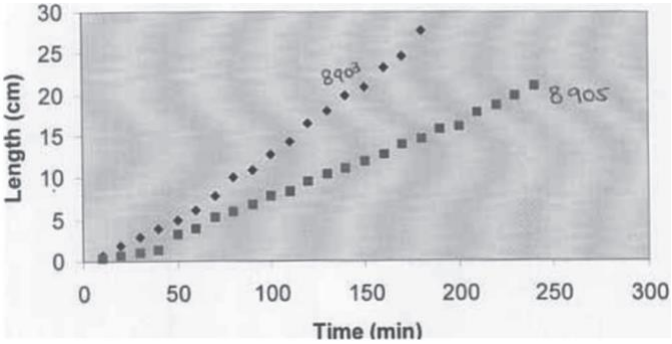
(c) (1 point) The given relationship is so weak because the first test score is not describing the final exam score. Also, student's study habits are not clearly established by first-test score.

33. (1 point)



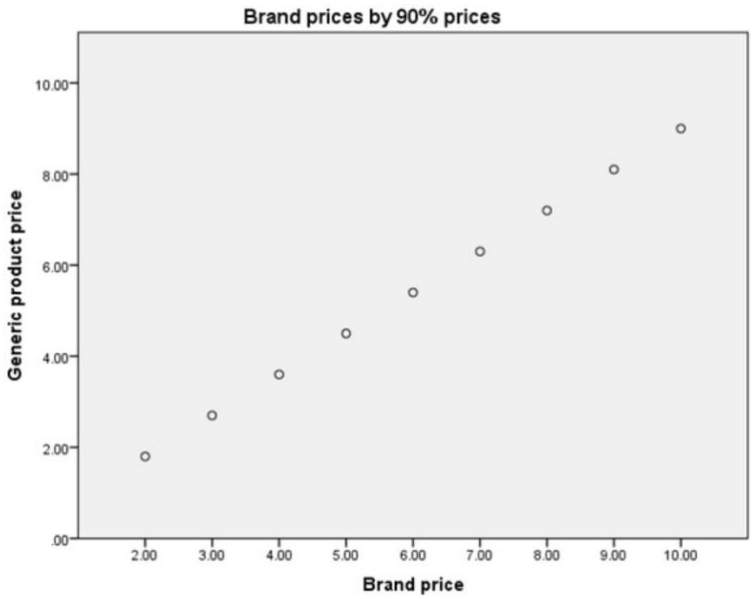
Direction: There is a positive relationship brain activity and social distress from the above plot.
From: The plot is roughly linear.
Strength: The strength of the relationship between brain activity and social distress is fairly strong.
Social exclusion does appear to trigger a pain response.

36. (a) (1 point) Given below the scatter plot of the length of the ICICLE versus time for both 8905 and 8903.



(b) (1 point) Increase in water flow rate decreases the growth of ICILES. Clearly from scatter plot growth rate for Run 8903 is higher than that of 8905.

Ch 2.2: 41, 42, 48, 53, 54, 60
41.(a) (1 point)

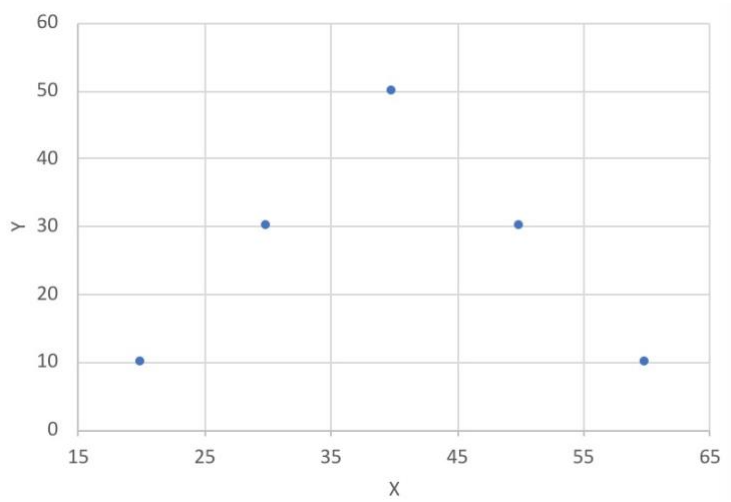


This is our scatterplot. Notice how the points directly follow a straight line. This represents a correlation of 1.

