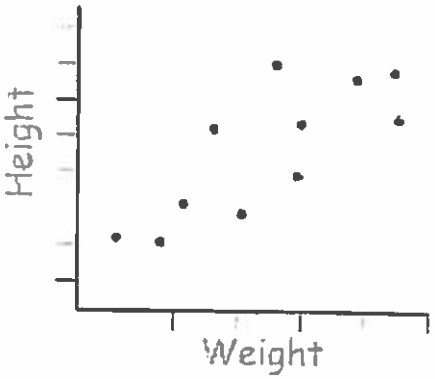
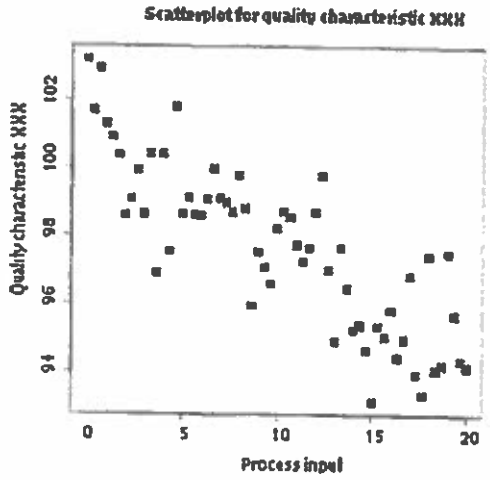
Displaying data visually can help you see relationships. A **scatter plot** is a graph with points plotted to show a possible relationship between two sets of data. A scatter plot is an effective way to display some types of data.

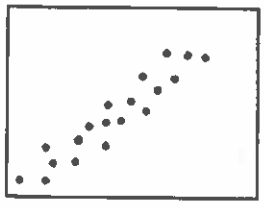




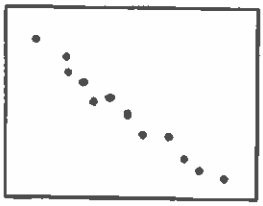
Is a scatter plot discrete or continuous?

discrete, the points aren't connected

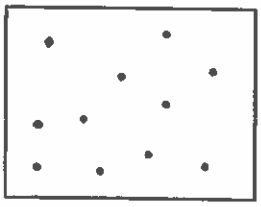
A *scatter plot* is helpful in understanding the form, direction, and strength of the relationship between two variables. **Correlation** is the strength and direction of the linear relationship between the two variables.



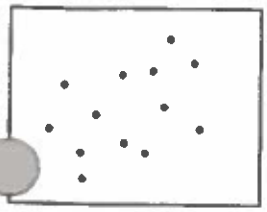
Strong Positive



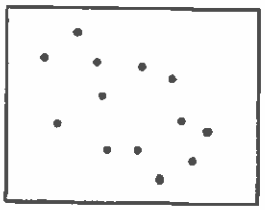
Strong Negative



None



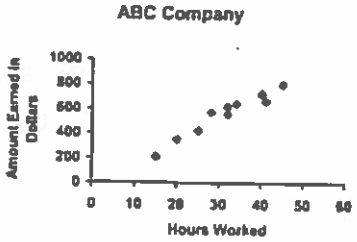
Weak Positive



Weak Negative

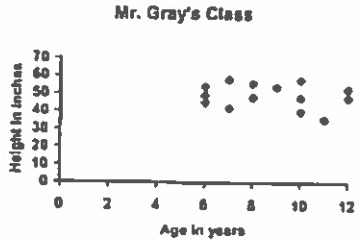
Scatter Plots

1) The scatter plot below shows a relationship between hours worked and money earned. Which best describes the relationship between the variables?



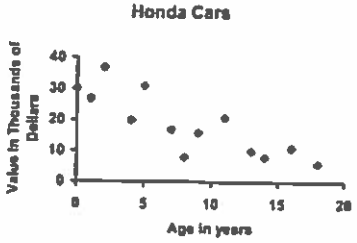
A) Strong positive correlation
B) Weak positive correlation
C) Strong negative correlation
D) Weak negative correlation

2) This scatter plot shows a relationship between age and height. Which best describes the relationship between the variables?



