

3

Descriptive Analysis and Presentation of Bivariate Data



3.3

Linear Regression

Linear Regression

Although the correlation coefficient measures the strength of a linear relationship, it does not tell us about the mathematical relationship between the two variables.

We know that, the correlation coefficient for the push-up/sit-up data was found to be 0.84.

This along with the pattern on the scatter diagram imply that there is a linear relationship between the number of push-ups and the number of sit-ups a student does.

Linear Regression

However, the correlation coefficient does not help us predict the number of sit-ups a person can do based on knowing that he or she can do 28 push-ups.

Regression analysis finds the equation of the line that best describes the relationship between two variables.

One use of this equation is to make predictions. We make use of these predictions regularly—for example, predicting the success a student will have in college based on high school results and predicting the distance required to stop a car based on its speed.

Linear Regression

Generally, the exact value of y is not predictable, and we are usually satisfied if the predictions are reasonably close.

The relationship between two variables will be an algebraic expression describing the mathematical relationship between x and y .

Linear Regression

Here are some examples of various possible relationships, called *models* or **prediction equations**:

Linear (straight-line): $\hat{y} = b_0 + b_1x$

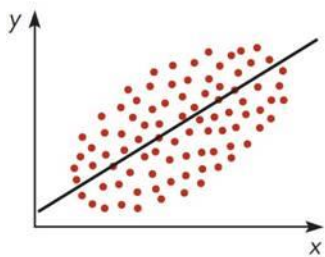
Quadratic: $\hat{y} = a + bx + cx^2$

Exponential: $\hat{y} = a(b^x)$

Logarithmic: $\hat{y} = a \log_b x$

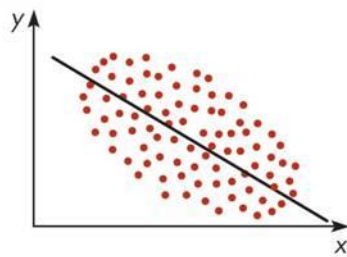
Linear Regression

Figures 3.17, 3.18, and 3.19 show patterns of bivariate data that appear to have a relationship,



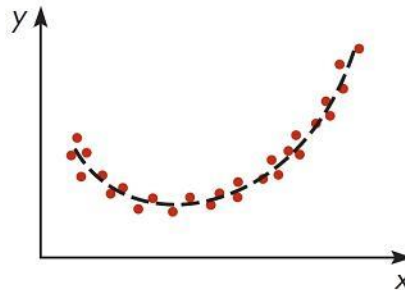
Linear Regression with Positive Slope

Figure 3.17



Linear Regression with Negative Slope

Figure 3.18



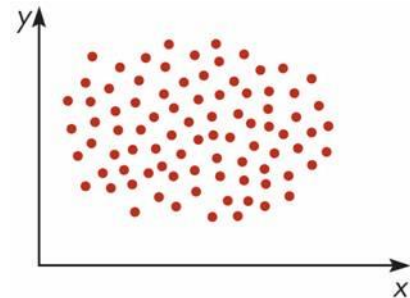
Curvilinear Regression (Quadratic)

Figure 3.19

Linear Regression

whereas in Figure 3.20 the variables do not seem to be related.

If a straight-line model seems appropriate, the best-fitting straight line is found by using the **method of least squares**.



No Relationship
Figure 3.20

Suppose that $\hat{y} = b_0 + b_1x$ is the equation of a straight line, where \hat{y} (read “y-hat”) represents the **predicted value of y** that corresponds to a particular value of x.

Linear Regression

The **least squares criterion** requires that we find the constants b_0 and b_1 such that $\sum (y - \hat{y})^2$ is as small as possible.

Figure 3.21 shows the distance of an observed value of y from a **predicted value of \hat{y}** .