(b) (1 point)

This is our scatterplot. Notice how the points directly follow a straight line. This represents a correlation of 1.

42. (a) (1 point)



(b) (1 point) It is a strong non-linear relationship between X and Y.

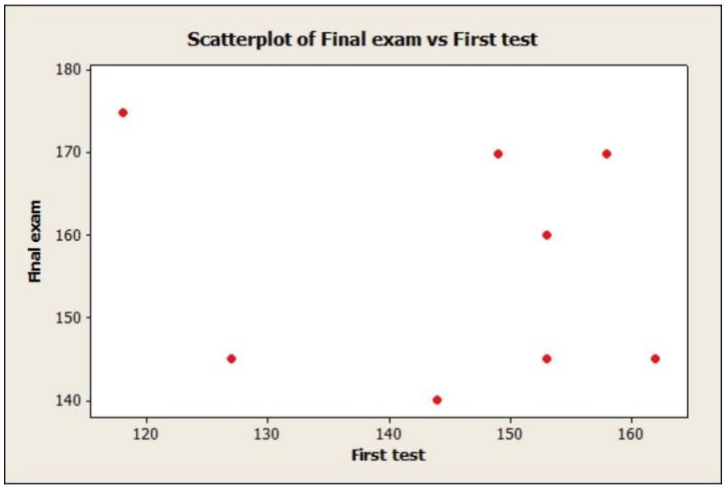
(c) (1 point) $r = 0$.

(d) (1 point) Correlation measures the strength of only the linear relationship between two variables. Correlation does not describe curved relationship between variables, no matter how strong they are.

48. (a) (1 point) We have data on First test (x) and Final exam (y). The formula for the correlation between x and y is given by

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = -0.201$$

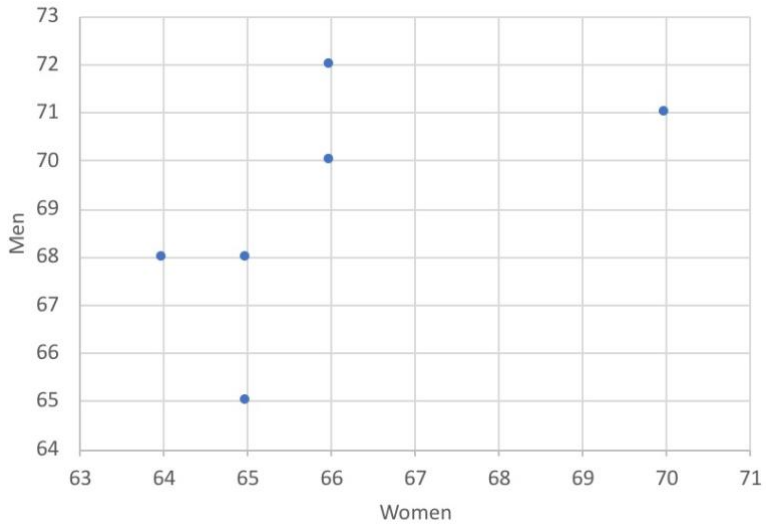
(b) (1 point) The scatter plot for the above data is as follows:



The scatter plot and the correlation support that the relationship between these two variables is weak.

53.

(a) (1 point)



Positive but not close to 1.

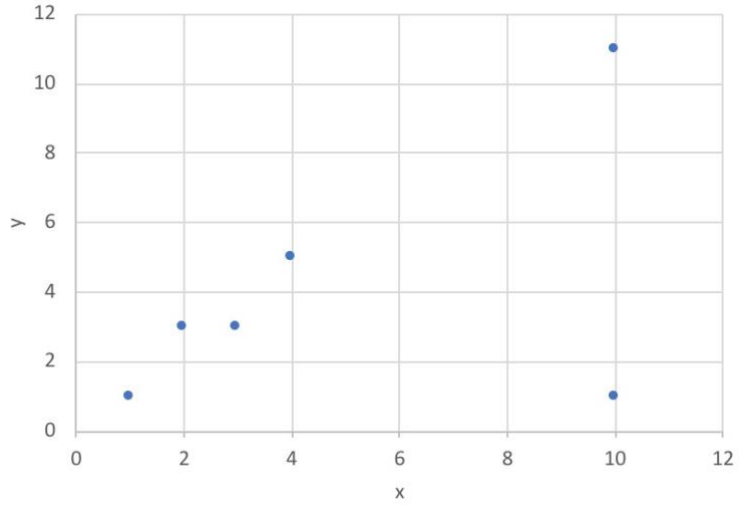
(b) (1 point) $r = 0.5653$.

