The Granger Collection, New York

SIR FRANCIS GALTON

Correlation, Regression, and Heredity

The least-squares method will happily fit a straight line to any two-variable data. It is an old method, going back to the French mathematician Legendre in about 1805. Legendre invented least squares for use on data from astronomy and surveying. It was *Sir Francis Galton (1822–1911)*, however, who turned "regression" into a general method for understanding relationships. He even invented the word. While he was at it, he also invented "correlation," both the word and the definition of *r*.

Galton was one of the last gentleman scientists, an upper-class Englishman who studied medicine at Cambridge and explored Africa before turning to the study of heredity. He was well connected here also: Charles Darwin, who published *The Origin of Species* in 1859, was his cousin.

Galton was full of ideas but was no mathematician. He didn't even use least squares, preferring to avoid unpleasant computations. But Galton was the first to apply regression ideas to biological and psychological data. He asked: If people's heights are distributed normally in every generation, and height is inherited, what is the relationship between generations? He discovered a straight-line relationship between the heights of parent and child and found that tall parents tended to have children who were taller than average but less tall than their parents. He called this "regression toward mediocrity." The name "regression" came to be applied to the statistical method.

Galton was full of ideas but was no mathematician. He didn't even use least squares, preferring to avoid unpleasant computations.

chapter 3

Examining Relationships

ACTIVITY 3 SAT/ACT Scores

Materials: Pencil, grid paper

Is there an association between SAT Math scores and SAT Verbal scores? If a student performs well on the Math part of the SAT exam, will he or she do well on the Verbal part, too? If a student performs well on one part, does that suggest that the student will not do as well on the other? Is it rare or fairly common for students to score about the same on both parts of the SAT? In this activity you will collect, anonymously of course, the SAT Math and SAT Verbal scores for each member of the class who has taken the SAT exam. You will then plot these data and inspect the graph to see if a pattern is evident. If your school is in a state where the ACT exam is the principal college placement test, then use ACT scores.

1. Begin by writing your Math score and Verbal score on an index card or similar uniform "ballot." Label your Math score M, and your Verbal score V. A selected student should collect the folded index cards in a box or other container. When all of the index cards have been placed in the box, mix them without looking, so that each student's privacy is protected.

If the size of your class is "small," then you may need to supplement your data with the scores of students in other classes. Perhaps your teacher can request that scores from other AP classes be provided to make a larger data set. Try to obtain data from at least 25 or 30 students.

2. The scores should be called out by the student who collects the data and recorded on the blackboard as ordered pairs in the form (Math, Verbal).

3. Each student should construct a plot of the data with pencil and paper. Since the Math scores appear first in the ordered pairs, label your horizontal axis "Math" and label the vertical axis "Verbal." Determine the range of the Math scores and the range of the Verbal scores, and then construct scales for both axes. Note that axes don't have to intersect at the point (0,0), but the scales on both axes should be uniform.

4. When you finish constructing your graph, look to see if there is any discernible pattern. If so, can you describe the pattern? Does the graph provide any insight into a possible association between SAT Math and SAT Verbal scores?

We will return to analyze these data in more detail after we develop some methodology.

INTRODUCTION

Most statistical studies involve more than one variable. Sometimes we want to compare the distributions of the same variable for several groups. For example, we might compare the distributions of SAT scores among students at several colleges. Side-by-side boxplots, stemplots, or histograms make the comparison visible. In this chapter, however, we concentrate on relationships among several variables for the same group of individuals. For example, Table 1.15 (page 71) records seven variables that describe education in the United States. We have already examined some of these variables one at a time. Now we might ask how SAT Mathematics scores are related to SAT Verbal scores or to the percent of a state's high school seniors who take the SAT or to what region a state is in.

When you examine the relationship between two or more variables, first ask the preliminary questions that are familiar from Chapters 1 and 2.

- What *individuals* do the data describe?
- What exactly are the *variables*? How are they measured?
- Are all the variables *quantitative* or is at least one a *categorical* variable?

We have concentrated on quantitative variables until now. When we have data on several variables, however, categorical variables are often present and help organize the data. Categorical variables will play a larger role in the next chapter. There is one more question you should ask when you are interested in relations among several variables:

- Do you want simply to explore the nature of the relationship, or do you think that some of the variables explain or even cause changes in others? That is, are some of the variables *response variables* and others *explanatory variables*?

RESPONSE VARIABLE, EXPLANATORY VARIABLE

A **response variable** measures an outcome of a study. An **explanatory variable** attempts to explain the observed outcomes.

independent variable
dependent variable

You will often find explanatory variables called ***independent variables***, and response variables called ***dependent variables.*** The idea behind this language is that the response variable depends on the explanatory variable. Because the words "independent" and "dependent" have other, unrelated meanings in statistics, we won't use them here.

It is easiest to identify explanatory and response variables when we actually set values of one variable in order to see how it affects another variable.

EXAMPLE 3.1 EFFECT OF ALCOHOL ON BODY TEMPERATURE

Alcohol has many effects on the body. One effect is a drop in body temperature. To study this effect, researchers give several different amounts of alcohol to mice, then measure the change in each mouse's body temperature in the 15 minutes after taking the alcohol. Amount of alcohol is the explanatory variable, and change in body temperature is the response variable.

When you don't set the values of either variable but just observe both variables, there may or may not be explanatory and response variables. Whether there are depends on how you plan to use the data.

EXAMPLE 3.2 ARE SAT MATH AND VERBAL SCORES LINKED?

Jim wants to know how the median SAT Math and Verbal scores in the 51 states (including the District of Columbia) are related to each other. He doesn't think that either score explains or causes the other. Jim has two related variables, and neither is an explanatory variable.

Julie looks at some data. She asks, "Can I predict a state's median SAT Math score if I know its median SAT Verbal score?" Julie is treating the Verbal score as the explanatory variable and the Math score as the response variable.

In Example 3.1 alcohol actually *causes* a change in body temperature. There is no cause-and-effect relationship between SAT Math and Verbal scores in Example 3.2. Because the scores are closely related, we can nonetheless use a state's SAT Verbal score to predict its Math score. We will learn how to do the prediction in Section 3.3. Prediction requires that we identify an explanatory variable and a response variable. Some other statistical techniques ignore this distinction. Do remember that calling one variable explanatory and the other response doesn't necessarily mean that changes in one *cause* changes in the other.

The statistical techniques used to study relations among variables are more complex than the one-variable methods in Chapters 1 and 2. Fortunately, analysis of several-variable data builds on the tools used for examining individual variables. The principles that guide examination of data are also the same:

- First plot the data, then add numerical summaries.
- Look for overall patterns and deviations from those patterns.
- When the overall pattern is quite regular, use a compact mathematical model to describe it.

EXERCISES

3.1 EXPLANATORY AND RESPONSE VARIABLES In each of the following situations, is it more reasonable to simply explore the relationship between the two variables or to view one

of the variables as an explanatory variable and the other as a response variable? In the latter case, which is the explanatory variable and which is the response variable?

(a) The amount of time a student spends studying for a statistics exam and the grade on the exam

(b) The weight and height of a person

(c) The amount of yearly rainfall and the yield of a crop

(d) A student's grades in statistics and in French

(e) The occupational class of a father and of a son

3.2 QUANTITATIVE AND CATEGORICAL VARIABLES How well does a child's height at age 6 predict height at age 16? To find out, measure the heights of a large group of children at age 6, wait until they reach age 16, then measure their heights again. What are the explanatory and response variables here? Are these variables categorical or quantitative?

3.3 GENDER GAP There may be a "gender gap" in political party preference in the United States, with women more likely than men to prefer Democratic candidates. A political scientist selects a large sample of registered voters, both men and women. She asks each voter whether they voted for the Democratic or for the Republican candidate in the last congressional election. What are the explanatory and response variables in this study? Are they categorical or quantitative variables?

3.4 TREATING BREAST CANCER The most common treatment for breast cancer was once removal of the breast. It is now usual to remove only the tumor and nearby lymph nodes, followed by radiation. The change in policy was due to a large medical experiment that compared the two treatments. Some breast cancer patients, chosen at random, were given each treatment. The patients were closely followed to see how long they lived following surgery. What are the explanatory and response variables? Are they categorical or quantitative?

3.5 What are the variables in Activity 3 (page 120)? Is there an explanatory/response relationship? If so, which is the explanatory variable and which is the response variable? Are the variables quantitative or categorical?

3.1 SCATTERPLOTS

The most effective way to display the relation between two quantitative variables is a *scatterplot*. Here is an example of a scatterplot.

EXAMPLE 3.3 STATE SAT SCORES

Some people use average SAT scores to rank state or local school systems. This is not proper, because the percent of high school students who take the SAT varies from place to place. Let us examine the relationship between the percent of a state's high school graduates who take the exam and the state average SAT Mathematics score, using data from Table 1.15 on page 70.

We think that "percent taking" will help explain "average score." Therefore, "percent taking" is the explanatory variable and "average score" is the response variable.

We want to see how average score changes when percent taking changes, so we put percent taking (the explanatory variable) on the horizontal axis. Figure 3.1 is the scatterplot. Each point represents a single state. In Alabama, for example, 9% take the SAT, and the average SAT Math score is 555. Find 9 on the x (horizontal) axis and 555 on the y (vertical) axis. Alabama appears as the point (9, 555) above 9 and to the right of 555. Figure 3.1 shows how to locate Alabama's point on the plot.

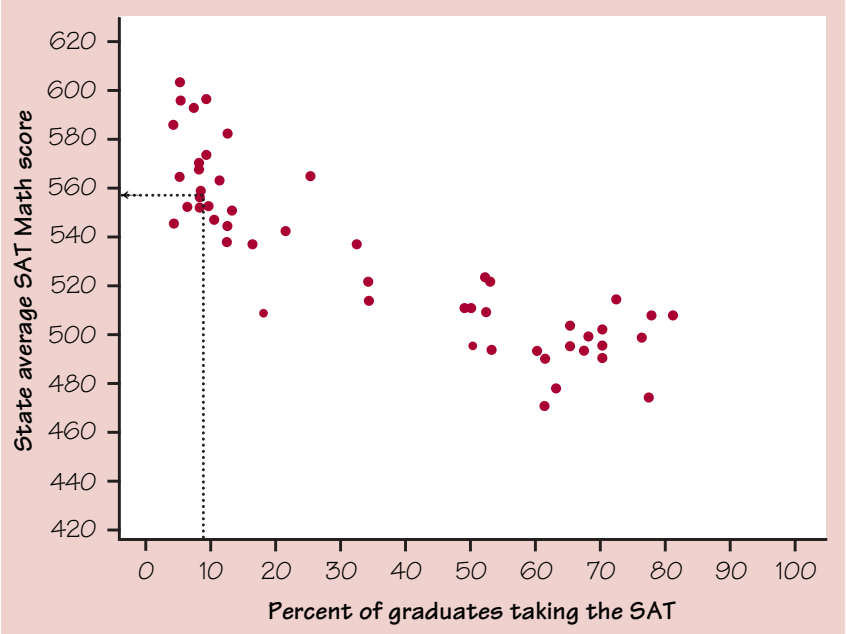


FIGURE 3.1 Scatterplot of the average SAT Math score in each state against the percent of that state's high school graduates who take the SAT, from Table 1.15. The dotted lines intersect at the point (9, 555), the data for Alabama.

SCATTERPLOT

A **scatterplot** shows the relationship between two quantitative variables measured on the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as the point in the plot fixed by the values of both variables for that individual.

Always plot the explanatory variable, if there is one, on the horizontal axis (the x axis) of a scatterplot. As a reminder, we usually call the explanatory variable x and the response variable y. If there is no explanatory-response distinction, either variable can go on the horizontal axis.

EXERCISES

3.6 THE ENDANGERED MANATEE Manatees are large, gentle sea creatures that live along the Florida coast. Many manatees are killed or injured by powerboats. Here are data on powerboat registrations (in thousands) and the number of manatees killed by boats in Florida in the years 1977 to 1990:

Year	Powerboat registrations (1000)	Manatees killed	Year	Powerboat registrations (1000)	Manatees killed
1977	447	13	1984	559	34
1978	460	21	1985	585	33
1979	481	24	1986	614	33
1980	498	16	1987	645	39
1981	513	24	1988	675	43
1982	512	20	1989	711	50
1983	526	15	1990	719	47

(a) We want to examine the relationship between number of powerboats and number of manatees killed by boats. Which is the explanatory variable?

(b) Make a scatterplot of these data. (Be sure to label the axes with the variable names, not just *x* and *y*.) What does the scatterplot show about the relationship between these variables?

3.7 ARE JET SKIS DANGEROUS? Propelled by a stream of pressurized water, jet skis and other so-called wet bikes carry from one to three people, retail for an average price of $5,700, and have become one of the most popular types of recreational vehicle sold today. But critics say that they're noisy, dangerous, and damaging to the environment. An article in the August 1997 issue of the *Journal of the American Medical Association* reported on a survey that tracked emergency room visits at randomly selected hospitals nationwide. Here are data on the number of jet skis in use, the number of accidents, and the number of fatalities for the years 1987–1996:[1]

Year	Number in use	Accidents	Fatalities
1987	92,756	376	5
1988	126,881	650	20
1989	178,510	844	20
1990	241,376	1,162	28
1991	305,915	1,513	26
1992	372,283	1,650	34
1993	454,545	2,236	35
1994	600,000	3,002	56
1995	760,000	4,028	68
1996	900,000	4,010	55

(a) We want to examine the relationship between the number of jet skis in use and the number of accidents. Which is the explanatory variable?

(b) Make a scatterplot of these data. (Be sure to label the axes with the variable names, not just *x* and *y*.) What does the scatterplot show about the relationship between these variables?

3.8 Make a scatterplot of the (Math SAT/ACT score, Verbal SAT/ACT score) data from Activity 3, if you haven't done so already. Does the scatterplot describe a strong association, a moderate association, a weak association, or no association between these variables?

Interpreting scatterplots

To interpret a scatterplot, apply the strategies of data analysis learned in Chapters 1 and 2.

EXAMINING A SCATTERPLOT

In any graph of data, look for the **overall pattern** and for striking **deviations** from that pattern.

You can describe the overall pattern of a scatterplot by the **form, direction,** and **strength** of the relationship.

An important kind of deviation is an **outlier,** an individual value that falls outside the overall pattern of the relationship.

clusters

Figure 3.1 shows a clear *form:* there are two distinct ***clusters*** of states with a gap between them. In the cluster at the right of the plot, 45% or more of high school graduates take the SAT, and the average scores are low. The states in the cluster at the left have higher SAT scores and lower percents of graduates taking the test. There are no clear outliers. That is, no points fall clearly outside the clusters.

What explains the clusters? There are two widely used college entrance exams, the SAT and the American College Testing (ACT) exam. Each state favors one or the other. The left cluster in Figure 3.1 contains the ACT states, and the SAT states make up the right cluster. In ACT states, most students who take the SAT are applying to a selective college that requires SAT scores. This select group of students has a higher average score than the much larger group of students who take the SAT in SAT states.