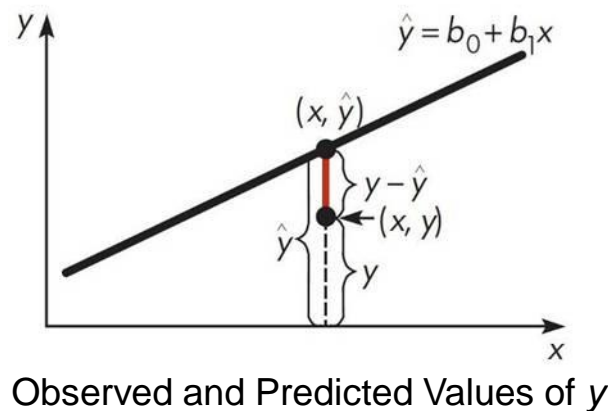


Figure 3.21

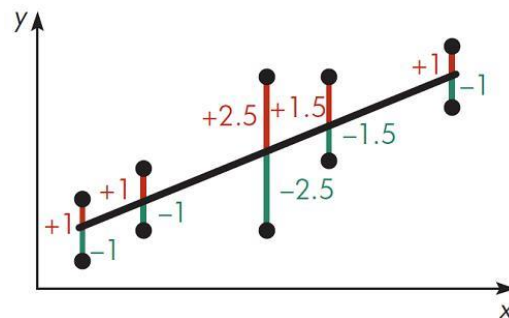
Linear Regression

The length of this distance represents the value $(y - \hat{y})$ (shown as the red line segment in Figure 3.21).

Note that $(y - \hat{y})$ is positive when the point (x, y) is above the line and negative when (x, y) is below the line (as shown on figure).

Linear Regression

Figure 3.22 shows a scatter diagram with what appears to be the **line of best fit**, along with 10 individual $(y - \hat{y})$ values. (Positive values are shown in red; negative, in green.)



$$\sum (y - \hat{y})^2 = (-1)^2 + (+1)^2 + \dots + (+1)^2 = 23.0$$

The Line of Best Fit

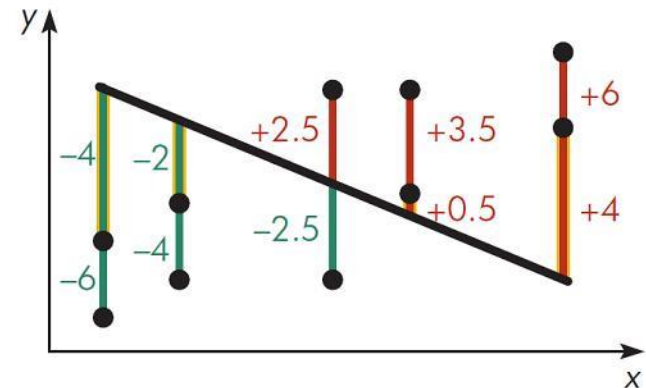
Figure 3.22

The sum of the squares of these differences is minimized (made as small as possible) if the line is indeed the line of best fit.

Linear Regression

Figure 3.23 shows the same data points as Figure 3.22. The 10 individual values $(y - \hat{y})$ are plotted with a line that is definitely not the line of best fit.

[The value $\sum (y - \hat{y})^2$ of is 149, much larger than the 23 from Figure 3.22.]



$$\sum (y - \hat{y})^2 = (-6)^2 + (-4)^2 + \dots + (+6)^2 = 149.0$$

Not the Line of Best Fit

Figure 3.23

Every different line drawn through this set of 10 points will result in a different value for $\sum (y - \hat{y})^2$.

Linear Regression

Our job is to find the one line that will make $\Sigma(y - \hat{y})^2$ the smallest possible value.

The equation of the line of best fit is determined by its **slope** (b_1) and its **y-intercept** (b_0).

The values of the constants—slope and y-intercept—that satisfy the least squares criterion are found by using the formulas presented next:

Definition Formula

$$\text{slope: } b_1 = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2} \quad (3.5)$$

Linear Regression

We will use a mathematical equivalent of formula (3.5) for the slope, b_1 , that uses the sums of squares found in the preliminary calculations for correlation:

Computational Formula

$$\text{slope: } b_1 = \frac{SS(xy)}{SS(x)} \quad (3.6)$$

Notice that the numerator of formula (3.6) is the $SS(xy)$ formula (3.4)

$$SS(xy) = \sum xy - \frac{\sum x \sum y}{n} \quad (3.4)$$

Linear Regression

and the denominator is formula (2.8)

$$SS(x) = \sum x^2 - \frac{(\sum x)^2}{n} \quad (2.8)$$

from the correlation coefficient calculations.

Thus, if you have previously calculated the linear correlation coefficient using the procedure outlined you can easily find the slope of the line of best fit.

Linear Regression

If you did not previously calculate r , set up a table similar to Table 3.12 and complete the necessary preliminary calculations.

