with a training set of 10,000 records
ent records need to be resampled if
the balanced data set to be 20%?

ormance, rather than merely citing

difference versus a relative differ-

baseline model take?

CHAPTER 8

SIMPLE LINEAR REGRESSION

Regression modeling represents a powerful and elegant method for estimating the value of a continuous target variable. In this chapter, we introduce regression modeling through simple linear regression, where a straight line is used to approximate the relationship between a single continuous predictor variable and a single continuous response variable. Later, in Chapter 9, we turn to multiple regression, where several predictor variables are used to estimate a single response.

8.1 AN EXAMPLE OF SIMPLE LINEAR REGRESSION

To develop the simple linear regression model, consider the *Cereals* data set,[1] an excerpt of which is presented in Table 8.1. The *Cereals* data set contains nutritional information for 77 breakfast cereals, and includes the following variables:

- Cereal name
- Cereal manufacturer
- Type (hot or cold)
- Calories per serving
- Grams of protein
- Grams of fat
- Milligrams of sodium
- Grams of fiber
- Grams of carbohydrates
- Grams of sugar
- Milligrams of potassium
- Percentage of recommended daily allowance of vitamins (0%, 25%, or 100%)

[1]Cereals data set, in Data and Story Library, http://lib.stat.cmu.edu/DASL. Also available at book web site www.DataMiningConsultant.com.

Data Mining and Predictive Analytics, First Edition. Daniel T. Larose and Chantal D. Larose.

TABLE 8.1 Excerpt from *Cereals* data set: eight fields, first 16 cereals

Cereal Name	Manufacture	Sugars	Calories	Protein	Fat	Sodium	Rating
100% Bran	N	6	70	4	1	130	68.4030
100% Natural Bran	Q	8	120	3	5	15	33.9837
All-Bran	K	5	70	4	1	260	59.4255
All-Bran Extra Fiber	K	0	50	4	0	140	93.7049
Almond Delight	R	8	110	2	2	200	34.3848
Apple Cinnamon Cheerios	G	10	110	2	2	180	29.5095
Apple Jacks	K	14	110	2	0	125	33.1741
	⋮	⋮	⋮	⋮	⋮	⋮	⋮

- Weight of one serving
- Number of cups per serving
- Shelf location (1 = bottom, 2 = middle, 3 = top)
- Nutritional rating, as calculated by Consumer Reports.

We are interested in estimating the nutritional *rating* of a cereal, given its *sugar* content. However, before we begin, it is important to note that this data set contains some missing data. The following four field values are missing:

- Potassium content of Almond Delight
- Potassium content of Cream of Wheat
- Carbohydrates and *sugars* content of Quaker Oatmeal.

We shall therefore not be able to use the sugar content of Quaker Oatmeal to help estimate nutrition rating using sugar content, and only 76 cereals are available for this purpose. Figure 8.1 presents a scatter plot of the nutritional rating versus the sugar content for the 76 cereals, along with the least-squares regression line.

The regression line is written in the form: $\hat{y} = b_0 + b_1 x$, called the *regression equation*, where:

- $\hat{y}$ is the estimated value of the response variable;
- b_0 is the *y-intercept* of the regression line;
- b_1 is the *slope* of the regression line;
- b_0 and b_1, together, are called the *regression coefficients*.

The regression equation for the relationship between sugars (x) and nutritional rating (y) for this sample of cereals is $\hat{y} = 59.853 - 2.4614x$. Below we demonstrate how this equation is calculated. This estimated regression equation can be interpreted as "the estimated cereal rating equals 59.953 minus 2.4614 times the sugar content in grams." The regression line and the regression equation are used as a *linear approximation* of the relationship between the x (predictor) and y (response) variables, that is, between sugar content and nutritional rating. We can then use the regression equation to make estimates or predictions.

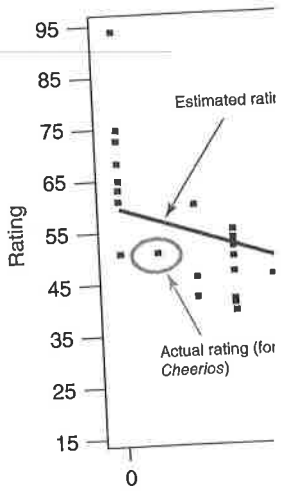


Figure 8.1 Scatter plot of

For example, sup
ing for a new cereal (
Using the regression e
with 1 gram of sugar t
mated value for the nutr
($x = 1, \hat{y} = 57.3916$),
content), the estimated
line.

Now, there is one
Cheerios. Its nutrition
for the new cereal with
($x = 1, y = 50.765$), w
is pointing to a locatic
is where the regressio
sugar content of 1 gra
rating points, which
point to the regress
general ($y - \hat{y}$), is kn
residual.

We of course s
squares regression w
sum of squared resi
choosing the line tha
such as median regre
Note that we say we
variable precedes th

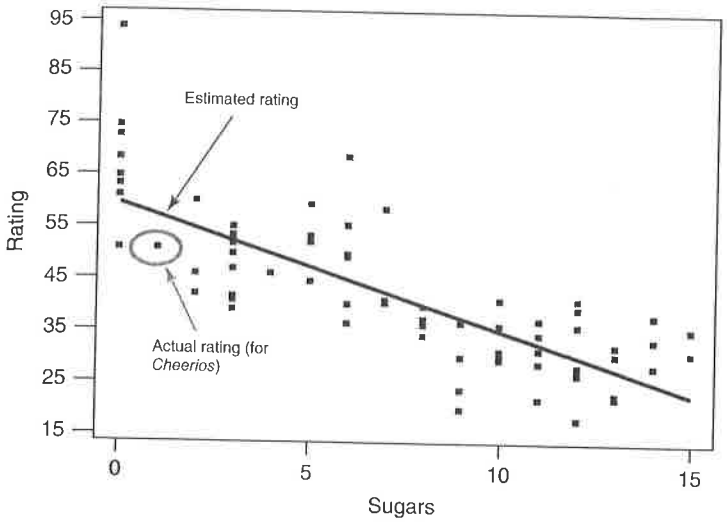


Figure 8.1 Scatter plot of nutritional rating versus sugar content for 77 cereals.

For example, suppose that we are interested in estimating the nutritional rating for a new cereal (not in the original data) that contains $x = 1$ gram of sugar. Using the regression equation, we find the estimated nutritional rating for a cereal with 1 gram of sugar to be $\hat{y} = 59.853 - 2.4614(1) = 57.3916$. Note that this estimated value for the nutritional rating lies directly on the regression line, at the location $(x = 1, \hat{y} = 57.3916)$, as shown in Figure 8.1. In fact, for any given value of x (sugar content), the estimated value for y (nutritional rating) lies precisely on the regression line.

Now, there is one cereal in our data set that does have a sugar content of 1 gram, Cheerios. Its nutrition rating, however, is 50.765, not 57.3916 as we estimated above for the new cereal with 1 gram of sugar. Cheerios' point in the scatter plot is located at $(x = 1, y = 50.765)$, within the oval in Figure 8.1. Now, the upper arrow in Figure 8.1 is pointing to a location on the regression line directly above the Cheerios point. This is where the regression equation predicted the nutrition rating to be for a cereal with a sugar content of 1 gram. The prediction was too high by $57.3916 - 50.765 = 6.6266$ rating points, which represents the vertical distance from the Cheerios data point to the regression line. This vertical distance of 6.6266 rating points, in general $(y - \hat{y})$, is known variously as the *prediction error*, *estimation error*, or *residual*.

We of course seek to minimize the overall size of our prediction errors. *Least squares* regression works by choosing the unique regression line that minimizes the sum of squared residuals over all the data points. There are alternative methods of choosing the line that best approximates the linear relationship between the variables, such as median regression, although least squares remains the most common method. Note that we say we are performing a "regression of *rating* on *sugars*," where the y variable precedes the x variable in the statement.

8.1.1 The Least-Squares Estimates

Now, suppose our data set contained a sample of 76 cereals different from the sample in our *Cereals* data set. Would we expect that the relationship between nutritional rating and sugar content to be exactly the same as that found above: Rating = 59.853 – 2.4614 Sugars? Probably not. Here, b_0 and b_1 are *statistics*, whose values differ from sample to sample. Like other statistics, b_0 and b_1 are used to estimate population parameters, in this case, β_0 and β_1, the y-intercept and slope of the true regression line. That is, the equation

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{8.1}$$

represents the true linear relationship between nutritional rating and sugar content for *all* cereals, not just those in our sample. The *error term* ε is needed to account for the indeterminacy in the model, because two cereals may have the same sugar content but different nutritional ratings. The residuals $(y_i - \hat{y})$ are estimates of the error terms, $\varepsilon_i, i = 1, \ldots, n$. Equation (8.1) is called the regression equation or the true population regression equation; it is associated with the true or population regression line.

Earlier, we found the estimated regression equation for estimating the nutritional rating from sugar content to be $\hat{y} = 59.853 - 2.4614(\text{sugars})$. Where did these values for b_0 and b_1 come from? Let us now derive the formulas for estimating the y-intercept and slope of the estimated regression line, given the data.[2]

Suppose we have n observations from the model in equation (8.1); that is, we have

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

The least-squares line is that line that minimizes the population sum of squared errors, $\text{SSE}_\text{p} = \sum_{i=1}^{n} \varepsilon_i^2$. First, we re-express the population SSEs as

$$\text{SSE}_\text{p} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{8.2}$$

Then, recalling our differential calculus, we may find the values of β_0 and β_1 that minimize $\sum_{i=1}^{n} \varepsilon_i^2$ by differentiating equation (8.2) with respect to β_0 and β_1, and setting the results equal to zero. The partial derivatives of equation (8.2) with respect to β_0 and β_1 are, respectively:

$$\begin{aligned} \frac{\partial \text{SSE}_\text{p}}{\partial \beta_0} &= -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \\ \frac{\partial \text{SSE}_\text{p}}{\partial \beta_1} &= -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) \end{aligned} \tag{8.3}$$

[2]These derivations assume calculus, but those whose calculus is rusty may skip ahead a couple of pages with little loss in understanding.

We are interested in the value
(8.3) equal to zero, we have

Distributing the summation

$$\sum_{i=}^{n}$$

which is re-expressed as

Solving equation (8.4) for b

$$b_1 =$$

$$b_0 =$$

where n is the total numbe
variable and $\bar{y}$ is the mean
$= 1$ to n. The equations in (
β_0 and β_1, the values that m

We now illustrate hov
using equations (8.5), (8.6)
the values for $x_i, y_i, x_i y_i$, a
of the 77 cereals are shown
3234.4309, $\sum x_i y_i = 19,1$

Plugging into formul

$$b_1 = \frac{\sum x_i y_i -}{\sum}$$

We are interested in the values for the estimates b_0 and b_1, so setting the equations in (8.3) equal to zero, we have

$$\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = 0$$
$$\sum_{i=1}^{n} x_i(y_i - b_0 - b_1 x_i) = 0$$

Distributing the summation gives us

$$\sum_{i=1}^{n} y_i - nb_0 - b_1 \sum_{i=1}^{n} x_i = 0$$
$$\sum_{i=1}^{n} x_i y_i - b_0 \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} x_i^2 = 0$$

which is re-expressed as

$$\begin{aligned} b_0 n + b_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i \end{aligned} \tag{8.4}$$

Solving equation (8.4) for b_1 and b_0, we have

$$b_1 = \frac{\sum x_i y_i - \left[\left(\sum x_i\right)\left(\sum y_i\right)\right]/n}{\sum x_i^2 - \left(\sum x_i\right)^2/n} \tag{8.5}$$

$$b_0 = \bar{y} - b_1 \bar{x} \tag{8.6}$$

where n is the total number of observations, $\bar{x}$ is the mean value for the predictor variable and $\bar{y}$ is the mean value for the response variable, and the summations are $i = 1$ to n. The equations in (8.5) and (8.6) are therefore the least squares estimates for β_0 and β_1, the values that minimize the SSEs.