The relationship in Figure 3.1 also has a clear *direction:* states in which a higher percent of students take the SAT tend to have lower average scores. This is a *negative association* between the two variables.

POSITIVE ASSOCIATION, NEGATIVE ASSOCIATION

Two variables are **positively associated** when above-average values of one tend to accompany above-average values of the other and below-average values also tend to occur together.

Two variables are **negatively associated** when above-average values of one tend to accompany below-average values of the other, and vice versa.

The *strength* of a relationship in a scatterplot is determined by how closely the points follow a clear form. The overall relationship in Figure 3.1 is not strong—states with similar percents taking the SAT show quite a bit of scatter in their average scores. Here is an example of a stronger relationship with a clearer form.

EXAMPLE 3.4 HEATING DEGREE-DAYS

The Sanchez household is about to install solar panels to reduce the cost of heating their house. In order to know how much the solar panels help, they record their consumption of natural gas before the panels are installed. Gas consumption is higher in cold weather, so the relationship between outside temperature and gas consumption is important.

Table 3.1 gives data for 16 months. The response variable y is the average amount of natural gas consumed each day during the month, in hundreds of cubic feet. The explanatory variable x is the average number of heating degree-days each day during the month. (Heating degree-days are the usual measure of demand for heating. One degree-day is accumulated for each degree a day's average temperature falls below 65° F. An average temperature of 20° F, for example, corresponds to 45 degree-days.)

TABLE 3.1 Average degree-days and natural gas consumption for the Sanchez household

Month	Degree-days	Gas (100 cu. ft.)	Month	Degree-days	Gas (100 cu. ft.)
Nov.	24	6.3	July	0	1.2
Dec.	51	10.9	Aug.	1	1.2
Jan.	43	8.9	Sept.	6	2.1
Feb.	33	7.5	Oct.	12	3.1
Mar.	26	5.3	Nov.	30	6.4
Apr.	13	4.0	Dec.	32	7.2
May	4	1.7	Jan.	52	11.0
June	0	1.2	Feb.	30	6.9

Source: Data provided by Robert Dale, Purdue University.

The scatterplot in Figure 3.2 shows a strong positive association. More degree-days means colder weather and so more gas consumed. The form of the relationship is ***linear.*** That is, the points lie in a straight-line pattern. It is a strong relationship because the points

linear

lie close to a line, with little scatter. If we know how cold a month is, we can predict gas consumption quite accurately from the scatterplot. That strong relationships make accurate predictions possible is an important point that we will soon discuss in more detail.

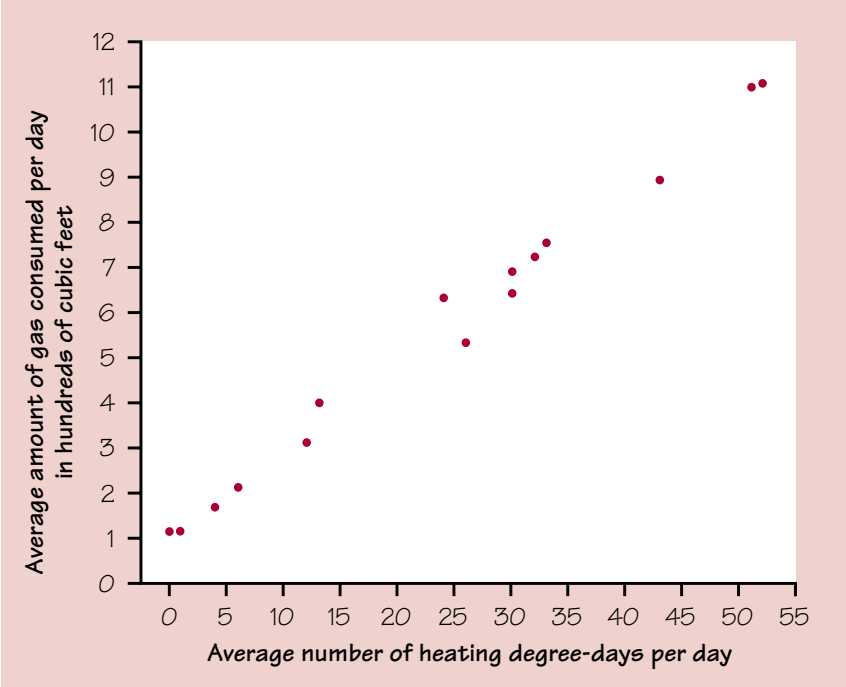


FIGURE 3.2 Scatterplot of the average amount of natural gas used per day by the Sanchez household in 16 months against the average number of heating degree-days per day in those months, from Table 3.1.

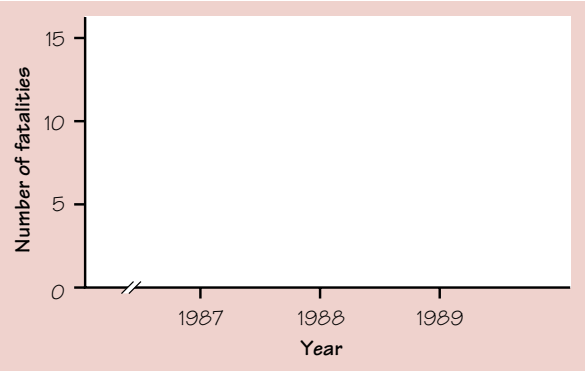
Of course, not all relationships are linear in form. What is more, not all relationships have a clear direction that we can describe as positive association or negative association. Exercise 3.11 gives an example that is not linear and has no clear direction.

Tips for drawing scatterplots

1. Scale the horizontal and vertical axes. The intervals must be uniform; that is, the distance between tick marks must be the same. If the scale does not begin at zero at the origin, then use the symbol shown to indicate a break.

2. Label both axes.

3. If you are given a grid, try to adopt a scale so that your plot uses the whole grid. Make your plot large enough so that the details can be easily seen. Don't compress the plot into one corner of the grid.



EXERCISES

3.9 MORE ON THE ENDANGERED MANATEE In Exercise 3.6 (page 125) you made a scatterplot of powerboats registered in Florida and manatees killed by boats.

(a) Describe the direction of the relationship. Are the variables positively or negatively associated?

(b) Describe the form of the relationship. Is it linear?

(c) Describe the strength of the relationship. Can the number of manatees killed be predicted accurately from powerboat registrations? If powerboat registrations remained constant at 719,000, about how many manatees would be killed by boats each year?

3.10 MORE JET SKIS In Exercise 3.7 (page 125) you made a scatterplot of jet skis in use and number of accidents.

(a) Describe the direction of the relationship. Are the variables positively or negatively associated?

(b) Describe the form of the association. Is it linear?

3.11 DOES FAST DRIVING WASTE FUEL? How does the fuel consumption of a car change as its speed increases? Here are data for a British Ford Escort. Speed is measured in kilometers per hour, and fuel consumption is measured in liters of gasoline used per 100 kilometers traveled.[2]

Speed (km/h)	Fuel used (liters/100 km)	Speed (km/h)	Fuel used (liters/100 km)
10	21.00	90	7.57
20	13.00	100	8.27
30	10.00	110	9.03
40	8.00	120	9.87
50	7.00	130	10.79
60	5.90	140	11.77
70	6.30	150	12.83
80	6.95		

(a) Make a scatterplot. (Which is the explanatory variable?)

(b) Describe the form of the relationship. Why is it not linear? Explain why the form of the relationship makes sense.

(c) It does not make sense to describe the variables as either positively associated or negatively associated. Why?

(d) Is the relationship reasonably strong or quite weak? Explain your answer.

Adding categorical variables to scatterplots

The South has long lagged behind the rest of the United States in the performance of its schools. Efforts to improve education have reduced the gap. We wonder if the South stands out in our study of state average SAT scores.

EXAMPLE 3.5 IS THE SOUTH DIFFERENT?

Figure 3.3 enhances the scatterplot in Figure 3.1 by plotting the southern states with plus signs. (We took the South to be the states in the East South Central and South Atlantic regions.) Most of the southern states blend in with the rest of the country. Several southern states do lie at the lower edges of their clusters, along with the District of Columbia, which is a city rather than a state. Georgia, South Carolina, and West Virginia have lower SAT scores than we would expect from the percent of their high school graduates who take the examination.

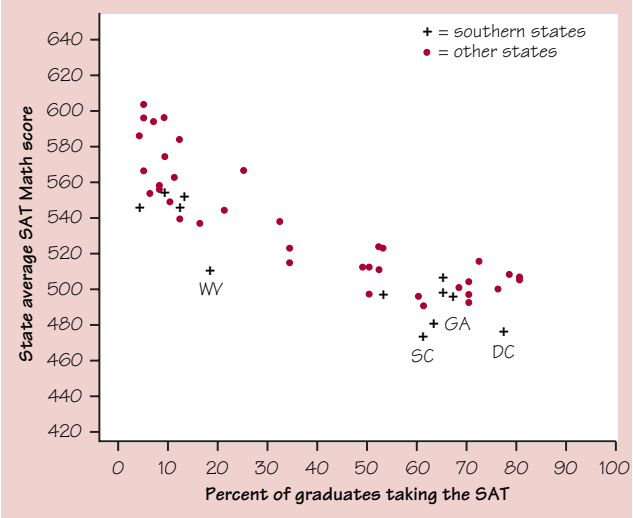


FIGURE 3.3 Average SAT Math score and percent of high school graduates who take the test, by state, with the southern states highlighted.

Dividing the states into "southern" and "nonsouthern" introduces a third variable into the scatterplot. This is a categorical variable that has only two values. The two values are displayed by the two different plotting symbols. Use different colors or symbols to plot points when you want to add a categorical variable to a scatterplot.[3]

EXAMPLE 3.6 DO SOLAR PANELS REDUCE GAS USAGE?

After the Sanchez household gathered the information recorded in Table 3.1 and Figure 3.2 (pages 127 and 128), they added solar panels to their house. They then measured their natural gas consumption for 23 more months. To see how the solar panels affected gas consumption, add the degree-days and gas consumption for these months to the scatterplot. Figure 3.4 is the result. We use different symbols to distinguish before from after. The "after" data form a linear pattern that is close to the "before" pattern in warm months (few degree-days). In colder months, with more degree-days, gas consumption after installing the solar panels is less than in similar months before the panels were added. The scatterplot shows the energy savings from the panels.

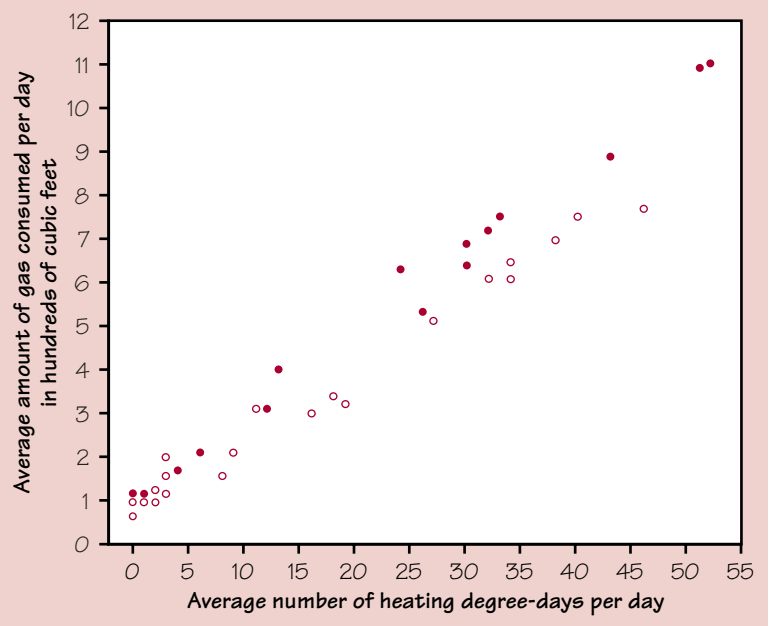


FIGURE 3.4 Natural gas consumption against degree-days for the Sanchez household. The observations indicated by filled circles are for 16 months before installing solar panels. The observations indicated by open circles are for 23 months with the panels in use.

Our gas consumption example suffers from a common problem in drawing scatterplots that you may not notice when a computer does the work. When several individuals have exactly the same data, they occupy the same point on the scatterplot. Look at June and July in Table 3.1. Table 3.1 contains data for 16 months, but there are only 15 points in Figure 3.2. June and July both occupy the same point. You can use a different plotting symbol to call attention to points that stand for more than one individual. Some computer software does this automatically, but some does not. We recommend that you do use a different symbol for repeated observations when you plot a small number of observations by hand.

EXERCISES

3.12 DO HEAVIER PEOPLE BURN MORE ENERGY? Metabolic rate, the rate at which the body consumes energy, is important in studies of weight gain, dieting, and exercise. Table 3.2 gives data on the lean body mass and resting metabolic rate for 12 women and 7 men who are subjects in a study of dieting. Lean body mass, given in kilograms, is a person's weight leaving out all fat. Metabolic rate is measured in calories burned per 24 hours, the same calories used to describe the energy content of foods. The researchers believe that lean body mass is an important influence on metabolic rate.

TABLE 3.2 Lean body mass and metabolic rate

Subject	Sex	Mass (kg)	Rate (cal)	Subject	Sex	Mass (kg)	Rate (cal)
1	M	62.0	1792	11	F	40.3	1189
2	M	62.9	1666	12	F	33.1	913
3	F	36.1	995	13	M	51.9	1460
4	F	54.6	1425	14	F	42.4	1124
5	F	48.5	1396	15	F	34.5	1052
6	F	42.0	1418	16	F	51.1	1347
7	M	47.4	1362	17	F	41.2	1204
8	F	50.6	1502	18	M	51.9	1867
9	F	42.0	1256	19	M	46.9	1439
10	M	48.7	1614				

(a) Make a scatterplot of the data for the female subjects. Which is the explanatory variable?

(b) Is the association between these variables positive or negative? What is the form of the relationship? How strong is the relationship?

(c) Now add the data for the male subjects to your graph, using a different color or a different plotting symbol. Does the pattern of relationship that you observed in **(b)** hold for men also? How do the male subjects as a group differ from the female subjects as a group?

TECHNOLOGY TOOLBOX *Making a calculator scatterplot*

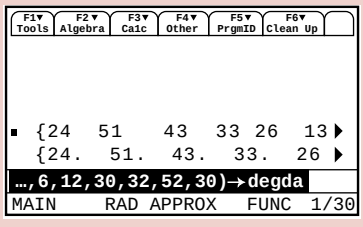
We will use the gas consumption data from Example 3.4 to show how to construct a scatterplot on the TI-83/89.

- Begin by entering the degree-days data and assigning the values to a list named DEGDA, as shown. Then press ENTER.

