



Scatter Plots, Lines of Regression and Residual Plots

Previously, you

- Graphed points on the coordinate plane
- Graphed linear equations
- Identified linear equations given its graph
- Used functions to solve problems in the context of the data

In this unit you will

- Represent data on a scatterplot
- Draw the line of best fit from a scatterplot
- Find the least squares regression line for a set of data
- Interpret the correlation coefficient
- Assess the fit of a function by using a residual plot

You can use the skills in this unit to

- Create a linear function to model the relationship between two quantitative variables
- Estimate the correlation coefficient by looking at a scatterplot
- Use technology to determine the least squares regression line, the correlation coefficient and a residual plot

Vocabulary

- **Bivariate data** – Data consisting of two variables.
- **Correlation** – Measures the strength and the direction of the linear relationship between two quantitative variables.
- **Correlation Coefficient** – A numerical value, represented by r , that describes the relationship between two numerical variables.
- **Dependent variable** – The response variable or the y -values of the function that depend on the values of the explanatory (independent) variable.
- **Independent variable** – The explanatory variable or the x -values of the function.
- **Line of best fit** – It is the line drawn through the data that best approximates the data trend (should include as many points above the line as below).
- **Least Squares Regression Line (LSRL)** – A linear equation that ensures minimum distance between the trend line and the data.
- **Negative correlation** – The association when one of the variables increase, the other variable decreases.
- **Positive correlation** – The association when one of the variables increases, the other increases as well.
- **Residual** – The difference between the observed value of the dependent variable and the value predicted by the regression line.
- **Residual plot** – A scatterplot plotting the explanatory variable against the residuals.
- **Scatterplot** – A collection of bivariate data graphed as ordered pairs on a coordinate plane



Essential Questions

- How can the relationship between two quantitative variables be described?
- How is a model created for linearly associated data?
- What is correlation? How can the strength of the linear relationship between two quantitative variables be measured?
- How is the fit of the model evaluated?

Overall Big Ideas

Mathematical functions can describe the relationship between two quantitative variables. Paired numerical data which has a linear relationship may be modeled with a linear equation. Correlation measures how strongly two numerical variables are linearly associated. To determine the fit of a function to data, its residual plot is analyzed.

Scatter plots are a collection of bivariate data that are plotted as ordered pairs on a coordinate plane. These points can be used to determine the trend of the data. If the trend models that of a non-curved line, then a line of best fit can be used to model the data as a linear equation. The linear equation (prediction equation) can be used to make predictions about future values. A linear regression is a more precise way of modeling data that utilizes a graphing calculator to write the linear equation.



Skills

To draw a line of best fit given a scatterplot.

To fit a linear regression to a data set.

To use a residual plot to select the best linear model.

To interpret the correlation coefficient for a linear model in context.

Related Standards

S.ID.B.6a

Fit a linear function to the data; use functions fitted to data to solve problems in the context of the data. Use given functions or choose a function suggested by the context. Emphasize linear, quadratic, and exponential models. *(Modeling Standard)

S.ID.B.6b

Informally assess the fit of a function by plotting and analyzing residuals. *(Modeling Standard)

S.ID.B.6c

Fit a linear function for a scatter plot that suggests a linear association. *(Modeling Standard)

S.ID.C.8

Compute (using technology) and interpret the correlation coefficient of a linear fit. *(Modeling Standard)



Notes, Examples, and Exam Questions

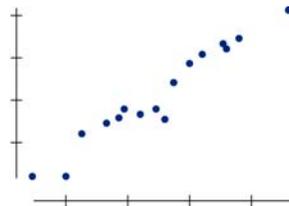
Scatter Plots and Line of Best Fit

A scatterplot is a collection of data graphed as points on a plane. The points are used to determine a relationship (correlation) between the data. Both of the variables must be quantitative (numerical) and cannot be categorical. The explanatory variable or independent variable attempts to “explain” the response or dependent variable. You would use the explanatory variable to predict the value of the response variable. In a scatterplot, the explanatory variable is always graphed on the horizontal (x) axis.

