Lesson 2-3

Linear Regression and Correlation

BIG IDEA The regression line is the line of best fit to data. The correlation coefficient measures the strength and direction of a linear pattern in data.

Vocabulary

method of least squares
line of best fit, least squares line, regression line
center of mass
correlation coefficient
perfect correlation
strong correlation
weak correlation

The sum of squared residuals can be used to determine which of two lines is the better fit to a specific set of data. When data are fairly linear, like the diamond weights and prices in Lesson 2-2, you are likely to find a good model quickly. When there is a large spread of data points on a scatterplot, you might have to try several equations. Even if you pick the equation with the smallest sum of squared residuals, there is likely another line with an even smaller sum of squared residuals that you did not try. How can you find the best model?

Mental Math

Suppose you thought the price of a new car was going to be $19,000. Instead, it was $20,000. By what percent of the actual price were you off?

The Line of Best Fit

The problem of finding the best linear model for a set of data emerged from studies of astronomy, geography, and navigation in the late 1700s and early 1800s. A method for finding the line of best fit was first published by the French mathematician Adrien Legendre in 1805. His approach to the problem is called the **method of least squares**, because he used the Sum of Squared Residuals to determine the linear equation of best fit.

A sextant is an instrument used to determine latitude and longitude, using angle measures.

The **line of best fit**, also known as the **least squares line** or **regression line**, has three important properties:

1. It is the line that minimizes the sum of squared residuals, and it is unique. There is only one line of best fit for a set of data.
2. It contains the **center of mass** of the data, that is, the point $(\bar{x}, \bar{y})$ whose coordinates are the mean of the x-values and the mean of the y-values.
3. Its slope and intercept can be computed directly from the coordinates of the given data points.

Although the formula for the slope of the least squares line uses only addition, subtraction, squaring, and division, it is too complex for computation by hand. Every statistics utility contains a *regression* routine that will take the coordinates of a set of data points and compute the slope and y-intercept of the line of best fit.

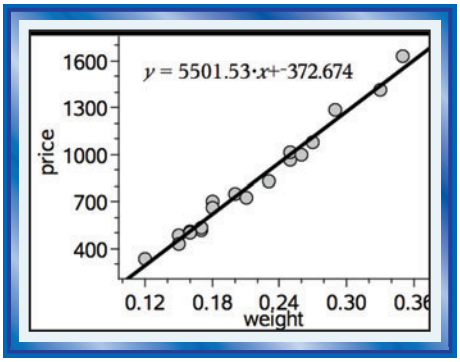
Example

a. Use a statistics utility to find a line of best fit for the data about the weight in carats and the price of diamond rings from Lesson 2-2.

b. According to the regression line, how much will a 0.5-carat diamond ring cost?

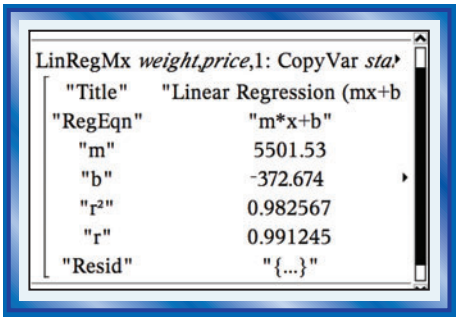
c. Verify that the center of mass (0.212, \$793.65) is on the line.

d. Find the sum of squared residuals for the linear regression.



Solution

a. Enter the entire data set from Lesson 2-2 into a statistics utility. Label the weight as x and the price as y. Use the linear regression feature to find an equation for y as a function of x. The linear regression model is $y = 5501.53x - 372.67$.



b. Substitute 0.5 for x and calculate y.

$y = 5501.53(0.5) - 372.67 \approx 2378.1$

According to the linear regression model, the price would be \$2378.10.

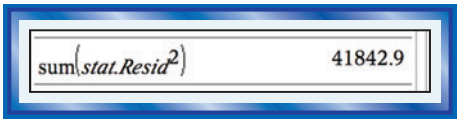
c. Substitute the coordinates of the center of mass for x and y in the regression equation.

$$y = 5501.53x - 372.67$$

Does $793.65 = 5501.53(0.212) - 372.67$?

