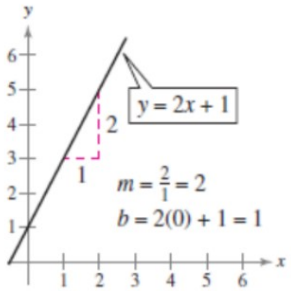
Linear Regression

OBJECTIVES

- How to find the equation of a regression line
- How to predict *y*-values using a regression equation

In algebra, you learned that you can write an equation of a line by finding its slope m and y-intercept b. The equation has the form Recall that the slope of a line is the ratio of its rise over its run and the y-intercept is the y-value of the point at which the line crosses the y-axis. It is the y-value when in algebra, you used two points to determine the equation of a line. In statistics, you will use every point in the data set to determine the equation of the regression line.



DEFINITION

A regression line, also called a **line of best fit,** is the line for which the sum of the squares of the residuals is a minimum.

After verifying that the linear correlation between two variables is significant, the next step is to determine the equation of the line that best models the data. This line is called a *regression line*, and its equation can be used to predict the value of y for a given value of x. Although many lines can be drawn through a set of points, a regression line is determined by specific criteria.

STUDY TIP

When determining the equation of a regression line, it is helpful to construct a scatter plot of the data to check for outliers, which can greatly influence a regression line. You should also check for gaps and clusters in the data.

Linear Regression

REGRESSION LINES

STUDY TIP

Notice that both the slope m and the y-intercept b in Example 1 are rounded to three decimal places. This *round-off rule* will be used throughout the text.

The equation of a regression line allows you to use the independent (explanatory) variable x to make predictions for the dependent (response) variable y.

THE EQUATION OF A REGRESSION LINE

The equation of a regression line for an independent variable x and a dependent variable y is

$$\hat{y} = mx + b$$

where $\hat{y}$ is the predicted y-value for a given x-value. The slope m and y-intercept b are given by

$$m = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} \quad \text{and} \quad b = \bar{y} - m\bar{x} = \frac{\Sigma y}{n} - m\frac{\Sigma x}{n}$$

where $\bar{y}$ is the mean of the y-values in the data set and $\bar{x}$ is the mean of the x-values. The regression line always passes through the point $(\bar{x}, \bar{y})$.

Linear Regression

EXAMPLE 1 Finding the Equation of a Regression Line

GDP (trillions of $), x	CO_2 emissions (millions of metric tons), y
1.6	428.2
3.6	828.8
4.9	1214.2
1.1	444.6
0.9	264.0
2.9	415.3
2.7	571.8
2.3	454.9
1.6	358.7
1.5	573.5

Find the equation of the regression line for the gross domestic products and carbon dioxide emissions data used in Section 9.1.

Solution

In Example 4 of Section 9.1, you found that $n = 10$, $\Sigma x = 23.1$, $\Sigma y = 5554$, $\Sigma xy = 15{,}573.71$, and $\Sigma x^2 = 67.35$. You can use these values to calculate the slope and y-intercept of the regression line as shown.

$$m = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{10(15{,}573.71) - (23.1)(5554)}{10(67.35) - 23.1^2}$$

$$= \frac{27{,}439.7}{139.89} \approx 196.151977$$

$$b = \overline{y} - m\overline{x} \approx \frac{5554}{10} - (196.151977)\frac{23.1}{10}$$

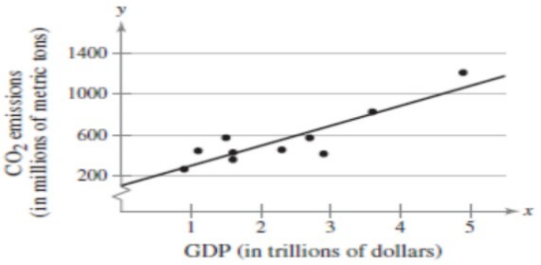
$$= 555.4 - (196.151977)(2.31)$$

$$\approx 102.2889$$

So, the equation of the regression line is

$$\hat{y} = 196.152x + 102.289.$$

To sketch the regression line, use any two x-values within the range of data and calculate their corresponding y-values from the regression line. Then draw a line through the two points. The regression line and scatter plot of the data are shown at the right. If you plot the point $(\overline{x}, \overline{y}) = (2.31, 555.4)$, you will notice that the line passes through this point.