Two variables have a positive association when the both increase or decrease together. Two variables have a negative association when an increase in one variable indicated a decrease in the other. Two variables have no association when the change in one variable cannot be determined from the change in the other. When describing scatterplots, we look for strength (strong, moderate or weak), direction (positive, negative, none), form (linear and nonlinear) and unusual features.

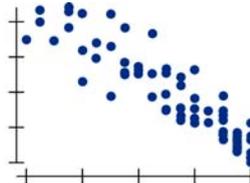
Strength:

a.



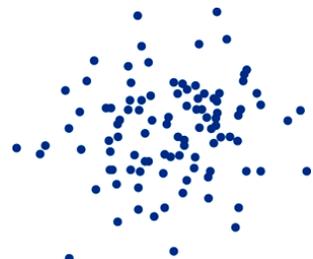
Strong

b.



Moderate

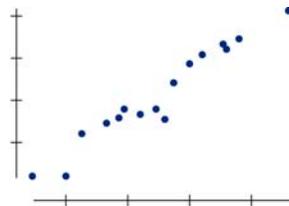
c.



Weak or None

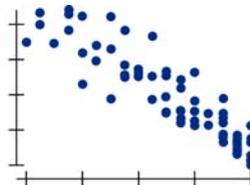
Direction:

a.



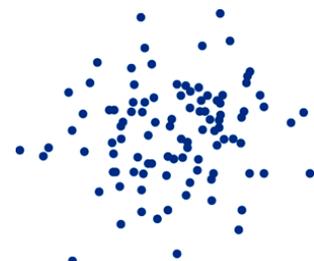
Positive

b.



Negative

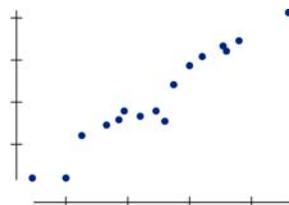
c.



No Direction

Form:

a.



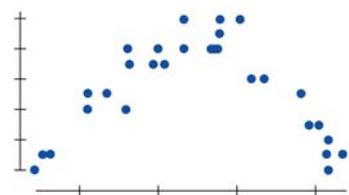
Linear

b.

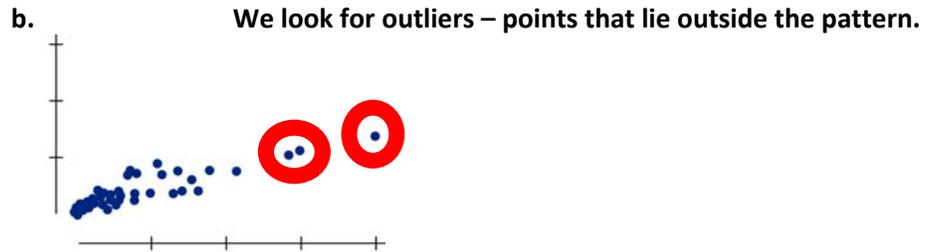
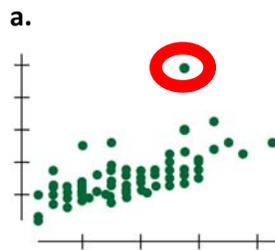


Nonlinear

c.



Nonlinear

**Unusual Features:**

Once correlation (if any) is determined, it is necessary to determine a best fit line and write an equation based on this line. A best fit line is a straight line drawn through the data points while keeping as many points above the line as below the line. This line is thought to model the behavior or trend of the data.

Steps to fitting and writing a trend line

1st - Plot the data as points on a plane.

2nd - Sketch the line (leaving as many points above as below the line).

3rd - Choose two lattice points (ordered pairs) on your line and estimate the coordinates (ordered pairs). These points **DO NOT** need to be part of the data set.

4th - Write the equation using the two points you found in step 3.