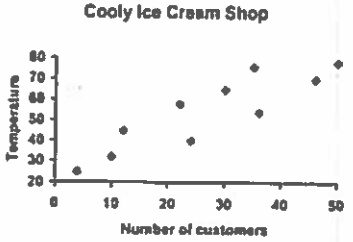
A) Strong positive correlation
B) Weak positive correlation
C) Strong negative correlation
D) No correlation

3) This scatter plot shows the relationship between the age of a car and its value. Which best describes the relationship between the variables?



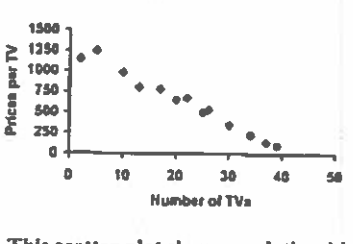
A) Strong positive correlation
B) Weak negative correlation
C) Strong negative correlation
D) No correlation

4) This scatter plot shows a relationship between the outdoor temperature and number of customers in an ice cream store. Which best describes the relationship between the variables?



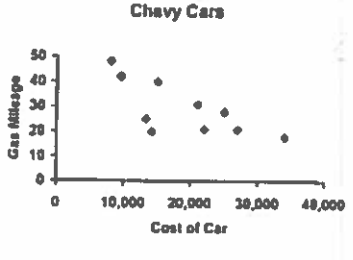
A) Strong positive correlation
B) Weak positive correlation
C) Weak negative correlation
D) No correlation

5) This scatter plot shows a relationship between the TVs purchased and prices. Which best describes the relationship between the variables?



A) Strong positive correlation
B) Weak positive correlation
C) Strong negative correlation
D) Weak negative correlation

6) This scatter plot shows a relationship between the cost of Chevy cars and their gas mileage. Which best describes the relationship between the variables?



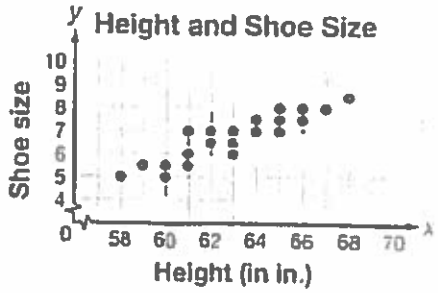
A) Strong positive correlation
B) Weak positive correlation
C) Weak negative correlation
D) No correlation

LESSON

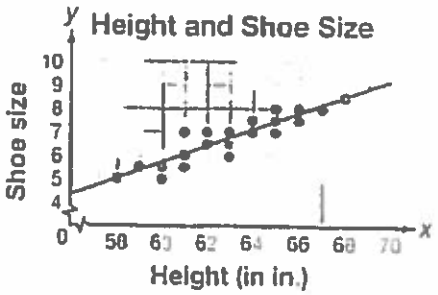
19

Constructing and Analyzing Scatter Plots

UNDERSTAND When you study the relationship between two variables—such as the heights and shoe sizes of a group of students—you are working with **bivariate data.** Bivariate data can be written as a set of (x, y) ordered pairs and graphed on a coordinate plane. This kind of graph is called a **scatter plot.** A scatter plot can help you interpret bivariate data. The scatter plot below shows a set of ordered pairs in which the x-values represent heights and the y-values represent shoe sizes.



Look at the shape formed by the plotted points. The shape resembles a straight line. This suggests a linear relationship between the variables. You can draw a line to fit, or model, the data. The line you draw represents a linear function. If the line is a good fit, you can use the graph and the equation of the line to interpret and make predictions about the data.



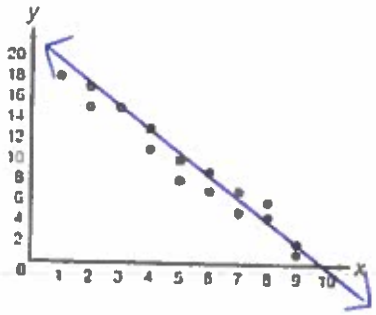
The line appears to be a good fit. The data points slant up from left to right, indicating a positive linear relationship. The line has a positive slope and is close to most data points.