(c) (1 point) r would not change; it does not tell us that the men were generally taller.

(d) (1 point) r would not change.

(e) (1 point) 1.

54. (2 points, 1 for plot and 1 for explanation)



$r = 0.4811$. Because of the outlier $(10, 1)$, the correlation is reduced.

60.

(a) (1 point) Correlation requires that both variables be quantitative but occupation is not quantitative variable.

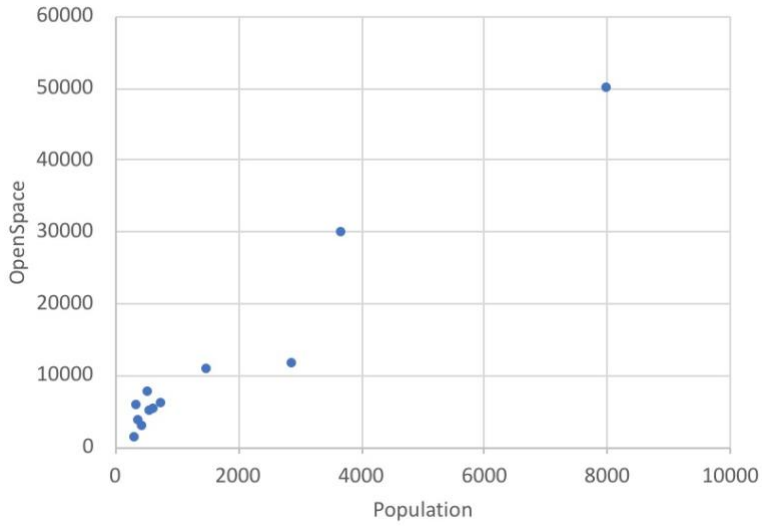
(b) (1 point) The correlation r is always a number between -1 and 1 so it cannot be 1.19.

(c) (1 point) Correlation requires that both variables be quantitative but gender and color are not quantitative variable.

Ch 2.3: 66, 77, 78, 79, 80, 81, 85

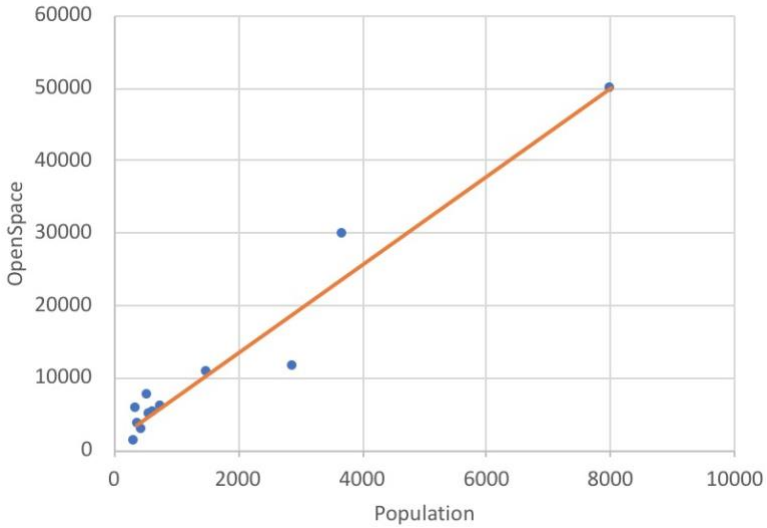
66.

(a) (1 point)



(b) (1 point) Yes. Because it is a positively linear relationship between population and open space.

(c) (1 point) $\hat{y} = 1248.197 + 6.105x$



(d) (1 point) $r^2 = 0.9519$. Thus, 95.19% of the variation in open space is explained by population.