Yes. It checks.

d. On many statistics utilities, once the regression equation is given, the residuals are calculated and stored until the next regression is found. In the screenshot at the right, the residuals are stored under the name Resid. So the sum of the squared residuals is 41842.9.



STOP QY1

▶ QY1

Use the least-squares regression line to estimate the price of a diamond ring weighing 0.17 carats.

Correlation

For the diamond ring data, the regression line is the best linear model for these data. But how good is "best?"

To measure the strength of the linear relation between two variables, a *correlation coefficient* is used. The correlation coefficient is often denoted by the letter r. The terms "co-relation" and "regression" were introduced by the English researcher Sir Francis Galton in the 1880s in a study comparing the heights of children with the heights of their parents. The statistic we now use for correlation was given a mathematical foundation by the English statistician Karl Pearson in 1896. Statisticians today use Pearson's formula, which can be evaluated automatically with a statistics utility.

Definition of Correlation Coefficient (Pearson's Formula)

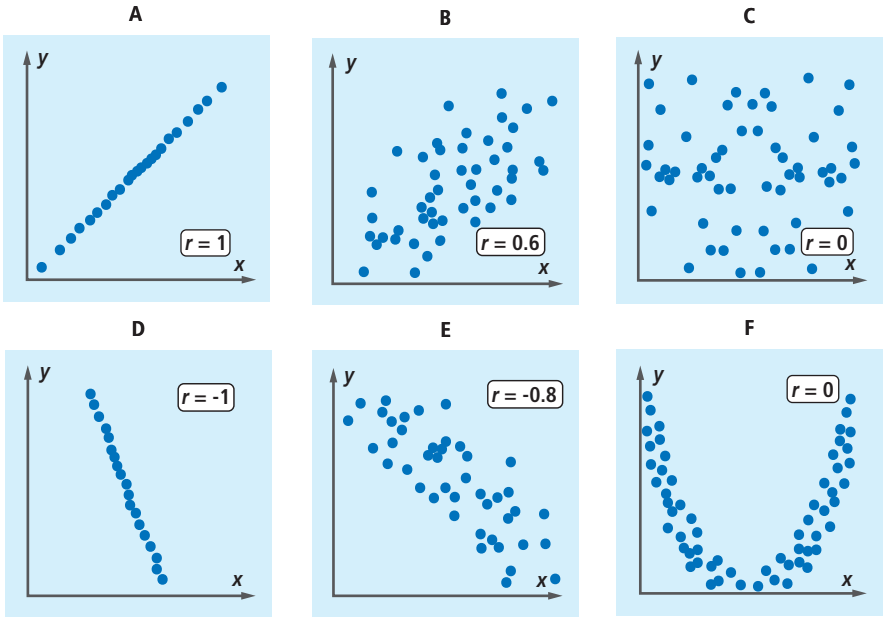
The **correlation coefficient** for a population with n elements is

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right).$$

$\frac{1}{n}$ is replaced by $\frac{1}{n-1}$ for data from a sample.

Unlike the sum of squared residuals, which can be measured for any line modeling a set of data, the correlation coefficient describes fit and direction for the regression line only.

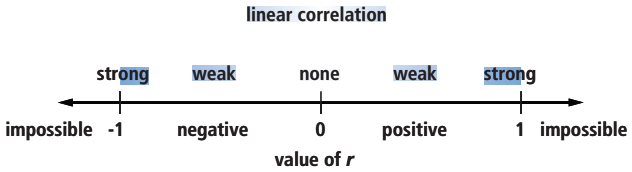
It can be proved that, regardless of the data set, the correlation coefficient r is always a number between –1 and 1. Some data sets and the corresponding values of r are shown in the scatterplots below.



You should use your statistics utility to calculate the value of r. However, you should be able to interpret the value of r in the context of your data. In general, the sign of r indicates the *direction* of the relation between the variables, and its magnitude indicates the *strength* of the relation. Positive values of r indicate a *positive association* between the variables. That is, larger values of one variable are associated with larger values of the other. Negative values of r indicate a *negative association* between the variables. That is, larger values of one variable are associated with smaller values of the other.