TI-83

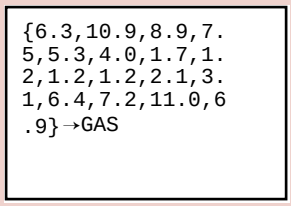
```
{24,51,43,33,26,
13,4,0,0,1,6,12,
30,32,52,30}→DEG
DA
```

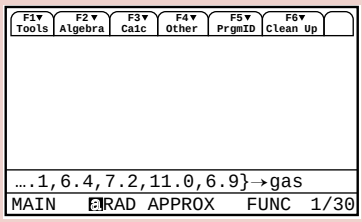
TI-89



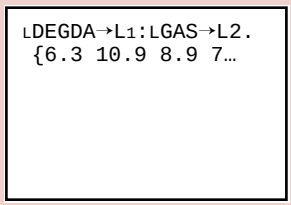
TECHNOLOGY TOOLBOX *Making a calculator scatterplot (continued)*

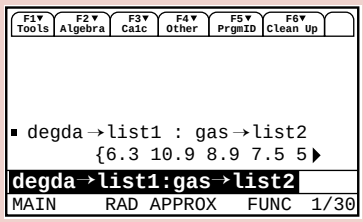
• Then enter the gas consumption data and assign them to the list GAS. Press ENTER.



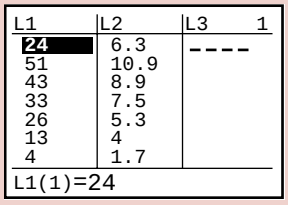


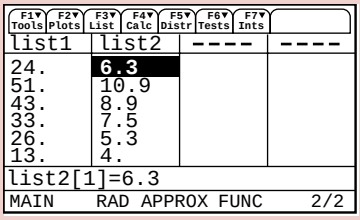
• These two lists are now saved in the calculator for later use. To make things easier, let's transfer the DEGDA data into list1 (L_1 on the TI-83) and the GAS data into list2. The named lists can be found in the LIST menu on the TI-83 and in the VAR-LINK menu on the TI-89.



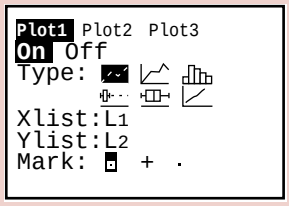


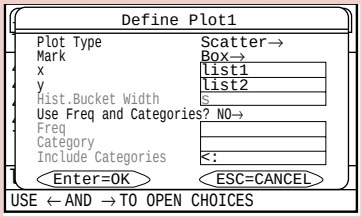
• You can verify that the two lists of data are now in L_1/list1 and L_2/list2 in the Statistics/List Editor.





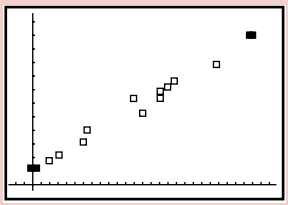
• Next, define a scatterplot in the statistics plot menu (press F2 on the TI-89). Specify the settings shown.

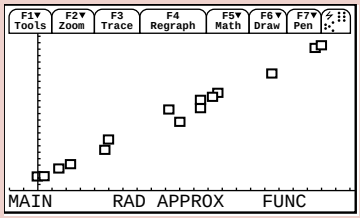




TECHNOLOGY TOOLBOX *Making a calculator scatterplot (continued)*

• Use ZoomStat (ZoomData on the TI-89) to obtain the graph. The calculator will set the window dimensions automatically by looking at the values in L_1/list1 and L_2/list2.





• Notice that there are no scales on the axes, and that the axes are not labeled. If you copy a scatterplot from your calculator onto your paper, make sure that you scale and label the axes. You can use TRACE to help you get started.

3.13 SCATTERPLOT BY CALCULATOR, I Rework Exercise 3.11 (page 129) using your calculator. The command `seq(10X,X,1,15)→SPEED` will create a list named SPEED and assign the numbers 10, 20, . . ., 150 to the list. (Note that `seq` is found under 2nd / LIST / OPS on the TI-83 and under CATALOG on the TI-89). Then assign the fuel data to the list FUEL, and copy the list SPEED to L_1/list1 and the list FUEL to L_2/list2. Define Plot 1 to be a scatterplot, and then ZOOM / 9:ZoomStat (ZoomData on the TI-89) to graph it. Verify your answers to Exercise 3.11.

3.14 SCATTERPLOT BY CALCULATOR, II Rework Exercise 3.12 (page 132) using your calculator. Verify your answers to Exercise 3.12.

SUMMARY

To study relationships between variables, we must measure the variables on the same group of individuals.

If we think that a variable x may explain or even cause changes in another variable y, we call x an **explanatory variable** and y a **response variable.**

A **scatterplot** displays the relationship between two quantitative variables measured on the same individuals. Mark values of one variable on the horizontal axis (x axis) and values of the other variable on the vertical axis (y axis). Plot each individual's data as a point on the graph.

Always plot the explanatory variable, if there is one, on the x axis of a scatterplot. Plot the response variable on the y axis.

Plot points with different colors or symbols to see the effect of a categorical variable in a scatterplot.

In examining a scatterplot, look for an overall pattern showing the **form, direction,** and **strength** of the relationship, and then for **outliers** or other deviations from this pattern.

Form: Linear relationships, where the points show a straight-line pattern, are an important form of relationship between two variables. Curved relationships and **clusters** are other forms to watch for.

Direction: If the relationship has a clear direction, we speak of either **positive association** (high values of the two variables tend to occur together) or **negative association** (high values of one variable tend to occur with low values of the other variable).

Strength: The **strength** of a relationship is determined by how close the points in the scatterplot lie to a simple form such as a line.

SECTION 3.1 EXERCISES

3.15 IQ AND SCHOOL GRADES Do students with higher IQ test scores tend to do better in school? Figure 3.5 is a scatterplot of IQ and school grade point average (GPA) for all 78 seventh-grade students in a rural Midwest school.[4]

(a) Say in words what a positive association between IQ and GPA would mean. Does the plot show a positive association?

(b) What is the form of the relationship? Is it roughly linear? Is it very strong? Explain your answers.

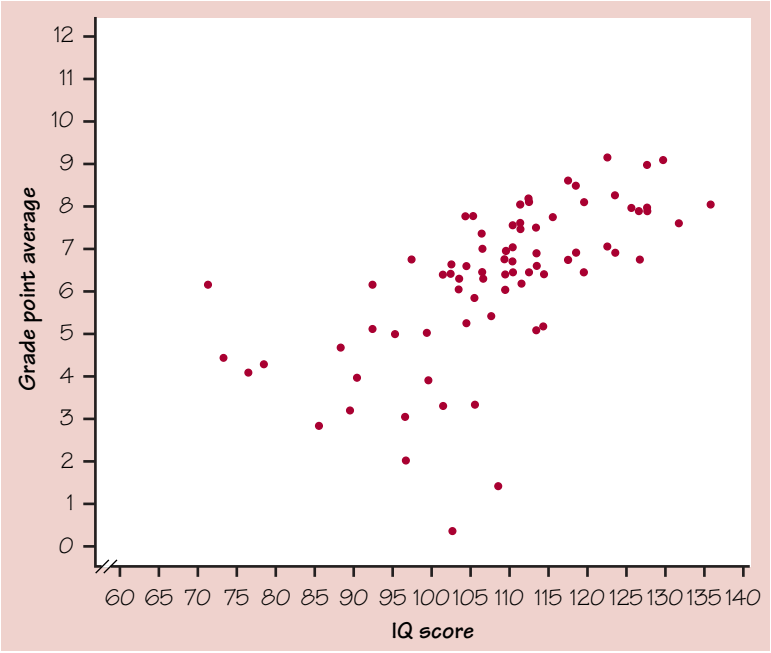


FIGURE 3.5 Scatterplot of school grade point average versus IQ test score for seventh-grade students.

(c) At the bottom of the plot are several points that we might call outliers. One student in particular has a very low GPA despite an average IQ score. What are the approximate IQ and GPA for this student?

3.16 CALORIES AND SALT IN HOT DOGS Are hot dogs that are high in calories also high in salt? Figure 3.6 is a scatterplot of the calories and salt content (measured as milligrams of sodium) in 17 brands of meat hot dogs.[5]

(a) Roughly what are the lowest and highest calorie counts among these brands? Roughly what is the sodium level in the brands with the fewest and with the most calories?

(b) Does the scatterplot show a clear positive or negative association? Say in words what this association means about calories and salt in hot dogs.

(c) Are there any outliers? Is the relationship (ignoring any outliers) roughly linear in form? Still ignoring outliers, how strong would you say the relationship between calories and sodium is?

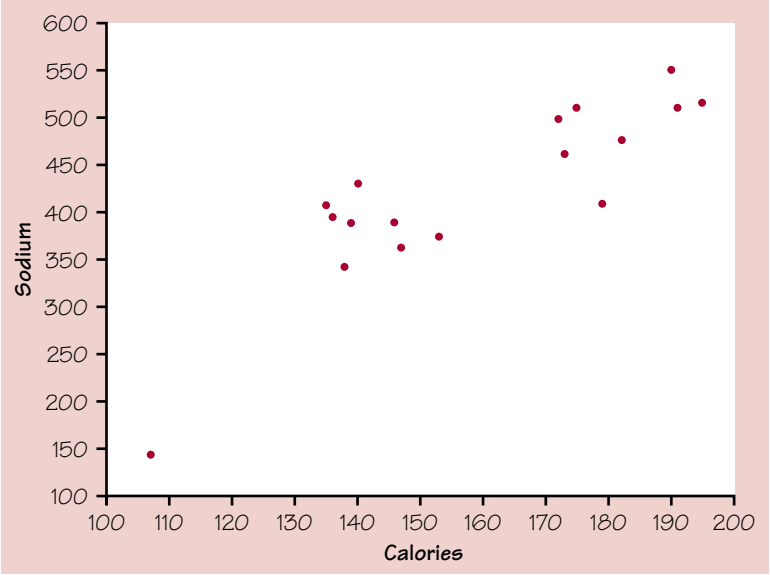


FIGURE 3.6 Scatterplot of milligrams of sodium and calories in each of 17 brands of meat hot dogs.

3.17 RICH STATES, POOR STATES One measure of a state's prosperity is the median income of its households. Another measure is the mean personal income per person in the state. Figure 3.7 is a scatterplot of these two variables, both measured in thousands of dollars. Because both variables have the same units, the plot uses equally spaced scales on both axes.[6]

(a) We have labeled the point for New York on the scatterplot. What are the approximate values of New York's median household income and mean income per person?

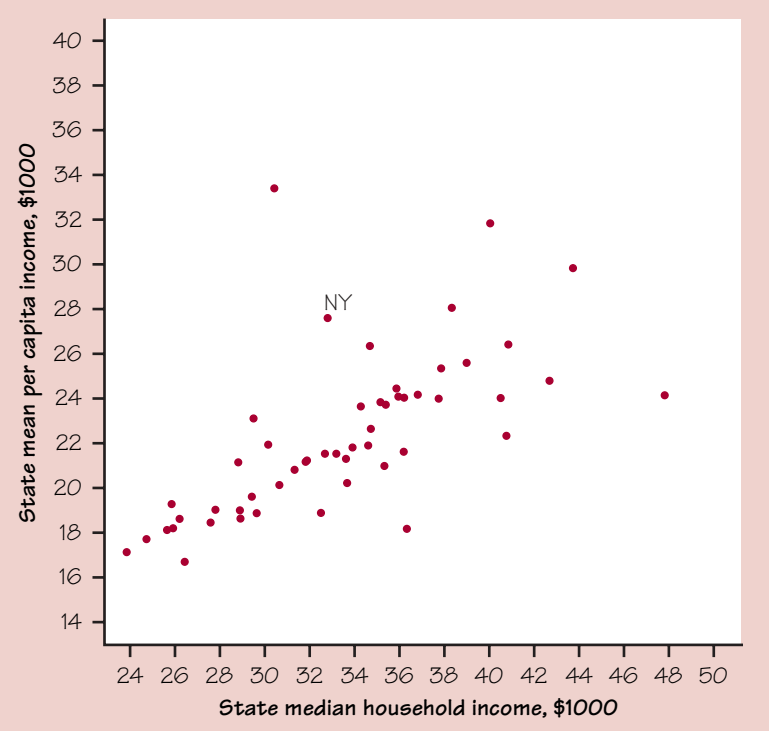


FIGURE 3.7 Scatterplot of mean income per person versus median household income for the states.

(b) Explain why you expect a positive association between these variables. Also explain why you expect household income to be generally higher than income per person.

(c) Nonetheless, the mean income per person in a state can be higher than the median household income. In fact, the District of Columbia has median income $30,748 per household and mean income $33,435 per person. Explain why this can happen.

(d) Alaska is the state with the highest median household income. What is the approximate median household income in Alaska? We might call Alaska and the District of Columbia outliers in the scatterplot.

(e) Describe the form, direction, and strength of the relationship, ignoring the outliers.

3.18 THE PROFESSOR SWIMS Professor Moore swims 2000 yards regularly in a vain attempt to undo middle age. Here are his times (in minutes) and his pulse rate after swimming (in beats per minute) for 23 sessions in the pool:

Time:	34.12	35.72	34.72	34.05	34.13	35.72	36.17	35.57	35.37
Pulse:	152	124	140	152	146	128	136	144	148
Time:	35.57	35.43	36.05	34.85	34.70	34.75	33.93	34.60	34.00
Pulse:	144	136	124	148	144	140	156	136	148
Time:	34.35	35.62	35.68	35.28	35.97				
Pulse:	148	132	124	132	139				

(a) Make a scatterplot. (Which is the explanatory variable?)

(b) Is the association between these variables positive or negative? Explain why you expect the relationship to have this direction.

(c) Describe the form and strength of the relationship.

3.19 MEET THE *ARCHAEOPTERYX* *Archaeopteryx* is an extinct beast having feathers like a bird but teeth and a long bony tail like a reptile. Only six fossil specimens are known. Because these specimens differ greatly in size, some scientists think they are different species rather than individuals from the same species. We will examine some data. If the specimens belong to the same species and differ in size because some are younger than others, there should be a positive linear relationship between the lengths of a pair of bones from all individuals. An outlier from this relationship would suggest a different species. Here are data on the lengths in centimeters of the femur (a leg bone) and the humerus (a bone in the upper arm) for the five specimens that preserve both bones:[7]

Femur:	38	56	59	64	74
Humerus:	41	63	70	72	84

Make a scatterplot. Do you think that all five specimens come from the same species?

3.20 DO YOU KNOW YOUR CALORIES? A food industry group asked 3368 people to guess the number of calories in each of several common foods. Here is a table of the average of their guesses and the correct number of calories:[8]

Food	Guessed calories	Correct calories
8 oz. whole milk	196	159
5 oz. spaghetti with tomato sauce	394	163
5 oz. macaroni with cheese	350	269
One slice wheat bread	117	61
One slice white bread	136	76
2-oz. candy bar	364	260
Saltine cracker	74	12
Medium-size apple	107	80
Medium-size potato	160	88
Cream-filled snack cake	419	160

(a) We think that how many calories a food actually has helps explain people's guesses of how many calories it has. With this in mind, make a scatterplot of these data. (Because both variables are measured in calories, you should use the same scale on both axes. Your plot will be square.)

(b) Describe the relationship. Is there a positive or negative association? Is the relationship approximately linear? Are there any outliers?