We now illustrate how we may find the values $b_0 = 59.853$ and $b_1 = -2.4614$, using equations (8.5), (8.6), and the summary statistics from Table 8.2, which shows the values for $x_i, y_i, x_i y_i$, and x_i^2, for the *Cereals* in the data set (note that only 16 of the 77 cereals are shown). It turns out that, for this data set, $\sum x_i = 534$, $\sum y_i = 3234.4309$, $\sum x_i y_i = 19,186.7401$, and $\sum x_i^2 = 5190$.

Plugging into formulas (8.5) and (8.6), we find:

$$b_1 = \frac{\sum x_i y_i - \left[\left(\sum x_i\right)\left(\sum y_i\right)\right]/n}{\sum x_i^2 - \left(\sum x_i\right)^2/n}$$

TABLE 8.2 Summary statistics for finding b_0 and b_1

Cereal Name	X = Sugars	Y = Rating	$X*Y$	X^2
100% Bran	6	68.4030	410.418	36
100% Natural Bran	8	33.9837	271.870	64
All-Bran	5	59.4255	297.128	25
All-Bran Extra Fiber	0	93.7049	0.000	0
Almond Delight	8	34.3848	275.078	64
Apple Cinnamon Cheerios	10	29.5095	295.095	100
Apple Jacks	14	33.1741	464.437	196
Basic 4	8	37.0386	296.309	64
Bran Chex	6	49.1203	294.722	36
Bran Flakes	5	53.3138	266.569	25
Cap'n Crunch	12	18.0429	216.515	144
Cheerios	1	50.7650	50.765	1
Cinnamon Toast Crunch	9	19.8236	178.412	81
Clusters	7	40.4002	282.801	49
Cocoa Puffs	13	22.7364	295.573	169
⋮	⋮	⋮		
Wheaties Honey Gold	8	36.1876	289.501	64
	$\sum x_i = 534$ $\bar{x} = 534/76$ $= 7.0263$	$\sum y_i = 3234.4309$ $\bar{y} = 3234.4309/76$ $= 42.5583$	$\sum x_i y_i$ $= 19,186.7401$	$\sum x_i^2 = 5190$

$$= \frac{19,186.7401 - (534)(3234.4309)/76}{5190 - (534)^2/76} = \frac{-3539.3928}{1437.9474}$$
$$= -2.4614 \quad (8.7)$$

and

$$b_0 = \bar{y} - b_1\bar{x} = 42.5583 + 2.4614(7.0263) = 59.853 \quad (8.8)$$

These values for the slope and y-intercept provide us with the estimated regression line indicated in Figure 8.1.

The y-intercept b_0 is the location on the y-axis where the regression line intercepts the y-axis; that is, the estimated value for the response variable when the predictor variable equals zero. The interpretation of the value of the y-intercept b_0 is as the estimated value of y, given $x = 0$. For example, for the *Cereals* data set, the y-intercept $b_0 = 59.853$ represents the estimated nutritional rating for cereals with zero sugar content. Now, in many regression situations, a value of zero for the predictor variable would not make sense. For example, suppose we were trying to predict elementary school students' weight (y) based on the students' height (x). The meaning of *height* = 0 is unclear, so that the denotative meaning of the y-intercept would not make interpretive sense in this case. However, for our data set, a value

of zero for the sugar co
of sugar.

The slope of the r
increase in x. We interpr
of 1 gram in sugar conte
ing points." For exampl
would have an estimatec
than Cereal B.

8.2 DANGERS O

Suppose that a new cerea
the comic strip characte
very high sugar content
equation to estimate the

$$\hat{y} = 59.853 -$$

In other words, C
is actually a negative nu
mum = 18) and analogou
going on here? The negat
Bombs is an example of

Analysts should c
sion equation to values
values of x in the data s
content is 0 grams and t
ing for any value of x (su
However, *extrapolation*,
be dangerous, because v
response and predictor v

Extrapolation shou
range of x must be perfor
no x-data is available to
that the relationship betv
the data set, may no long

Consider Figure 8.
in black but that the tru
(observed) and the gray
on the available (black d
line indicated. Suppose
x-value located at the tria
be represented by the dot
this prediction has failed
huge prediction error. Of
hidden data, he or she w

$X*Y$	X^2
410.418	36
271.870	64
297.128	25
0.000	0
275.078	64
295.095	100
464.437	196
296.309	64
294.722	36
266.569	25
216.515	144
50.765	1
178.412	81
282.801	49
295.573	169
289.501	64
$\sum x_i y_i$	$\sum x_i^2 = 5190$
9,186.7401	

$$\frac{-3539.3928}{1437.9474} \qquad (8.7)$$

$= 59.853$ (8.8)

the estimated regression

the regression line inter-
e variable when the pre-
of the y-intercept b_0 is
or the *Cereals* data set,
tional rating for cereals
s, a value of zero for the
ppose we were trying to
students' height (x). The
aning of the y-intercept
for our data set, a value

of zero for the sugar content does make sense, as several cereals contain 0 grams of sugar.

The slope of the regression line indicates the estimated change in y per unit increase in x. We interpret $b_1 = -2.4614$ to mean the following: "For each increase of 1 gram in sugar content, the estimated nutritional rating *decreases* by 2.4614 rating points." For example, Cereal A with five more grams of sugar than Cereal B would have an estimated nutritional rating $5(2.4614) = 12.307$ ratings points lower than Cereal B.

8.2 DANGERS OF EXTRAPOLATION

Suppose that a new cereal (say, the Chocolate Frosted Sugar Bombs loved by Calvin, the comic strip character written by Bill Watterson) arrives on the market with a very high sugar content of 30 grams per serving. Let us use our estimated regression equation to estimate the nutritional rating for Chocolate Frosted Sugar Bombs:

$$\hat{y} = 59.853 - 2.4614(\text{sugars}) = 59.4 - 2.4614(30) = -13.989.$$

In other words, Calvin's cereal has so much sugar that its nutritional rating is actually a negative number, unlike any of the other cereals in the data set (minimum = 18) and analogous to a student receiving a negative grade on an exam. What is going on here? The negative estimated nutritional rating for Chocolate Frosted Sugar Bombs is an example of the dangers of *extrapolation*.

Analysts should confine the estimates and predictions made using the regression equation to values of the predictor variable contained within the range of the values of x in the data set. For example, in the *Cereals* data set, the lowest sugar content is 0 grams and the highest is 15 grams, so that predictions of nutritional rating for any value of x (sugar content) between 0 and 15 grams would be appropriate. However, *extrapolation*, making predictions for x-values lying outside this range, can be dangerous, because we do not know the nature of the relationship between the response and predictor variables outside this range.

Extrapolation should be avoided if possible. If predictions outside the given range of x must be performed, the end-user of the prediction needs to be informed that no x-data is available to support such a prediction. The danger lies in the possibility that the relationship between x and y, which may be linear within the range of x in the data set, may no longer be linear outside these bounds.

Consider Figure 8.2. Suppose that our data set consisted only of the data points in black but that the true relationship between x and y consisted of both the black (observed) and the gray (unobserved) points. Then, a regression line based solely on the available (black dot) data would look approximately similar to the regression line indicated. Suppose that we were interested in predicting the value of y for an x-value located at the triangle. The prediction based on the available data would then be represented by the dot on the regression line indicated by the upper arrow. Clearly, this prediction has failed spectacularly, as shown by the vertical line indicating the huge prediction error. Of course, as the analyst would be completely unaware of the hidden data, he or she would hence be oblivious to the massive scope of the error

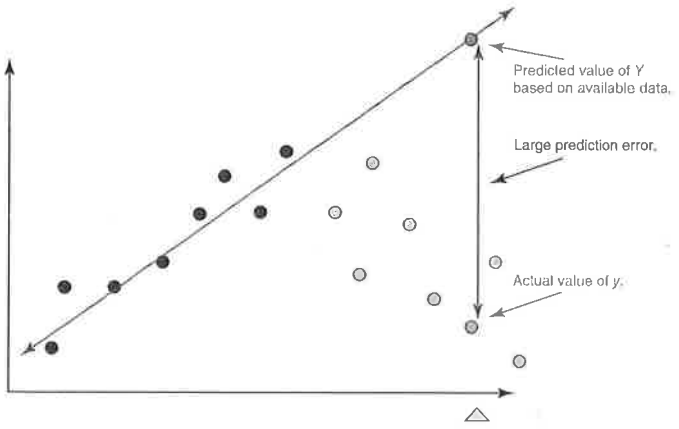


Figure 8.2 Dangers of extrapolation.

in prediction. Policy recommendations based on such erroneous predictions could certainly have costly results.

EXTRAPOLATION

- *Extrapolation* refers to estimates and predictions of the target variable made using the regression equation with values of the predictor variable outside of the range of the values of x in the data set.
- The analyst does not know the shape of the relationship between x and y in areas beyond the range of x. It may no longer be linear.
- Extrapolation should be avoided. If unable to avoid extrapolation, inform the end-user of the analysis that no x-data is available to support such a prediction.

8.3 HOW USEFUL IS THE REGRESSION? THE COEFFICIENT OF DETERMINATION, r^2

Of course, a least-squares regression line could be found to approximate the relationship between any two continuous variables, regardless of the quality of the relationship between them, but this does not guarantee that the regression will therefore be useful. The question therefore arises as to how we may determine whether a particular estimated regression equation is useful for making predictions.

We shall work toward developing a statistic, r^2, for measuring the goodness of fit of the regression. That is, r^2, known as the *coefficient of determination*, measures how well the linear approximation produced by the least-squares regression line actually fits the observed data.

Recall that $\hat{y}$ repres
$(y - \hat{y})$ represents the *pred*
which shows the distanc
competitors, along with t
traveled 10 kilometers in
regression takes the form
kilometers plus twice the
this estimated regression
(8.7) and (8.8).

This estimated regr
distance traveled for a gi
in the Predicted Score co
tion error may then be c
sum of squares error, SS
prediction resulting from
SSE = 12. Is this value l
is large, because at this p

TABLE 8.3 Calculation of

Subject	X = Time	Y = Di
1	2	1
2	2	1
3	3	1
4	4	1
5	4	1
6	5	1
7	6	2
8	7	1
9	8	2
10	9	2

Now, imagine for
traveled *without knowle*
have access to the x-var
our estimates of the dis
information usually res

Because we lack
simply $\bar{y}$, the sample me
$\bar{y} = 16$ to estimate the r
of the number of hours

Consider Figure
information is shown b

Recall that $\hat{y}$ represents the estimated value of the response variable, and that $(y - \hat{y})$ represents the *prediction error* or *residual*. Consider the data set in Table 8.3, which shows the distance in kilometers traveled by a sample of 10 orienteering competitors, along with the elapsed time in hours. For example, the first competitor traveled 10 kilometers in 2 hours. On the basis of these 10 competitors, the estimated regression takes the form $\hat{y} = 6 + 2x$, so that the estimated distance traveled equals 6 kilometers plus twice the number of hours. You should verify that you can calculate this estimated regression equation, either using software, or using the equations in (8.7) and (8.8).

This estimated regression equation can be used to make predictions about the distance traveled for a given number of hours. These estimated values of y are given in the Predicted Score column in Table 8.3. The prediction error and squared prediction error may then be calculated. The sum of the squared prediction errors, or the sum of squares error, $\text{SSE} = \sum(y - \hat{y})^2$, represents an overall measure of the error in prediction resulting from the use of the estimated regression equation. Here we have $\text{SSE} = 12$. Is this value large? We are unable to state whether this value, $\text{SSE} = 12$, is large, because at this point we have no other measure to compare it to.

TABLE 8.3 Calculation of the SSE for the orienteering example

Subject	X = Time	Y = Distance	Predicted Score $\hat{y} = 6 + 2x$	Error in Prediction $(y - \hat{y})$	(Error in Prediction)^2 $(y - \hat{y})^2$
1	2	10	10	0	0
2	2	11	10	1	1
3	3	12	12	0	0
4	4	13	14	−1	1
5	4	14	14	0	0
6	5	15	16	−1	1
7	6	20	18	2	4
8	7	18	20	−2	4
9	8	22	22	0	0
10	9	25	24	1	1
					$\text{SSE} = \sum(y - \hat{y})^2 = 12$

Now, imagine for a moment that we were interested in estimating the distance traveled *without knowledge of the number of hours*. That is, suppose that we did not have access to the x-variable information for use in estimating the y-variable. Clearly, our estimates of the distance traveled would be degraded, on the whole, because less information usually results in less accurate estimates.