The extreme values of 1 and -1 indicate a perfect linear relation, as in scatterplots A and D. That is, all data points lie on a line. Thus, a situation in which $r = \pm 1$ is sometimes called a **perfect correlation**.

A relation for which most of the data fall close to a line (scatterplot E) is called a **strong correlation**. A **weak correlation** is one for which, although a linear trend can be seen, many points are not very close to the line (scatterplot B). A value of r close or equal to 0 (scatterplots C and F) indicates that the variables are not related by a linear model. Note, however, that as indicated in scatterplot F, if $r = 0$, the variables might be strongly related in some other way as denoted by the pattern. A number line below summarizes these relations.



There are no strict rules about how large a correlation must be to be considered strong. In some cases, $|r| = 0.5$ is considered fairly strong, and in others it might be considered moderate or weak.

STOP QY2

Without calculating the correlation coefficient, you can get a sense of its value by looking at the numerical data or at a scatterplot of the data.

Activity

Set 1		Set 2		Set 3		Set 4		Set 5	
x	y	x	y	x	y	x	y	x	y
10	30	4	10	250	3	3	250	–3	9
11	40	8	9	300	9	9	300	–2	4
15	80	13	2	500	11	11	500	0	0
12	50	11	5	750	10	10	750	2	4
14	70	8	4	600	12	12	600	3	9

Step 1 Look at the pattern in each data table and predict what the correlation coefficient might be for that data set.

Step 2 Draw a scatterplot of each data set. Use the scatterplot to predict the correlation coefficient, altering your prediction from Step 1 if necessary.

Step 3 Use a statistics utility to determine the regression line and correlation coefficient for each data set. Compare your prediction from Step 2 to the actual correlation coefficient.

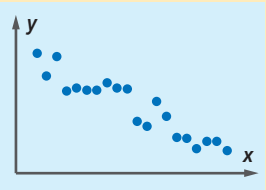
Some statistics utilities give values of r^2 rather than of r. This is because r^2 is used in advanced statistical techniques. You can calculate r by taking the square root of r^2, and then determine the sign of correlation according to the direction in the scatterplot.

STOP QY3

▶ **QY2**

Multiple Choice

In the graph below, which of the values is a reasonable value for r?



A –1 **B** –0.9
C –0.1 **D** 0.5
E 1.25

▶ **QY3**

An r^2 value of 0.6 yields what possible values for r, to the nearest tenth?

Cautions about Correlation

It is important to note that while r provides a mathematical measure of *linearity*, it does not provide information about *cause and effect*. For instance, there is a large positive correlation between shoe size and reading level of children. But this does not mean that learning to read better causes your feet to grow or that wearing bigger shoes improves your reading.

It is up to the people who analyze and interpret the data to determine why two variables might be related. In the case of shoe size and reading level, the correlation is strong because each variable is related to age. Older children generally have both larger feet and higher reading skills than younger children. Similarly, the data in the Activity in Lesson 2-2 give a relationship between TVs and unemployment. The correlation is –0.8, which is strong, but it does not imply that unemployment can be reduced by providing a country with more TVs. This idea is sometimes summarized as *correlation does not imply causation*.

Another caution about correlation and regression: watch out for influential points. Outliers in either the x- or y-coordinate can have a strong impact on the values of slope, y-intercept, and correlation of the least squares line.

Questions

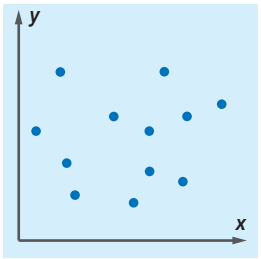
COVERING THE IDEAS

1. Give two other names for the regression line.

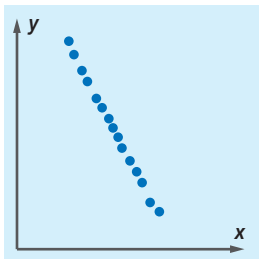
In 2–5, match the scatterplot with the best description.