Student	Push-ups, x	x^2	Sit-ups, y	y^2	xy
1	27	729	30	900	810
2	22	484	26	676	572
3	15	225	25	625	375
4	35	1,225	42	1,764	1,470
5	30	900	38	1,444	1,140
6	52	2,704	40	1,600	2,080
7	35	1,225	32	1,024	1,120
8	55	3,025	54	2,916	2,970
9	40	1,600	50	2,500	2,000
10	40	1,600	43	1,849	1,720
<hr/>					
$\Sigma x = 351$		$\Sigma x^2 = 13,717$	$\Sigma y = 380$	$\Sigma y^2 = 15,298$	$\Sigma xy = 14,257$
sum of x		sum of x^2	sum of y	sum of y^2	sum of xy

Extensions Table for Finding Five Summations [TA03-10]

Table 3.12

Linear Regression

For the y-intercept, we have:

Computational Formula

$$\begin{aligned} \text{y-intercept} &= \frac{(\text{sum of } y) - [(\text{slope})(\text{sum of } x)]}{\text{number}} \\ b_0 &= \frac{\sum y - (b_1 \cdot \sum x)}{n} \end{aligned} \quad (3.7)$$

Alternative Computational Formula

$$\begin{aligned} \text{y-intercept} &= y\text{-bar} - (\text{slope} \cdot x\text{-bar}) \\ b_0 &= \bar{y} - (b_1 \cdot \bar{x}) \end{aligned} \quad (3.7a)$$

Linear Regression

Now let's consider the following data and the question of predicting a student's number of sit-ups based on the number of push-ups.

Student	1	2	3	4	5	6	7	8	9	10
Push-ups, x	27	22	15	35	30	52	35	55	40	40
Sit-ups, y	30	26	25	42	38	40	32	54	50	43

Data for Push-ups and Sit-ups [TA03-10]

We want to find the line of best fit, $\hat{y} = b_0 + b_1x$.

To calculate the slope, b_1 , using formula (3.6), we know that $SS(xy) = 919.0$ and $SS(x) = 1396.9$.

Linear Regression

Therefore,

$$\text{slope: } b_1 = \frac{SS(xy)}{SS(x)} = \frac{919.0}{1396.9} = 0.6579 = \mathbf{0.66}$$

To calculate the y -intercept, b_0 , using formula (3.7), we know that $\Sigma x = 351$ and $\Sigma y = 380$ from the extensions table. We have

$$\begin{aligned} \text{y-intercept: } b_0 &= \frac{\Sigma y - (b_1 \cdot \Sigma x)}{n} \\ &= \frac{380 - (0.6579)(351)}{10} \end{aligned}$$

Linear Regression

$$= \frac{380 - 230.9229}{10}$$

$$= 14.9077$$

$$= \mathbf{14.9}$$

By placing the two values just found into the model $\hat{y} = b_0 + b_1x$, we get the equation of the line of best fit:

$$\hat{y} = \mathbf{14.9 + 0.66x}$$

Linear Regression

Notes

1. Remember to keep at least three extra decimal places while doing the calculations to ensure an accurate answer.
2. When rounding off the calculated values of b_0 and b_1 , always keep at least two significant digits in the final answer.

Linear Regression

Now that we know the equation for the line of best fit, let's draw the line on the scatter diagram so that we can see the relationship between the line and the data.

We need two points in order to draw the line on the diagram.

Select two convenient x values, one near each extreme of the domain ($x = 10$ and $x = 60$ are good choices for this illustration), and find their corresponding y values.

$$\begin{aligned}\text{For } x = 10: \hat{y} &= 14.9 + 0.66x \\ &= 14.9 + 0.66(10)\end{aligned}$$

Linear Regression

$$= 21.5; \mathbf{(10, 21.5)}$$

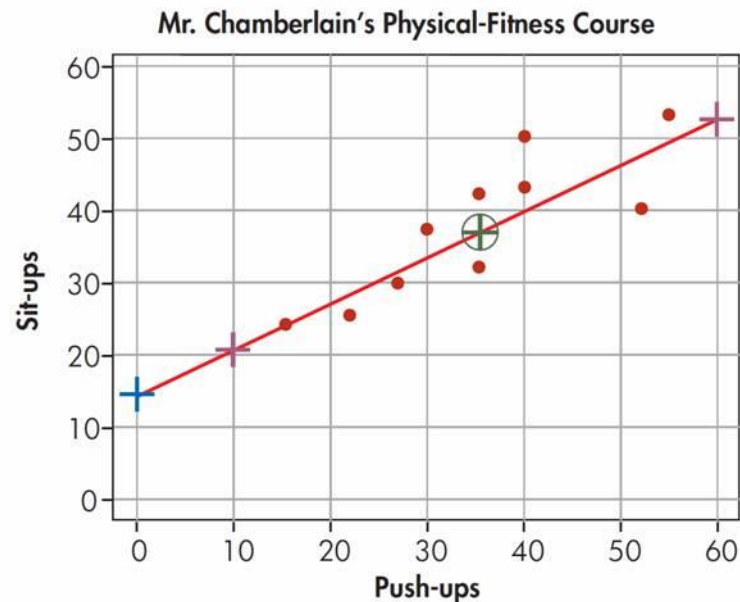
$$\text{For } x = 60: \hat{y} = 14.9 + 0.66x$$