3.21 MAXIMIZING CORN YIELDS How much corn per acre should a farmer plant to obtain the highest yield? Too few plants will give a low yield. On the other hand, if there are

too many plants, they will compete with each other for moisture and nutrients, and yields will fall. To find the best planting rate, plant at different rates on several plots of ground and measure the harvest. (Be sure to treat all the plots the same except for the planting rate.) Here are the data from such an experiment:[9]

Plants per acre	Yield (bushels per acre)			
12,000	150.1	113.0	118.4	142.6
16,000	166.9	120.7	135.2	149.8
20,000	165.3	130.1	139.6	149.9
24,000	134.7	138.4	156.1	
28,000	119.0	150.5		

(a) Is yield or planting rate the explanatory variable?

(b) Make a scatterplot of yield and planting rate.

(c) Describe the overall pattern of the relationship. Is it linear? Is there a positive or negative association, or neither?

(d) Find the mean yield for each of the five planting rates. Plot each mean yield against its planting rate on your scatterplot and connect these five points with lines. This combination of numerical description and graphing makes the relationship clearer. What planting rate would you recommend to a farmer whose conditions were similar to those in the experiment?

3.22 TEACHERS' PAY Table 1.15 (page 70) gives data for the states. We might expect that states with less educated populations would pay their teachers less, perhaps because these states are poorer.

(a) Make a scatterplot of average teachers' pay against the percent of state residents who are not high school graduates. Take the percent with no high school degree as the explanatory variable.

(b) The plot shows a weak negative association between the two variables. Why do we say that the association is negative? Why do we say that it is weak?

(c) Circle on the plot the point for the state your school is in.

(d) There is an outlier at the upper left of the plot. Which state is this?

(e) We wonder about regional patterns. There is a relatively clear cluster of nine states at the lower right of the plot. These states have many residents who are not high school graduates and pay low salaries to teachers. Which states are these? Are they mainly from one part of the country?

3.23 CATEGORICAL EXPLANATORY VARIABLE A scatterplot shows the relationship between two quantitative variables. Here is a similar plot to study the relationship between a categorical explanatory variable and a quantitative response variable.

The presence of harmful insects in farm fields is detected by putting up boards covered with a sticky material and then examining the insects trapped on the board. Which colors attract insects best? Experimenters placed six boards of each of four colors in a field of oats and measured the number of cereal leaf beetles trapped.[10]

Board color	Insects trapped					
Lemon yellow	45	59	48	46	38	47
White	21	12	14	17	13	17
Green	37	32	15	25	39	41
Blue	16	11	20	21	14	07

(a) Make a plot of the counts of insects trapped against board color (space the four colors equally on the horizontal axis). Compute the mean count for each color, add the means to your plot, and connect the means with line segments.

(b) Based on the data, what do you conclude about the attractiveness of these colors to the beetles?

(c) Does it make sense to speak of a positive or negative association between board color and insect count?

3.2 CORRELATION

A scatterplot displays the direction, form, and strength of the relationship between two quantitative variables. Linear relations are particularly important because a straight line is a simple pattern that is quite common. We say a linear relation is strong if the points lie close to a straight line, and weak if they are widely scattered about a line. Our eyes are not good judges of how strong a linear relationship is. The two scatterplots in Figure 3.8 depict exactly the same data, but the lower plot is drawn smaller in a large field. The lower plot seems to show a stronger linear relationship. Our eyes can be fooled by changing the plotting scales or the amount of white space around the cloud of points in a scatterplot.[11] We need to follow our strategy for data analysis by using a numerical measure to supplement the graph. *Correlation* is the measure we use.

CORRELATION *r*

The **correlation** measures the direction and strength of the linear relationship between two quantitative variables. Correlation is usually written as r.

Suppose that we have data on variables x and y for n individuals. The values for the first individual are x_1 and y_1, the values for the second individual are x_2 and y_2, and so on. The means and standard deviations of the two variables are $\bar{x}$ and s_x for the x-values, and $\bar{y}$ and s_y for the y-values. The correlation r between x and y is

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

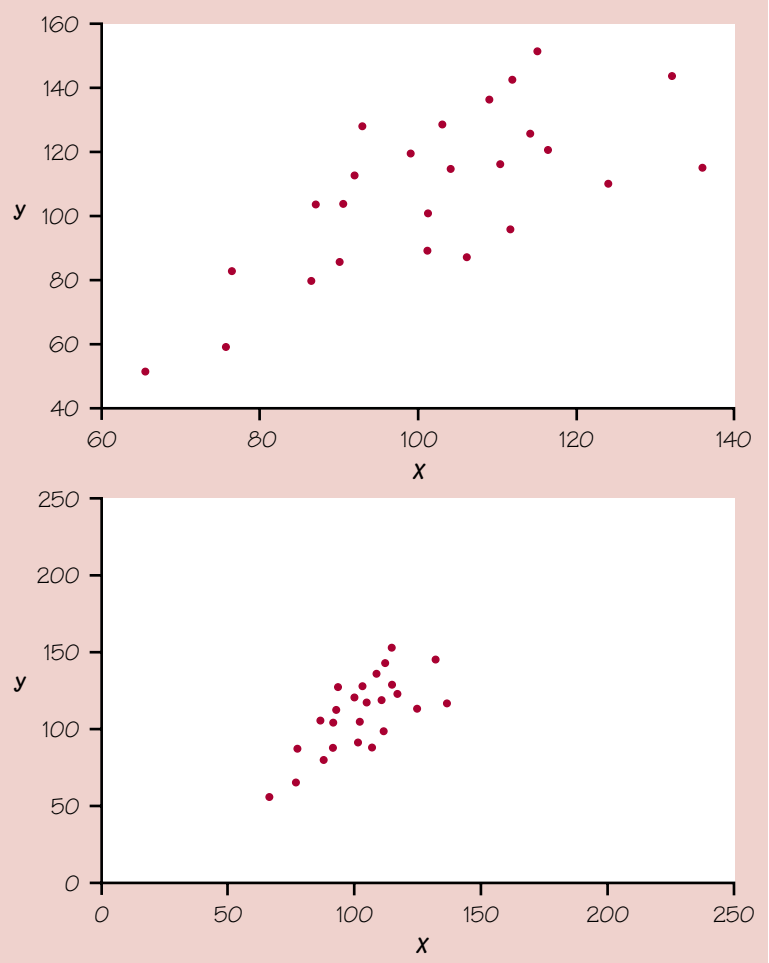


FIGURE 3.8 Two scatterplots of the same data; the straight-line pattern in the lower plot appears stronger because of the surrounding white space.

As always, the summation sign Σ means "add these terms for all the individuals." The formula for the correlation r is a bit complex. It helps us see what correlation is, but in practice you should use software or a calculator that finds r from keyed-in values of two variables x and y. Exercise 3.24 asks you to calculate a correlation step-by-step from the definition to solidify its meaning.

The formula for r begins by standardizing the observations. Suppose, for example, that x is height in centimeters and y is weight in kilograms and that we have height and weight measurements for n people. Then $\bar{x}$ and s_x are the mean and standard deviation of the n heights, both in centimeters. The value

$$\frac{x_i - \bar{x}}{s_x}$$

is the standardized height of the ith person, familiar from Chapter 2. The standardized height says how many standard deviations above or below the mean a person's height lies. Standardized values have no units—in this example, they are no longer measured in centimeters. Standardize the weights also. The correlation r is an average of the products of the standardized height and the standardized weight for the n people.

EXERCISE

3.24 CLASSIFYING FOSSILS Exercise 3.19 (page 138) gives the lengths of two bones in five fossil specimens of the extinct beast *Archaeopteryx:*

Femur:	38	56	59	64	74
Humerus:	41	63	70	72	84

(a) Find the correlation r step-by-step. That is, find the mean and standard deviation of the femur lengths and of the humerus lengths. Then find the five standardized values for each variable and use the formula for r.

(b) Duplicate the steps in the Technology Toolbox below to obtain the correlation for the *Archaeopteryx* data, and compare your result with that calculated by hand in (a).

TECHNOLOGY TOOLBOX *Using the definition to calculate correlation*

We will use the *Archaeopteryx* data to show how to calculate the correlation using the definition and the list features of the TI-83/89.

• Begin by entering the femur lengths (x-values) in L_1/list1 and the humerus lengths (y-values) in L_2/list2. Then calculate two-variable statistics for the x- and y-values. The calculator will remember all of the computed statistics until the next time you calculate one- or two-variable statistics.

TI-83

• Press STAT, choose CALC, then `2:2-Var Stats`.

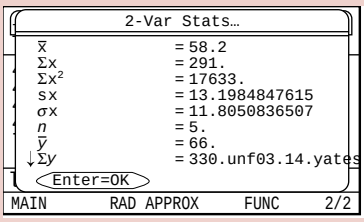
• Complete the command `2-Var Stats` L_1, L_2, and press ENTER.

```
2-Var Stats
 x̄=58.2
 Σx=291
 Σx²=17633
 Sx=13.19848476
 σx=11.80508365
↓n=5
```

TI-89

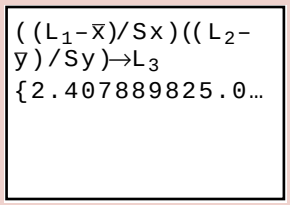
• In the Statistics/List Editor, press F4 and choose `2:2-Var Stats`.

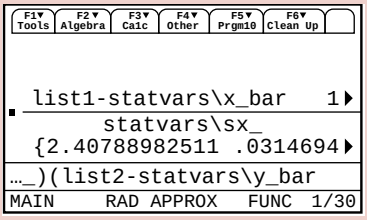
• In the new window, enter list1 as the Xlist and list2 as the Ylist, then press ENTER.



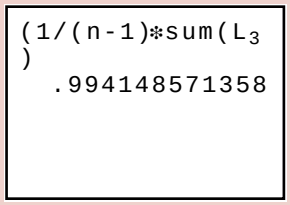
TECHNOLOGY TOOLBOX *Using the definition to calculate correlation (continued)*

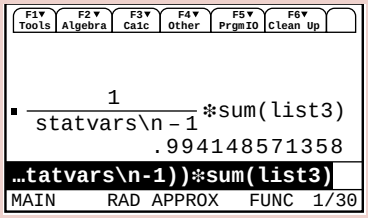
• Next, define L_3/list3 = ((list1 – $\bar{x}$)/s_x)((list2 – $\bar{y}$)/s_y) from the home screen as shown. Note that $\bar{x}$, $\bar{y}$, s_x, and s_y can be found under VARS/5:Statistics (in the VAR-LINK menu on the TI-89).





• To complete the formula for the correlation $r = \frac{1}{n-1}\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$, enter the command shown in the (two) calculator screens. Press ENTER to see the correlation.





Facts about correlation

The formula for correlation helps us see that r is positive when there is a positive association between the variables. Height and weight, for example, have a positive association. People who are above average in height tend to also be above average in weight. Both the standardized height and the standardized weight are positive. People who are below average in height tend to also have below-average weight. Then both standardized height and standardized weight are negative. In both cases, the products in the formula for r are mostly positive and so r is positive. In the same way, we can see that r is negative when the association between x and y is negative. More detailed study of the formula gives more detailed properties of r. Here is what you need to know in order to interpret correlation.

1. Correlation makes no distinction between explanatory and response variables. It makes no difference which variable you call x and which you call y in calculating the correlation.

2. Correlation requires that both variables be quantitative, so that it makes sense to do the arithmetic indicated by the formula for r. We cannot calculate

a correlation between the incomes of a group of people and what city they live in, because city is a categorical variable.

3. Because r uses the standardized values of the observations, r does not change when we change the units of measurement of x, y, or both. Measuring height in inches rather than centimeters and weight in pounds rather than kilograms does not change the correlation between height and weight. The correlation r itself has no unit of measurement; it is just a number.

4. Positive r indicates positive association between the variables, and negative r indicates negative association.

5. The correlation r is always a number between –1 and 1. Values of r near 0 indicate a very weak linear relationship. The strength of the linear relationship increases as r moves away from 0 toward either –1 or 1. Values of r close to –1 or 1 indicate that the points in a scatterplot lie close to a straight line. The extreme values $r = -1$ and $r = 1$ occur only in the case of a perfect linear relationship, when the points lie exactly along a straight line.

6. Correlation measures the strength of only a linear relationship between two variables. Correlation does not describe curved relationships between variables, no matter how strong they are.

7. Like the mean and standard deviation, the correlation is not resistant: r is strongly affected by a few outlying observations. The correlation for Figure 3.7 (page 137) is $r = 0.634$ when all 51 observations are included, but rises to $r = 0.783$ when we omit Alaska and the District of Columbia. Use r with caution when outliers appear in the scatterplot.

The scatterplots in Figure 3.9 illustrate how values of r closer to 1 or –1 correspond to stronger linear relationships. To make the meaning of r clearer, the standard deviations of both variables in these plots are equal and the horizontal and vertical scales are the same. In general, it is not so easy to guess the value of r from the appearance of a scatterplot. Remember that changing the plotting scales in a scatterplot may mislead our eyes, but it does not change the correlation.

The real data we have examined also illustrate how correlation measures the strength and direction of linear relationships. Figure 3.2 (page 128) shows a very strong positive linear relationship between degree-days and natural gas consumption. The correlation is $r = 0.9953$. Check this on your calculator using the data in Table 3.1. Figure 3.1 (page 124) shows a clear but weaker negative association between percent of students taking the SAT and the median SAT Math score in a state. The correlation is $r = -0.868$.

Do remember that **correlation is not a complete description of two-variable data,** even when the relationship between the variables is linear. You should give the means and standard deviations of both x and y along with the correlation. (Because the formula for correlation uses the means and standard deviations, these measures are the proper choice to accompany a correlation.) Conclusions based on correlations alone may require rethinking in the light of a more complete description of the data.

FIGURE 3.9 How correlation measures the strength of a linear relationship. Patterns closer to a straight line have correlations closer to 1 or –1.

EXAMPLE 3.7 SCORING DIVERS

Competitive divers are scored on their form by a panel of judges who use a scale from 1 to 10. The subjective nature of the scoring often results in controversy. We have the scores awarded by two judges, Ivan and George, on a large number of dives. How well do they agree? We do some calculation and find that the correlation between their scores is $r = 0.9$. But the mean of Ivan's scores is 3 points lower than George's mean.

These facts do not contradict each other. They are simply different kinds of information. The mean scores show that Ivan awards much lower scores than George. But because Ivan gives *every* dive a score about 3 points lower than George, the correlation remains high. Adding or subtracting the same number to all values of either x or y does not change the correlation. If Ivan and George both rate several divers, the contest is fairly scored because Ivan and George agree on which dives are better than others. The high r shows their agreement. But if Ivan scores one diver and George another, we must add 3 points to Ivan's scores to arrive at a fair comparison.

EXERCISES

3.25 THINKING ABOUT CORRELATION Figure 3.5 (page 135) is a scatterplot of school grade point average versus IQ score for 78 seventh-grade students.

(a) Is the correlation r for these data near –1, clearly negative but not near –1, near 0, clearly positive but not near 1, or near 1? Explain your answer.

(b) Figure 3.6 (page 136) shows the calories and sodium content in 17 brands of meat hot dogs. Is the correlation here closer to 1 than that for Figure 3.5, or closer to zero? Explain your answer.