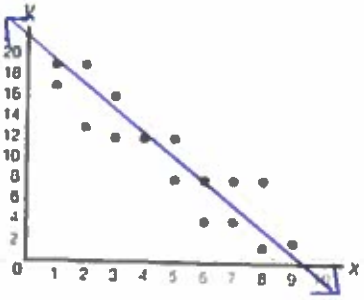
You could also show that the line is a good fit for the data by calculating **residuals.** For each point (x, y) on the scatter plot, there is a corresponding point $(x, \hat{y})$ on the line of fit. A residual is equal to the difference $y - \hat{y}$. Residuals measure the difference of each actual y-value and the expected y-value $(\hat{y})$, which is based on the equation of the line of fit.

Residuals help you determine how accurately the linear function could predict actual points on the scatter plot. That is, if the values of the residuals are relatively small, the linear function is a good fit for the data. So, for any value of x, you could use the equation of the line to make a good prediction about what the value of y would be, and vice versa.

20

Draw a line of fit for each of the scatter plots. Determine how well each fits the data.

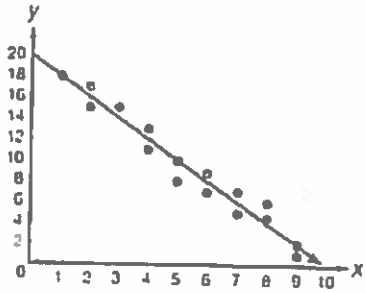


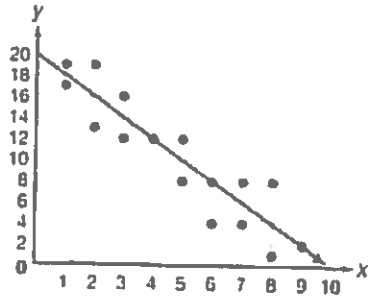


1

Draw a line to model the data for each scatter plot.

For each plot, draw a line that has about as many points above it as below it.





Are the lines drawn the only possible lines of fit that could have been drawn for these scatter plots? Why or why not?

2

Use residuals to determine how well the lines fit the data in the first plot.

Pick several data points, such as (1, 18), (4, 11), (6, 7), and (8, 6). Find the corresponding points, $(x, \hat{y})$, on the line for those x-values: (1, 18), (4, 12), (6, 8), and (8, 4). Calculate the residuals.

(1, 18): $y - \hat{y} = 18 - 18 = 0$

(4, 11): $y - \hat{y} = 11 - 12 = -1$

(6, 7): $y - \hat{y} = 7 - 8 = -1$

(8, 6): $y - \hat{y} = 6 - 4 = 2$

None of the residuals have large values. The line fits the first data set well.

3

Use residuals to determine how well the line fits the data in the second plot.

Pick several data points: (1, 19), (4, 12), (6, 4), and (8, 8). Find the corresponding points on the line: (1, 18), (4, 12), (6, 8), and (8, 4). Calculate the residuals.

(1, 19): $y - \hat{y} = 19 - 18 = 1$

(4, 12): $y - \hat{y} = 12 - 12 = 0$

(6, 4): $y - \hat{y} = 4 - 8 = -4$

(8, 8): $y - \hat{y} = 8 - 4 = 4$

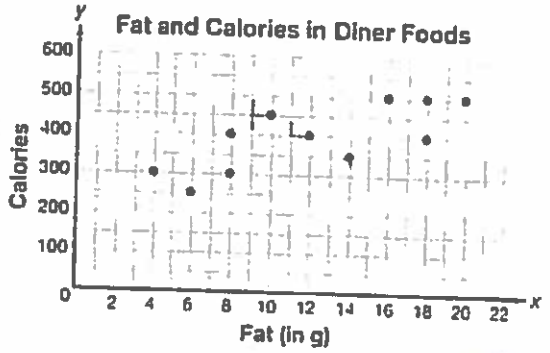
Some of the residuals have large values. The line does not fit the second data set well.

EXAMPLE For a health project, Dylan recorded the number of grams of fat and the number of calories in lunch entrees sold at his favorite diner.

