Monday, January 15, 2018
4:43 PM

Do Now

Influential point or Outlier?

1.) An observation that lies outside the overall pattern.

Outliers are points that are far away in the (y) vertical direction.
Outliers have large residuals.

2.) If removed, there would be a significant change in the position of the LSRL.

Influential points are far away in the (x) horizontal direction and significantly change the LSRL if removed from the data set.

True or False?

3.) Points that lie close to a straight line always have an r value close to 1.

FALSE - an r value of -1 will also have points close to a straight line.

STOP

Class Work

Directions: Answer each question. Use your notebook if you need more room.

4.) What is the difference between r & r^2?

The corellation coefficient, r, measures the strength and direction of a linear relationship.

r squared measures the percent of variation in the response variable (y) that can be explained by its linear relationship with the explanatory variaqble (x).

5.) What do residuals tell us?

The residuals tell us how far each data point falls from the LSRL.

residual = $y - \hat{y}$ = actual y value - predicted y value (from LSRL).

6.) If a residual plot has a uniform scatter, what does that reveal about the data?

If there is a uniform scatter, with an equal number of points above and below the line, then the LSRL is a good fit for the data.

7.) If a residual plot is curved, what does that reveal about the data?

The original data in not linear and the LSRL is a not good summary for the data.

8.) If a residual plot has an increasing spread, what does that reveal about the data?

The LSRL will not be accurate for larger x-values.

9.) A researcher found that the correlation *(r)* between the color of a person's eyes and their GPA in college was 0.57. What do we know about the study?

The correlation coefficient, r, is not valid. The correlation coefficient, r, is only valid for quantitative variables. Since eye color is categorical, r cannot be used.

10.) A researcher found that the correlation *(r)* between the number of hours a person studied for a test and the person's grade on the test was 0.92. What ***percentage of the variation*** in a person's grade can be explained by the relationship with the number of hours they studied?

$r^2 = (.92)^2 = 84.64\ \%$

84.64 % of the variation in their test grade can be explained by the linear relationship with the number of hours studied.

11.) It is usual in finance to describe the returns from investing in a single stock by regressing the stock's returns on the returns from the stock market as a whole. This helps us see how closely the stock follows the market. We analyzed the monthly percent total return y on Philip Morris common stock and the monthly return x on the Standard & Poor's 500 Index, which represents the market, for the period between July 1990 and May 1997. Here are the results:

$\bar{x} = 1.304$ $s_x = 3.392$ $r = 0.5251$

$\bar{y} = 1.878$ $s_y = 7.554$

A scatterplot shows no very influential observations

$$a = \bar{y} - b\bar{x} \qquad b = r \cdot \frac{S_y}{S_x} \qquad \hat{y} = a + bx$$

Use the formulas to find the equation of the LSRL.

$b = r \cdot \frac{Sy}{Sx} = .5251\left(\frac{7.554}{3.392}\right) = 1.1694$

$a = \bar{y} - b\bar{x} = 1.878 - 1.1694(1.304) = .3531$

$\hat{y} = .3531 + 1.1694x$

Multiple Choice: Circle the letter that best answers the question.

12.) A study is conducted to determine if one can predict the yield of a crop based on the amount of yearly rainfall. The response variable in this study is:

(a) the yield of crop
(b) the amount of yearly rainfall
(c) the experimenter
(d) either bushels or inches of water

13.) When creating a scatterplot, one should:

(a) use the horizontal axis for the response variable
(b) use the horizontal axis for the explanatory variable
(c) use a different plotting symbol depending on whether the explanatory variable is categorical or the response variable is categorical
(d) use a plotting scale that makes the overall trend roughly linear

14.) A school guidance counselor examines the number of extracurricular activities of students and their grade point average (GPA). The guidance counselor says, "The evidence indicates that the correlation between the number of extracurricular activities a student participates in and his or her GPA is close to zero." A correct interpretation of this statement would be that:

(a) active students tend to be students with poor grades, and vice – versa
(b) students with good grades tend to be students that are not involved in many activities, and vice – versa
(c) students involved in many extracurricular activities are just as likely to get good grades as bad grades and the same is true for students involved in few extracurricular activities
(d) involvement in many extracurricular activities and good grades go hand – in – hand

15.) At a large university, the office responsible for scheduling classes notices that demand is low for classes that meet between 10:00 AM and 3:00 PM. Which of the following may we conclude?

(a) There is an association between demand for classes and the time the classes meet.
(b) There is a *positive* association between demand for classes and the time the classes meet.
(c) There is a *negative* association between demand for classes and the time the classes meet.
(d) There is a no association between demand for classes and the time the classes meet.

16.) Which of the following statements is true?

(a) The correlation coefficient equals the proportion of times two variable lie on a straight line.
(b) The correlation coefficient will be +1.0 only if all the data lie on a perfectly horizontal straight line.
(c) The correlation coefficient measures the fraction of outliers that appear in a scatterplot.
(d) The correlation coefficient has no unit of measurement and must always lie between -1.0 and +1.0 inclusive.