A strong negative correlation
B weak negative correlation
C strong positive correlation
D almost no correlation

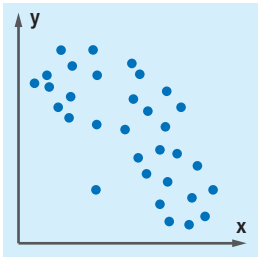
2.



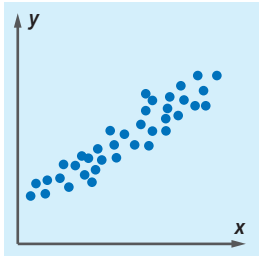
3.



4.



5.

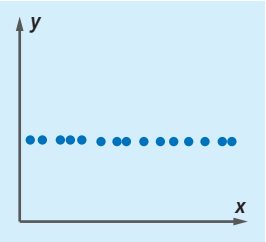


In 6–8, refer to the Activity.

6. Why is the correlation coefficient 1 for Set 1?
7. The x- and y-values in Sets 3 and 4 are swapped. How does this affect r? Does this make sense?
8. The line of best fit for Set 5 has a slope of 0 and r of 0. However, there is a definitive pattern. What other type of model would fit these data?
9. What point must be on the line of best fit?
10. Draw a scatterplot showing perfect positive correlation.
11. Suppose for some data set, $r^2 = 0.4$. Find all possible values of r.
12. **True or False** A negative value of r implies a negative slope for a linear regression line.

APPLYING THE MATHEMATICS

13. Make up a data set in which all the data lie on a single horizontal line, as shown at the right. Calculate the correlation coefficient for your set. Explain the result that you get.



In 14 and 15, state whether you think the correlation coefficient is positive, negative, or almost zero. Explain your answer.

14. number of putts sunk in golf and the distance of ball from hole
15. a person's height and the distance he/she lives from school
16. Heavy metals can enter the food chain when metal-rich discharges from mines contaminate streams, rivers, and lakes. The table shows the lead and zinc contents in milligrams of metal per kilogram of fish (mg/kg) for 10 whole fish (4 rainbow trout, 4 large scale suckers, and 2 mountain whitefish) taken from the Spokane River during July, August, and October of 1999.

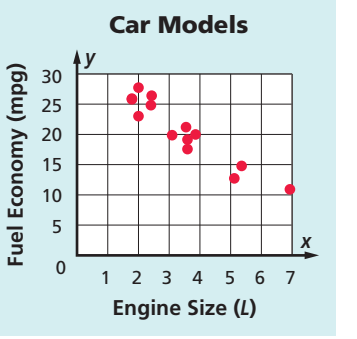
	Rainbow Trout				Large Scale Sucker				Mountain Whitefish	
Lead (mg/kg)	0.73	1.14	0.60	1.59	4.34	1.98	3.12	1.80	0.65	0.56
Zinc (mg/kg)	45.3	50.8	40.2	64.0	150.0	106.0	90.8	58.8	35.4	28.4

Source: Quantitative Environmental Learning Project, Seattle Central Community College

   a. Use a statistics utility to find an equation of the least squares line to predict the amount of zinc from the amount of lead.
   b. Use a statistics utility to find an equation of the least squares line to predict the amount of lead from the amount of zinc.
   c. Are your answers to Parts a and b the same?
   d. Use an appropriate equation to predict the amount of zinc in a fish that has 2 $\frac{\text{mg}}{\text{kg}}$ of lead.

17. A regression line for a set of data is $y = 10x + 4$. If the sum of squared residuals is 0, what is the correlation coefficient?

18. The scatterplot at the right shows the engine size (in liters) of 14 models of cars and their respective fuel economies (in miles per gallon).



Source: United States Department of Energy

a. Is the association positive or negative? Explain your answer.

b. Is the correlation coefficient positive or negative? Explain your answer.

19. There is a 0.8 correlation between the total sales tax collected in Florida each year from 1960–2000 and the numbers of shark attacks in those years. Does this mean the more sales tax, the more shark attacks? Explain why or why not.