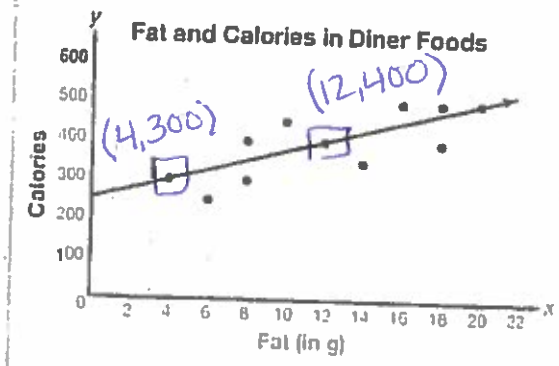
Fat (in grams)	4	6	8	8	10	12	14	16	18	18	20
Calories	300	250	300	400	450	400	350	500	400	500	500

Create a scatter plot for the data. Draw a line to fit the data. Find the equation of the line.

1 Use the ordered pairs of data items to create a scatter plot.



2 Draw a line to fit the data.



*PICK 2 points on the line

3 Write an equation for the line of fit.

The points (4, 300) and (12, 400) are on the line. Use those points to find the slope.

$$m = \frac{400 - 300}{12 - 4} = \frac{100}{8} = 12.5$$

The y-intercept is at (0, 250), so $b = 250$.

▶ The equation of the line is $y = 12.5x + 250$.

Slope: $m = \frac{Y_2 - Y_1}{X_2 - X_1}$

DISCUSS

Explain what the slope of the line tells you in this context. Do the data show a positive linear relationship or a negative linear relationship?

how much calories increase for each gram of fat, positive linear relationship

22

Problem Solving

READ

The scatter plot shows the ages of various Model Z smartphones, in months, and the prices for which they sold. Predict how much Trent will pay if he buys a Model Z smartphone that is 5 years old.

PLAN

Draw a line to fit the data. Write the equation of the line, and use it to predict the price for a phone that is 5 years, or 60 months, old.

SOLVE

On the scatter plot, draw a line that fits the data.

$\frac{400-175}{6-48} = \frac{225}{-42} = -5.4$

Choose two points on the line, (6, 400) and (48, 175).

Use the points to find the slope of the line. m = -5.4

In this context, the slope represents decrease in price per month.

Find the y-intercept of the line. Extend the line to the y-axis if necessary. b = 425

The equation for the line is y = -5.4x + 425.

In this context, the y-intercept represents starting price per month.

To predict the cost of a 5-year-old smartphone, substitute 60 for x in the equation. $ 101

-5.4(60) + 425

CHECK

Pick three data points from the scatter plot: (6, 350), (30, 250), (48, 225).

Find the points with corresponding x-values on the line of fit: (6, 400), (30, 250), (48, 175).

350 - 400 = -50

250 - 250 = 0

225 - 175 = 50

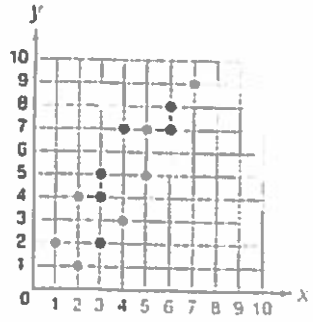
Calculate the residual for each point. Each residual is relatively small.

Does the line fit the data well? yes Is your answer a reasonable prediction? yes

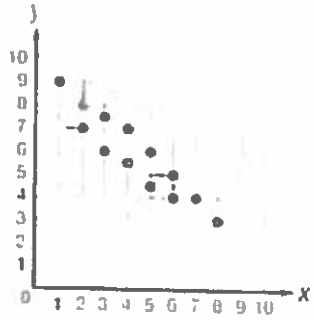
A good prediction is that Trent will pay about 100 for a Model Z smartphone that is 5 years old.

Practice

Describe the relationship shown in each scatter plot as either *positive* or *negative*.

1. 

positive

2. 

negative

HINT: A line that slants from lower left to upper right has a positive slope

Use the information and table below for questions 3 and 4.

The table below shows T-shirt sales data for a store one weekend.