Linear Regression

Try It Yourself 1

Find the equation of the regression line for the number of years out of school and annual contribution data used in Section 9.1.

a. *Identify* n, $\sum x$, $\sum y$, $\sum xy$, and $\sum x^2$ from Try It Yourself 4 in Section 9.1.
b. Calculate the *slope* m and the *y-intercept* b.
c. Write the *equation* of the regression line.

STUDY TIP

Notice that both the slope m and the y-intercept b in Example 1 are rounded to three decimal places. This *round-off rule* will be used throughout the text.

Linear Regression

EXAMPLE 2 Using Technology to Find a Regression Equation

Duration, x	Time, y	Duration, x	Time, y
1.80	56	3.78	79
1.82	58	3.83	85
1.90	62	3.88	80
1.93	56	4.10	89
1.98	57	4.27	90
2.05	57	4.30	89
2.13	60	4.43	89
2.30	57	4.47	86
2.37	61	4.53	89
2.82	73	4.55	86
3.13	76	4.60	92
3.27	77	4.63	91
3.65	77		

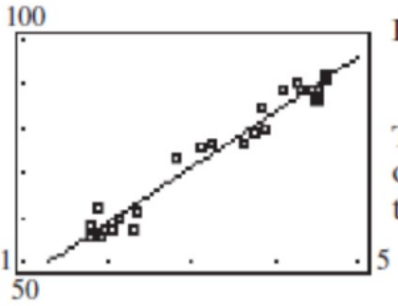
Use a technology tool to find the equation of the regression line for the Old Faithful data used in Section 9.1.

Solution

Geogebra, Excel, and the TI-83/84 Plus each have features that automatically calculate a regression equation. Try using this technology to find the regression equation. You should obtain results similar to the following.

TI-83/84 PLUS

LinReg
y=ax+b
a=12.48094391
b=33.68290034
r^2=.9577738551
r=.9786592129



From the displays, you can see that the regression equation is

$$\hat{y} = 12.481x + 33.683.$$

The TI-83/84 Plus display at the left shows the regression line and a scatter plot of the data in the same viewing window. To do this, use *Stat Plot* to construct the scatter plot and enter the regression equation as y_1.

Linear Regression

Try It Yourself 2

Use a technology tool to find the equation of the regression line for the salaries and average attendances at home games for the teams in Major League Baseball given in the Chapter Opener on page 483.

a. *Enter* the data.
b. Perform the necessary steps to calculate the *slope* and *y-intercept*.
c. Specify the *regression equation*.

STUDY TIP

If the correlation between x and y is not significant, the best predicted y-value is $\bar{y}$, the mean of the y-values in the data set.

Linear Regression

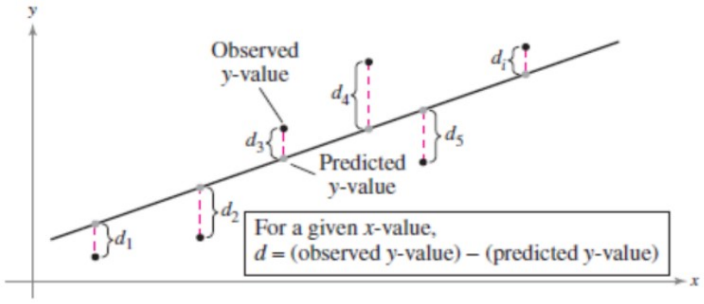
APPLICATIONS OF REGRESSION LINES

After finding the equation of a regression line, you can use the equation to predict y-values over the range of the data *if the correlation between x and y is significant*. For instance, an environmentalist could forecast carbon dioxide emissions on the basis of gross domestic products. To predict y-values, substitute the given x-value into the regression equation, then calculate $\hat{y}$, the predicted y-value.

Consider the scatter plot and the line shown below. For each data point, d_i represents the difference between the observed y-value and the predicted y-value for a given x-value on the line. These differences are called **residuals** and can be positive, negative, or zero. When the point is above the line, d_i is positive. When the point is below the line, d_i is negative. If the observed y-value equals the predicted y-value, $d_i = 0$. Of all possible lines that can be drawn through a set of points, the regression line is the line for which the sum of the squares of all the residuals

$$\Sigma d_i^2$$

is a minimum.



Linear Regression

EXAMPLE 3 Predicting y-Values Using Regression Equations

The regression equation for the gross domestic products (in trillions of dollars) and carbon dioxide emissions (in millions of metric tons) data is

$$\hat{y} = 196.152x + 102.289.$$

Use this equation to predict the *expected* carbon dioxide emissions for the following gross domestic products. (Recall from Section 9.1, Example 7, that x and y have a significant linear correlation.)