Because we lack access to the predictor information, our best estimate for y is simply $\bar{y}$, the sample mean of the number of hours traveled. We would be forced to use $\bar{y} = 16$ to estimate the number of kilometers traveled for every competitor, regardless of the number of hours that person had traveled.

Consider Figure 8.3. The estimates for distance traveled when ignoring the time information is shown by the horizontal line $\bar{y} = 16$ Disregarding the time information

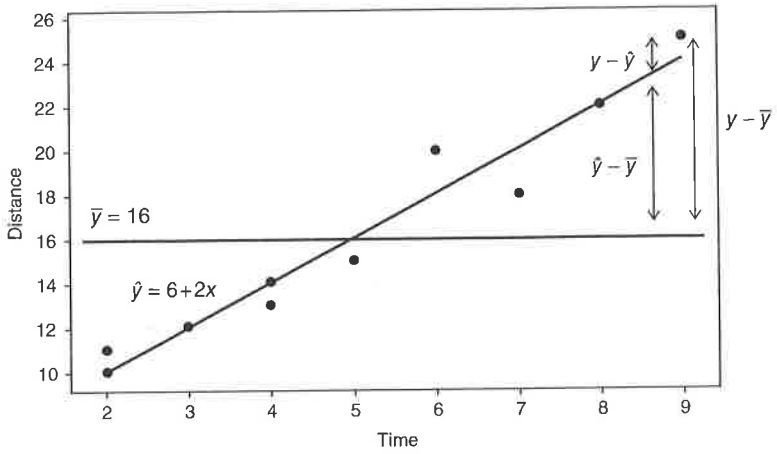


Figure 8.3 Overall, the regression line has smaller prediction error than the sample mean.

entails predicting $\bar{y} = 16$ kilometers for the distance traveled, for orienteering competitors who have been hiking only 2 or 3 hours, as well as for those who have been out all day (8 or 9 hours). This is clearly not optimal.

The data points in Figure 8.3 seem to "cluster" tighter around the estimated regression line than around the line $\bar{y} = 16$, which suggests that, overall, the prediction errors are smaller when we use the x-information than otherwise. For example, consider competitor #10, who hiked $y = 25$ kilometers in $x = 9$ hours. If we ignore the x-information, then the estimation error would be $(y - \bar{y}) = 25 - 16 = 9$ kilometers. This prediction error is indicated as the vertical line between the data point for this competitor and the horizontal line; that is, the vertical distance between the observed y and the predicted $\bar{y} = 16$.

Suppose that we proceeded to find $(y - \bar{y})$ for every record in the data set, and then found the sum of squares of these measures, just as we did for $(y - \hat{y})$ when we calculated the SSE. This would lead us to SST, the *sum of squares total*:

$$\text{SST} = \sum_{i=1}^{n} (y - \bar{y})^2$$

SST, also known as the *total sum of squares*, is a measure of the total variability in the values of the response variable alone, without reference to the predictor. Note that SST is a function of the *sample variance* of y, where the variance is the square of the standard deviation of y:

$$\text{SST} = \sum_{i=1}^{n} (y - \bar{y})^2 = (n-1)s_y^2 = (n-1)(s_y)^2$$

Thus, all three of these measures—SST, variance, and standard deviation—are univariate measures of the variability in y alone (although of course we could find the variance and standard deviation of the predictor as well).

Would we expect SST to
tions shown in Table 8.4, we h
We now have something to co
SST, this indicates that using
much tighter estimates overall
of squares measure errors in pre
the regression improves our est

TABLE 8.4 Finding SST for the or

Student	X = Time
1	2
2	2
3	3
4	4
5	4
6	5
7	6
8	7
9	8
10	9

Next, what we would li
equation improves the estima
estimation error when using
mation error when ignoring th
the amount of *improvement* (

Once again, we may pr
$(\hat{y} - \bar{y})$. Such a statistic is knc
the overall improvement in pr
to ignoring the predictor info

Observe from Figure 8.2 that
"pieces," $(\hat{y} - \bar{y})$ and $(y - \hat{y})$.

Now, suppose we square eac

$$\sum (y_i -$$

[3]The cross-product term $2 \cdot \sum(\hat{y}_i$
Regression Analysis, 3rd edition, V

Would we expect SST to be larger or smaller than SSE? Using the calculations shown in Table 8.4, we have SST = 228, which is much larger than SSE = 12. We now have something to compare SSE against. As SSE is so much smaller than SST, this indicates that using the predictor information in the regression results in much tighter estimates overall than ignoring the predictor information. These sums of squares measure errors in prediction, so that smaller is better. In other words, using the regression improves our estimates of the distance traveled.

TABLE 8.4 Finding SST for the orienteering example

Student	X = Time	Y = Distance	$\bar{y}$	$(y-\bar{y})$	$(y-\bar{y})^2$
1	2	10	16	−6	36
2	2	11	16	−5	25
3	3	12	16	−4	16
4	4	13	16	−3	9
5	4	14	16	−2	4
6	5	15	16	−1	1
7	6	20	16	4	16
8	7	18	16	2	4
9	8	22	16	6	36
10	9	25	16	9	81
					$\text{SST} = \sum(y-\bar{y})^2 = 228$

Next, what we would like is a measure of how much the estimated regression equation improves the estimates. Once again examine Figure 8.3. For hiker #10, the estimation error when using the regression is $(y-\hat{y}) = 25-24 = 1$, while the estimation error when ignoring the time information is $(y-\bar{y}) = 25-16 = 9$. Therefore, the amount of *improvement* (reduction in estimation error) is $(\hat{y}-\bar{y}) = 24-16 = 8$.

Once again, we may proceed to construct a sum of squares statistic based on $(\hat{y}-\bar{y})$. Such a statistic is known as *SSR*, the *sum of squares regression*, a measure of the overall improvement in prediction accuracy when using the regression as opposed to ignoring the predictor information.

$$\text{SSR} = \sum_{i=1}^{n} (\hat{y}-\bar{y})^2$$

Observe from Figure 8.2 that the vertical distance $(y-\bar{y})$ may be partitioned into two "pieces," $(\hat{y}-\bar{y})$ and $(y-\hat{y})$. This follows from the following identity:

$$(y-\bar{y}) = (\hat{y}-\bar{y}) + (y-\hat{y}) \tag{8.9}$$

Now, suppose we square each side, and take the summation. We then obtain[3]:

$$\sum(y_i-\bar{y})^2 = \sum(\hat{y}_i-\bar{y})^2 + \sum(y_i-\hat{y}_i)^2 \tag{8.10}$$

[3]The cross-product term $2\cdot\sum(\hat{y}_i-\bar{y})(y_i-\hat{y}_i)$ cancels out. For details, see Draper and Smith, *Applied Regression Analysis*, 3rd edition, Wiley Publishers, Hoboken, New Jersey, 1998.

We recognize from equation (8.8) the three sums of squares we have been developing, and can therefore express the relationship among them as follows:

$$\text{SST} = \text{SSR} + \text{SSE} \tag{8.11}$$

We have seen that SST measures the total variability in the response variable. We may then think of SSR as the amount of variability in the response variable that is "explained" by the regression. In other words, SSR measures that portion of the variability in the response variable that is accounted for by the linear relationship between the response and the predictor.

However, as not all the data points lie precisely on the regression line, this means that there remains some variability in the y-variable that is not accounted for by the regression. SSE can be thought of as measuring all the variability in y from all sources, including random error, after the linear relationship between x and y has been accounted for by the regression.

Earlier, we found SST = 228 and SSE = 12. Then, using equation (8.11), we can find SSR to be SSR = SST − SSE = 228 − 12 = 216. Of course, these sums of squares must always be nonnegative. We are now ready to introduce the *coefficient of determination*, r^2, which measures the goodness of fit of the regression as an approximation of the linear relationship between the predictor and response variables.

$$r^2 = \frac{\text{SSR}}{\text{SST}}$$

As r^2 takes the form of a ratio of SSR to SST, we may interpret r^2 to represent the proportion of the variability in the y-variable that is explained by the regression; that is, by the linear relationship between the predictor and response variables.

What is the maximum value that r^2 can take? The maximum value for r^2 would occur when the regression is a perfect fit to the data set, which takes place when each of the data points lies precisely on the estimated regression line. In this optimal situation, there would be no estimation errors from using the regression, meaning that each of the residuals would equal zero, which in turn would mean that SSE would equal zero. From equation (8.11), we have that SST = SSR + SSE. If SSE = 0, then SST = SSR, so that r^2 would equal SSR/SST = 1. Thus, the maximum value for r^2 is 1, which occurs when the regression is a perfect fit.

What is the minimum value that r^2 can take? Suppose that the regression showed no improvement at all, that is, suppose that the regression explained none of the variability in y. This would result in SSR equaling zero, and consequently, r^2 would equal zero as well. Thus, r^2 is bounded between 0 and 1, inclusive.

How are we to interpret the value that r^2 takes? Essentially, the higher the value of r^2, the better the fit of the regression to the data set. Values of r^2 near one denote an extremely good fit of the regression to the data, while values near zero denote an extremely poor fit. In the physical sciences, one encounters relationships that elicit very high values of r^2, while in the social sciences, one may need to be content with lower values of r^2, because of person-to-person variability. As usual, the analyst's judgment should be tempered with the domain expert's experience.

8.4 STANDARD ER

We have seen how the r^2 sta
data set. Next, the s statistic,
of the accuracy of the estin
most important statistics to
the value of s, we first find

where m indicates the num
ear regression case, and gr
Like SSE, MSE represents
unexplained by the regress
Then, *the standard e*

The value of s provides an
value of the standard devia
of the typical deviation. I
mation, the typical differe
response value. In this wa
sion of the predictions gen
of s are better, and s has t
variable y.
For the orienteering

s :

Thus, the typical estimatic
is 1.2 kilometers. That is,
estimate of the distance cc
1.2 kilometers. Note fron
absolute value, so that 1.
residual. (Other measures
also be considered, but ar
We may compare
obtained from ignoring t
the response,

8.4 STANDARD ERROR OF THE ESTIMATE, s

We have seen how the r^2 statistic measures the goodness of fit of the regression to the data set. Next, the s statistic, known as the *standard error of the estimate*, is a measure of the accuracy of the estimates produced by the regression. Clearly, s is one of the most important statistics to consider when performing a regression analysis. To find the value of s, we first find the *mean square error* (MSE):

$$\text{MSE} = \frac{\text{SSE}}{(n - m - 1)}$$

where m indicates the number of predictor variables, which is 1 for the simple linear regression case, and greater than 1 for the multiple regression case (Chapter 9). Like SSE, MSE represents a measure of the variability in the response variable left unexplained by the regression.

Then, *the standard error of the estimate* is given by

$$s = \sqrt{\text{MSE}} = \sqrt{\frac{\text{SSE}}{(n - m - 1)}}$$

The value of s provides an estimate of the size of the "typical" residual, much as the value of the standard deviation in univariate analysis provides an estimate of the size of the typical deviation. In other words, s is a measure of the typical error in estimation, the typical difference between the predicted response value and the actual response value. In this way, the standard error of the estimate s represents the precision of the predictions generated by the estimated regression equation. Smaller values of s are better, and s has the benefit of being expressed in the units of the response variable y.

For the orienteering example, we have

$$s = \sqrt{\text{MSE}} = \sqrt{\frac{12}{(10 - 1 - 1)}} = 1.2$$

Thus, the typical estimation error when using the regression model to predict distance is 1.2 kilometers. That is, if we are told how long a hiker has been traveling, then our estimate of the distance covered will typically differ from the actual distance by about 1.2 kilometers. Note from Table 8.3 that all of the residuals lie between 0 and 2 in absolute value, so that 1.2 may be considered a reasonable estimate of the typical residual. (Other measures, such as the mean absolute deviation of the residuals, may also be considered, but are not widely reported in commercial software packages.)