77. Given that the equation of a Least-square regression line is $y = 12 + 6x$.

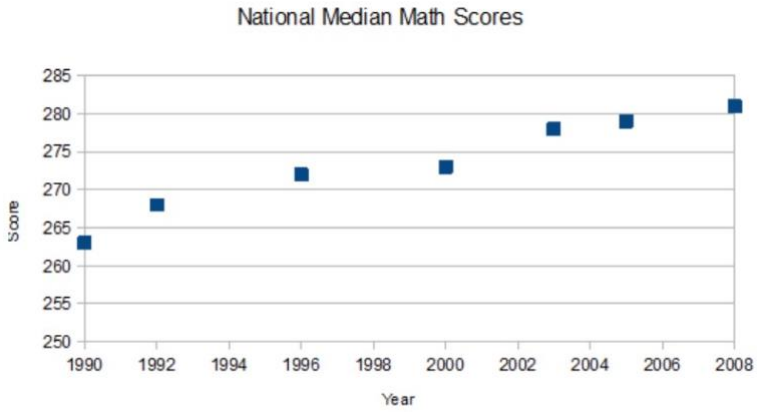
(a) (1 point) To find the value of y for $x = 5$, by substituting $x = 5$ in the regression line, we have $y = 42$.

(b) (1 point) Let's suppose $x = x_0$ the y value is $y = 12 + 6x_0$. If there is increasing in x by one unit $x = x_0 + 1$. The corresponding y-value is $y = 12 + 6(x_0 + 1) = 12 + 6x_0 + 6$. If there is increasing in x by one unit, the corresponding increasing in y by 6 units.

(c) (1 point) Then intercept is the value of y when $x = 0$. $y = 12 + 6(0) = 12$. So the intercept for this equation is 12.

78.

(a) (1 point)



There is an increasing trend and therefore positive relationship,

(b) (1 point) The sample mean of X is $\bar{x} = \frac{1}{n}\sum x_i = \frac{1}{7}(13994) = 1999.143$. The sample mean of Y is $\bar{y} = \frac{1}{n}\sum y_i = \frac{1}{7}(1914) = 273.43$.

The standard deviations of X and Y are,

$$s_x = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2} = \sqrt{\frac{1}{7-1}\sum(x_i - 1999.143)^2} = 6.744,$$

$$s_y = \sqrt{\frac{1}{n-1}\sum(y_i - \bar{y})^2} = \sqrt{\frac{1}{7-1}\sum(y_i - 273.429)^2} = 6.451.$$

The correlation coefficient r is $r = \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) = 0.979$.

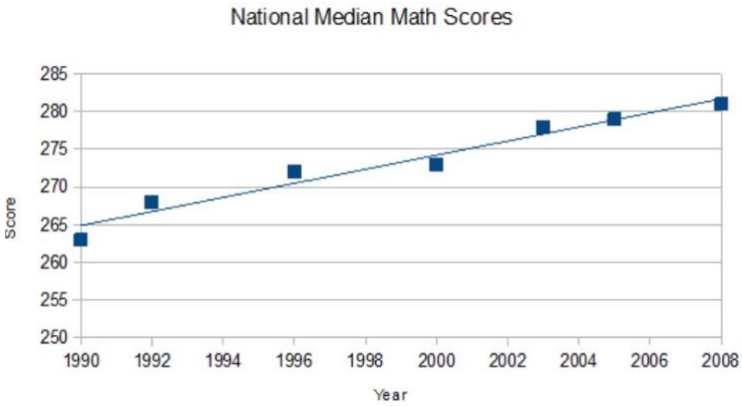
The slope of the regression line is $b_1 = \frac{rs_y}{s_x} = 0.936$.

The intercept is $b_0 = \bar{y} - b_1\bar{x} = 273.429 - (0.936)(1999.143) = -1597.769$

Therefore, the equation of the least squares regression line is

$$\hat{y} = b_0 + b_1 x = -1597.769 + 0.936x$$

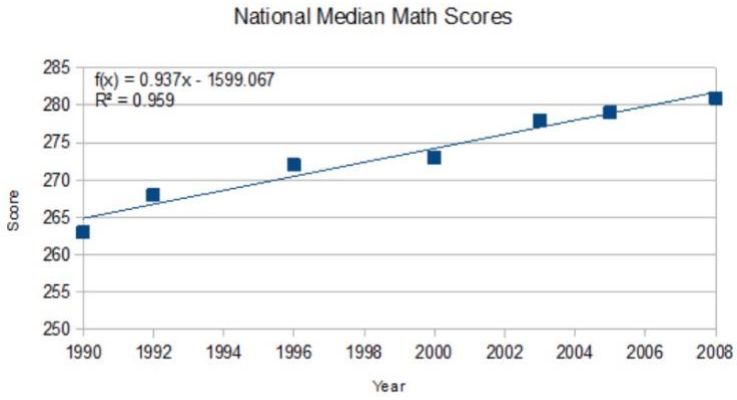
The plot is



The data fits the line quite will with no obvious outliers.