20. The data set below gives the greenhouse emissions and fuel economy for five car models.

Car Model	City MPG	Greenhouse Emissions (tons per year)
A	21	8.0
B	11	14.1
C	19	8.7
D	16	9.6
E	24	6.8

a. Use a statistics utility to calculate the regression line and the correlation coefficient. Use city mpg as the independent variable.

b. It can be proven that the slope of the regression line is given by the formula slope $= r \cdot \frac{s_y}{s_x}$, where r is the correlation coefficient, s_y is the standard deviation of the dependent variable n, and s_x is the standard deviation of the independent variable. Confirm that this formula gives you the same answer as you found in Part a.

c. Compute the values of $\bar{x}$ and $\bar{y}$.

d. Show that the point $(\bar{x}, \bar{y})$ lies on the least squares line.

REVIEW

21. A rule of thumb on body measurement is "arm span is equal to height." Selena and Luis used data on 15 adult men to write an equation expressing arm span y as a function of height x. Selena used the rule of thumb. She found the sum of squared residuals for the model $y = x$ was 3050. Luis fit a line on the scatterplot. His line had equation $y = 0.9x + 0.2$. His sum of squared residuals was 4109. Which line had the better fit? **(Lesson 2-2)**

22. Suppose $h(x) = \sqrt{x + 9}$. **(Lesson 2-1)**

a. Find $h(16)$.

b. What is the domain of h?

c. Give the range of h.

23. Suppose $g(t) = t^2 - 6t - 6$. For what value(s) of t does $g(t) = 0$? **(Lesson 2-1)**

24. Let $y = 5x^2 + 2x$. Is y a function of x? Why or why not? **(Previous Course)**

25. ***Skill Sequence*** Rewrite each expression without fractions. **(Previous Course)**

a. $\frac{t}{\frac{1}{5t}}$ **b.** $\frac{4x}{\frac{x}{y}}$ **c.** $\left(\frac{r}{\frac{1}{5}}\right)^2$

EXPLORATION

26. In Parts a–d, use or modify the data as indicated. Then use a statistics utility to compute the regression line and the correlation coefficient for the data. Record your results.

a. Use the unemployment and TV data from Lesson 2-2.

b. Use the data in Part a but replace the data for South Africa with that of Nicaragua: 6.8 TVs per 100 people and 3.9 unemployed per 100 people.

c. Leave Nicaragua in the data set. Replace the Netherlands with Spain, which has 40.2 TVs per 100 people and 13.9 unemployed per 100 people.

d. Change the Spain data to the extreme situation of 80 TVs per 100 people and unemployment of 50 per 100 people. (No country has these statistics.)

e. Write a paragraph summarizing what you have found.

QY ANSWERS

1. $562.59

2. B

3. $r \approx 0.8$ or $r \approx -0.8$

d. The average gestation period increases by about 12.8 days for every increase of one year in an animal's expected life span. e. −55.35 f. about 986 days; extrapolation g. about 211,453 **13.** a. about 10.59 b. Because there are only five data points and the range is 29, the spread is large.

Lesson 2-3 (pp. 094-101)

Questions

1. least squares line, line of best fit **3.** A **5.** C **7.** This does not affect r because swapping the data does not affect its degree of correlation. **9.** $(\bar{x}, \bar{y})$ **11.** $r \approx \pm 0.6325$ **13.** Answers vary. Sample: {(1, 4), (3, 4), (4, 4), (5, 4), (5.5, 4), (6, 4), (9, 4)}. The correlation is undefined. This is because in the equation for computing the correlation coefficient, s_y appears in the denominator, and for the chosen data set $s_y = 0$ and division by 0 is undefined. **15.** zero; There is no correlation between a person's height and how far he/she lives from school. **17.** $r = 1$ **19.** No; both likely correlate with Florida's increasing population, but an increase in the rate of sales tax would not cause an increase in shark attacks. **21.** $y = x$ **23.** $t = 3 \pm \sqrt{15}$ **25.** a. $5t^2$ b. $4y$ c. $25r^2$