$$= 14.9 + 0.66(60)$$

$$= 54.5; \mathbf{(60, 54.5)}$$

Linear Regression

These two points, (10, 21.5) and (60, 54.5), are then located on the scatter diagram (we use a purple + to distinguish them from data points) and the line of best fit is drawn (shown in red in Figure 3.24).



Line of Best Fit for Push-ups versus Sit-ups

Figure 3.24

Linear Regression

There are some additional facts about the least squares method that we need to discuss.

1. The slope, b_1 , represents the predicted change in y per unit increase in x .

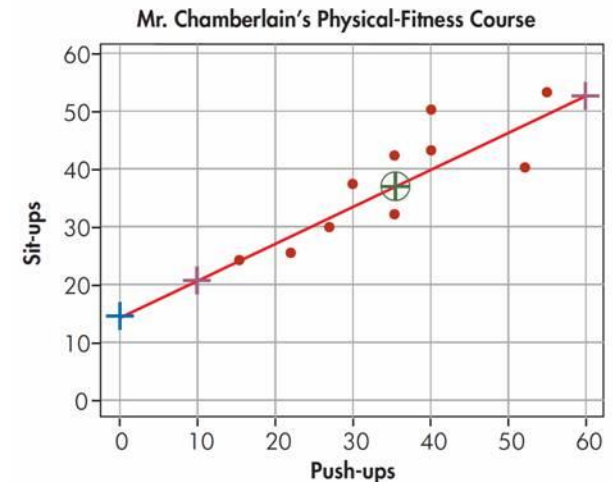
In our example, where $b_1 = 0.66$, if a student can do an additional 10 push-ups (x), we predict that he or she would be able to do approximately 7 (0.66×10) additional sit-ups (y).

Linear Regression

2. The y -intercept is the value of y where the line of best fit intersects the y -axis.

(When the vertical scale is located above $x = 0$, the y -intercept is easily seen on the scatter diagram, shown as a blue $+$ in Figure 3.24.)

First, however, in interpreting b_0 , you must consider whether $x = 0$ is a realistic x value before you can conclude that you would predict if $\hat{y} = b_0$ if $x = 0$.



Line of Best Fit for Push-ups
versus Sit-ups

Figure 3.24

Linear Regression

To predict that if a student did no push-ups, he or she would still do approximately 15 sit-ups ($b_0 = 14.9$) is probably incorrect.

Second, the x value of zero may be outside the domain of the data on which the regression line is based.

In predicting y based on an x value, check to be sure that the x value is within the domain of the x values observed.

Linear Regression

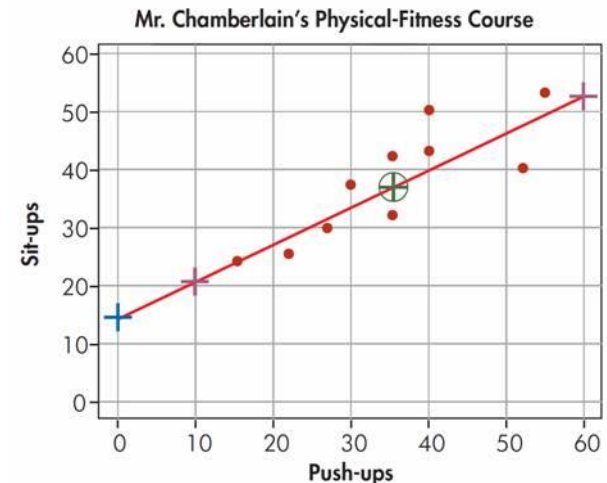
3. The line of best fit will always pass through the *centroid*, the point (\bar{x}, \bar{y}) .

When drawing the line of best fit on your scatter diagram, use this point as a check. For our illustration,

$$\bar{x} = \frac{\sum x}{n} = \frac{351}{10} = 35.1,$$

$$\bar{y} = \frac{\sum y}{n} = \frac{380}{10} = 38.0$$

We see that the line of best fit does pass through $(\bar{x}, \bar{y}) = (35.1, 38.0)$, as shown in green in \oplus Figure 3.24.