We may compare $s = 1.2$ kilometers against the typical estimation error obtained from ignoring the predictor data, obtained from the standard deviation of the response,

$$\sigma_y = \sqrt{\frac{\sum_{i=1}^{n}(y - \bar{y})^2}{n - 1}} = 5.0$$

The typical prediction error when ignoring the time data is 5 kilometers. Using the regression has reduced the typical prediction error from 5 to 1.2 kilometers.

In the absence of software, one may use the following computational formulas for calculating the values of SST and SSR. The formula for SSR is exactly the same as for the slope b_1, except that the numerator is squared.

$$\text{SST} = \sum y^2 - \frac{\left(\sum y\right)^2}{n}$$

$$\text{SSR} = \frac{\left[\sum xy - \left(\sum x\right)\left(\sum y\right)/n\right]^2}{\sum x^2 - \left(\sum x\right)^2/n}$$

Let us use these formulas for finding the values of SST and SSR for the orienteering example. You should verify that we have $\sum x = 50$, $\sum y = 160$, $\sum xy = 908$, $\sum x^2 = 304$, and $\sum y^2 = 2788$.

Then, $\text{SST} = \sum y^2 - \left(\sum y\right)^2/n = 2788 - (160)^2/10 = 2478 - 2560 = 228.$

And, $\text{SSR} = \dfrac{\left[\sum xy - \left(\sum x\right)\left(\sum y\right)/n\right]^2}{\sum x^2 - \left(\sum x\right)^2/n} = \dfrac{[908 - (50)(160)/10]^2}{304 - (50)^2/10} = \dfrac{108^2}{54} = 216.$

Of course, these are the same values found earlier using the more onerous tabular method. Finally, we calculate the value of the coefficient of determination r^2 to be

$$r^2 = \frac{\text{SSR}}{\text{SST}} = \frac{216}{228} = 0.9474$$

In other words, the linear relationship between time and distance accounts for 94.74% of the variability in the distances traveled. The regression model fits the data very nicely.

8.5 CORRELATION COEFFICIENT *r*

A common measure used to quantify the linear relationship between two quantitative variables is the *correlation coefficient*. The correlation coefficient r (also known as the *Pearson product moment correlation coefficient*) is an indication of the strength of the linear relationship between two quantitative variables, and is defined as follows:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{(n-1)s_x s_y}$$

where s_x and s_y represent the sample standard deviations of the x and y data values, respectively.

INTERPRETING CORRELA

- When x and y are *positiv* tends to increase as well.
- When x and y are *negati* tends to decrease.
- When x and y are *uncorr* remain unaffected.

The correlation coefficie
sive. Values of r close to 1 indi
of r close to -1 indicate that
of the large sample sizes ass
small in absolute value may b
a relatively modest-sized data
$r = 0.07$ would be considered
how to construct a confidence
the correlation coefficient r.

The definition formula
because the numerator woulc
x-data and the y-data. We the
following computational forn

$$r = \frac{}{\sqrt{\sum x}}$$

For the orienteering example

$$r = \frac{}{\sqrt{\sum x}}$$

$$= \frac{}{\sqrt{304}}$$

$$= \frac{108}{\sqrt{54}\sqrt{}}$$

We would say that the time
positively correlated. As the
to increase.

INTERPRETING CORRELATIONS

- When x and y are *positively correlated*, as the value of x increases, the value of y tends to increase as well.
- When x and y are *negatively correlated*, as the value of x increases, the value of y tends to decrease.
- When x and y are *uncorrelated*, as the value of x increases, the value of y tends to remain unaffected.

The correlation coefficient r always takes on values between 1 and -1, inclusive. Values of r close to 1 indicate that x and y are *positively correlated*, while values of r close to -1 indicate that x and y are *negatively correlated*. However, because of the large sample sizes associated with data mining, even values of r relatively small in absolute value may be considered statistically significant. For example, for a relatively modest-sized data set of about 1000 records, a correlation coefficient of $r = 0.07$ would be considered statistically significant. Later in this chapter, we learn how to construct a confidence interval for determining the statistical significance of the correlation coefficient r.

The definition formula for the correlation coefficient above may be tedious, because the numerator would require the calculation of the deviations for both the x-data and the y-data. We therefore have recourse, in the absence of software, to the following computational formula for r:

$$r = \frac{\sum xy - \left(\sum x\right)\left(\sum y\right)/n}{\sqrt{\sum x^2 - \left(\sum x\right)^2/n}\sqrt{\sum y^2 - \left(\sum y\right)^2/n}}$$

For the orienteering example, we have

$$r = \frac{\sum xy - \left(\sum x\right)\left(\sum y\right)/n}{\sqrt{\sum x^2 - \left(\sum x\right)^2/n}\sqrt{\sum y^2 - \left(\sum y\right)^2/n}}$$

$$= \frac{908 - (50)(160)/10}{\sqrt{304 - (50)^2/10}\sqrt{2788 - (160)^2/10}}$$

$$= \frac{108}{\sqrt{54}\sqrt{228}} = 0.9733$$

We would say that the time spent traveling and the distance traveled are strongly positively correlated. As the time spent hiking increases, the distance traveled tends to increase.

However, it is more convenient to express the correlation coefficient r as $r = \pm\sqrt{r^2}$. When the slope b_1 of the estimated regression line is positive, then the correlation coefficient is also positive, $r = \sqrt{r^2}$; when the slope is negative, then the correlation coefficient is also negative, $r = -\sqrt{r^2}$ In the orienteering example, we have $b_1 = 2$. This is positive, which means that the correlation coefficient will also be positive, $r = \sqrt{r^2} = \sqrt{0.9474} = 0.9733$.

It should be stressed here that the correlation coefficient r measures only the *linear* correlation between x and y. The predictor and target may be related in a curvilinear manner, for example, and r may not uncover the relationship.

8.6 ANOVA TABLE FOR SIMPLE LINEAR REGRESSION

Regression statistics may be succinctly presented in an analysis of variance (ANOVA) table, the general form of which is shown here in Table 8.5. Here, m represents the number of predictor variables, so that, for simple linear regression, $m = 1$.

TABLE 8.5 The ANOVA table for simple linear regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Regression	SSR	m	$\text{MSR} = \frac{\text{SSR}}{m}$	$F = \frac{\text{MSR}}{\text{MSE}}$
Error (or residual)	SSE	$n - m - 1$	$\text{MSE} = \frac{\text{SSE}}{n - m - 1}$	
Total	SST = SSR + SSE	$n - 1$		

The ANOVA table conveniently displays the relationships among several statistics, showing, for example, that the sums of squares add up to SST. The *mean squares* are presented as the ratios of the items to their left, and, for inference, the test statistic F is represented as the ratio of the mean squares. Tables 8.6 and 8.7 show the Minitab regression results, including the ANOVA tables, for the orienteering example and the cereal example, respectively.

8.7 OUTLIERS, HIGH LEVERAGE POINTS, AND INFLUENTIAL OBSERVATIONS

Next, we discuss the role of three types of observations that may or may not exert undue influence on the regression results. These are as follows:

- Outliers
- High leverage points
- Influential observations.

An *outlier* is an observation that has a very large standardized residual in absolute value. Consider the scatter plot of nutritional rating against sugars in Figure 8.4.

TABLE 8.6 Results for regressi orienteering example

```
The regression equatic
Distance = 6.00 + 2.0C

Predictor      Coef   SE
Constant     6.0000   0.
Time         2.0000   0.

S = 1.22474    R-Sq =

Analysis of Variance

Source           DF
Regression        1  2
Residual Error    8
Total             9  2
```

TABLE 8.7 Results for regres

```
The regression equat
Rating = 59.9 - 2.4(

Predictor      Coef
Constant     59.853
Sugars      -2.4614

S = 9.16616    R-Sq

Analysis of Varianc

Source           DF
Regression        1
Residual Error   74
Total            75

Unusual Observatior

Obs   Sugars   Ratinc
  1      6.0   68.4(
  4      0.0   93.7(

R denotes an obser
```

ation coefficient r as $r =$
is positive, then the cor-
ope is negative, then the
rienteering example, we
ation coefficient will also

ient r measures only the
may be related in a curvi-
tionship.

EGRESSION

sis of variance (ANOVA)
5. Here, m represents the
ression, $m = 1$.

Mean Square	F
$R = \frac{SSR}{m}$	$F = \frac{MSR}{MSE}$
$E = \frac{SSE}{n-m-1}$	

hips among several statis-
o SST. The *mean squares*
nference, the test statistic
and 8.7 show the Minitab
nteering example and the

AND

at may or may not exert
ows:

dardized residual in abso-
ainst sugars in Figure 8.4.

TABLE 8.6 Results for regression of *distance* versus *time* for the orienteering example

```
The regression equation is
Distance = 6.00 + 2.00 Time


Predictor      Coef  SE Coef       T       P
Constant     6.0000   0.9189    6.53   0.000
Time         2.0000   0.1667   12.00   0.000


S = 1.22474    R-Sq = 94.7%    R-Sq(adj) = 94.1%


Analysis of Variance

Source           DF      SS      MS       F       P
Regression        1  216.00  216.00  144.00   0.000
Residual Error    8   12.00    1.50
Total             9  228.00
```

TABLE 8.7 Results for regression of *nutritional rating* versus *sugar content*

```
The regression equation is
Rating = 59.9 - 2.46 Sugars


Predictor      Coef  SE Coef       T       P
Constant     59.853    1.998   29.96   0.000
Sugars      -2.4614   0.2417  -10.18   0.000


S = 9.16616    R-Sq = 58.4%    R-Sq(adj) = 57.8%


Analysis of Variance

Source           DF      SS      MS       F       P
Regression        1  8711.9  8711.9   103.69   0.000
Residual Error   74  6217.4    84.0
Total            75 14929.3


Unusual Observations

Obs  Sugars  Rating     Fit  SE Fit  Residual  St Resid
  1     6.0   68.40   45.08    1.08     23.32      2.56R
  4     0.0   93.70   59.85    2.00     33.85      3.78R

R denotes an observation with a large standardized residual.
```

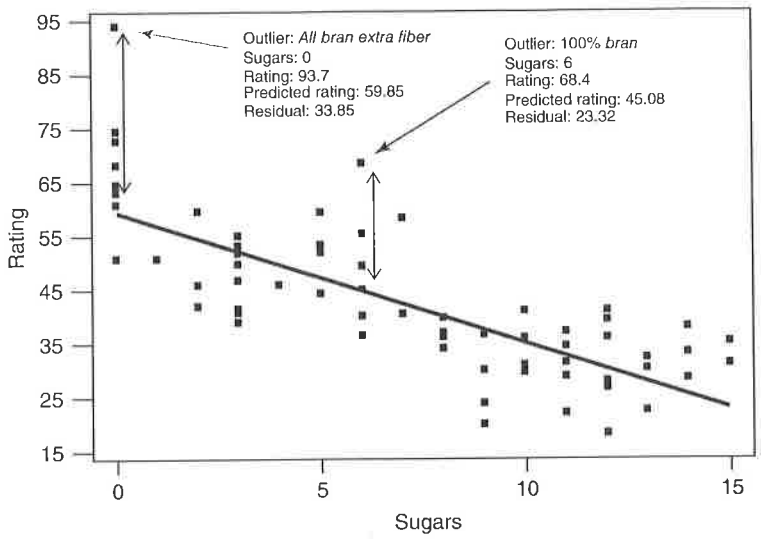


Figure 8.4 Identifying the outliers in regression of *nutritional rating* versus *sugars*.

The two observations with the largest absolute residuals are identified as *All Bran Extra Fiber* and *100% Bran*. Note that the vertical distance away from the regression line (indicated by the vertical arrows) is greater for these two observations than for any other cereals, indicating the largest residuals.

For example, the nutritional rating for *All Bran Extra Fiber* (93.7) is much higher than predicted (59.85), based on its sugar content alone (0 grams). Similarly, the nutritional rating for *100% Bran* (68.4) is much higher than would have been estimated (45.08) based on its sugar content alone (6 grams).

Residuals may have different variances, so that it is preferable to use the standardized residuals in order to identify outliers. Standardized residuals are residuals divided by their standard error, so that they are all on the same scale. Let $s_{i,\text{resid}}$ denote the standard error of the ith residual. Then

$$s_{i,\text{resid}} = s\sqrt{1 - h_i}$$

where h_i refers to the *leverage* of the ith observation (see below).