1. 1.2 trillion dollars **2.** 2.0 trillion dollars **3.** 2.5 trillion dollars

Solution

To predict the expected carbon dioxide emissions, substitute each gross domestic product for x in the regression equation. Then calculate $\hat{y}$.

1. $\hat{y} = 196.152x + 102.289$
$= 196.152(1.2) + 102.289$
≈ 337.671

Interpretation When the gross domestic product is \$1.2 trillion, the CO_2 emissions are about 337.671 million metric tons.

2. $\hat{y} = 196.152x + 102.289$
$= 196.152(2.0) + 102.289$
$= 494.593$

Interpretation When the gross domestic product is \$2.0 trillion, the CO_2 emissions are 494.593 million metric tons.

3. $\hat{y} = 196.152x + 102.289$
$= 196.152(2.5) + 102.289$
$= 592.669$

Interpretation When the gross domestic product is \$2.5 trillion, the CO_2 emissions are 592.669 million metric tons.

Prediction values are meaningful only for x-values in (or close to) the range of the data. The x-values in the original data set range from 0.9 to 4.9. So, it would not be appropriate to use the regression line $\hat{y} = 196.152x + 102.289$ to predict carbon dioxide emissions for gross domestic products such as \$0.2 or \$14.5 trillion dollars.

Linear Regression

Try It Yourself 3

The regression equation for the Old Faithful data is $\hat{y} = 12.481x + 33.683$. Use this to predict the time until the next eruption for each of the following eruption durations. (Recall from Section 9.1, Example 6, that x and y have a significant linear correlation.)

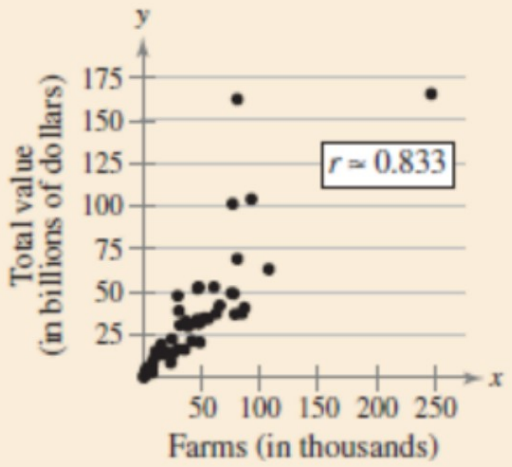
1. 2 minutes
2. 3.32 minutes

a. *Substitute* each value of x into the regression equation.
b. *Calculate* $\hat{y}$.
c. *Specify* the predicted time until the next eruption for each eruption duration.

Linear Regression

PICTURING THE WORLD

The following scatter plot shows the relationship between the number of farms (in thousands) in a state and the total value of the farms (in billions of dollars).
(Source: U.S. Department of Agriculture and National Agriculture Statistics Service)



PICTURING THE WORLD

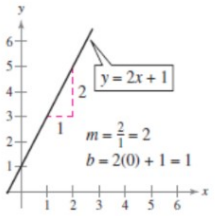
Describe the correlation between these two variables in words. Use the scatter plot to predict the total value of farms in a state that has 150,000 farms. The regression line for this scatter plot is $\hat{y} = 0.714x + 3.367$. Use this equation to make a prediction. (Assume x and y have a significant linear correlation.) How does your algebraic prediction compare with your graphical one?

Linear Regression

OBJECTIVES

- How to find the equation of a regression line
- How to predict *y*-values using a regression equation

In algebra, you learned that you can write an equation of a line by finding its slope m and y-intercept b. The equation has the form Recall that the slope of a line is the ratio of its rise over its run and the y-intercept is the y-value of the point at which the line crosses the y-axis. It is the y-value when in algebra, you used two points to determine the equation of a line. In statistics, you will use every point in the data set to determine the equation of the regression line.



Classwork

Online text, P. 505, #1 - 16

DEFINITION

A **regression line,** also called a **line of best fit,** is the line for which the sum of the squares of the residuals is a minimum.

After verifying that the linear correlation between two variables is significant, the next step is to determine the equation of the line that best models the data. This line is called a *regression line*, and its equation can be used to predict the value of y for a given value of x. Although many lines can be drawn through a set of points, a regression line is determined by specific criteria.

STUDY TIP

When determining the equation of a regression line, it is helpful to construct a scatter plot of the data to check for outliers, which can greatly influence a regression line. You should also check for gaps and clusters in the data.