Line of Best Fit for Push-ups
versus Sit-ups

Figure 3.24

Example 7 – *Calculating the Line of Best Fit Equation*

In a random sample of eight college women, each woman was asked her height (to the nearest inch) and her weight (to the nearest 5 pounds). The data obtained are shown in Table 3.14.

	1	2	3	4	5	6	7	8
Height, x	65	65	62	67	69	65	61	67
Weight, y	105	125	110	120	140	135	95	130

College Women's Heights and Weights [TA03-14]

Table 3.14

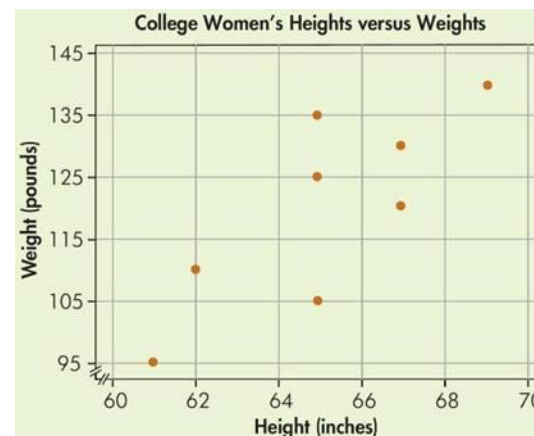
Find an equation to predict the weight of a college woman based on her height (the equation of the line of best fit), and draw it on the scatter diagram in Figure 3.25.

Example 7 – *Solution*

cont'd

Before we start to find the equation for the line of best fit, it is often helpful to draw the scatter diagram, which provides visual insight into the relationship between the two variables.

The scatter diagram for the data on the heights and weights of college women, shown in Figure 3.25, indicates that the linear model is appropriate.



Scatter Diagram
Figure 3.25

Example 7 – Solution

cont'd

To find the equation for the line of best fit, we first need to complete the preliminary calculations, as shown in Table 3.15. The other preliminary calculations include finding $SS(x)$ from formula (2.8) and $SS(xy)$ from formula (3.4):

Student	Height, x	x^2	Weight, y	xy
1	65	4225	105	6825
2	65	4225	125	8125
3	62	3844	110	6820
4	67	4489	120	8040
5	69	4761	140	9660
6	65	4225	135	8775
7	61	3721	95	5795
8	67	4489	130	8710
$\Sigma x = 521$		$\Sigma x^2 = 33,979$	$\Sigma y = 960$	$\Sigma xy = 62,750$

Preliminary calculations needed to find b_1 and b_0

Table 3.15

Example 7 – *Solution*

cont'd

$$\begin{aligned}SS(x) &= \sum x^2 - \frac{(\sum x)^2}{n} \\&= 33,979 - \frac{(521)^2}{8} \\&= 48.875\end{aligned}$$

$$\begin{aligned}SS(xy) &= \sum xy - \frac{\sum x \sum y}{n} \\&= 62,750 - \frac{(521)(960)}{8} \\&= 230.0\end{aligned}$$

Example 7 – *Solution*

cont'd

Second, we need to find the slope and the y -intercept using formulas (3.6) and (3.7):

$$\begin{aligned}\text{slope: } b_1 &= \frac{SS(xy)}{SS(x)} \\ &= \frac{230.0}{48.875} \\ &= 4.706 \\ &= 4.71\end{aligned}$$

Example 7 – *Solution*

cont'd

$$\begin{aligned} \text{y-intercept: } b_0 &= \frac{\Sigma y - (b_1 \cdot \Sigma x)}{n} \\ &= \frac{960 - (4.706)(521)}{8} \\ &= -186.478 \\ &= -186.5 \end{aligned}$$

Thus, the equation of the line of best fit is

$$\hat{y} = -186.5 + 4.71x.$$

Example 7 – *Solution*

cont'd

To draw the line of best fit on the scatter diagram, we need to locate two points. Substitute two values for x —for example, 60 and 70—into the equation for the line of best fit to obtain two corresponding values for \hat{y} :

$$\begin{aligned}\hat{y} &= -186.5 + 4.71x \\ &= -186.5 + (4.71)(60) \\ &= -186.5 + 282.6 \\ &= 96.1 \\ &\approx 96\end{aligned}$$