Ex 1

Draw a scatter plot of the data. Then approximate the best fitting line for the data

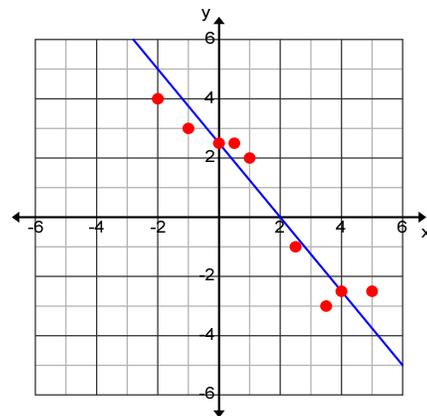
X	-2	-1	0	0.5	1	2.5	3.5	4	5
Y	4	3	2.5	2.5	2	-1	-3	-2.5	-2.5

The best fit line was drawn in blue – approx. half the points above and below the line. Then, write the equation of the line using your two points.

Using the points (0, 2.5) and (4, -2.5), find the slope.

$$m = \frac{-2.5 - 2.5}{4 - 0} = -1.25 \quad \text{and the y-intercept is } 2.5.$$

My line of best fit is: $y = -1.25x + 2.5$





Correlation

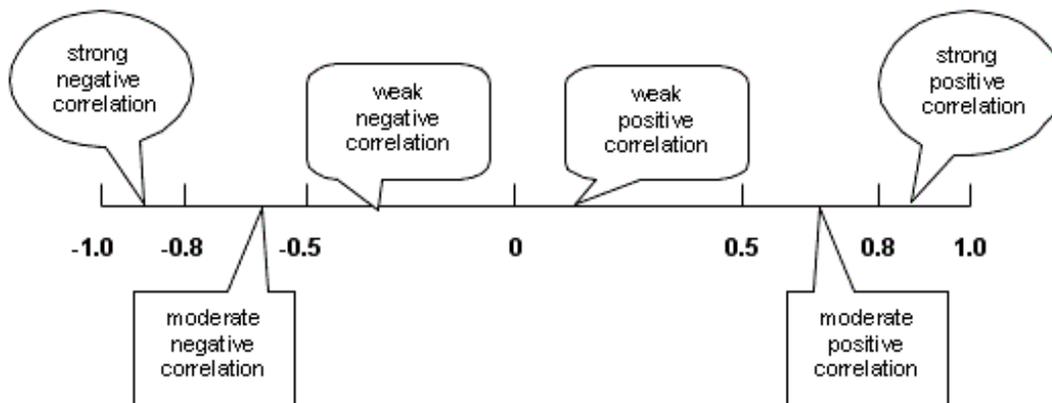
Correlation measures the strength and direction of the linear relationship between two quantitative variables. The numerical measure we use is called the correlation coefficient and its symbol is r . It measures how tightly the points on a

scatterplot cluster about a straight line. The formula is $r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$. The good news is that

technology can calculate this value for us and the formula will not be used.

Facts about correlation:

- Correlation makes no distinction between explanatory and response variables. It makes no difference which variable is graphed on the x-axis and which is on the y-axis.
- Correlation requires that both variables be quantitative (not categorical).
- r does not change if the values of either variable are converted to a different scale. Example: if all the measures are changed from feet to inches, r doesn't change.
- r has no units of measure. It is just a number.
- r is always a value between -1 and $+1$, inclusive.
- Extreme values ($r = -1$ or $r = 1$) indicate a perfect straight line association.
- Values of r near 0 indicate a very weak linear relationship. The strength of the relationship increases as r moves away from 0 toward -1 or $+1$.
- Correlation measures the strength of only a LINEAR relationship between two variables. It does not describe nonlinear relationships.
- Correlation is not resistant. That means it is strongly affected by outliers.
- If your $r = 0$, it does not mean that the variables are not related. It just means that there is not a LINEAR relationship. They could be related with a nonlinear relationship.
- If the correlation is strong, it does not mean there is a cause and effect relationship between the two variables. We are just saying they are associated, not showing causation.