(c) Both Figures 3.5 and 3.6 contain outliers. Removing the outliers will *increase* the correlation r in one figure and *decrease* r in the other figure. What happens in each figure, and why?

3.26 If women always married men who were 2 years older than themselves, what would be the correlation between the ages of husband and wife? (*Hint:* Draw a scatterplot for several ages.)

3.27 RETURN OF THE *ARCHAEOPTERYX* Exercise 3.19 (page 138) gives the lengths of two bones in five fossil specimens of the extinct beast *Archaeopteryx*. You found the correlation r in Exercise 3.24 (page 142).

(a) Make a scatterplot if you did not do so earlier. Explain why the value of r matches the scatterplot.

(b) The lengths were measured in centimeters. If we changed to inches, how would r change? (There are 2.54 centimeters in an inch.)

3.28 STRONG ASSOCIATION BUT NO CORRELATION The gas mileage of an automobile first increases and then decreases as the speed increases. Suppose that this relationship is very regular, as shown by the following data on speed (miles per hour) and mileage (miles per gallon):

Speed:	20	30	40	50	60
MPG:	24	28	30	28	24

Make a scatterplot of mileage versus speed. Show that the correlation between speed and mileage is $r = 0$. Explain why the correlation is 0 even though there is a strong relationship between speed and mileage.

SUMMARY

The **correlation** r measures the strength and direction of the linear association between two quantitative variables x and y. Although you can calculate a correlation for any scatterplot, r measures only straight-line relationships.

Correlation indicates the direction of a linear relationship by its sign: $r > 0$ for a positive association and $r < 0$ for a negative association.

Correlation always satisfies $-1 \le r \le 1$ and indicates the strength of a relationship by how close it is to –1 or 1. Perfect correlation, $r = \pm 1$, occurs only when the points on a scatterplot lie exactly on a straight line.

Correlation ignores the distinction between explanatory and response variables. The value of r is not affected by changes in the unit of measurement of either variable. Correlation is not resistant, so outliers can greatly change the value of r.

SECTION 3.2 EXERCISES

3.29 THE PROFESSOR SWIMS Exercise 3.18 (page 137) gives data on the time to swim 2000 yards and the pulse rate after swimming for a middle-aged professor.

(a) If you did not do Exercise 3.18, do it now. Find the correlation r. Explain from looking at the scatterplot why this value of r is reasonable.

(b) Suppose that the times had been recorded in seconds. For example, the time 34.12 minutes would be 2047 seconds. How would the value of r change?

3.30 BODY MASS AND METABOLIC RATE Exercise 3.12 (page 132) gives data on the lean body mass and metabolic rate for 12 women and 7 men.

(a) Make a scatterplot if you did not do so in Exercise 3.12. Use different symbols or colors for women and men. Do you think the correlation will be about the same for men and women or quite different for the two groups? Why?

(b) Calculate r for women alone and also for men alone. (Use your calculator.)

(c) Calculate the mean body mass for the women and for the men. Does the fact that the men are heavier than the women on the average influence the correlations? If so, in what way?

(d) Lean body mass was measured in kilograms. How would the correlations change if we measured body mass in pounds? (There are about 2.2 pounds in a kilogram.)

3.31 HOW MANY CALORIES? Exercise 3.20 (page 138) gives data on the true calorie counts in ten foods and the average guesses made by a large group of people.

(a) Make a scatterplot if you did not do so in Exercise 3.20. Then calculate the correlation r (use your calculator). Explain why your r is reasonable based on the scatterplot.

(b) The guesses are all higher than the true calorie counts. Does this fact influence the correlation in any way? How would r change if every guess were 100 calories higher?

(c) The guesses are much too high for spaghetti and snack cake. Circle these points on your scatterplot. Calculate r for the other eight foods, leaving out these two points. Explain why r changed in the direction that it did.

3.32 BRAIN SIZE AND IQ SCORE Do people with larger brains have higher IQ scores? A study looked at 40 volunteer subjects, 20 men and 20 women. Brain size was measured

by magnetic resonance imaging. Table 3.3 gives the data. The MRI count is the number of "pixels" the brain covered in the image. IQ was measured by the Wechsler test.[13]

TABLE 3.3 Brain size (MRI count) and IQ score

Men				Women			
MRI	IQ	MRI	IQ	MRI	IQ	MRI	IQ
1,001,121	140	1,038,437	139	816,932	133	951,545	137
965,353	133	904,858	89	928,799	99	991,305	138
955,466	133	1,079,549	141	854,258	92	833,868	132
924,059	135	945,088	100	856,472	140	878,897	96
889,083	80	892,420	83	865,363	83	852,244	132
905,940	97	955,003	139	808,020	101	790,619	135
935,494	141	1,062,462	103	831,772	91	798,612	85
949,589	144	997,925	103	793,549	77	866,662	130
879,987	90	949,395	140	857,782	133	834,344	83
930,016	81	935,863	89	948,066	133	893,983	88

Source: There are some of the data from the EESEE story "Brain Size and Intelligence." The study is described in L. Willerman, R. Schultz, J.N. Rutledge, and E. Bigler, "In vivo brain size and intelligence," *Intelligence*, 15 (1991), pp. 223–228.

(a) Make a scatterplot of IQ score versus MRI count, using distinct symbols for men and women. In addition, find the correlation between IQ and MRI for all 40 subjects, for the men alone, and for the women alone.

(b) Men are larger than women on the average, so they have larger brains. How is this size effect visible in your plot? Find the mean MRI count for men and women to verify the difference.

(c) Your result in **(b)** suggests separating men and women in looking at the relationship between brain size and IQ. Use your work in **(a)** to comment on the nature and strength of this relationship for women and for men.

3.33 Changing the units of measurement can dramatically alter the appearance of a scatterplot. Consider the following data:

x	−4	−4	−3	3	4	4
y	0.5	−0.6	−0.5	0.5	0.5	−0.6

(a) Enter the data into L_1/list1 and L_2/list2. Then use Plot1 to define and plot the scatterplot. Use the box (□) as your plotting symbol.

(b) Use L_3/list3 and the technique described in the Technology Toolbox on page 142 to calculate the correlation.

(c) Define new variables $x^* = x/10$ and $y^* = 10y$, and enter these into L_4/list4 and L_5/list5 as follows: list4 = list1/10 and list5 = 10 × list2. Define Plot2 to be a scatterplot with Xlist: list4 and Ylist: list5, and Mark: +. Plot both scatterplots at the same time, and on the same axes, using ZoomStat/ZoomData. The two plots are very different in appearance.

(d) Use L_6/list6 and the technique described in the Technology Toolbox to calculate the correlation between x^* and y^*. How are the two correlations related? Explain why this isn't surprising.

3.34 TEACHING AND RESEARCH A college newspaper interviews a psychologist about student ratings of the teaching of faculty members. The psychologist says, "The evidence indicates that the correlation between the research productivity and teaching rating of faculty members is close to zero." The paper reports this as "Professor McDaniel said that good researchers tend to be poor teachers, and vice versa." Explain why the paper's report is wrong. Write a statement in plain language (don't use the word "correlation") to explain the psychologist's meaning.

3.35 INVESTMENT DIVERSIFICATION A mutual fund company's newsletter says, "A well-diversified portfolio includes assets with low correlations." The newsletter includes a table of correlations between the returns on various classes of investments. For example, the correlation between municipal bonds and large-cap stocks is 0.50 and the correlation between municipal bonds and small-cap stocks is 0.21.[12]

(a) Rachel invests heavily in municipal bonds. She wants to diversify by adding an investment whose returns do not closely follow the returns on her bonds. Should she choose large-cap stocks or small-cap stocks for this purpose? Explain your answer.

(b) If Rachel wants an investment that tends to increase when the return on her bonds drops, what kind of correlation should she look for?

3.36 DRIVING SPEED AND FUEL CONSUMPTION The data in Exercise 3.28 were made up to create an example of a strong curved relationship for which, nonetheless, $r = 0$. Exercise 3.11 (page 129) gives actual data on gas used versus speed for a small car. Make a scatterplot if you did not do so in Exercise 3.11. Calculate the correlation, and explain why r is close to 0 despite a strong relationship between speed and gas used.

3.37 SLOPPY WRITING ABOUT CORRELATION Each of the following statements contains a blunder. Explain in each case what is wrong.

(a) "There is a high correlation between the gender of American workers and their income."

(b) "We found a high correlation ($r = 1.09$) between students' ratings of faculty teaching and ratings made by other faculty members."

(c) "The correlation between planting rate and yield of corn was found to be $r = 0.23$ bushel."

3.3 LEAST-SQUARES REGRESSION

Correlation measures the strength and direction of the linear relationship between any two quantitative variables. If a scatterplot shows a linear relationship, we would like to summarize this overall pattern by drawing a line through the scatterplot. *Least-squares regression* is a method for finding a line that summarizes the relationship between two variables, but only in a specific setting.

REGRESSION LINE

A **regression line** is a straight line that describes how a response variable y changes as an explanatory variable x changes. We often use a regression line to predict the value of y for a given value of x. Regression, unlike correlation, requires that we have an explanatory variable and a response variable.

model

The least-squares regression line, which we will occasionally abbreviate LSRL, is a ***model***—or more formally, a ***mathematical model***—for the data. If we believe that the data show a linear trend, then it would be appropriate to try to fit an LSRL to the data. In the next chapter, we will explore data that are not linear and for which a curve is a more appropriate model. At the beginning, though, we will focus our discussion on linear trends.

EXAMPLE 3.8 PREDICTING NATURAL GAS CONSUMPTION

A scatterplot shows that there is a strong linear relationship between the average outside temperature (measured by heating degree-days) in a month and the average amount of natural gas that the Sanchez household uses per day during the month. The Sanchez household wants to use this relationship to predict their natural gas consumption. "If a month averages 20 degree-days per day (that's 45° F), how much gas will we use?

prediction

In Figure 3.10 we have drawn a regression line on the scatterplot. To use this line to ***predict*** gas consumption at 20 degree-days, first locate 20 on the x axis. Then go "up

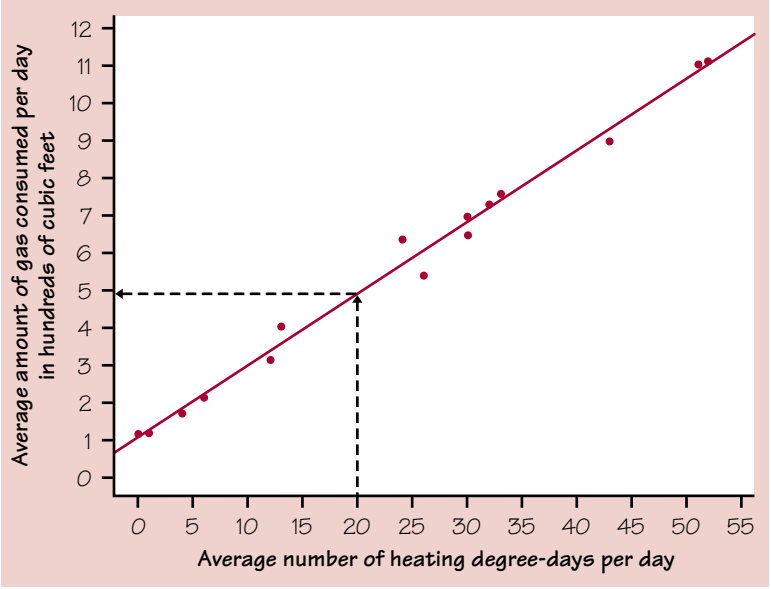


FIGURE 3.10 The Sanchez household gas consumption data, with a regression line for predicting gas consumption from degree-days. The dashed lines illustrate how to use the regression line to predict gas consumption for a month averaging 20 degree-days per day.

and over" as in the figure to find the gas consumption y that corresponds to $x = 20$. We predict that the Sanchez household will use about 4.9 hundreds of cubic feet of gas each day in such a month.

The least-squares regression line

Different people might draw different lines by eye on a scatterplot. This is especially true when the points are more widely scattered than those in Figure 3.10. We need a way to draw a regression line that doesn't depend on our guess as to where the line should go. No line will pass exactly through all the points, so we want one that is as close as possible. We will use the line to predict y from x, so we want a line that is as close as possible to the points in the *vertical* direction. That's because the prediction errors we make are errors in y, which is the vertical direction in the scatterplot. If we predict 4.9 hundreds of cubic feet for a month with 20 degree-days and the actual usage turns out to be 5.1 hundreds of cubic feet, our error is

$$\begin{aligned}\text{error} &= \text{observed} - \text{predicted}\\ &= 5.1 - 4.9 = 0.2\end{aligned}$$

We want a regression line that makes the vertical distances of the points in a scatterplot from the line as small as possible. Figure 3.11(a) illustrates the idea. For clarity, the plot shows only three of the points from Figure 3.10, along with the line, on an expanded scale. The line passes above two of the points and below one of them. The vertical distances of the data points from the line appear as vertical line segments. A "good" regression line makes these distances as small as possible. There are many ways to make "as small as possible" precise. The most common is the *least-squares* idea.

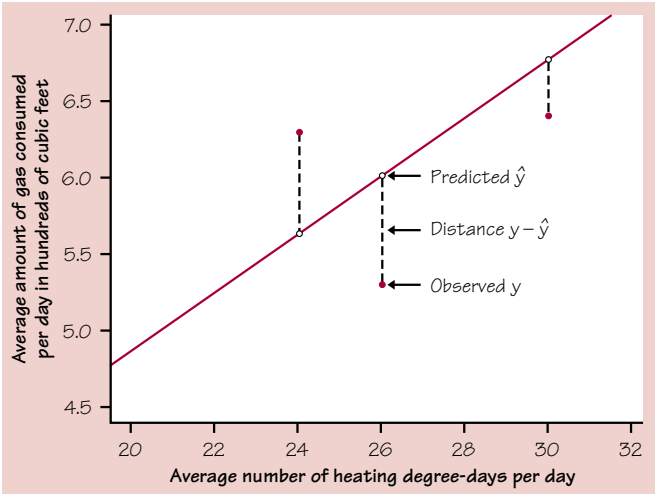


FIGURE 3.11(a) The least-squares idea. For each observation, find the vertical distance of each point on the scatterplot from a regression line. The least-squares regression line makes the sum of the squares of these distances as small as possible.

LEAST-SQUARES REGRESSION LINE

The **least-squares regression line** of y on x is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.

Figure 3.11(b) gives a geometric interpretation to the phrase "sum of the squares of the vertical distances of the data points from the line."