Example 7 – *Solution*

cont'd

$$\hat{y} = -186.5 + 4.71x$$

$$= -186.5 + (4.71)(70)$$

$$= -186.5 + 329.7$$

$$= 143.2$$

$$\approx 143$$

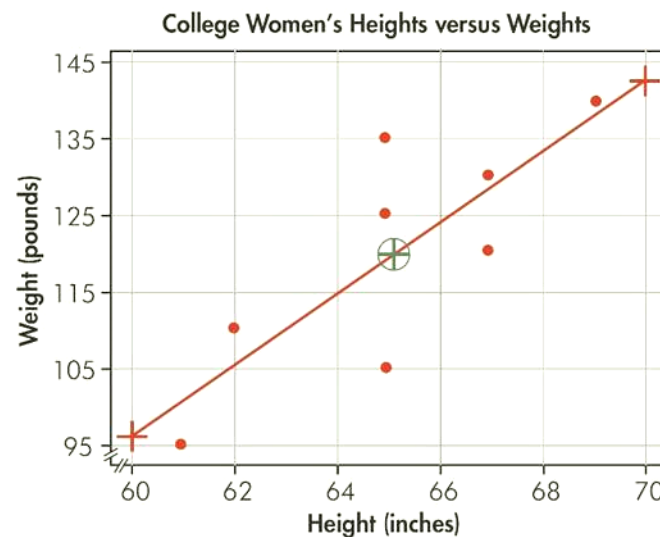
Example 7 – Solution

cont'd

The values (60, 96) and (70, 143) represent two points (designated by a red + in Figure 3.26) that enable us to draw the line of best fit.

Note

In Figure 3.26, $(\bar{x}, \bar{y}) = (65.1, 120)$ is also on the line of best fit. It is the green \oplus .



Scatter Diagram with Line of Best Fit

Figure 3.26



Making Predictions

Making Predictions

One of the main reasons for finding a regression equation is to make predictions. Once a linear relationship has been established and the value of the input variable x is known, we can predict a value of y , \hat{y} .

Consider the equation $\hat{y} = -186.5 + 4.71x$ relating the height and weight of college women.

If a particular female college student is 66 inches tall, what do you predict her weight to be?

Making Predictions

The predicted value is

$$\hat{y} = -186.5 + 4.71x$$

$$= -186.5 + (4.71)(66)$$

$$= -186.5 + 310.86$$

$$= 124.36$$

$$\approx 124 \text{ lb}$$

Making Predictions

You should not expect this predicted value to occur exactly; rather, it is the average weight you would expect for all female college students who are 66 inches tall.

When you make predictions based on the line of best fit, observe the following restrictions:

1. The equation should be used to make predictions only about the population from which the sample was drawn.

For example, using our relationship between the height and the weight of college women to predict the weight of professional athletes given their height would be questionable.

Making Predictions

2. The equation should be used only within the sample domain of the input variable.

We know that the data demonstrate a linear trend within the domain of the x data, but we do not know what the trend is outside this interval.

Hence, predictions can be very dangerous outside the domain of the x data. For instance, in Example 7 it is nonsense to predict that a college woman of height zero will weigh -186.5 pounds.

Making Predictions

Do not use a height outside the sample domain of 61 to 69 inches to predict weight. On occasion you might wish to use the line of best fit to estimate values outside the domain interval of the sample.

This can be done, but you should do it with caution and only for values close to the domain interval.

3. If the sample was taken in 2010, do not expect the results to have been valid in 1929 or to hold in 2020.

The women of today may be different from the women of 1929 and the women in 2020.



Understanding the Line of Best Fit

Understanding the Line of Best Fit

The following method will create

- (1) a visual meaning for the line of best fit,
- (2) a visual meaning for what the line of best fit is describing, and
- (3) an estimate for the slope and y -intercept of the line of best fit.

As with the approximation of r , estimations of the slope and y -intercept of the line of best fit should be used only as a mental estimate or check.

Understanding the Line of Best Fit

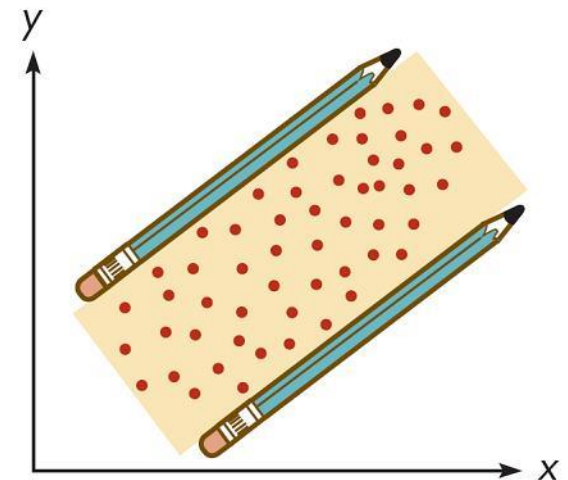
Note

This estimation technique *does not* replace the calculations for b_1 and b_0 .