Ex 2 Technology

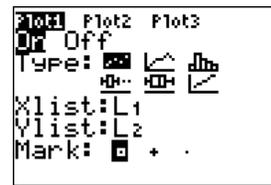
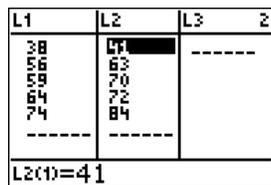
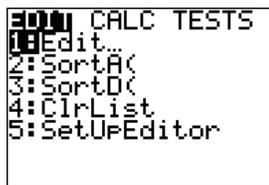
Archaeopteryx is an extinct beast having feathers like a bird but teeth and a long bony tail like a reptile. Only six fossil specimens are known. Because these specimens differ greatly in size, some scientists think they are different species rather than individuals from the same species. If the specimens belong to the same species and differ in size because some are younger than others, there should be a positive linear relationship between the lengths of a pair of bones from all individuals. An outlier would suggest different species. Here are data on the lengths in centimeters of the femur and the humerus for the specimens that preserve both bones:

Femur	38	56	59	64	74
Humerus	41	63	70	72	84

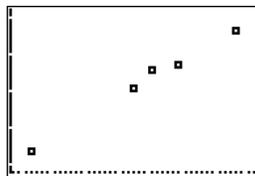
We will use the TI-84 graphing calculator to graph the scatterplot and find the correlation coefficient.

We need to enter the data into two lists. Hit STAT, ENTER and enter the Femur data in L₁ and put the Humerus data in L₂.

After putting the data in the two lists, press “2nd, Y=” to get to STAT PLOT. Go in to Plot1 and set up the following screen.

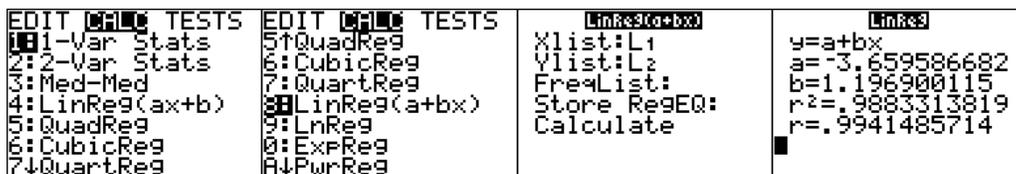


The first “Type” is a scatterplot, the plot we want graphed. To get a nice formatted window, hit “ZOOM 9” and the following scatterplot should be on the device. very strong, positive, linear correlation between in a very linear pattern. To find the correlation regression.



Looking at this scatterplot, we see that there is a femur length and humerus length. The data lies value on the calculator, we have to run a linear

Hit STAT and arrow right to CALC. There are lots of different regressions on this screen and there are two for linear. Number four and number eight both do linear regression. The only difference between the two is the way the linear equation is written. In four, it is written as $\hat{y} = ax + b$, in slope-intercept form, and in eight, it is $\hat{y} = a + bx$, the form used in statistics where the y-intercept is written before the slope. Press 8 or arrow down to 8 and hit ENTER.



Arrow down to “Calculate” and hit Enter. The r shows up at bottom of the screen. $r = .994$ is very close to positive one and tells us that there is a very strong, positive, linear relationship. If the r and r^2 do not show up, they can be turned on using the “CATALOG” button. Provided is a link to a video as a reference for using the calculator to create a scatterplot and write the regression equation or help with getting the the r and r^2 on the screen.