Price, x (in dollars)	4	8	8	12	12	16	20	20	24	24
Number Sold, y	32	26	30	22	26	20	12	20	14	10

3. Create a scatter plot for the data. Then draw a line of fit for the data.

(8, 30)(16, 20)

$\frac{30-20}{8-16} = \frac{10}{-8} = -1.25$

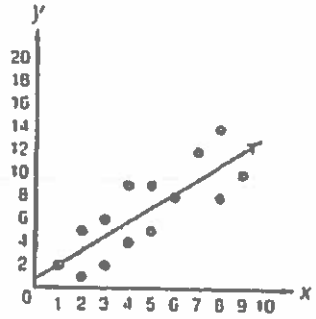
4. Find the slope of the line of fit. What does it represent in the context of this problem?

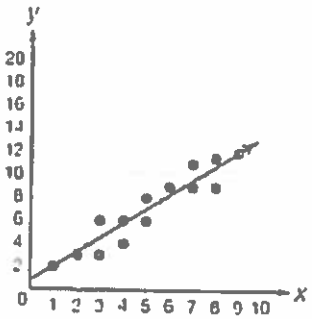
$m = -1.25$, as the price increases, numbers sold decreases by 1.25 per dollar

22

Assess the fit of the lines to the data.

5. The lines of fit in the scatter plots below are identical.



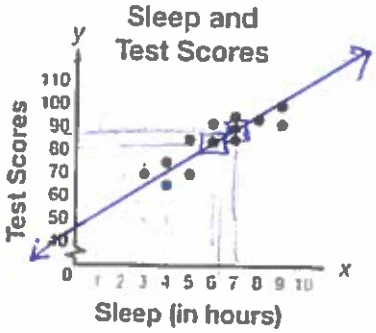


Which line better fits the data in its scatter plot? How did you determine your answer?

second one - points are closer to the line

Use the information and scatter plot below for questions 6 and 7.

The scatter plot shows the number of hours of sleep that students got the night before a test and their scores on the test.



6. INTERPRET Draw a line of fit for the scatter plot. Identify the slope and y-intercept of the line. What does each represent in the context of this problem?

m = 5 (increase in score per hr of sleep)

y-int: 45 (score with no sleep)

(6, 80) (7, 85)

$\frac{85-80}{7-6} = \frac{5}{1} = 5$

7. PREDICT Write the equation of the line. Then use the equation to predict a student's test score if she gets only 2 hours of sleep before the next test.

$y = 5x + 45$

$y = 5(2) + 45$

$y = 55$

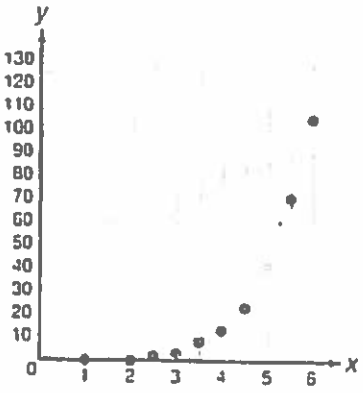
LESSON

23

Best Fit and Correlation

UNDERSTAND You can draw a line of fit for a scatter plot by analyzing the data visually. Someone else, however, could look at the same data and draw a slightly different line. To find the line that best fits the data, you need to use a process called regression analysis. Regression analysis helps you find the function that minimizes residuals.

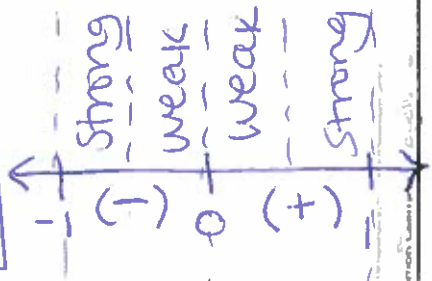
When there seems to be a linear relationship in the data, regression analysis can find the equation of a **line of best fit**. But not all bivariate data show a linear association. In some cases, the relationship between the variables is better modeled by a curve, as in the scatter plot shown. For data that do not have a linear association, you will need to find a **curve of best fit**. To find the equation of either a line of best fit or a curve of best fit, you can use a graphing calculator to perform a regression analysis.



UNDERSTAND Once you have determined the line of best fit for bivariate data, you can use the **correlation coefficient**, r, to describe the strength and direction of the relationship between the two variables.

These statements will help you interpret a correlation coefficient.