And the standardized residual equals:

$$\text{residual}_{i,\text{standardized}} = \frac{y_i - \hat{y}_i}{s_{i,\text{resid}}}$$

A rough rule of thumb is to flag observations whose standardized residuals exceed 2 in absolute value as being outliers. For example, note from Table 8.7 that *Minitab* identifies observations 1 and 4 as outliers based on their large standardized residuals; these are *All Bran Extra Fiber* and *100% Bran*.

In general, if the residual is *positive*, we may say that the observed y-value is *higher* than the regression estimated, given the x-value. If the residual is *negative*, we may say that the observed y-value is *lower* than the regression estimated,

given the x-value. For exa
ual), we would say that th
estimated, given its sugar
fiber.)

A *high leverage poi*
In other words, a high lev
without reference to the
x-variables, and ignores th
concept of the lever, whic
it were long enough.

The leverage h_i for

For a given data set, the
constants, so that the lev
the squared distance bet
predictor. The farther the
the x-space, the greater th
the upper bound is 1.0. A
$3(m + 1)/n$ may be cons
of predictors).

For example, in th
vation, a real hard-core c
39 kilometers. Figure 8.

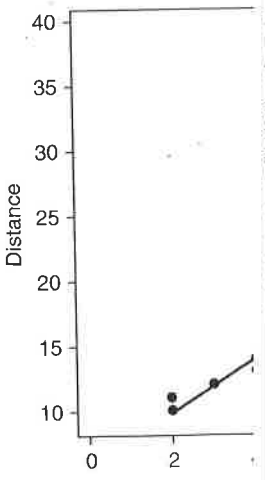


Figure 8.5 Scatter plot o

given the x-value. For example, for *All Bran Extra Fiber* (which has a positive residual), we would say that the observed nutritional rating is higher than the regression estimated, given its sugars value. (This may presumably be because of all that extra fiber.)

A *high leverage point* is an observation that is extreme in the predictor space. In other words, a high leverage point takes on extreme values for the x-variable(s), without reference to the y-variable. That is, leverage takes into account only the x-variables, and ignores the y-variable. The term *leverage* is derived from the physics concept of the lever, which Archimedes asserted could move the Earth itself if only it were long enough.

The leverage h_i for the ith observation may be denoted as follows:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}$$

For a given data set, the quantities $1/n$ and $\sum (x_i - \bar{x})^2$ may be considered to be constants, so that the leverage for the ith observation depends solely on $(x_i - \bar{x})^2$, the squared distance between the value of the predictor and the mean value of the predictor. The farther the observation differs from the mean of the observations, in the x-space, the greater the leverage. The lower bound on leverage values is $1/n$, and the upper bound is 1.0. An observation with leverage greater than about $2(m + 1)/n$ or $3(m + 1)/n$ may be considered to have high leverage (where m indicates the number of predictors).

For example, in the orienteering example, suppose that there was a new observation, a real hard-core orienteering competitor, who hiked for 16 hours and traveled 39 kilometers. Figure 8.5 shows the scatter plot, updated with this 11th hiker.

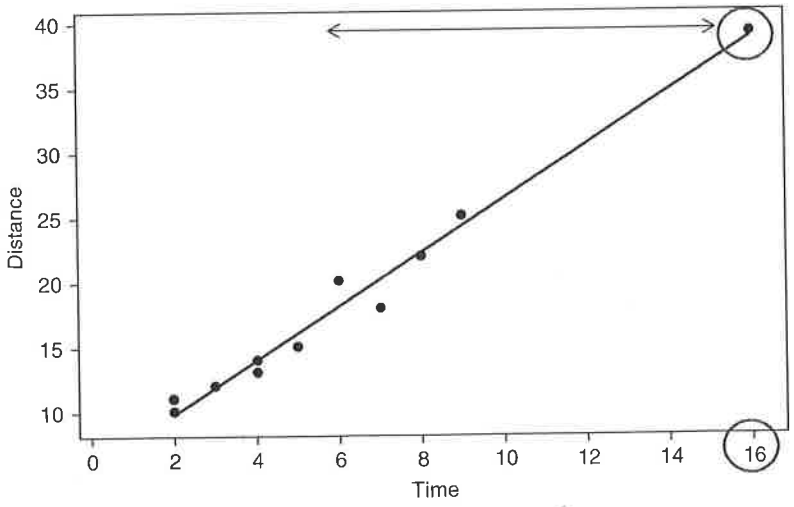
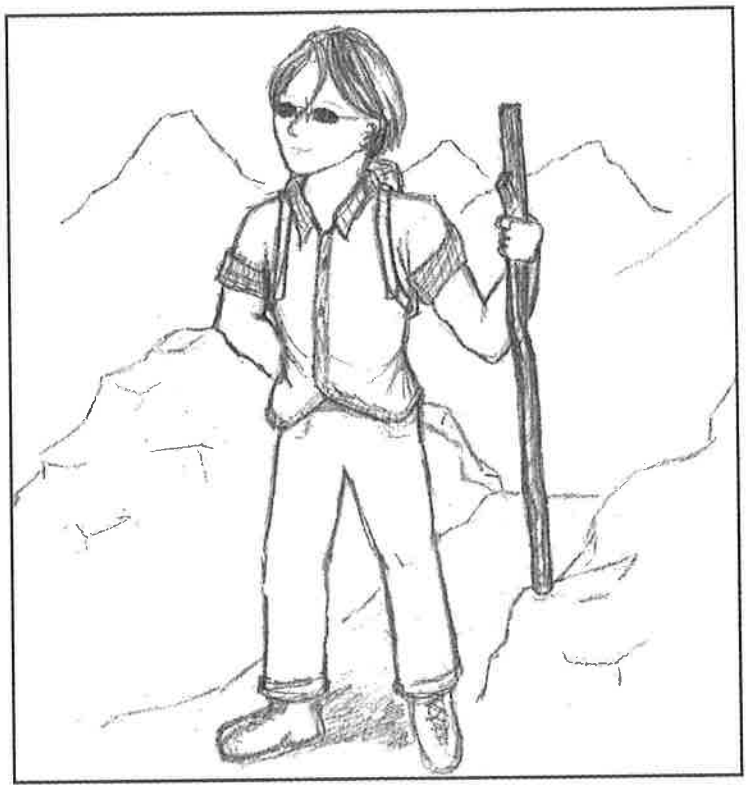


Figure 8.5 Scatter plot of distance versus time, with new competitor who hiked for 16 hours.



The Hard-Core Orienteer hiked 39 kilometers in 16 hours. Does he represent an outlier or a high-leverage point?

Note from Figure 8.5 that the time traveled by the new hiker (16 hours) is extreme in the x-space, as indicated by the horizontal arrows. This is sufficient to identify this observation as a high leverage point, without reference to how many kilometers he or she actually traveled. Examine Table 8.8, which shows the updated regression results for the 11 hikers. Note that *Minitab* correctly points out that the extreme orienteer does indeed represent an unusual observation, because its x-value gives it large leverage. That is, Minitab has identified the hard-core orienteer as a high leverage point, because he hiked for 16 hours. It correctly did not consider the distance (y-value) when considering leverage.

However, the hard-core orienteer is not an outlier. Note from Figure 8.5 that the data point for the hard-core orienteer lies quite close to the regression line, meaning that his distance of 39 kilometers is close to what the regression equation would have predicted, given the 16 hours of hiking. Table 8.8 tells us that the standardized residual is only $\text{residual}_{i,\text{standardized}} = 0.47$, which is less than 2, and therefore not an outlier.

Next, we consider what it means to be an influential observation. In the context of history, what does it mean to be an influential person? A person is influential if

TABLE 8.8 Updated reg

```
The regression equ
Distance = 5.73 +

Predictor      Coef
Constant     5.725
Time         2.0609

S = 1.16901    R-S

Analysis of Varia

Source           D
Regression
Residual Error
Total            1

Unusual Observati

Obs  Time  Distar
 11  16.0    39.(

X denotes an obse

Predicted Values

New Obs      Fit
      1   18.091
```

their presence or abs
of Bedford Falls (fr
(played by James S
shows him how diff
born. Similarly, in r
eters alter significar
data set.

An outlier ma
or may not be influ
acteristics of large r
not-quite flagged as
be influential throug

First let us co
influential. Suppose
with someone who l
the regression result

TABLE 8.8 Updated regression results, including the hard-core hiker

```
The regression equation is
Distance = 5.73 + 2.06 Time


Predictor       Coef  SE Coef      T      P
Constant      5.7251   0.6513   8.79  0.000
Time         2.06098  0.09128  22.58  0.000


S = 1.16901   R-Sq = 98.3%   R-Sq(adj) = 98.1%


Analysis of Variance

Source           DF      SS      MS       F      P
Regression        1  696.61  696.61  509.74  0.000
Residual Error    9   12.30    1.37
Total            10  708.91

Unusual Observations

Obs  Time  Distance     Fit  SE Fit  Residual  St Resid
 11  16.0    39.000  38.701   0.979     0.299      0.47 X

X denotes an observation whose X value gives it large leverage.


Predicted Values for New Observations

New Obs      Fit  SE Fit        95% CI             95% PI
      1   18.091   0.352  (17.294, 18.888)  (15.329, 20.853)
```

their presence or absence significantly changes the history of the world. In the context of Bedford Falls (from the Christmas movie *It's a Wonderful Life*), George Bailey (played by James Stewart) discovers that he really was influential when an angel shows him how different (and poorer) the world would have been had he never been born. Similarly, in regression, an observation is *influential* if the regression parameters alter significantly based on the presence or absence of the observation in the data set.

An outlier may or may not be influential. Similarly, a high leverage point may or may not be influential. Usually, influential observations combine both the characteristics of large residual and high leverage. It is possible for an observation to be not-quite flagged as an outlier, and not-quite flagged as a high leverage point, but still be influential through the combination of the two characteristics.

First let us consider an example of an observation that is an outlier but is not influential. Suppose that we replace our 11th observation (no more hard-core guy) with someone who hiked 20 kilometers in 5 hours. Examine Table 8.9, which presents the regression results for these 11 hikers. Note from Table 8.9 that the new observation

TABLE 8.9 Regression results including person who hiked 20 kilometers in 5 hours

```
The regression equation is
Distance = 6.36 + 2.00 Time


Predictor     Coef  SE Coef      T      P
Constant     6.364    1.278   4.98  0.001
Time        2.0000   0.2337   8.56  0.000


S = 1.71741   R-Sq = 89.1%   R-Sq(adj) = 87.8%


Analysis of Variance

Source           DF      SS      MS      F      P
Regression        1  216.00  216.00  73.23  0.000
Residual Error    9   26.55    2.95
Total            10  242.55


Unusual Observations

Obs  Time  Distance     Fit  SE Fit  Residual  St Resid
 11  5.00    20.000  16.364   0.518     3.636      2.22R

R denotes an observation with a large standardized residual.
```

is flagged as an outlier (unusual observation with large standardized residual). This is because the distance traveled (20 kilometers) is higher than the regression predicted (16.364 kilometers), given the time (5 hours).

Now, would we consider this observation to be influential? Overall, probably not. Compare Table 8.9 (the regression output for the new hiker with 5 hours/20 kilometers) and Table 8.6 (the regression output for the original data set) to assess the effect the presence of this new observation has on the regression coefficients. The y-intercept changes from $b_0 = 6.00$ to $b_0 = 6.36$, but the slope does not change at all, remaining at $b_1 = 2.00$, regardless of the presence of the new hiker.

Figure 8.6 shows the relatively mild effect this outlier has on the estimated regression line, shifting it vertically a small amount, without affecting the slope at all. Although it is an outlier, this observation is not influential because it has very low leverage, being situated exactly on the mean of the x-values, so that it has the minimum possible leverage for a data set of size $n = 11$.