<http://www.youtube.com/watch?v=nw6GOUtC2jY>



Linear Regression

A regression line is a straight line that describes how a response variable, y , changes as an explanatory variable, x , changes. We often use a regression line to PREDICT the value of y for a given value of x . Regression, unlike correlation, requires that we have an explanatory and response variable. The Least-Squares Regression Line (LSRL) is a mathematical model to fit a linear trend. Since we predict y from x (and not the other way around), we want a line that is as close as possible to the points in the vertical direction.

$$\text{Error} = \text{observed} - \text{predicted}$$

$$\text{residual} = y - \hat{y}$$

What does the least square formula do? We find the “error” or residual, which is the vertical distance from each point to the line. The LSRL make the sum of the squares of these distances as small as possible. The equation $\hat{y} = a + bx$, where a is the y -intercept, b is the slope and \hat{y} means the predicted value of y (this is called y -hat). Although formulas can be used to find the equation of the line, the TI-84 can find the model for us by using STAT, CALC, 8.

When discussing the slope of the least squares regression line, it is the average amount of change in the y -variable when the x -variable increases by one unit. Use this definition in the context of the problem. The y -intercept is the value of the prediction y value when x is equal to zero. Again, make sure to use the context of the problem when defining the variables.

Ex 3

Consider the following data from the article, “The Carbonation of Concrete Structures in the Tropical Environment of Singapore”. The explanatory variable is carbonation depth in concrete (mm) and the response variable is the strength of the concrete (Mpa).

x	8	20	20	30	35	40	50	55	65
y	22.8	17.1	21.5	16.1	13.4	12.4	11.4	9.7	6.8

First, input the data into the calculator and draw the scatterplot.

L1	L2	L3	3
8	22.8		
20	17.1		
20	21.5		
30	16.1		
35	13.4		
40	12.4		
50	11.4		

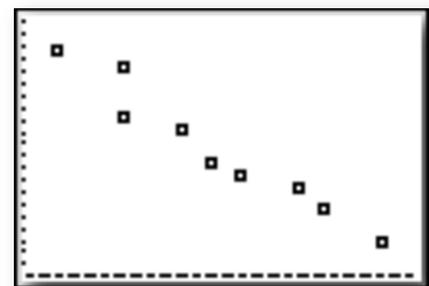
Plot1	Plot2	Plot3
On	Off	Off
Type:		
Xlist:	L1	
Ylist:	L2	
Mark:		

MEMORY
4:ZDecimal
5:ZSquare
6:ZStandard
7:ZTrig
8:ZInteger
9:ZoomStat
0:ZoomFit

Interpretation: There is a strong, negative linear relationship between depth of corrosion and concrete strength.

In simple, everyday language, this says as the depth increases, the strength decreases at a fairly constant rate.

Strength
(Mpa)



Depth (mm)



Next, find the equation of the LSRL that models the relationship between corrosion and strength. This was shown in the last example.

```

EDIT [2ND] TESTS
4↑LinReg(ax+b)
5:QuadReg
6:CubicReg
7:QuartReg
8⇄LinReg(a+bx)
9:LnReg
0↓ExpReg
  
```

```

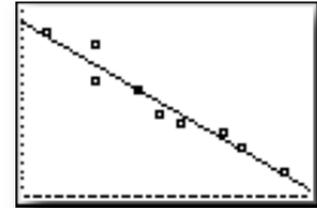
LinReg(a+bx) L1,
L2
  
```

```

LinReg
y=a+bx
a=24.51683116
b=-.276939568
r²=.9375144639
r=-.9682533056
  
```

$\hat{y} = 24.517 - .277x$, where x is the depth and \hat{y} is the predicted strength.

$r = -.968$, telling us that there is a strong, negative linear relationship between depth of corrosion and strength of concrete.



What does the slope tell us? For every increase of 1 mm in depth of corrosion, we predict on average, a 0.277 Mpa decrease in strength of the concrete. Let's define the y-intercept. When the carbonation depth in the concrete is 0 mm, the predicted strength of the concrete would be 24.517 Mpa.

Use the LSRL to find the predicted strength of concrete with a corrosion depth of 25 mm.

$$\hat{y} = 24.517 - 0.277(25) = 17.592$$

$$\hat{y} = 17.592 \text{ Mpa}$$

What is the predicted strength of concrete with a corrosion depth of 40 mm?