- The value of r is always in the range $-1 \leq r \leq 1$.
- If r is close to 1, the data show a strong positive correlation.
- If r is close to -1, the data show a strong negative correlation.
- If $r = 0$, the data do not show a linear correlation.



When bivariate data have a strong correlation, the predictions you make by using the line of best fit are likely to be very accurate. When there is a weak correlation, these predictions will tend to be less accurate. A positive correlation means that as one variable increases, the other variable tends to increase also. A negative correlation means that as one variable increases, the other tends to decrease.

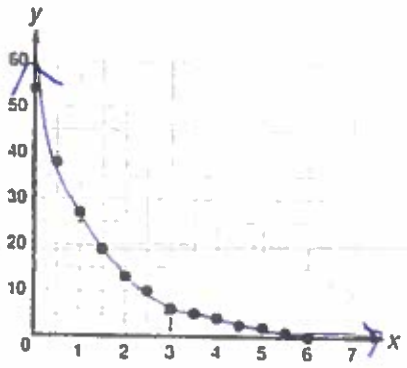
The correlation coefficient, r, is calculated using a rather complex formula involving residuals. Fortunately, you can use a calculator to do that work for you!

Keep in mind that there is a crucial difference between correlation and causation. A strong correlation does not tell you that x is the cause of y. For example, buying lemonade and going to the beach might be strongly correlated, but one does not cause the other.

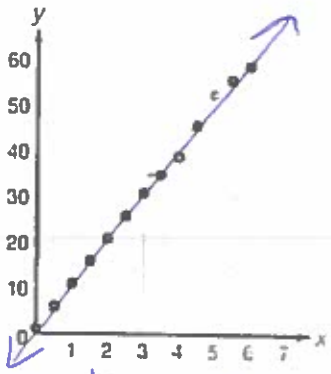
24

Practice

Use *line* or *curve* to tell which kind of model best fits each data set.

1. 

exponential

2. 

linear

Use *strong*, *weak*, *positive*, *negative*, or *no linear correlation* to describe what each correlation coefficient, *r*, tells you about a bivariate data set.

3. $r = 0$ — no correlation

4. $r = 0.250$ — weak positive

5. $r = -0.895$ — Strong negative

REMEMBER The closer r is to 1 or -1, the stronger the correlation.

Write *true* or *false* for each statement. If false, rewrite the statement so it is true.

6. A line of best fit will help you predict values for variables with complete accuracy.

false - with some accuracy

7. Not all bivariate data show a linear correlation, so sometimes data are better modeled by a curve than a line.

true

8. If regression analysis shows that there is a strong correlation between two variables, x and y, then x must cause y.

Use the information, table, and scatter plot below for question 14.

Mrs. Chen started a business 20 years ago. The table and scatter plot show the number of employees her growing business has had over a period of 20 years.

Years in Business	Number of Employees
0	3
5	7
10	19
15	46
20	115

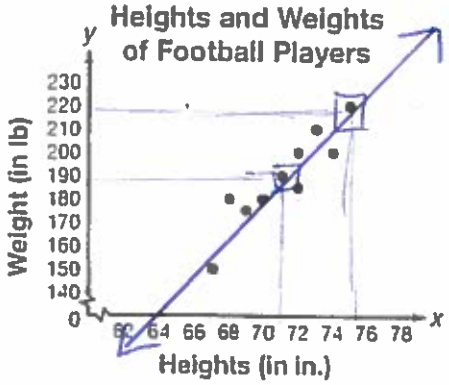
14. Use your calculator to perform an exponential regression for the data. What is the equation of the curve of best fit? Graph that curve on the scatter plot above.

$y = 2.9(1.2)^x$

Use the information, table, and scatter plot below for questions 15 and 16.

The table and scatter plot both show the heights and weights of a randomly selected sample of football players from an all-star team.

Height (in inches)	Weight (in pounds)
67	150
68	180
69	175
70	180
71	190
72	185
72	200
73	210
74	200
75	220



$\frac{220-190}{75-71} = \frac{30}{4} = 7.5$

15. Draw the line of best fit. . What is the equation of the line of best fit? What is the correlation coefficient?

$m = 7.5$ Weak positive

16. Graph that line on the scatter plot above. How good a fit is the line?

2Ce