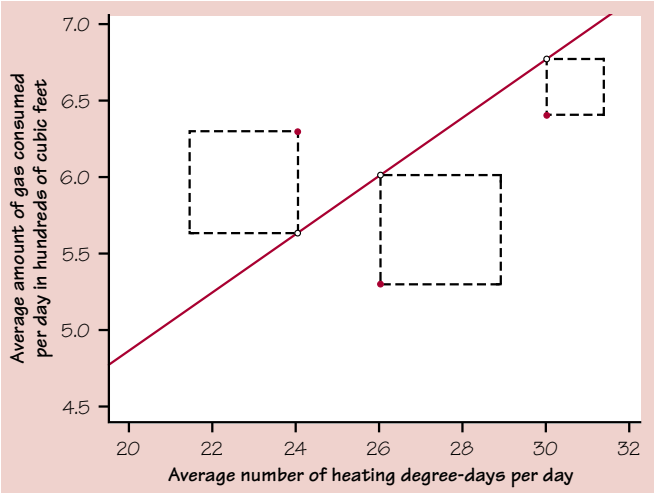


FIGURE 3.11(b) Equivalently, the least-squares regression line is the line that minimizes the total *area* in the squares.

One reason for the popularity of the least-squares regression line is that the problem of finding the line has a simple answer. We can give the recipe for the least-squares line in terms of the means and standard deviations of the two variables and their correlation.

EQUATION OF THE LEAST-SQUARES REGRESSION LINE

We have data on an explanatory variable x and a response variable y for n individuals. From the data, calculate the means $\bar{x}$, and $\bar{y}$ and the standard deviations s_x and s_y of the two variables, and their correlation r. The least-squares regression line is the line

$$\hat{y} = a + bx$$

EQUATION OF THE LEAST-SQUARES REGRESSION LINE *(continued)*

with **slope**

$$b = r\frac{s_y}{s_x}$$

and **intercept**

$$a = \bar{y} - b\bar{x}$$

Although you are probably used to the form $y = mx + b$ for the equation of a line from your study of algebra, statisticians have adopted $\hat{y} = a + bx$ as the form for the equation of the least-squares line. We will adopt this form, too, in the interest of good communication. The variable y denotes the *observed* value of y, and the term $\hat{y}$ means the *predicted* value of y. We write $\hat{y}$ (read "y hat") in the equation of the regression line to emphasize that the line gives a predicted response $\hat{y}$ for any x. When you are solving regression problems, make sure you are careful to distinguish between y and $\hat{y}$.

To determine the equation of a least-squares line, we need to solve for the intercept a and the slope b. Since there are two unknowns, we need two conditions in order to solve for the two unknowns. It can be shown that *every* least-squares regression line passes through the point $(\bar{x}, \bar{y})$. This is one important piece of information about the least-squares line. The other fact that is known is that the slope of the least-squares line is equal to the product of the correlation and the quotient of the standard deviations:

$$b = r\frac{s_y}{s_x}$$

Commit these two facts to memory, and you will be able to find equations of least-squares lines.

EXAMPLE 3.9 CONSTRUCTING THE LEAST-SQUARES EQUATION

Suppose we have explanatory and response variables and we know that $\bar{x} = 17.222$, $\bar{y} = 161.111$, $s_x = 19.696$, $s_y = 33.479$, and the correlation $r = 0.997$. Even though we don't know the actual data, we can still construct the equation for the least-squares line and use it to make predictions. The slope and intercept can be calculated as

$$b = r\frac{s_y}{s_x} = 0.997\frac{33.479}{19.696} = 1.695$$

$$a = \bar{y} - b\bar{x} = 161.111 - (1.695)(17.222) = 131.920$$

so that the least-squares line has equation $\hat{y} = 131.920 + 1.695x$

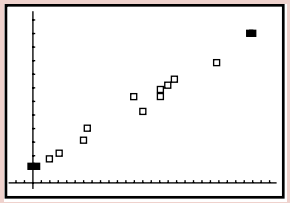
In practice, you don't need to calculate the means, standard deviations, and correlation first. Statistical software or your calculator will give the slope b and intercept a of the least-squares line from keyed-in values of the variables x and y. You can then concentrate on understanding and using the regression line.

TECHNOLOGY TOOLBOX *Least-squares lines on the calculator*

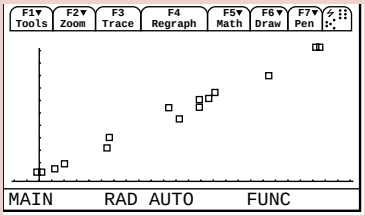
We will use the gas consumption and degree-days data from Example 3.8 to show how to use the TI-83/89 to determine the equation of the least-squares line.

- Enter the degree-days data into L_1/list1 and the gas consumption data into L_2/list2. (Recall that you saved these lists as DEGDA and GAS, respectively.) Refer to the Technology Toolbox on page 132 for details on copying these lists of data into L_1/list1 and L_2/list2.
- Define a scatterplot using L_1/list1 and L_2/list2, and then use ZoomStat (ZoomData) to plot the scatterplot.

TI-83

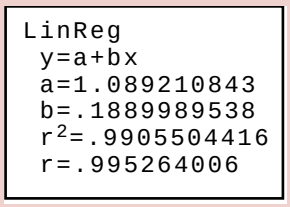


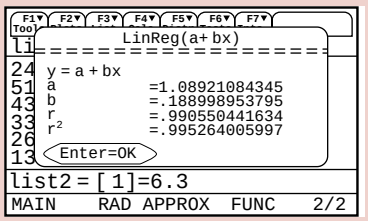
TI-89



To determine the LSRL:

- Press STAT, choose CALC, then `8:LinReg (a+bx).` Finish the command to read `LinReg (a+bx)L1, L2, Y1.` (Y_1 is found under VARS/Y-VARS/1:Function.)
- In the Statistics/ListEditor, press F4 (CALC), choose `3: Regressions,` then `1:LinReg(a+bx).`
- Enter list1 for the Xlist, list2 for the Ylist, choose to store the RegEqn to y1(x) and press ENTER.

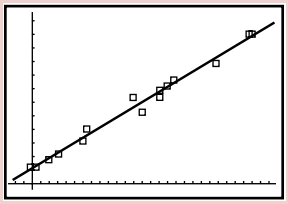


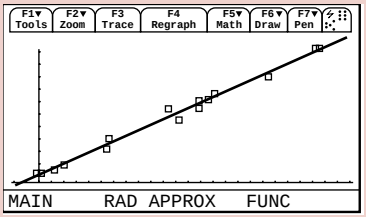


Note: If r^2 and r do not appear on your TI-83 screen, then do this one-time series of keystrokes: Press 2nd 0 (CATALOG), scroll down to `DiagnosticOn` and press ENTER. Press ENTER again to execute the command. The screen should say "Done." Then press 2nd ENTER (ENTRY) to recall the regression command and ENTER again to calculate the LSRL. The r^2- and r-values should now appear.

TECHNOLOGY TOOLBOX *Least-squares lines on the calculator (continued)*

• Deselect all other equations in the Y=screen and press GRAPH (◆ F3 on the TI-89) to overlay the LSRL on the scatterplot.





Although the calculator will report the values for a and b to nine decimal places, we usually round off to four decimal places. You would write the LSRL equation as

$$\hat{y} = 1.0892 + 0.1890x$$

When you write the equation, don't forget the hat symbol over the y; this means *predicted value*.

Figure 3.12 displays the regression output for the gas consumption data from two statistical software packages. Each output records the slope and intercept of the least-squares line, calculated to more decimal places than we need. The software also provides information that we do not yet need—part of the art of using software is to ignore the extra information that is almost always present. We will make use of other parts of the output in Chapters 14 and 15.

slope

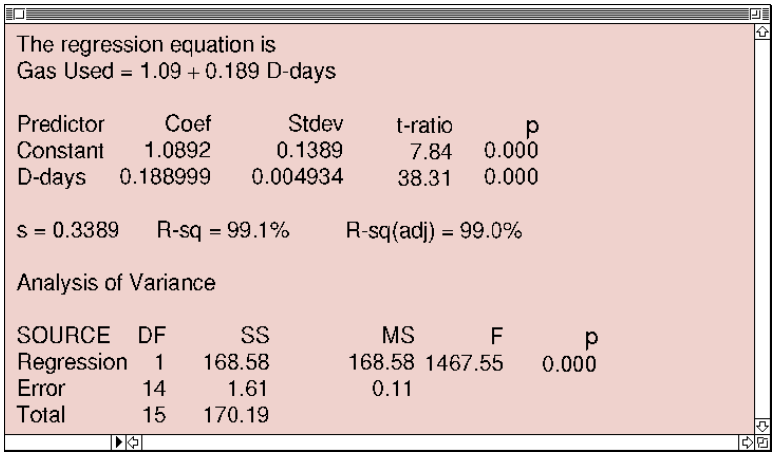
The ***slope*** of a regression line is usually important for the interpretation of the data. The slope is the rate of change, the amount of change in $\hat{y}$ when x increases by 1. The slope $b = 0.1890$ in this example says that, on the average, each additional degree-day predicts consumption of 0.1890 more hundreds of cubic feet of natural gas per day.

intercept

The ***intercept*** of the regression line is the value of $\hat{y}$ when $x = 0$. Although we need the value of the intercept to draw the line, it is statistically meaningful only when x can actually take values close to zero. In our example, $x = 0$ occurs when the average outdoor temperature is at least 65° F. We predict that the Sanchez household will use an average of $a = 1.0892$ hundreds of cubic feet of gas per day when there are no degree-days. They use this gas for cooking and heating water, which continue in warm weather.

The equation of the regression line makes prediction easy. Just substitute an x-value into the equation. To predict gas consumption at 20 degree-days, substitute $x = 20$.

$$\begin{aligned}\hat{y} &= 1.0892 + (0.1890)(20) \\ &= 1.0892 + 3.78 = 4.869\end{aligned}$$



The regression equation is
Gas Used = 1.09 + 0.189 D-days

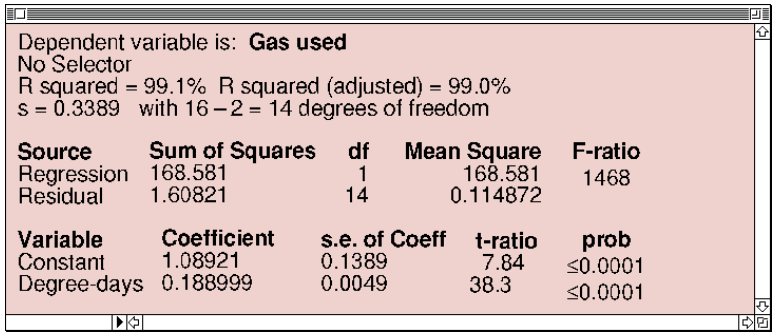
Predictor	Coef	Stdev	t-ratio	p
Constant	1.0892	0.1389	7.84	0.000
D-days	0.188999	0.004934	38.31	0.000

s = 0.3389 R-sq = 99.1% R-sq(adj) = 99.0%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	168.58	168.58	1467.55	0.000
Error	14	1.61	0.11		
Total	15	170.19			

(a)



Dependent variable is: **Gas used**
No Selector
R squared = 99.1% R squared (adjusted) = 99.0%
s = 0.3389 with 16 – 2 = 14 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	168.581	1	168.581	1468
Residual	1.60821	14	0.114872	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	1.08921	0.1389	7.84	≤0.0001
Degree-days	0.188999	0.0049	38.3	≤0.0001

(b)

FIGURE 3.12 Least-squares regression output for the gas consumption data from two statistical software packages: (a) Minitab and (b) Data Desk.

plot the line

To ***plot the line*** on the scatterplot by hand, use the equation to find $\hat{y}$ for two values of x, one near each end of the range of x in the data. Plot each $\hat{y}$ above its x and draw the line through the two points.

EXERCISES

3.38 GAS CONSUMPTION The Technology Toolbox (page 154) gives the equation of the regression line of gas consumption y on degree-days x for the data in Table 3.1 as

$$\hat{y} = 1.0892 + 0.1890x$$

Use your calculator to find the mean and standard deviation of both x and y and their correlation r. Find the slope b and the intercept a of the regression line from these, using the facts in the box *Equation of the least-squares regression line.* (page 152) Verify that you get the equation above. (Results may differ slightly because of rounding off.)

3.39 ARE SAT SCORES CORRELATED? If you previously plotted a scatterplot for the ordered-pairs (Math SAT scores, Verbal SAT scores) data collected by the class in Activity 3, then ask yourself, "Do these data describe a linear trend?" If so, then use your calculator to determine the LSRL equation and correlation coefficient. Overlay this regression line on your scatterplot. Considering the appearance of the scatterplot, the regression line, and the correlation, write a brief statement about the appropriateness of this regression line to model the data. Is the line useful?

3.40 ACID RAIN Researchers studying acid rain measured the acidity of precipitation in a Colorado wilderness area for 150 consecutive weeks. Acidity is measured by pH. Lower pH values show higher acidity. The acid rain researchers observed a linear pattern over time. They reported that the least-squares regression line

$$\text{pH} = 5.43 - (0.0053 \times \text{weeks})$$

fit the data well.[13]

(a) Draw a graph of this line. Is the association positive or negative? Explain in plain language what this association means.

(b) According to the regression line, what was the pH at the beginning of the study (weeks = 1)? At the end (weeks = 150)?

(c) What is the slope of the regression line? Explain clearly what this slope says about the change in the pH of the precipitation in this wilderness area.

3.41 THE ENDANGERED MANATEE Exercise 3.6 (page 125) gives data on the number of powerboats registered in Florida and the number of manatees killed by boats in the years from 1977 to 1990.

(a) Use your calculator to make a scatterplot of these data.

(b) Find the equation of the least-squares line and overlay that line on your scatterplot.

(c) Predict the number of manatees that will be killed by boats in a year when 716,000 powerboats are registered.

(d) Here are four more years of manatee data, in the same form as in Exercise 3.6:

1991	716	53	1993	716	35
1992	716	38	1994	735	49

Add these points to your scatterplot. Florida took stronger measures to protect manatees during these years. Do you see any evidence that these measures succeeded?

(e) In part **(c)** you predicted manatee deaths in a year with 716,000 powerboat registrations. In fact, powerboat registrations were 716,000 for three years. Compare the mean manatee deaths in these three years with your prediction from part **(c)**. How accurate was your prediction?

The role of r^2 in regression

Calculator and computer output for regression report a quantity called r^2. Some computer packages call it "R-sq." For examples, look at the calculator

screen shots in the Technology Toolbox on page 154 and the computer output in Figure 3.12(a) on page 156. Although it is true that this quantity is equal to the square of r, there is much more to this story.

To illustrate the meaning of r^2 in regression, the next two examples use two simple data sets and in each case calculate the quantity r^2. In the first example, a line would be a poor model, and the r^2-value turns out to be small (closer to 0). In the second example, a straight line would fit the data fairly well, and the r^2 value is larger (closer to 1).