We can calculate the leverage for this observation ($x = 5$, $y = 20$) as follows. As $\bar{x} = 5$, we have

$$\sum (x_i - \bar{x})^2 = (2-5)^2 + (2-5)^2 + (3-5)^2 + \cdots + (9-5)^2 + (5-5)^2 = 54.$$

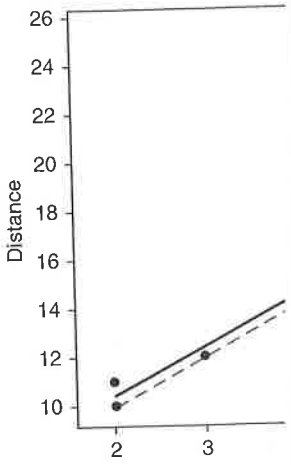


Figure 8.6 The mild outlier

Then the leverage for the

Now that we have the leve
residual, as follows. First

$s_{(5,20)}$

So that the standardized r

$\text{residual}_{(5,20)}$

as shown in Table 8.9. N
slightly larger than 2.0, s
a mild outlier.

Cook's distance is
tion. It works by taking
leverage for that observ
observation:

where $(y_i - \hat{y}_i)$ represen
mate, h_i represents the l
of predictors.

The left-hand rati
resenting the residual, v
Thus Cook's distance c

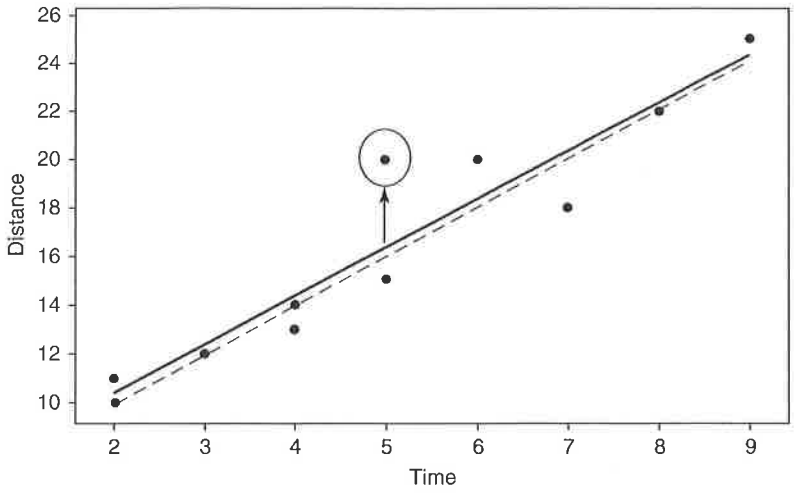


Figure 8.6 The mild outlier shifts the regression line only slightly.

Then the leverage for the new observation is

$$h_{(5,20)} = \frac{1}{11} + \frac{(5-5)^2}{54} = 0.0909.$$

Now that we have the leverage for this observation, we may also find the standardized residual, as follows. First, we have the standard error of the residual:

$$s_{(5,20),\text{resid}} = 1.71741\sqrt{1 - 0.0909} = 1.6375$$

So that the standardized residual equals:

$$\text{residual}_{(5,20),\text{standardized}} = \frac{y_i - \hat{y}_i}{s_{(5,20),\text{resid}}} = \frac{20 - 16.364}{1.6375} = 2.22,$$

as shown in Table 8.9. Note that the value of the standardized residual, 2.22, is only slightly larger than 2.0, so by our rule of thumb this observation may considered only a mild outlier.

Cook's distance is the most common measure of the influence of an observation. It works by taking into account both the size of the residual and the amount of leverage for that observation. Cook's distance takes the following form, for the ith observation:

$$D_i = \frac{(y_i - \hat{y}_i)^2}{(m+1)s^2}\left[\frac{h_i}{(1-h_i)^2}\right]$$

where $(y_i - \hat{y}_i)$ represents the ith residual, s represents the standard error of the estimate, h_i represents the leverage of the ith observation, and m represents the number of predictors.

The left-hand ratio in the formula for Cook's distance contains an element representing the residual, while the right-hand ratio contains functions of the leverage. Thus Cook's distance combines the two concepts of outlier and leverage into a single

measure of influence. The value of the Cook's distance measure for the hiker who traveled 20 kilometers in 5 hours is as follows:

$$D_i = \frac{(20 - 16.364)^2}{(1+1)1.71741^2}\left[\frac{0.0909}{(1-0.0909)^2}\right] = 0.2465$$

A rough rule of thumb for determining whether an observation is influential is if its Cook's distance exceeds 1.0. More accurately, one may also compare the Cook's distance against the percentiles of the F-distribution with $(m, n-m-1)$ degrees of freedom. If the observed value lies within the first quartile of this distribution (lower than the 25th percentile), then the observation has little influence on the regression; however, if the Cook's distance is greater than the median of this distribution, then the observation is influential. For this observation, the Cook's distance of 0.2465 lies within the 22nd percentile of the $F_{2,9}$ distribution, indicating that while the influence of the observation is small.

What about the hard-core hiker we encountered earlier? Was that observation influential? Recall that this hiker traveled 39 kilometers in 16 hours, providing the 11th observation in the results reported in Table 8.8. First, let us find the leverage.

We have $n = 11$ and $m = 1$, so that observations having $h_i > \frac{2(m+1)}{n} = 0.36$ or $h_i > \frac{3(m+1)}{n} = 0.55$ may be considered to have high leverage. This observation has $h_i = 0.7007$, which indicates that this durable hiker does indeed have high leverage, as mentioned earlier with reference to Figure 8.5. Figure 8.5 seems to indicate that this hiker ($x = 16$, $y = 39$) is not however an outlier, because the observation lies near the regression line. The standardized residual supports this, having a value of 0.46801. The reader will be asked to verify these values for leverage and standardized residual in the exercises. Finally, the Cook's distance for this observation is 0.2564, which is about the same as our previous example, indicating that the observation is not influential. Figure 8.7 shows the slight change in the regression with (solid line) and without (dotted line) this observation.

So we have seen that an observation that is an outlier with low influence, or an observation that is a high leverage point with a small residual may not be particularly influential. We next illustrate how a data point that has a moderately high residual and moderately high leverage may indeed be influential. Suppose that our 11th hiker had instead hiked for 10 hours, and traveled 23 kilometers. The regression analysis for these 11 hikers is then given in Table 8.10.

Note that *Minitab* does not identify the new observation as either an outlier or a high leverage point. This is because, as the reader is asked to verify in the exercises, the leverage of this new hiker is $h_i = 0.36019$, and the standardized residual equals -1.70831.

However, despite lacking either a particularly large leverage or large residual, this observation is nevertheless influential, as measured by its Cook's distance of $D_i = 0.821457$, which is in line with the 62nd percentile of the $F_{1,10}$ distribution.

The influence of this observation stems from the combination of its moderately large residual with its moderately large leverage. Figure 8.8 shows the influence this single hiker has on the regression line, pulling down on the right side to decrease the slope (from 2.00 to 1.82), and thereby increase the y-intercept (from 6.00 to 6.70).

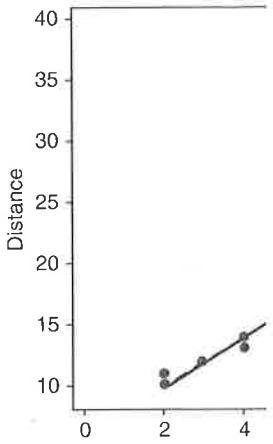


Figure 8.7 Slight change

TABLE 8.10 Regression re

```
The regression e
Distance = 6.70

Predictor      Coe
Constant     6.696
Time         1.822

S = 1.40469    R-

Analysis of Vari

Source
Regression
Residual Error
Total
```

8.8 POPULATIC

Least squares regressio
assumptions of the regr
and model building are
unverified assumptions

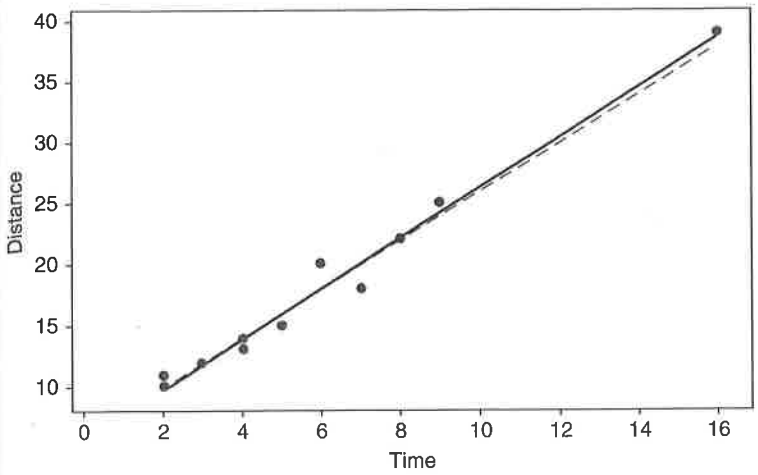


Figure 8.7 Slight change in regression line when hard-core hiker added.

TABLE 8.10 Regression results for new observation with *time* = 10, *distance* = 23

```
The regression equation is
Distance = 6.70 + 1.82 Time


Predictor     Coef  SE Coef      T      P
Constant    6.6967   0.9718   6.89  0.000
Time        1.8223   0.1604  11.36  0.000


S = 1.40469   R-Sq = 93.5%   R-Sq(adj) = 92.8%


Analysis of Variance

Source           DF      SS      MS       F      P
Regression        1  254.79  254.79  129.13  0.000
Residual Error    9   17.76    1.97
Total            10  272.55
```

8.8 POPULATION REGRESSION EQUATION

Least squares regression is a powerful and elegant methodology. However, if the assumptions of the regression model are not validated, then the resulting inference and model building are undermined. Deploying a model whose results are based on unverified assumptions may lead to expensive failures later on. The simple linear

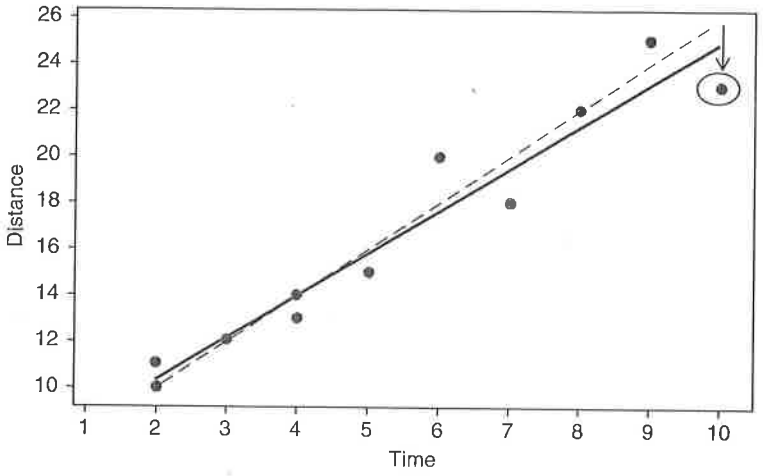


Figure 8.8 Moderate residual plus moderate leverage = influential observation.

regression model is given as follows. We have a set of n bivariate observations, with response value y_i related to predictor value x_i through the following linear relationship.

THE POPULATION REGRESSION EQUATION

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where

- β_0 and β_1 represent the model parameters, for the y-intercept and slope respectively. These are constants, whose true value remains unknown, and which are estimated from the data using the least squares estimates.
- ε represents the error term. As most predictor–response relationships are not deterministic, a certain amount of error will be introduced by any linear approximation of the actual relationship. Therefore, an error term, modeled by a random variable, is needed.

THE ASSUMPTIONS ABOUT THE ERROR TERM

- **Zero-Mean Assumption.** The error term ε is a random variable, with mean or expected value equal to zero. In other words, $E(\varepsilon) = 0$.
- **Constant Variance Assumption.** The variance of ε, denoted by σ^2, is constant, regardless of the value of x.
- **Independence Assumption.** The values of ε are independent.
- **Normality Assumption.** The error term ε is a normally distributed random variable.

In other words, the values of the error term ε_i are independent normal random variables, with mean 0 and variance σ^2.

On the basis of th
behavior of the respons

**IMPLICATIONS OF T
RESPONSE VARIABI**

1. On the basis of th
   $E(y) =$
   That is, for each
2. On the basis of
   Var(y), given as
   which value taker
3. On the basis of the
   of x, the values of
4. Based on the nor
   random variable.