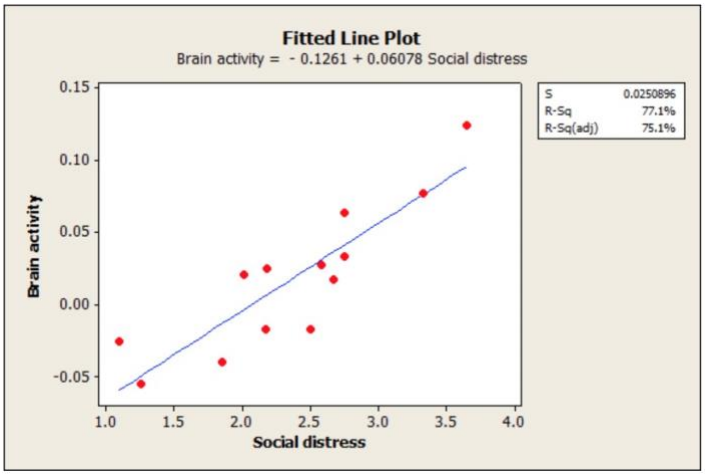
Since, $r = 0.979$ then the variation in scores from year to year is $r^2 = 0.958 = 95.8\%$.

(c) (1 point)



79. (a) (1 point) The mean and standard deviation of the variable social distress score is $\bar{x} = 2.369,\ s_x = 0.726$. The mean and standard deviation of the variable brain activity is $\bar{y} = 0.0179,\ s_y = 0.0502$. The correlation, $r = 0.878$. $r^2 = 0.771$. The slope of the least squares regression line for predicting brain activity from social distress score is $b_1 = \frac{rs_y}{s_x} = 0.06071$. The intercept is $b_0 = \bar{y} - b_1\bar{x} = -0.12592$. Therefore, the equation of the least squares regression line for predicting brain activity from social distress score is

$$\hat{y} = -0.12592 + 0.06071x$$



(b) (1 point) The predicting brain activity for social distress score 2.0 is

$$\text{Brain activity} = -0.126 + (0.0607 \times 2) = -0.0046.$$

(c) (1 point) Since $r^2 = 0.770884$, therefore 77% of the variation in brain activity among these subjects is explained by the straight-line relationship with social distress score.

80. (1 point)

The mean and standard deviation of the variable time for slower flow rate (run 8903) is $\bar{x}_1 = 95,\ s_{x_1} = 53.4$. Then mean and deviation of the variable length for slower flow rate

(run 8903) is $\bar{y}_1 = 12.66$, $s_{y_1} = 8.5$. The correlation $r_1 = 0.996$. The slope of the least squares regression of icicles growth at slower flow rate (run 8903) is $b = \frac{r_1 s_{y_1}}{s_{x_1}} = 0.158$.

The mean and standard deviation of the variable time for flow rate (run 8905) is $\bar{x}_2 = 125$, $s_{x_2} = 70.7$. Then mean and deviation of the variable length for slower flow rate (run 8905) is $\bar{y}_2 = 9.94$, $s_{y_2} = 6.46$. The correlation $r_2 = 0.998$. The slope of the least squares regression of icicles growth at slower flow rate (run 8905) is $b = \frac{r_2 s_{y_2}}{s_{x_2}} = 0.091$.

For the slower flow rate (run 8903) icicles grow at 0.158 cm/min and for the flow rate (run 8905) icicles grow at 0.091 cm/min. Decrease in the flow rate increases the growth of icicles.