Procedure

1. On the scatter diagram of the data, draw the straight line that appears to be the line of best fit.

(*Hint:* If you draw a line parallel to and halfway between the two pencils described in Figure 3.13, you will have a reasonable estimate for the line of best fit.)



Focusing on Pattern

Figure 3.13

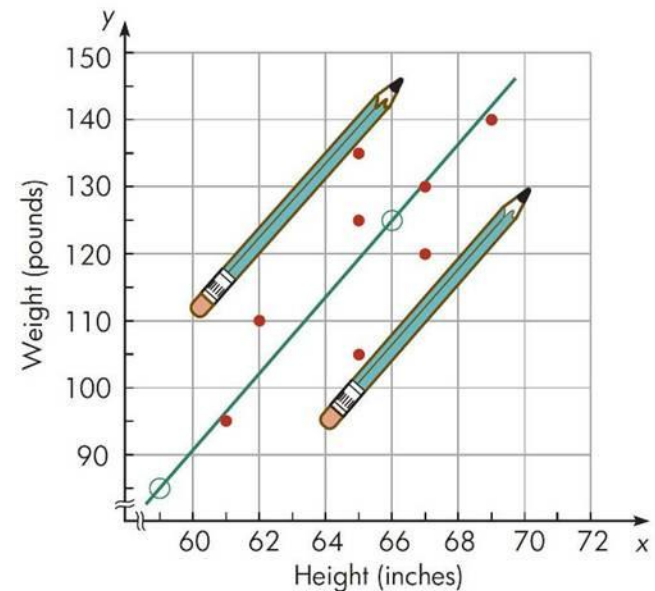
Understanding the Line of Best Fit

The two pencils border the “path” demonstrated by the ordered pairs, and the line down the center of this path approximates the line of best fit.

Figure 3.27 shows the pencils and the resulting estimated line for Example 7.

2. This line can now be used to approximate the equation.

First, locate any two points (x_1, y_1) and (x_2, y_2) along the line and determine their coordinates.



Estimate the Line of Best Fit for the College Women Data

Figure 3.27

Understanding the Line of Best Fit

Two such points, circled in Figure 3.27, have the coordinates (59, 85) and (66, 125).

These two pairs of coordinates can now be used in the following formula to estimate the slope b_1 :

estimate of the slope, b_1 :

$$\begin{aligned} b_1 &\approx \frac{y_2 - y_1}{x_2 - x_1} = \frac{125 - 85}{66 - 59} \\ &= \frac{40}{7} \\ &= 5.7 \end{aligned}$$

Understanding the Line of Best Fit

3. Using this result, the coordinates of one of the points, and the following formula, we can determine an estimate for the y -intercept, b_0 :

estimate of the y -intercept, b_0 :

$$b_0 \approx y - b_1 \cdot x$$

$$x = 85 - (5.7)(59)$$

$$= 85 - 336.3$$

$$= -251.3$$

Understanding the Line of Best Fit

Thus, b_0 is approximately -250 .

4. We now can write the estimated equation for the line of best fit:

$$\hat{y} = -250 + 5.7x$$

This should serve as a crude estimate.

The actual equation calculated using all of the ordered pairs was $\hat{y} = -186.5 + 4.71x$.

Example 8 – *Viewing an Old Faithful Eruption*

Old Faithful very faithfully erupts for a short period of time (1.5 to 5 minutes) periodically throughout every day (every 35 to 120 minutes) and has been doing so since 1870, when such records began being kept; thus its name. It is not the most regular, nor the biggest, but is the biggest regular geyser in Yellowstone.

If your luck is like that of many and you traveled to see one of these famous eruptions, you probably arrived minutes after an eruption had stopped. “When will it erupt again?” and “How long will it last?” are typical questions.

Example 8 – *Viewing an Old Faithful Eruption* cont'd

What you are really asking is, “How long do I have to wait for the next show?” and “Will it be worth the wait?” Since Old Faithful is one of the most studied geysers, the Park Rangers are able to predict the next eruption with reasonable accuracy(± 10 minutes).

They are able to predict only the next eruption, so you had better stick around for it.