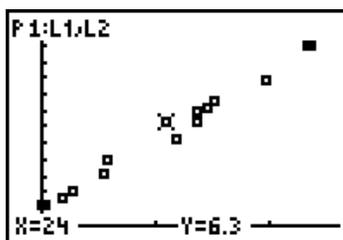
$$\hat{y} = 24.517 - 0.277(40) = 13.437$$

$$\hat{y} = 13.437 \text{ Mpa}$$

Residual Plot

A residual plot is a scatterplot of the regression residuals against the explanatory variable (x). Residual plots help us to assess the fit of a regression line. There should be NO PATTERN in the residual plot. We are looking for uniform (random) scatter – this means the line fits the data well and that a linear model is appropriate. If there is curvature, increasing or decreasing spread, or lots of points with large residuals, this may not be a good fit and is an indicator that the linear regression is not a good fit for the data.

Scatterplot of x vs. y



Residual plot



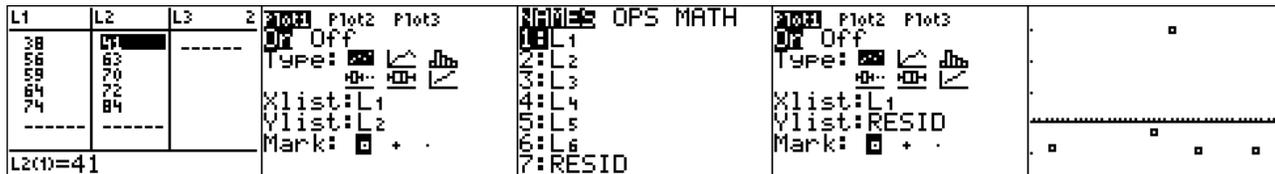
****Note the uniform scatter in this residual plot.**

This illustrates a linear model is appropriate.



Ex 4 Technology

To graph a residual plot, we need to do a scatterplot of the explanatory variable and the residuals. Every time the calculator does a linear regression, it finds the residual for each ordered pair ($\text{residual} = y - \hat{y}$) and puts them in a list. The calculator-created list is called RESID. Going back to the data on femur and humerus from example 2, place the data in L_1 and L_2 , press STAT, CALC, 8, to run the linear regression and then go to STAT PLOT and change the L_2 to RESID. The list name RESID can be found under “2nd” “STAT” or the LIST menu. It is usually number 7 unless you have other named lists in the calculator.



Enter “ZOOM 9” to get a nice screen for the residual plot. It is hard to see pattern or curvature when they are only five ordered pairs, but there appears to be random scatter in this residual plot. This means that a linear model is appropriate.

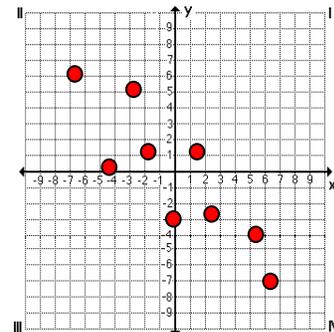
SAMPLE QUESTIONS

Ex 5

Using the scatterplot at the right, describe the correlation.

Remember to discuss strength, direction and form.

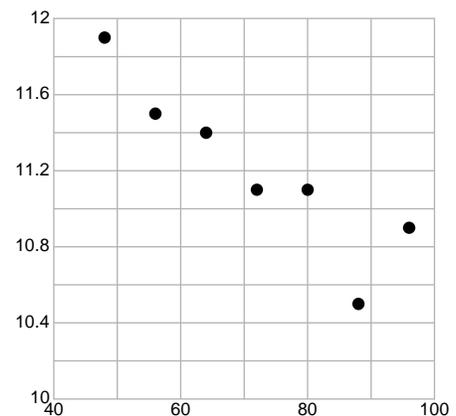
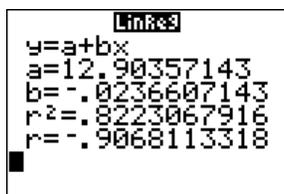
There is a strong, negative linear correlation between the two variables. I would estimate the r to be around 0.82.