In other words, th
variables, with mean β_0

Figure 8.9 illustra
constant variance σ^2. S

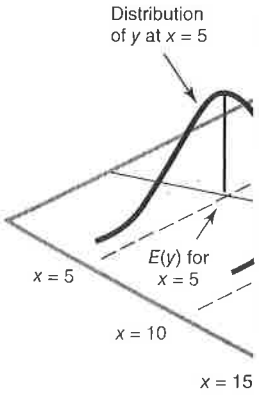


Figure 8.9 For each valu
sion line, and constant var

On the basis of these four assumptions, we can derive four implications for the behavior of the response variable, y, as follows.

IMPLICATIONS OF THE ASSUMPTIONS FOR THE BEHAVIOR OF THE RESPONSE VARIABLE y

1. On the basis of the Zero-Mean Assumption, we have
$$E(y) = E(\beta_0 + \beta_1 x + \varepsilon) = E(\beta_0) + E(\beta_1 x) + E(\varepsilon) = \beta_0 + \beta_1 x$$
That is, for each value of x, the mean of the y's lies on the regression line.
2. On the basis of the Constant Variance Assumption, we have the variance of y, $\text{Var}(y)$, given as $\text{Var}(y) = \text{Var}(\beta_0 + \beta_1 x + \varepsilon) = \text{Var}(\varepsilon) = \sigma^2$. That is, regardless of which value taken by the predictor x, the variance of the y's is always constant.
3. On the basis of the Independence Assumption, it follows that, for any particular value of x, the values of y are independent as well.
4. Based on the normality assumption, it follows that y is also a normally distributed random variable.

In other words, the values of the response variable y_i are independent normal random variables, with mean $\beta_0 + \beta_1 x$ and variance σ^2.

Figure 8.9 illustrates graphically the normality of the y_i, with mean $\beta_0 + \beta_1 x$ and constant variance σ^2. Suppose we have a data set which includes predictor values at

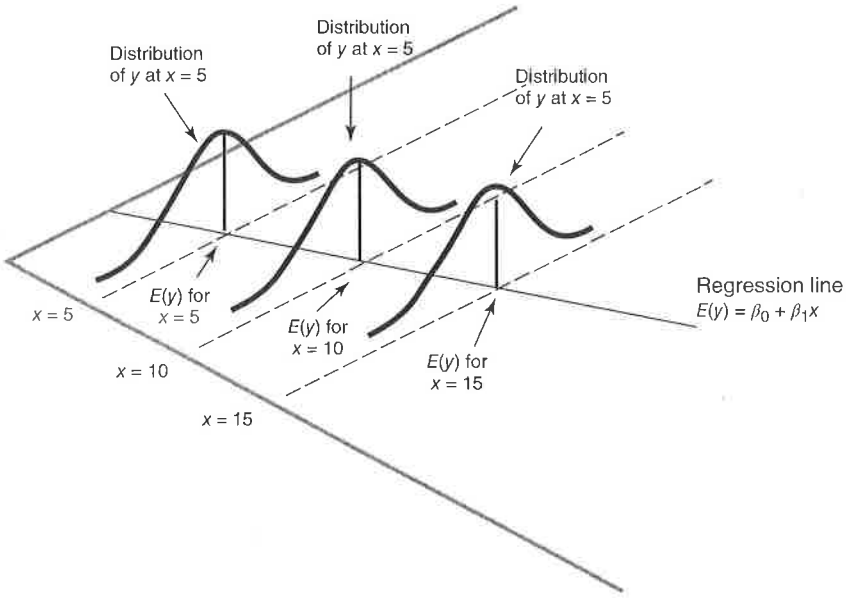


Figure 8.9 For each value of x, the y_i are normally distributed, with mean on the true regression line, and constant variance.

$x = 5$, 10, and 15, among other values. Then, at each of these values of x, the regression assumptions assert that observed values of y are samples from a normally distributed population with a mean on the regression line ($E(y) = \beta_0 + \beta_1 x$), and constant standard deviation σ^2. Note from Figure 8.9 that each of the normal curves has precisely the same shape, which indicates that the variance is constant for each value of x.

If one is interested in using regression analysis in a strictly descriptive manner, with no inference and no model building, then one need not worry quite so much about assumption validation. This is because the assumptions are about the error term. If the error term is not involved, then the assumptions are not needed. However, if one wishes to do inference or model building, then the assumptions must be verified.

8.9 VERIFYING THE REGRESSION ASSUMPTIONS

So, how does one go about verifying the regression assumptions? The two main graphical methods used to verify regression assumptions are as follows:

a. A normal probability plot of the residuals.

b. A plot of the standardized residuals against the fitted (predicted) values.

A *normal probability plot* is a quantile–quantile plot of the quantiles of a particular distribution against the quantiles of the standard normal distribution, for the purposes of determining whether the specified distribution deviates from normality. (Similar to a percentile, a *quantile* of a distribution is a value x_p such that $p\%$ of the distribution values are less than or equal to x_p.) In a normality plot, the observed values of the distribution of interest are compared against the same number of values that would be expected from the normal distribution. If the distribution is normal, then the bulk of the points in the plot should fall on a straight line; systematic deviations from linearity in this plot indicate non-normality.

To illustrate the behavior of the normal probability plot for different kinds of data distributions, we provide three examples. Figures 8.10–8.12 contain the normal probability plots for 10,000 values drawn from a uniform (0, 1) distribution, a chi-square (5) distribution, and a normal (0, 1) distribution, respectively.

Note in Figure 8.10 that the bulk of the data do not line up on the straight line, and that a clear pattern (reverse S curve) emerges, indicating systematic deviation from normality. The uniform distribution is a rectangular-shaped distribution, whose tails are much heavier than the normal distribution. Thus, Figure 8.10 is an example of a probability plot for a distribution with heavier tails than the normal distribution.

Figure 8.11 also contains a clear curved pattern, indicating systematic deviation from normality. The chi-square (5) distribution is right-skewed, so that the curve pattern apparent in Figure 8.11 may be considered typical of the pattern made by right-skewed distributions in a normal probability plot.

Finally, in Figure 8.12, the points line up nicely on a straight line, indicating normality, which is not surprising because the data are drawn from a normal (0, 1) distribution. It should be remarked that we should not expect real-world data to behave

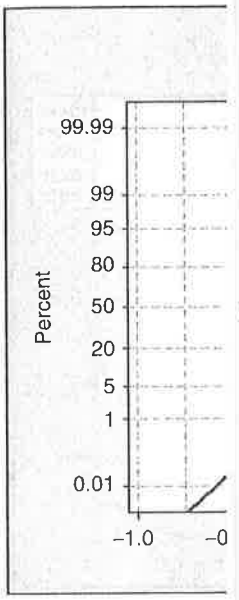


Figure 8.10 Normal p

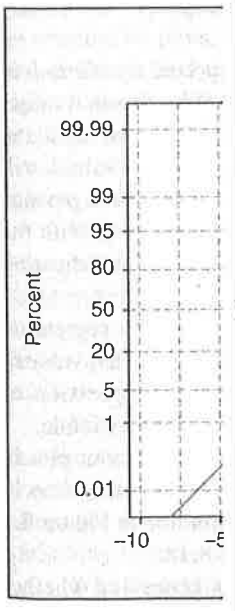


Figure 8.11 Probabil

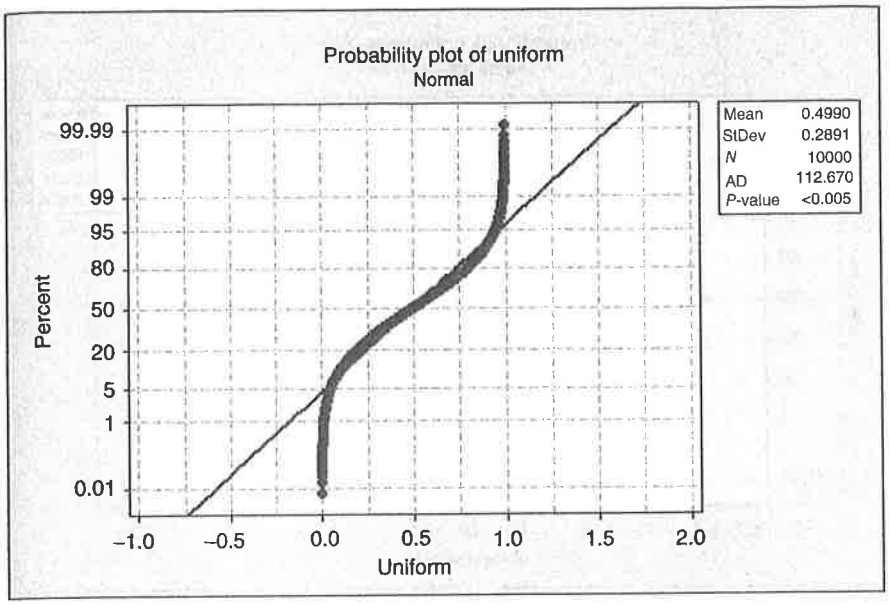


Figure 8.10 Normal probability plot for a uniform distribution: heavy tails.

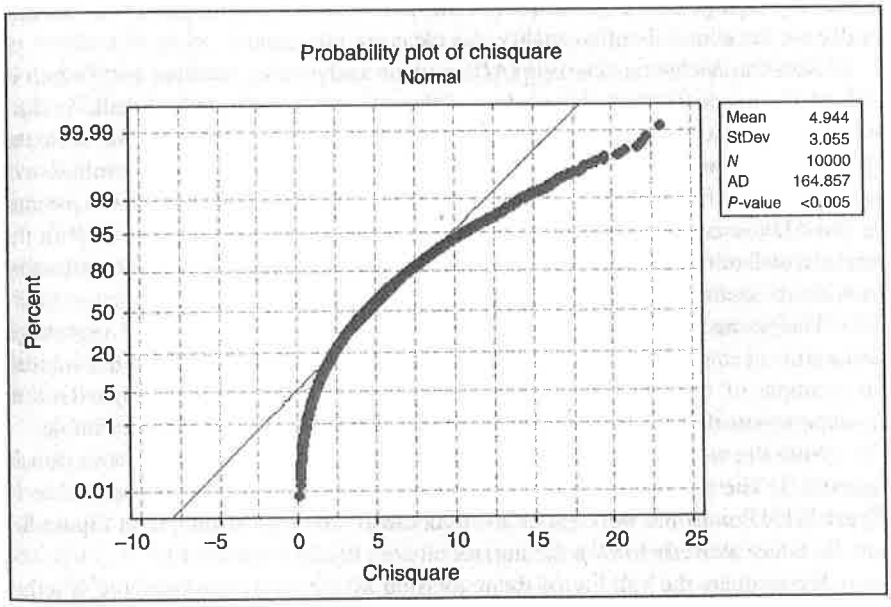


Figure 8.11 Probability plot for a chi-square distribution: right-skewed.

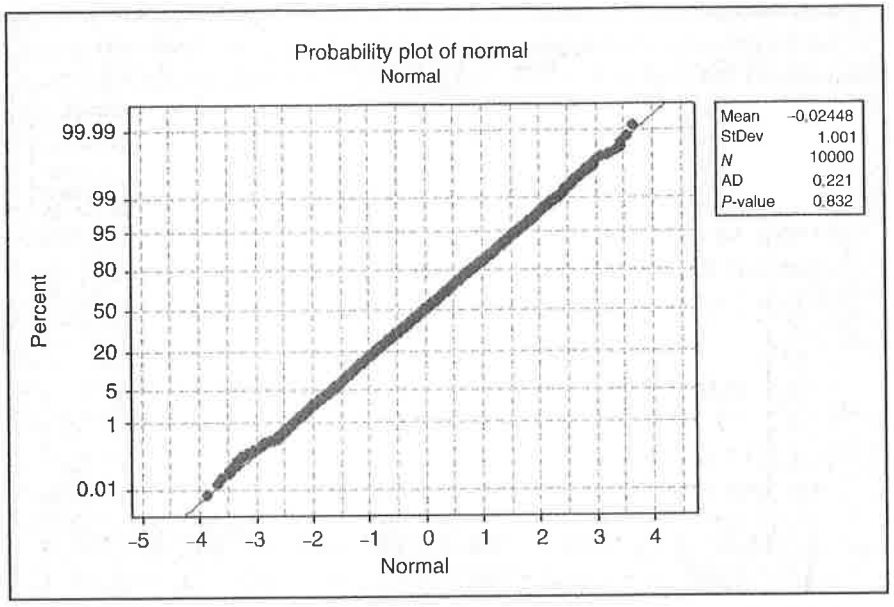


Figure 8.12 Probability plot for a normal distribution: Do not expect real-world data to behave this nicely.