81. (1 point) By the question 80, we know for run 8908, $\bar{x}_1 = 95$, $s_{x_1} = 53.4$, $r_1 = 0.996$, $b = \frac{r_1 s_{y_1}}{s_{x_1}} = 0.158$. Then the intercept is $a = \bar{y}_1 - b\bar{x}_1 = -2.39$. Therefore, the equation of the regression line is $\hat{y} = -2.39 + 0.158x$. Now slope for predicting time from growing length is $b = \frac{r_1 s_{x_1}}{s_{y_1}} = 0.9958 \times \frac{53.385}{8.497} = 6.256$, $a = 95 - 6.256 \times 12.66 = 15.8$. Thus, equation of regression line is $\hat{y} = 15.8 + 6.256x$. For slope 1, unit is cm/minute. For slope 2, unit is minute/cm.

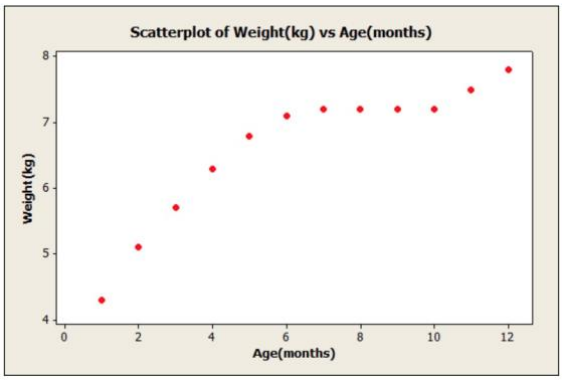
85. (1 point) When $x = \bar{x}$, $\hat{y} = a + b\bar{x} = (\bar{y} - b\bar{x}) + b\bar{x} = \bar{y}$.

Ch 2.4: 98, 99, 102, 103, 110

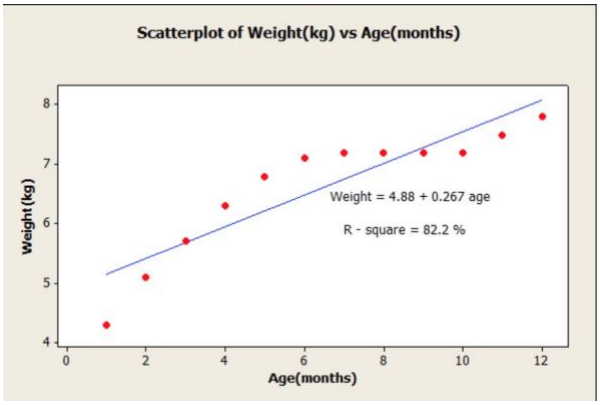
98. (1 point) No. The more serious the patient's illness is, the more likely he/she goes to a bigger hospital, which has better equipment and more beds. Obviously, the number of days that patients with serious illness remain in the hospital are more.

99. (1 point) Weekly visit were made to a local nursing home by college students where thy visit with the residents and serve them herbal tea. The nursing home staff reports that after several months many of the residents are healthier and more cheerful. The explanatory variable in this case is the consumption of herbal tea and the response variable is cheerfulness. A lurking variable could be that the residents were really happy to be visited by people and so became more cheerful the more visits that occurred.

102.(a) (1 point) Scatter plot of the weight against time

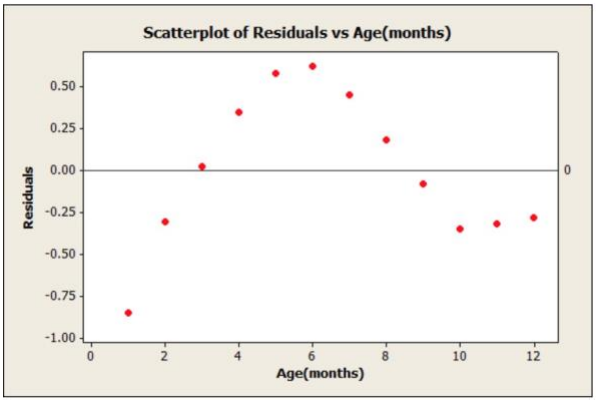


(b) (1 point) Scatter plot of the weight against time with fitted least squares line



The above plot explains clearly that the given pattern of growth is not being explained by the regression equation, Weight = 4.88 +0.267 age. Since the pattern visible to be non-linear from the above plot, to which linear fit is not suitable. Since $R^2 = 82.2\%$ variation. The straight line is only poor approximation of the relationship.