Ex 6

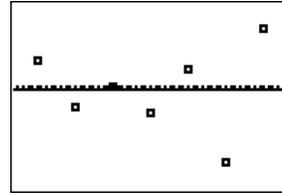
Draw a scatter plot for the Women’s Olympic 100 meter times below. Find the equation of a line that closely fits the data, determine the correlation and assess the residual plot.

Year (19_) y	48	56	64	72	80	88	96
Time (sec) t	11.9	11.5	11.4	11.1	11.1	10.5	10.9





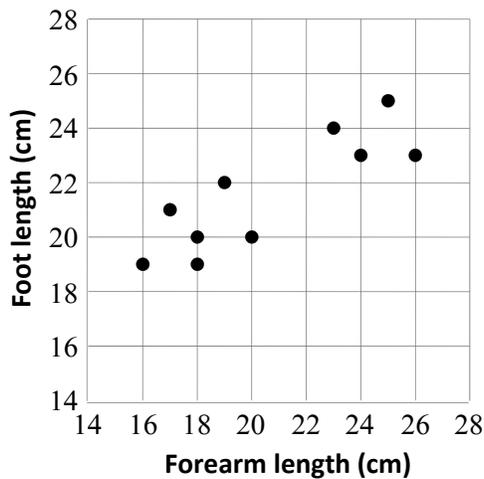
There is a strong, negative linear correlation between year and Women's Olympic 100 meter times. The equation is $\hat{y} = 12.903 - 0.0237x$, where x is year and \hat{y} is the predicted 100 meter time. The slope tells us that as years increase by one, on average, the 100 meter times decrease by 0.0237 seconds. The residual plot below does not illustrate a pattern or curvature indicating that a linear model is appropriate.



QOD: The points of a scatterplot lie on a horizontal line. What type of correlation does this show? Explain your answer.

SAMPLE EXAM QUESTIONS

1. The scatterplot below represents the forearm lengths and foot lengths of 10 people.



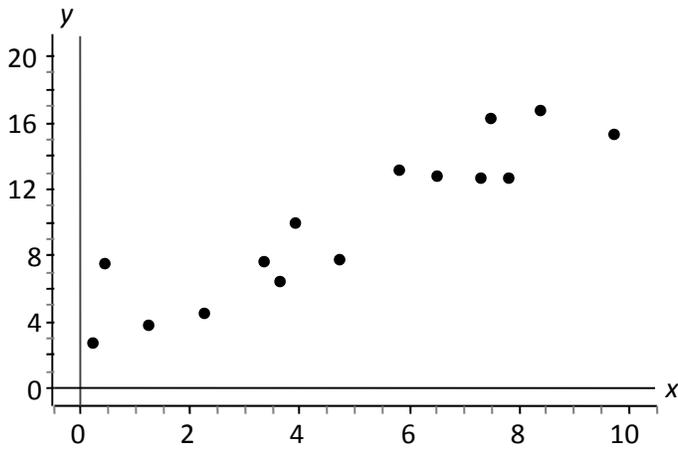
Based on a linear model of the data, which is the best prediction for the length of a person's foot if his/her forearm length is 21 centimeters?

- A. 19 cm
- B. 20 cm
- C. 22 cm
- D. 24 cm

Ans: C



2. The line of best fit for the scatterplot below is $\hat{y} = 1.4x + 2.9$

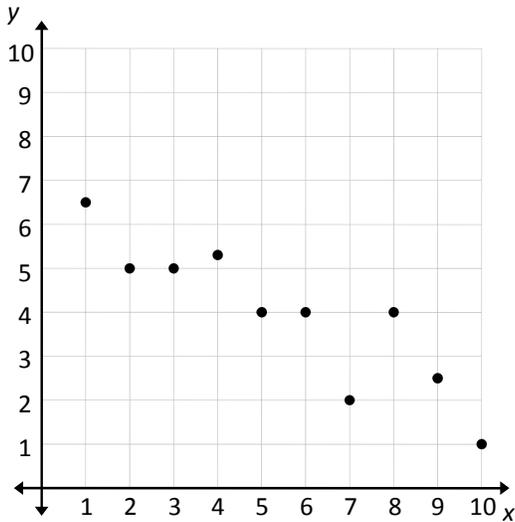


Predict y when $x = 6$.