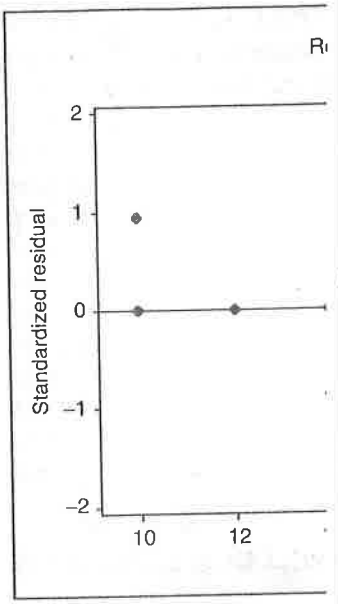


Figure 8.13 Plot of standardize

this nicely. The presence of sampling error and other sources of noise will usually render our decisions about normality less clear-cut than this.

Note the Anderson–Darling (*AD*) statistic and *p*-value reported by *Minitab* in each of Figures 8.10–8.12. This refers to the AD test for normality. Smaller values of the AD statistic indicate that the normal distribution is a better fit for the data. The null hypothesis is that the normal distribution fits, so that small *p*-values will indicate lack of fit. Note that for the uniform and chi-square examples, the *p*-value for the AD test is less than 0.005, indicating strong evidence for lack of fit with the normal distribution. However, the *p*-value for the normal example is 0.832, indicating no evidence against the null hypothesis that the distribution is normal.

The second graphical method used to assess the validity of the regression assumptions is a plot of the standardized residuals against the fits (predicted values). An example of this type of graph is given in Figure 8.13, for the regression of *distance* versus *time* for the original 10 observations in the orienteering example.

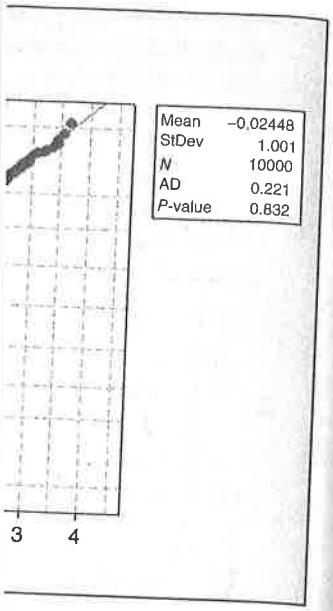
Note the close relationship between this graph and the original scatter plot in Figure 8.3. The regression line from Figure 8.3 is now the horizontal zero line in Figure 8.13. Points that were either above/below/on the regression line in Figure 8.3 now lie either above/below/on the horizontal zero line in Figure 8.13.

We evaluate the validity of the regression assumptions by observing whether certain patterns exist in the plot of the residuals versus fits, in which case one of the assumptions has been violated, or whether no such discernible patterns exists, in which case the assumptions remain intact. The 10 data points in Figure 8.13 are really too few to try to determine whether any patterns exist. In data mining applications, of course, paucity of data is rarel watch out for. Figure 8.14 sl in residual-fit plots. Plot (a) are observed, and the points right. Plot (b) exhibits curvat (c) displays a "funnel" patte Finally, plot (d) exhibits a p the zero-mean assumption.

Why does plot (b) viola assumed to be independent, independent behavior as wel for a given residual, we may within a certain margin of er prediction would not be poss

Why does plot (c) viol that the variability in the res stant, regardless of the value is smaller for smaller value variability is non-constant, v

Why does plot (d) viol tion states that the mean of th plot (d) shows that, for small for large values of x, the mea the zero-mean assumption, a



t expect real-world data to behave

sources of noise will usually
this.
value reported by *Minitab* in
or normality. Smaller values
n is a better fit for the data.
, so that small p-values will
quare examples, the p-value
lence for lack of fit with the
example is 0.832, indicating
ion is normal.
validity of the regression
st the fits (predicted values).
8.13, for the regression of
ne orienteering example.
l the original scatter plot in
the horizontal zero line in
egression line in Figure 8.3
Figure 8.13.
ions by observing whether
fits, in which case one of
scernible patterns exists, in
nts in Figure 8.13 are really
ata mining applications, of

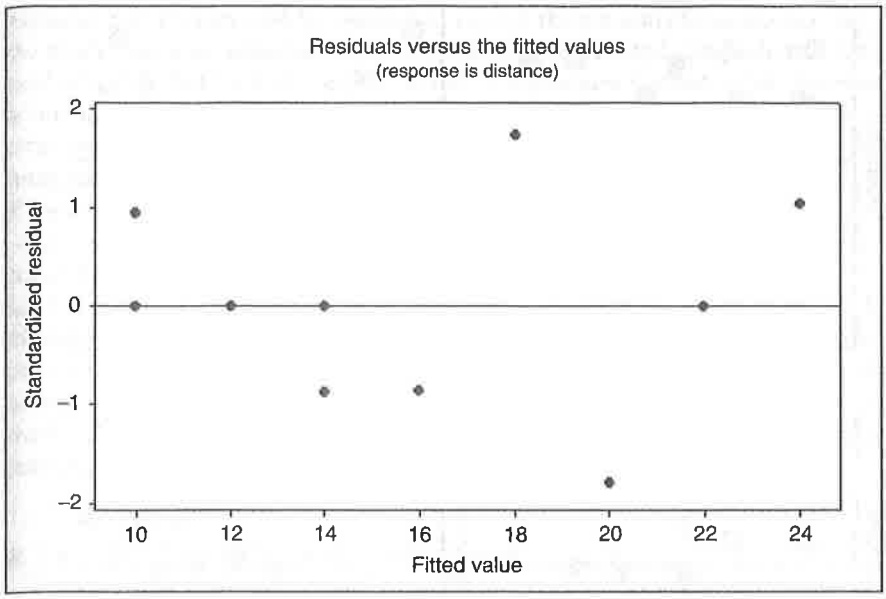


Figure 8.13 Plot of standardized residuals versus predicted values for orienteering example.

course, paucity of data is rarely the issue. Let us see what types of patterns we should watch out for. Figure 8.14 shows four pattern "archetypes" that may be observed in residual-fit plots. Plot (a) shows a "healthy" plot, where no noticeable patterns are observed, and the points display an essentially rectangular shape from left to right. Plot (b) exhibits curvature, which violates the independence assumption. Plot (c) displays a "funnel" pattern, which violates the constant variance assumption. Finally, plot (d) exhibits a pattern that increases from left to right, which violates the zero-mean assumption.

Why does plot (b) violate the independence assumption? Because the errors are assumed to be independent, the residuals (which estimate the errors) should exhibit independent behavior as well. However, if the residuals form a curved pattern, then, for a given residual, we may predict where its neighbors to the left and right will fall, within a certain margin of error. If the residuals were truly independent, then such a prediction would not be possible.

Why does plot (c) violate the constant variance assumption? Note from plot (a) that the variability in the residuals, as shown by the vertical distance, is fairly constant, regardless of the value of x. However, in plot (c), the variability of the residuals is smaller for smaller values of x, and larger for larger values of x. Therefore, the variability is non-constant, which violates the constant variance assumption.

Why does plot (d) violate the zero-mean assumption? The zero-mean assumption states that the mean of the error term is zero, regardless of the value of x. However, plot (d) shows that, for small values of x, the mean of the residuals is less than 0, while, for large values of x, the mean of the residuals is greater than 0. This is a violation of the zero-mean assumption, as well as a violation of the independence assumption.

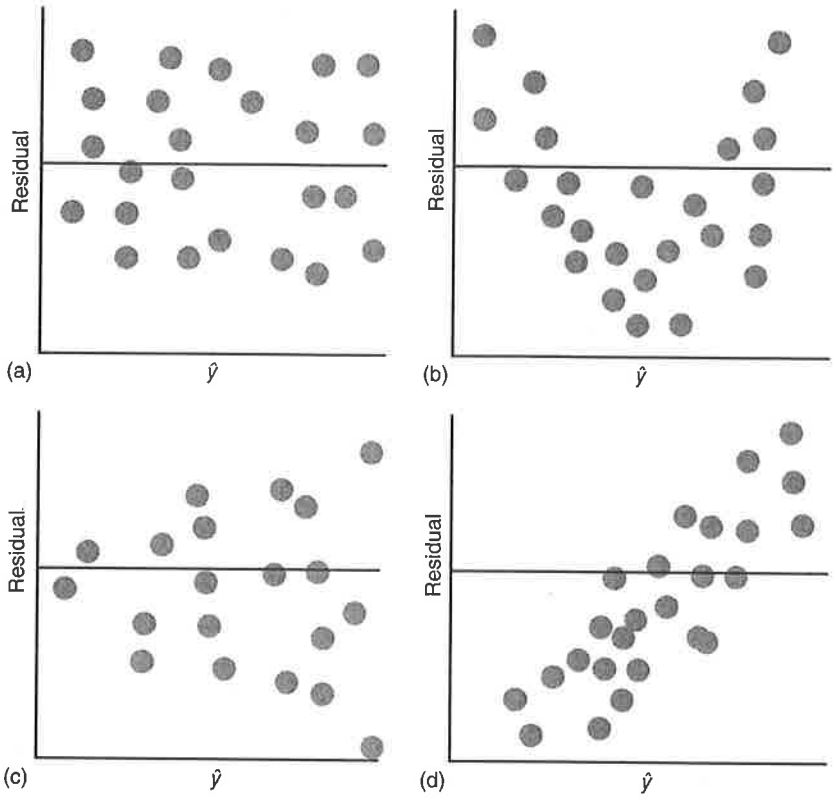


Figure 8.14 Four possible patterns in the plot of residuals versus fits.

When examining plots for patterns, beware of the "Rorschach effect" of seeing patterns in randomness. The null hypothesis when examining these plots is that the assumptions are intact; only systematic and clearly identifiable patterns in the residuals plots offer evidence to the contrary.

Apart from these graphical methods, there are also several diagnostic hypothesis tests that may be carried out to assess the validity of the regression assumptions. As mentioned above, the AD test may be used to indicate fit of the residuals to a normal distribution. For assessing whether the constant variance assumption has been violated, either Bartlett's test or Levene's test may be used. For determining whether the independence assumption has been violated, either the Durban–Watson test or the runs test may be applied. Information about all of these diagnostic tests may be found in Draper and Smith (1998).[4]

Note that these assumptions represent the structure needed to perform inference in regression. Descriptive methods in regression, such as point estimates, and simple reporting of such statistics, as the slope, correlation, standard error of the

[4]Draper and Smith, *Applied Regression Analysis*, 3rd edition, Wiley Publishers, Hoboken, New Jersey, 1998.

estimate, and r^2, may still be u
the results are cross-validated. W
tical inference. But we as data mi
is not our primary *modus opera*
cross-validation of the results a
ining the relationship between
if the training data set and test
-0.7, and there is graphical evi
reporting to our client *in a descr*
lated, even if both variables are
the correlation test). We just can
cally significant negative value,
to the realm of inference. So, f
results across partitions, and (ii)
language, and avoid inferential t

8.10 INFERENCE IN R

Inference in regression offers a
of linear association between tw
the usual caveats regarding the
For very large sample sizes, ev
significant, even when their prac

We shall examine five infe

1. The *t-test* for the relations variable.
2. The correlation coefficient
3. The confidence interval fo
4. The confidence interval fo value of the predictor.
5. The prediction interval fo particular value of the pre

In Chapter 9, we also inv
sion as a whole. However, for s
equivalent.

How do we go about perf
consider the form of the true (po

This equation asserts that there
and some function of x on the
constant whose value is unknow
β_1 took that value, there would n