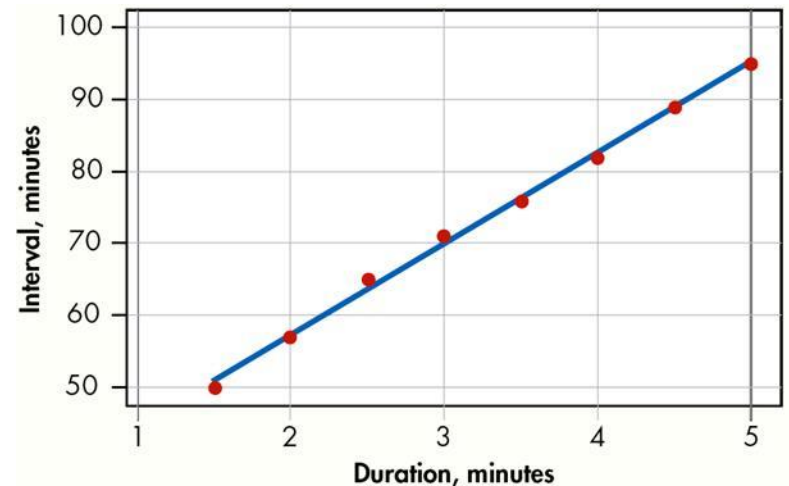
The time until the next eruption, the interval, is predicted based on the length of the previous eruption, the duration. It is not possible to predict the time of occurrence for more than one eruption in advance.

Example 8 – *Viewing an Old Faithful Eruption* cont'd

Here is a chart summarizing the predicted interval based on the previous duration.

By looking at the chart, it appears that the time to the next show interval increases 5 to 7 minutes for each extra half minute of eruption.

The information on the chart can also be seen on the scatter diagram with the line of best fit.



Old Faithful Geyser Interval
 $\text{min} = 32.04 + 12.64 \text{ Duration min}$

Figure 3.28

Example 8 – *Viewing an Old Faithful Eruption* cont'd

The slope for the line of best fit is 12.64, implying that every extra minute of eruption results in an additional 12.6 minutes in wait time for the next eruption, or about 6.3 minutes for every half minute of eruption, as in the information given.

Duration	1.5 min	2.0 min	2.5 min	3.0 min	3.5 min	4.0 min	4.5 min	5.0 min
Interval	50 min	57 min	65 min	71 min	76 min	82 min	89 min	95 min

Table 3.16

The ordered pairs from Table 3.16 and on the scatter diagram, Figure 3.28, are not data values; they are the result of an averaging effect as hundreds of recorded values were summarized.

Example 8 – *Viewing an Old Faithful Eruption* cont'd

Old Faithful data will not result in points being exactly distributed along the line of best fit like those shown in Figure 3.28, they will instead show a substantial amount of variability.

Table 3.17 contains data collected by a visitor over a weekend. It is arranged in sequential order.

Duration, min	1.7	1.9	2.0	2.3	3.1	3.4	3.5	4.0	4.3	4.5	4.7	4.9
Interval, min	55	49	51	53	57	75	80	76	84	76	93	76

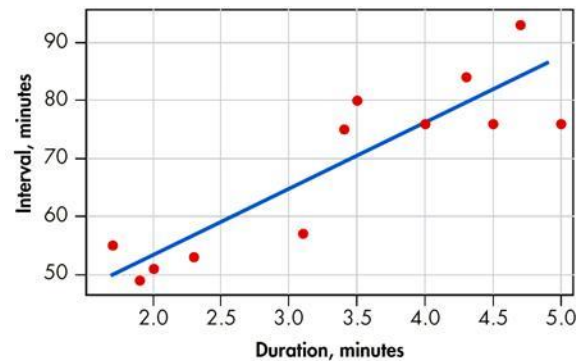
Table 3.17

Example 8 – *Viewing an Old Faithful Eruption* cont'd

The 12 duration and interval times listed in Table 3.17 and shown on Figure 3.29 offer a different impression than the eight points listed in Table 3.16.

Duration	1.5 min	2.0 min	2.5 min	3.0 min	3.5 min	4.0 min	4.5 min	5.0 min
Interval	50 min	57 min	65 min	71 min	76 min	82 min	89 min	95 min

Table 3.16



Old Faithful Geyser Eruption Data
Interval min = $30.33 + 11.44 \text{ Duration min}$

Figure 3.29

Example 8 – *Viewing an Old Faithful Eruption* cont'd

These data look more realistic, with points scattered above and below the line of best fit. A comparison of the two lines of best fit shows very similar results.