EXAMPLE 3.10 SMALL r^2

One way to determine the usefulness of the least-squares regression model is to measure the contribution of x in predicting y. A simple example will help clarify the reasoning. Consider data set A:

x	0	3	6
y	0	10	2

and its scatterplot in Figure 3.13(a). The association between x and y appears to be positive but weak. The sample means are easily calculated to be $\bar{x} = 3$ and $\bar{y} = 4$. Knowing that x is 0 or 3 or 6 gives us very little information to predict y, and so we have to fall back to $\bar{y}$ as a predictor of y. The deviations of the three points about the mean $\bar{y}$ are shown in Figure 3.13(b). The horizontal line in Figure 3.13(b) is at height $\bar{y} = 4$. The sum of the squares of the deviations for the prediction equation $\hat{y} = \bar{y}$ is

$$\text{SST} = \sum (y - \bar{y})^2$$

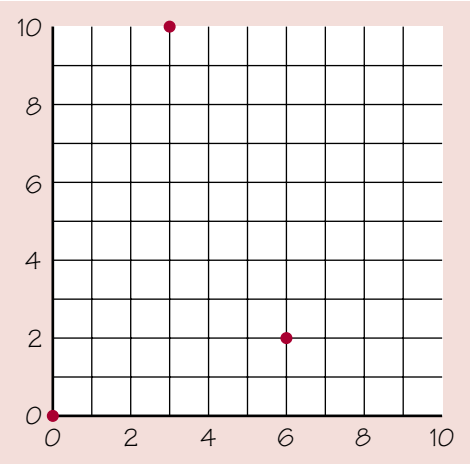


FIGURE 3.13(a) Scatterplot for data set A.

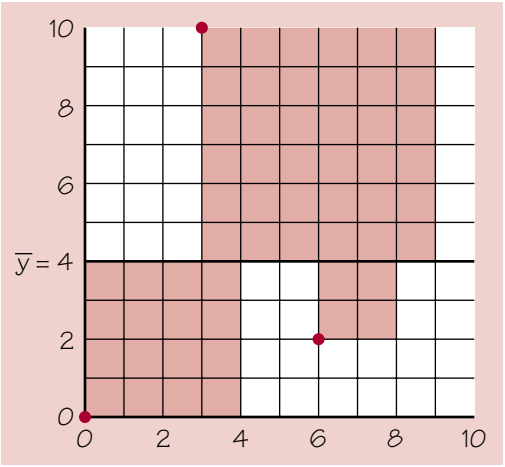


FIGURE 3.13(b) Squares of deviations about $\bar{y}$.

Geometric squares have been constructed on the graph with the deviations from the mean as one side. The total area of these three squares is a measure of the total sample variability. So we call this quantity SST for "total sum of squares about the mean $\overline{y}$."

The LSRL has equation $\hat{y} = 3 + (1/3)x$; see Figure 3.13(c). It has y intercept 3 and passes through the point $(\overline{x}, \overline{y}) = (3, 4)$.

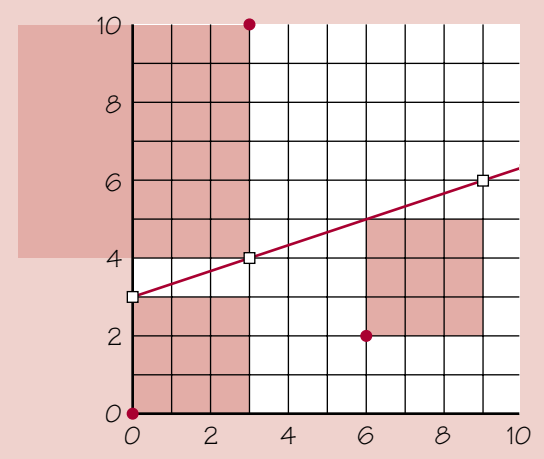


FIGURE 3.13(c) Squares of deviations about $\hat{y}$.

Now we want to consider the sum of the squares of the deviations of the points about this regression line. We call this SSE for "sum of squares for error."

$$\text{SSE} = \sum (y - \hat{y})^2$$

Figure 3.13(c) also shows geometric squares with deviations from the regression line as one side. The calculations can be summarized in a table:

x	y	$(y-\bar{y})^2$	$(y-\hat{y})^2$
0	0	16	9
3	10	36	36
6	2	4	9
		56	54
		SST	SSE

If x is a poor predictor of y, then the sum of squares of deviations about the mean If x is a poor predictor of y, then the sum of squares of deviations about the mean $\bar{y}$ and the sum of squares of deviations about the regression line $\hat{y}$ would be approximately the same. This is the case in our example. If SST = 56 measures the total sample variation of the observations about the mean $\bar{y}$, then SSE = 54 is the remaining "unexplained sample variability" after fitting the regression line. The difference, SST – SSE, measures the amount of variation of y that can be explained by the regression line of y on x. The ratio of these two quantities

$$\frac{\text{SST}-\text{SSE}}{\text{SST}}$$

is interpreted as *the proportion of the total sample variability that is explained by the least-squares regression of y on x*. It can be shown algebraically that this fraction is equal to the square of the correlation coefficient. For this reason, we call this fraction r^2 and refer to it as the ***coefficient of determination.*** For data set A,

coefficient of determination

$$r^2 = \frac{\text{SST}-\text{SSE}}{\text{SST}} = \frac{56-54}{56} = 0.0357$$

We say that 3.57% of the variation in y is explained by least-squares regression of y on x.

For contrast, the next example shows a simple data set where the least-squares line is a much better model.

EXAMPLE 3.11 LARGE r^2

Consider data set B and its accompanying scatterplot in Figure 3.14(a):

x	0	5	10
y	0	7	8

The association between x and y appears to be positive and strong. The sample means are $\bar{x} = 5$ and $\bar{y} = 5$. The squares of the deviations about the mean $\bar{y}$ are shown in Figure 3.14(b), and the squares of the deviations about the regression line $\hat{y}$ are shown in Figure 3.14(c).

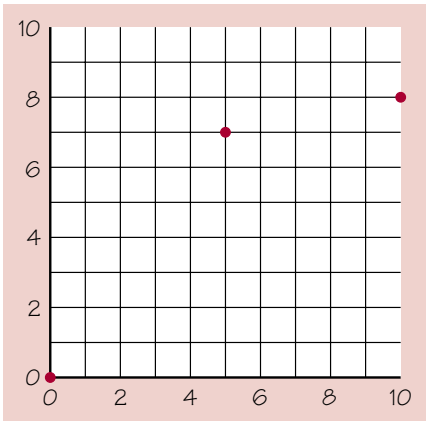


FIGURE 3.14(a) Scatterplot for data set B.

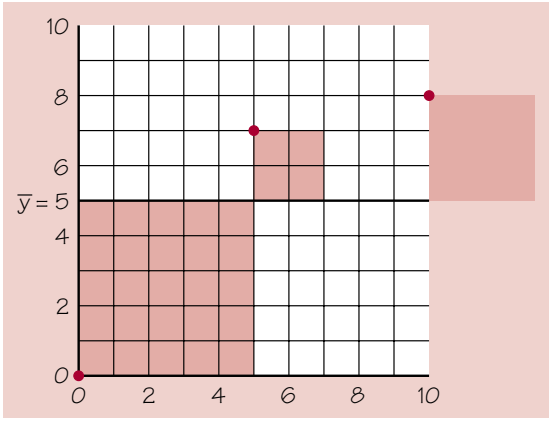


FIGURE 3.14(b) Squares of deviations about $\bar{y}$.

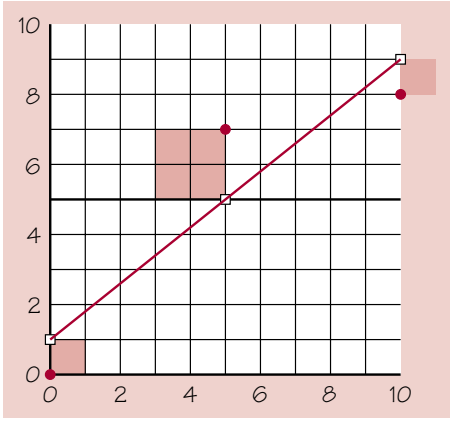


FIGURE 3.14(c) Squares of deviations about $\hat{y}$.

The LSRL has equation $\hat{y} = 1 + 0.8x$. It has y intercept 1 and passes through the points $(\bar{x}, \bar{y}) = (5,5)$ and (10,9). Here are the calculations:

x	y	$(y - \bar{y})^2$	$(y - \hat{y})^2$
0	0	25	1
5	7	4	4
10	8	9	1
		38	6
		SST	SSE

If x is a good predictor of y, then the deviations and hence the SSE would be small; in fact, if all of the points fell exactly on the regression line, SSE would be 0. For data set B, we have

$$r^2 = \frac{\text{SST} - \text{SSE}}{\text{SST}} = \frac{38 - 6}{38} = 0.842$$

We say that 84% of the variation in y is explained by least-squares regression of y on x.

r^2 IN REGRESSION

The **coefficient of determination, r^2**, is the fraction of the variation in the values of y that is explained by least-squares regression of y on x.

Facts about least-squares regression

Regression is one of the most common statistical settings, and least-squares is the most common method for fitting a regression line to data. Here are some facts about least-squares regression lines.

Fact 1. The distinction between explanatory and response variables is essential in regression. Least-squares regression looks at the distances of the data points from the line only in the y direction. If we reverse the roles of the two variables, we get a different least-squares regression line.

EXAMPLE 3.12 THE EXPANDING UNIVERSE

Figure 3.15 is a scatterplot of data that played a central role in the discovery that the universe is expanding. They are the distances from earth of 24 spiral galaxies and the speed at which these galaxies are moving away from us, reported by the astronomer Edwin Hubble in 1929.[14] There is a positive linear relationship, $r = 0.7842$, so that more distant galaxies are moving away more rapidly. Astronomers believe that there is in fact a perfect linear relationship, and that the scatter is caused by imperfect measurements.

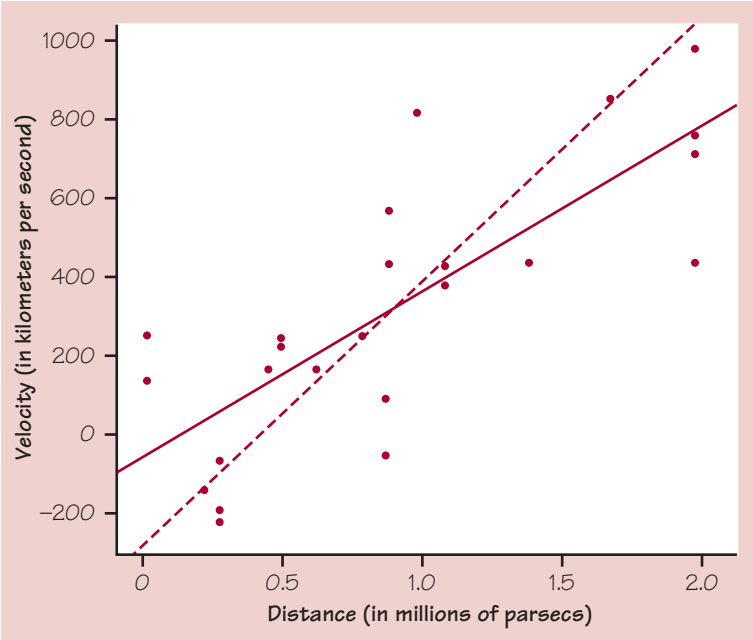


FIGURE 3.15 Scatterplot of Hubble's data on the distance from earth of 24 galaxies and the velocity at which they are moving away from us. The two lines are the two least-squares regression lines: of velocity on distance (solid) and of distance on velocity (dashed).

The two lines on the plot are the two least-squares regression lines. The regression line of velocity on distance is solid. The regression line of distance on velocity is dashed. *Regression of velocity on distance and regression of distance on velocity give different lines.* In the regression setting you must know clearly which variable is explanatory.

Fact 2. There is a close connection between correlation and the slope of the least-squares line. The slope is

$$b = r\frac{s_y}{s_x}$$

This equation says that along the regression line, **a change of one standard deviation in x corresponds to a change of r standard deviations in y.** When the variables are perfectly correlated ($r = 1$ or $r = -1$), the change in the predicted response $\hat{y}$ is the same (in standard deviation units) as the change in x. Otherwise, because $-1 \le r \le 1$, the change in $\hat{y}$ is less than the change in x. As the correlation grows less strong, the prediction $\hat{y}$ moves less in response to changes in x.

Fact 3. The least-squares regression line always passes through the point $(\bar{x}, \bar{y})$ on the graph of y against x. So the least-squares regression line of y on x is the line with slope rs_y/s_x that passes through the point $(\bar{x}, \bar{y})$. We can describe regression entirely in terms of the basic descriptive measures $\bar{x}$, s_x, $\bar{y}$, s_y, and r.

Fact 4. The correlation r describes the strength of a straight-line relationship. In the regression setting, this description takes a specific form: **the square of the correlation, r^2, is the fraction of the variation in the values of y that is explained by the least-squares regression of y on x.**

EXAMPLE 3.13 COMPARING r^2 VALUES

First consider the Sanchez gas consumption data in Figure 3.16(a). There is a lot of variation in the observed y's, the gas consumption data. They range from a low of about 1 to a high of 11. The scatterplot shows that most of this variation in y is accounted for by the fact that outdoor temperature (measured by degree-days x) was changing and pulled gas consumption along with it. There is only a little remaining variation in y, which appears in the scatter of points about the line.The correlation is very strong: $r = 0.9953$, and $r^2 = 0.9906$. Our interpretation is that over 99% of the variation in gas consumption is accounted for by the linear relationship with degree-days.

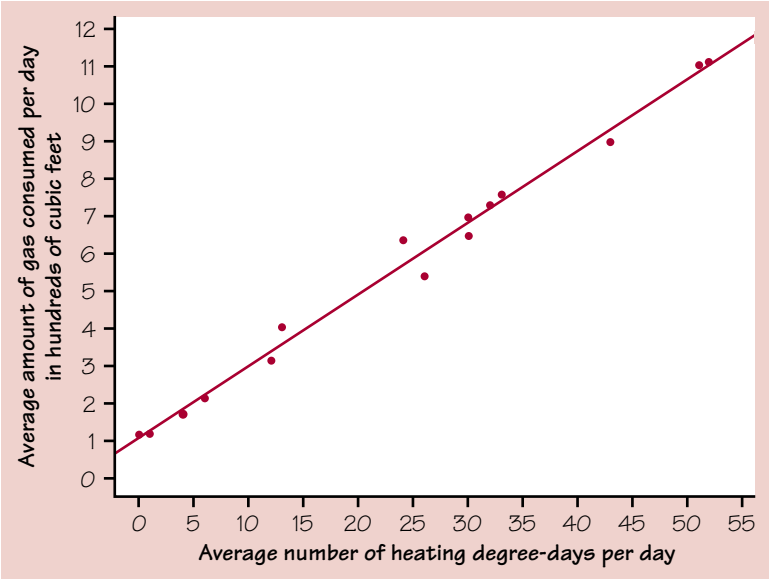


FIGURE 3.16(a) The Sanchez household gas consumption data.

The points in Figure 3.16(b), on the other hand, are more scattered. Linear dependence on distance does explain some of the observed variation in velocity. You would guess a higher value for the velocity y knowing that $x = 2$ than you would if you were told that $x = 0$. But there is still considerable variation in y even when x is held fixed—look at the four points in Figure 3.16(b) with $x = 2$. For the Hubble data, $r = 0.7842$ and $r^2 = 0.6150$. The linear relationship between distance and velocity explains 61.5% of the variation *in either variable*. There are two regression lines, but just one correlation, and r^2 helps interpret both regressions.