- A. 2.2 B. 10.5 C. 11.3 D. 18.8

Ans: C

3. Which equation best describes fits the data shown in the scatterplot?

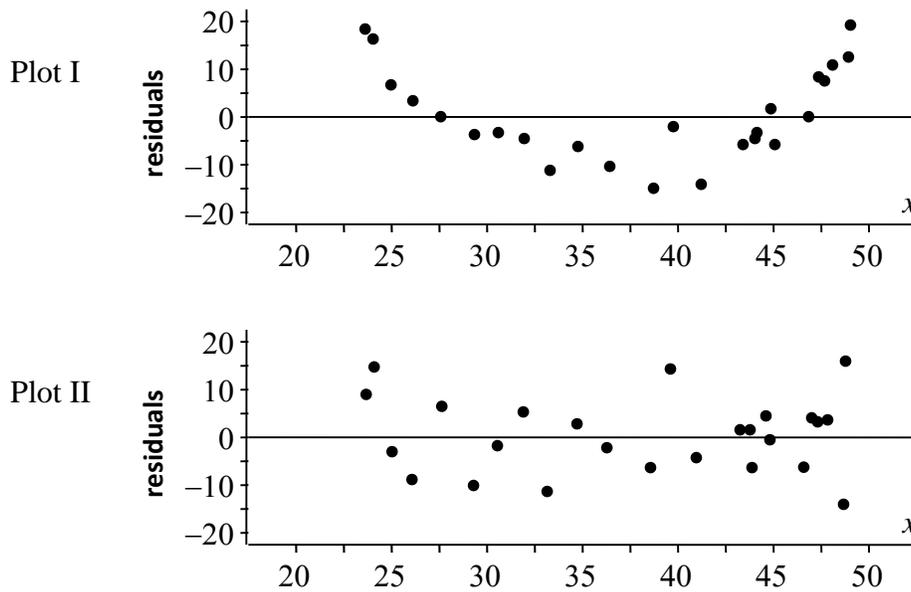


- A. $y = -\frac{3}{5}x + 7$ C. $y = -\frac{1}{3}x + 8$
 B. $y = x + 8$ D. $y = 4$

Ans: A



4. Two residual plots are shown below.



Which residual plot(s) would indicate a linear model is appropriate?

- A. Plot I only C. Plot II only
 B. Both Plot I and Plot II D. Neither Plot I nor Plot II

Ans: C

5. A scatterplot is made of a city's population over time. The equation of the line of best fit is $\hat{p} = 629t + 150,000$ where \hat{p} is the city's predicted population size and t is the number of years since 2000. What is the meaning of the slope of this line?

- A. In 2000, the city's population was about 629 people.
 B. In 2000, the city's population was about 150,000 people.
 C. The city's population increases by about 629 people each year.
 D. The city's population increases by about 150,000 people each year.

Ans: C

6. A value of r near zero indicates there is a weak linear relationship between x and y .

- A. True
 B. False

Ans: A

**7. The data below comes from a scatterplot.**

x	2	3	4	5	6	7	8	8	8	9	10	10
y	2	8	4	1	10	4	6	10	2	7	3	9

Which best describes the linear relationship between x and y ?

- A. weak or no correlation
- B. strong positive correlation
- C. strong negative correlation

Ans: A

8. A value of $r = -0.5$ indicates a weaker linear relationship between x and y than a value of $r = 0.5$.

- A. True
- B. False

Ans: B

9. A value of $r = 1$ indicates that there is a cause-and-effect relationship between x and y .

- A. True
- B. False

Ans: B