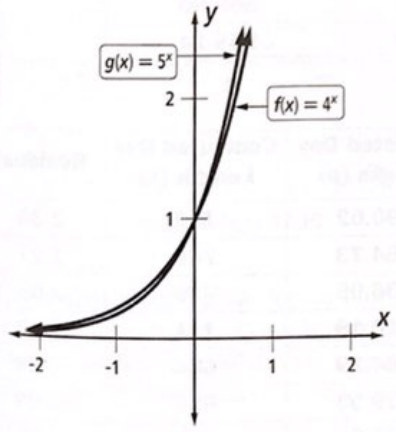
Lesson 2-4 (pp. 102-109)

Guided Example 2 a. 1.053 b. 2500; 1.053; $2500(1.053)^t$ c. 1.053; $3,588.71 d. 1.053; 2254.67; $2,254.67

Guided Example 3 a. all real numbers; positive real numbers b. 8 c. $y = 0$ d. increasing

Questions

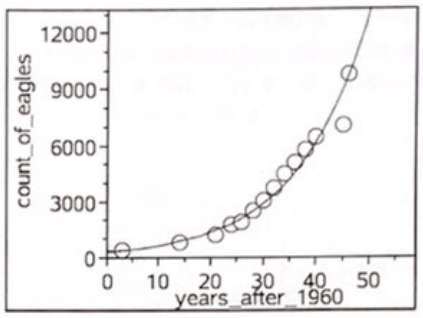
1. a. 4,203,000 b. $P(n) = 4156119(1.0113)^n$ c. 4,756,000 **3.** a. $A = 4{,}000(1.08)^t$ b. about $118,223.89 c. about $2,940.12 **5.** a. false b. decay **7.** a. g b. f c.



9. a. B b. $y = 12$ when $x = 0$ is the initial value. The growth factor is $\frac{1}{2}$ because as x increases by 1, y is divided by 2. **11.** a. 18, 16.2, and 14.53 cubic feet b. $20(0.9)^n$ c. false **13.** False, correlation does not mean causation. For instance, both a and b may be causally related to a third variable c, and thus correlate to each other without being directly related. **15.** a. −1 b. 2 c. False; $f(1) + f(2) = -1 + 0 = -1$, but $f(1 + 2) = f(3) = \frac{3}{5}$ d. False; the domain is $\{x | x \neq -2\}$, while the range is $\{y | y \neq 3\}$ e. $\frac{3p - 15}{p - 1}$ **17.** $r = \pm\frac{20}{27}, s = \pm 3$

Lesson 2-5 (pp. 110-116)

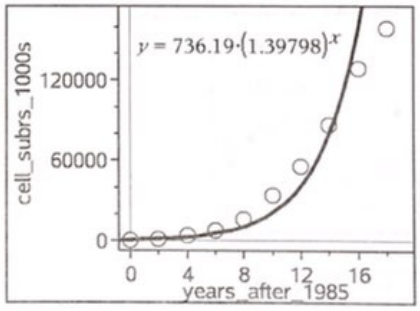
Guided Example 2 a. $296.177(1.079)^x$ b.



c. 296; 1960; 1.079; 7.9 d. 6200; 6471; 6200; 271; 9068; 7066; 9068; −2002

Questions

1. a. $\begin{cases} 20 = ab^3 \\ 156 = ab^{10} \end{cases}$ b. $f(t) = 8.293(1.341)^t$ c. $8.293 \cdot 1.341^3 \approx 20$, $8.293 \cdot 1.341^{10} \approx 156$ **3.** a. $y = 47.979(1.372)^x$ b. about 48 c. 233 **5.** a. $y = 10(0.766)^x$ b. about 0.000001 units **7.** a. $a \approx 736$, b ≈ 1.398; $y = 736(1.398)^t$



b. The model increases much more quickly than the data in the years after 1997, but otherwise fits well. c. 2003

9. a.

Number of Half-Lives	t	$f(t)$
0	0	3
1	1620	1.5
2	3240	0.75
3	4860	0.375

b. $y = 3(0.9995722)^x$ c. about 542 mg **11.** a. g b. f **13.** x-axis **15.** Answers vary. Sample: $y = -3^x$ domain: the set of all real numbers **17.** a. about 30,400 thousand