17.) A study found a correlation of $r = -0.61$ between the gender of a worker and his or her income. You may correctly conclude that:

(a) women earn more than men on average
(b) women earn less than men on average
(c) an arithmetic mistake was made. Correlation must be positive.
(d) this is incorrect because *r* makes no sense here. Gender is categorical.

18.) The fraction of the variation in the values of y that is explained the least – squares regression of y on x is:

(a) the correlation coefficient *(r)*
(b) the slope of the least – squares regression line *(b)*
(c) the square of the correlation coefficient *(r^2)*
(d) the intercept of the least – squares regression line *(a)*

19.) Which of the following is true of the least – squares regression line?

(a) The slope is the change in the response variable (y) that would be predicted by a unit change in the explanatory variable (x).
(b) It always passes through the point $(\overline{x}, \overline{y})$, the means of the explanatory and response variables, respectively.
(c) It will only pass through all of the data points if $r = \pm 1$
(d) All of the above.

20). A researcher wishes to determine whether the rate of water flow (in liters per second) over an experimental soil bed can be used to predict the amount of soil washed away (in kilograms). The researcher measures the amount of soil washed away for various flow rates and from these data calculates the least – squares regression line to be:

$$\hat{y} = a + bx$$

$$\text{amount of eroded soil} = 0.4 + 1.3(\text{flow rate})$$

One of the flow rates used by the researcher was 0.3 liters per second; for this flow rate, the amount of eroded soil was 0.8 kilograms. These values were used in the calculation of the least – squares regression line. The residual corresponding to these values is:

$$Residual = y - \hat{y}$$

(a) 0.01
(b) -0.01
(c) 0.5
(d) -0.5

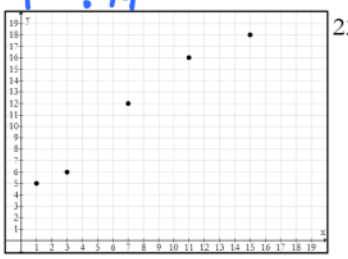
21.) The equation $y = -.43x + 1.25$, with $r = -.72$ is given. Find the sum of the residuals.

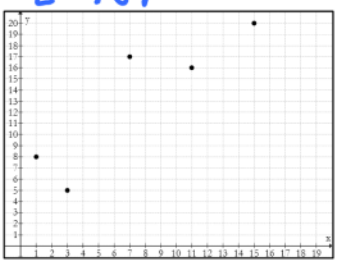
(a) $-.43$
(b) 1.25
(c) $-.72$
(d) 0
(e) 1

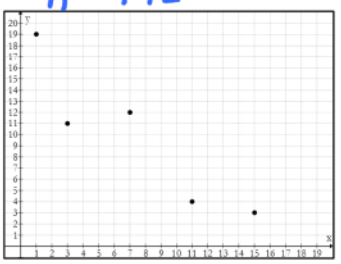
Directions: Match each graph to the corresponding correlation.

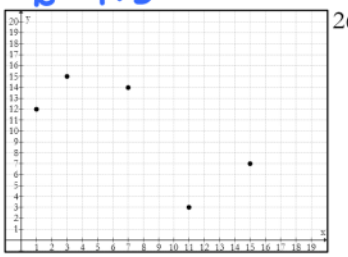
(a) $r = -0.92$ (b) $r = -0.73$ (c) $r = -0.41$

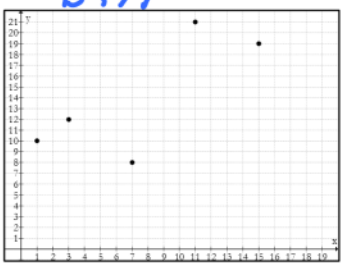
(d) $r = 0.77$ (e) $r = 0.89$ (f) $r = 0.99$

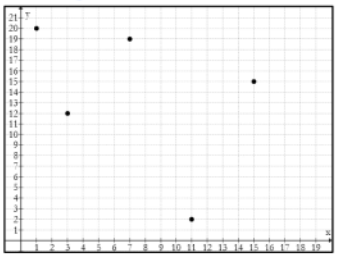
22.) F .99 

23.) E .89 

24.) A -.92 

25.) B -.73 

26.) D .77 

27.) C -.41 

28) **FIGHTING FIRES:** Someone says, "There is a strong positive correlation between the number of firefighters at a fire and the amount of damage the fire does. So sending lots of firefighters just causes more damage." Another possible reason could be the size of the fire.

a) Identify the variables, x, y & z.

X - Number of firefighters at a fire Y - Amount of damage the fire does Z - The size of the fire

b) Draw a diagram of the relationship in which each circle represents a variable. Write a brief description of the variable by each circle.

c) State whether the relationship between the two variables involves causation, common response, or confounding.

Common Response