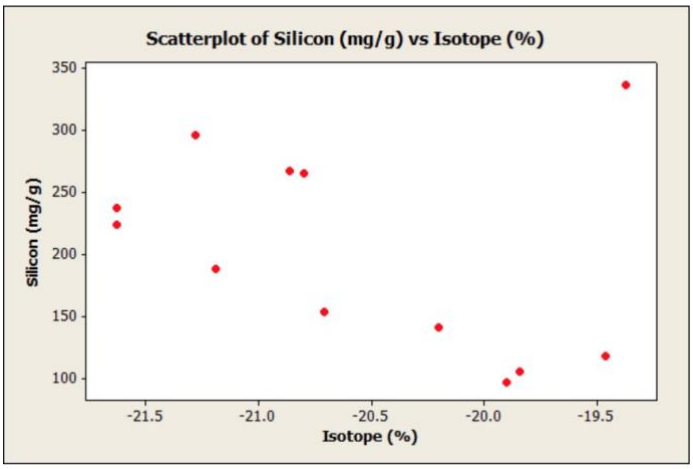
(c) (1 point) Plot of the residuals against age



The first two and last four residuals are below the value zero and the middle six residuals are above the zero. Pattern is visible is curved indicates that the relationship between age and weight is curved not linear.

103. (1 point) The correlation between age and weight for the 170 individual children would surely is much smaller because the variation of individual data increases the scatter; thus decreasing the strength of relationship.

110. (a) (1 point)

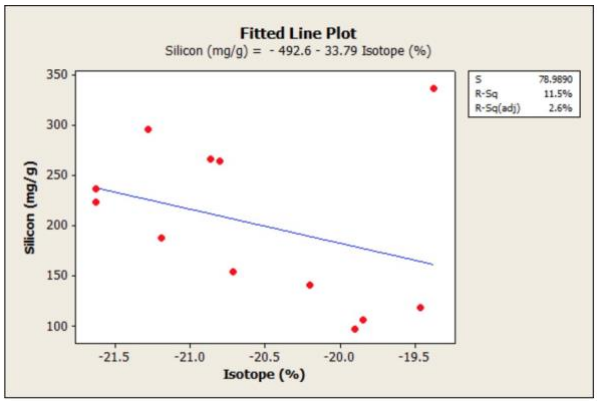


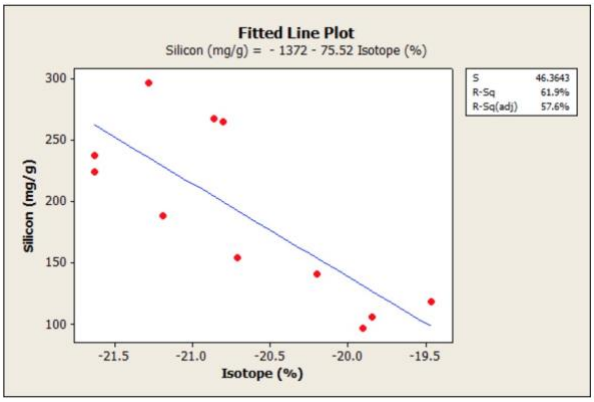
Ignoring the outlier from the above plot; there is a negative association between isotope and silicon. The points in the above plot show a straight pattern. The points in the above scatter plot are not much scattered.

(b) (1 point) Correlation with the outlier is as follows:
Let x be the variable of isotope, then mean of the isotope is $\bar{x} = -20.573$ and the standard deviation of the isotope is $s_x = 0.802$. Let y be the variable of silicon, then the mean and standard deviation of the isotope is $\bar{y} = 202.5$, $s_y = 80$. Correlation between isotope and silicon is $r_{with} = \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) = -0.339$.

Correlation without the outlier can be calculated similarly, $r_{without} = -0.787$.

(c) (1 point)





The slope of the regression line of silicon on isotope with the outlier is $b = -33.815$.
The intercept of the regression line of silicon on isotope with the outlier is $a = -493.185$.
The regression line of silicon on isotope with the outlier is $\hat{y} = -493.185 - 33.815x$.

The slope of the regression line of silicon on isotope without the outlier is $b = -75.518$.
The intercept of the regression line of silicon on isotope without the outlier is $a = -1371.56$. The regression line of silicon on isotope with the outlier is $\hat{y} = -1371.56 - 75.518x$. Due to the outlier the regression line has changed a lot. The regression line without outlier is a good fit to predict silicon from isotope.