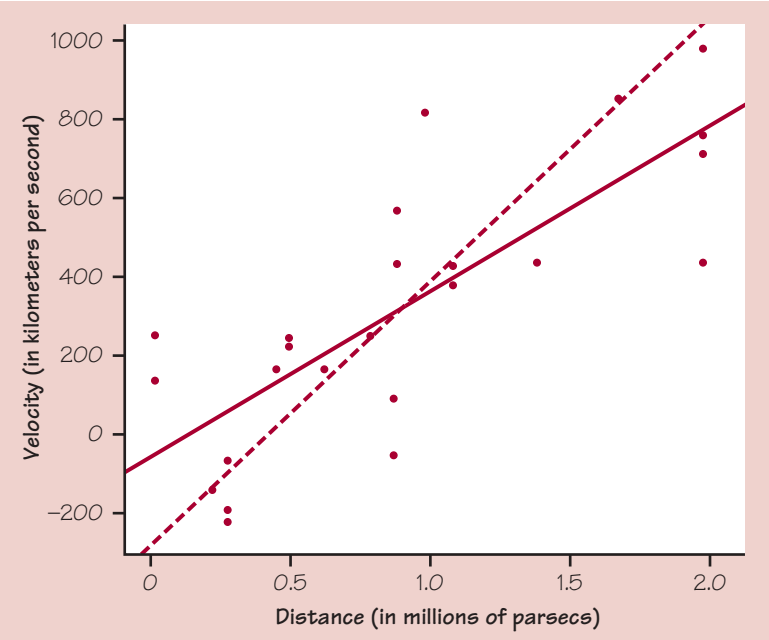


FIGURE 3.16(b) Hubble's data on the distance from earth of 24 galaxies and the velocity at which they are moving away from us.

When you report a regression, give r^2 as a measure of how successful the regression was in explaining the response. When you see a correlation, square it to get a better feel for the strength of the association. Perfect correlation ($r = -1$ or $r = 1$) means the points lie exactly on a line. Then $r^2 = 1$ and all of the variation in one variable is accounted for by the linear relationship with the other variable. If $r = -0.7$ or $r = 0.7$, $r^2 = 0.49$ and about half the variation is accounted for by the linear relationship. In the r^2 scale, correlation ± 0.7 is about halfway between 0 and ± 1.

These connections with correlation are special properties of least-squares regression. They are not true for other methods of fitting a line to data. Another reason that least-squares is the most common method for fitting a regression line to data is that it has many of these convenient special properties.

EXERCISES

3.42 CLASS ATTENDANCE AND GRADES A study of class attendance and grades among first-year students at a state university showed that in general students who attended a higher percent of their classes earned higher grades. Class attendance explained 16% of the variation in grade index among the students. What is the numerical value of the correlation between percent of classes attended and grade index?

3.43 THE PROFESSOR SWIMS Here are Professor Moore's times (in minutes) to swim 2000 yards and his pulse rate after swimming (in beats per minute) for 23 sessions in the pool:

Time:	34.12	35.72	34.72	34.05	34.13	35.72	36.17	35.57
Pulse:	152	124	140	152	146	128	136	144
Time:	35.37	35.57	35.43	36.05	34.85	34.70	34.75	33.93
Pulse:	148	144	136	124	148	144	140	156
Time:	34.60	34.00	34.35	35.62	35.68	35.28	35.97	
Pulse:	136	148	148	132	124	132	139	

(a) A scatterplot shows a moderately strong negative linear relationship. Use your calculator or software to verify that the least-squares regression line is

$$\text{pulse} = 479.9 - (9.695 \times \text{time})$$

(b) The next day's time is 34.30 minutes. Predict the professor's pulse rate. In fact, his pulse rate was 152. How accurate is your prediction?

(c) Suppose you were told only that the pulse rate was 152. You now want to predict swimming time. Find the equation of the least-squares regression line that is appropriate for this purpose. What is your prediction, and how accurate is it?

(d) Explain clearly, to someone who knows no statistics, why there are two different regression lines.

3.44 PREDICTING THE STOCK MARKET Some people think that the behavior of the stock market in January predicts its behavior for the rest of the year. Take the explanatory variable x to be the percent change in a stock market index in January and the response variable y to be the change in the index for the entire year. We expect a positive correlation between x and y because the change during January contributes to the full year's change. Calculation from data for the years 1960 to 1997 gives

$$\bar{x} = 1.75\% \qquad s_x = 5.36\% \qquad r = 0.596$$
$$\bar{y} = 9.07\% \qquad s_y = 15.35\%$$

(a) What percent of the observed variation in yearly changes in the index is explained by a straight-line relationship with the change during January?

(b) What is the equation of the least-squares line for predicting full-year change from January change?

(c) The mean change in January is $\bar{x} = 1.75\%$. Use your regression line to predict the change in the index in a year in which the index rises 1.75% in January. Why could you have given this result (up to roundoff error) without doing the calculation?

3.45 BEAVERS AND BEETLES Ecologists sometimes find rather strange relationships in our environment. One study seems to show that beavers benefit beetles. The researchers laid out 23 circular plots, each four meters in diameter, in an area where beavers were cutting down cottonwood trees. In each plot, they counted the number of stumps from trees cut by beavers and the number of clusters of beetle larvae. Here are the data:[15]

Stumps:	2	2	1	3	3	4	3	1	2	5	1	3
Beetle larvae:	10	30	12	24	36	40	43	11	27	56	18	40
Stumps:	2	1	2	2	1	1	4	1	2	1	4	
Beetle larvae:	25	8	21	14	16	6	54	9	13	14	50	

(a) Make a scatterplot that shows how the number of beaver-caused stumps influences the number of beetle larvae clusters. What does your plot show? (Ecologists think that the new sprouts from stumps are more tender than other cottonwood growth, so that beetles prefer them.)

(b) Find the least-squares regression line and draw it on your plot.

(c) What percent of the observed variation in beetle larvae counts can be explained by straight-line dependence on stump counts?

Residuals

A regression line is a mathematical model for the overall pattern of a linear relationship between an explanatory variable and a response variable. Deviations from the overall pattern are also important. In the regression setting, we see deviations by looking at the scatter of the data points about the regression line. The vertical distances from the points to the least-squares regression line are as small as possible, in the sense that they have the smallest possible sum of squares. Because they represent "left-over" variation in the response after fitting the regression line, these distances are called *residuals*.

RESIDUALS

A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line. That is,

$$\text{residual} = \text{observed } y - \text{predicted } y$$
$$= y - \hat{y}$$

EXAMPLE 3.14 GESELL SCORES

Does the age at which a child begins to talk predict later score on a test of mental ability? A study of the development of young children recorded the age in months at which each of the 21 children spoke their first word and Gesell Adaptive Score, the result of an aptitude test taken much later. The data appear in Table 3.4.

TABLE 3.4 Age at first word and Gesell score

Child	Age	Score	Child	Age	Score	Child	Age	Score
1	15	95	8	11	100	15	11	102
2	26	71	9	8	104	16	10	100
3	10	83	10	20	94	17	12	105
4	9	91	11	7	113	18	42	57
5	15	102	12	9	96	19	17	121
6	20	87	13	10	83	20	11	86
7	18	93	14	11	84	21	10	100

Source: These data were originally collected by L. M. Linde of UCLA but were first published by M. R. Mickey, O. J. Dunn, and V. Clark, "Note on the use of stepwise regression in detecting outliers," *Computers and Biomedical Research*, 1 (1967), pp. 105–111. The data have been used by several authors. We found them in N. R. Draper and J. A. John, "Influential observations and outliers in regression," *Technometrics*, 23 (1981), pp. 21–26.

Figure 3.17 is a scatterplot, with age at first word as the explanatory variable x and Gesell score as the response variable y. Children 3 and 13, and also Children 16 and 21, have identical values of both variables. We use a different plotting symbol to show that one point stands for two individuals. The plot shows a negative association. That is, children who begin to speak later tend to have lower test scores than early talkers. The overall pattern is moderately linear. The correlation describes both the direction and strength of the linear relationship. It is $r = -0.640$.

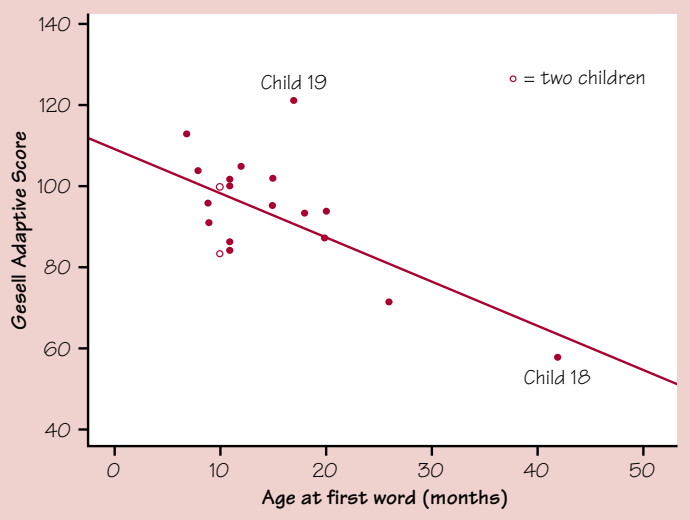


FIGURE 3.17 Scatterplot of Gesell Adaptive Scores versus the age at first word for 21 children from Table 3.4. The line is the least-squares regression line for predicting Gesell score from age at first word.

The line on the plot is the least-squares regression line of Gesell score on age at first word. Its equation is

$$\hat{y} = 109.8738 - 1.1270x$$

For Child 1, who first spoke at 15 months, we predict the score

$$\hat{y} = 109.8738 - (1.1270)(15) = 92.97$$

This child's actual score was 95. The residual is

$$\text{residual} = \text{observed } y - \text{predicted } y$$
$$= 95 - 92.97 = 2.03$$

The residual is positive because the data point lies above the line.

There is a residual for each data point. Here are the 21 residuals for the Gesell data, from Example 3.14, as output by a statistical software package:

```
residuals:
 2.0310   -9.5721  -15.6040  -8.7309   9.0310   -0.3341   3.4120
 2.5230    3.1421    6.6659  11.0151  -3.7309  -15.6040 -13.4770
 4.5230    1.3960    8.6500  -5.5403  30.2850  -11.4770   1.3960
```

Because the residuals show how far the data fall from our regression line, examining the residuals helps assess how well the line describes the data. Although residuals can be calculated from any model fitted to the data, the residuals from the least-squares line have a special property: **the mean of the least-squares residuals is always zero.** You can check that the sum of the residuals above is –0.0002. The sum is not exactly 0 because the software rounded the residuals to four decimal places. This is ***roundoff error.***

roundoff error

Compare the scatterplot in Figure 3.17 with the *residual plot* for the same data in Figure 3.18. The horizontal line at zero in Figure 3.18 helps orient us. It corresponds to the regression line in Figure 3.17.

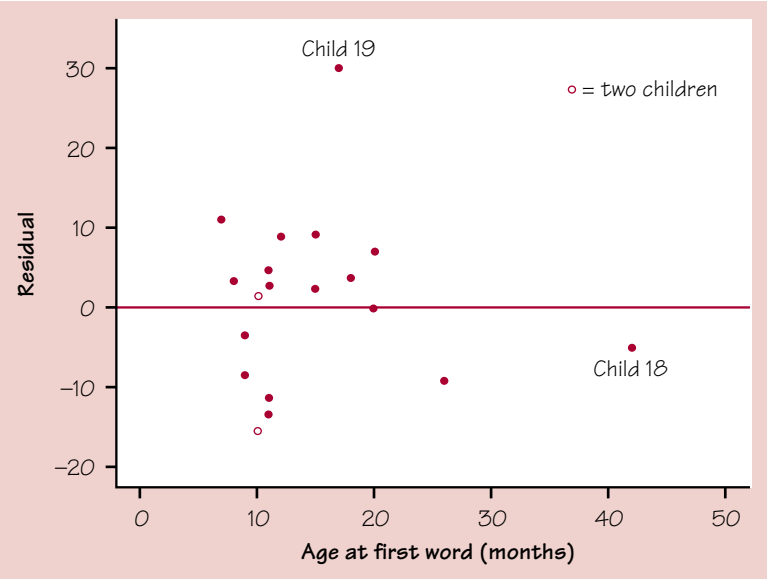


FIGURE 3.18 Residual plot for the regression of Gesell score on age at first word. Child 19 is an outlier, and Child 18 is an influential observation that does not have a large residual.

RESIDUAL PLOTS

A **residual plot** is a scatterplot of the regression residuals against the explanatory variable. Residual plots help us assess the fit of a regression line.

You should be aware that some computer utilities, such as Data Desk, prefer to plot the residuals against the fitted values $\hat{y}_i$ instead of against the values x_i of the explanatory variable. The information in the two plots is the same because $\hat{y}$ is linearly related to x.

If the regression line captures the overall relationship between x and y, the residuals should have no systematic pattern. The residual plot will look something like the simplified pattern in Figure 3.19(a). That plot shows a uniform scatter of the points about the fitted line, with no unusual individual observations.

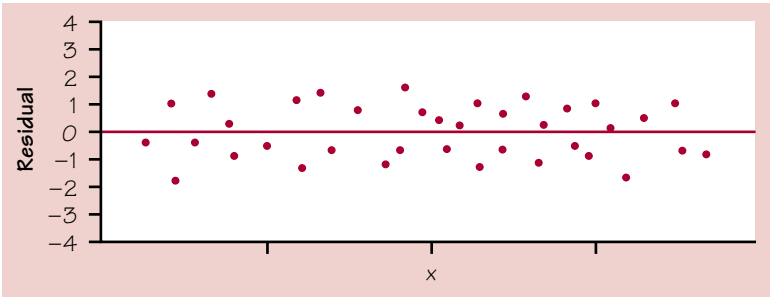


FIGURE 3.19(a) The uniform scatter of points indicates that the regression line fits the data well, so the line is a good model.

Here are some things to look for when you examine the residuals, using either a scatterplot of the data or a residual plot.

- **A curved pattern** shows that the relationship is not linear. Figure 3.19(b) is a simplified example. A straight line is not a good summary for such data.

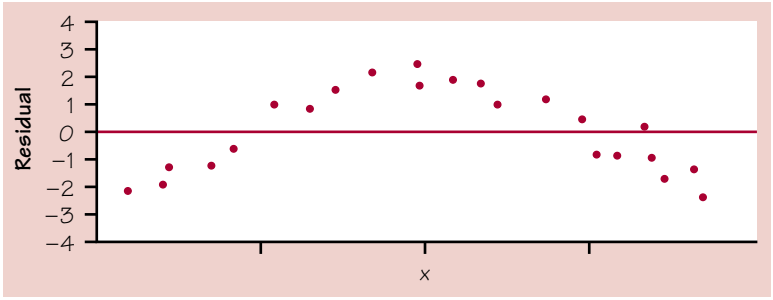


FIGURE 3.19(b) The residuals have a curved pattern, so a straight line is an inappropriate model.

• **Increasing or decreasing spread about the line** as x increases indicates that prediction of y will be less accurate for larger x. Figure 3.19(c) is a simplified example.

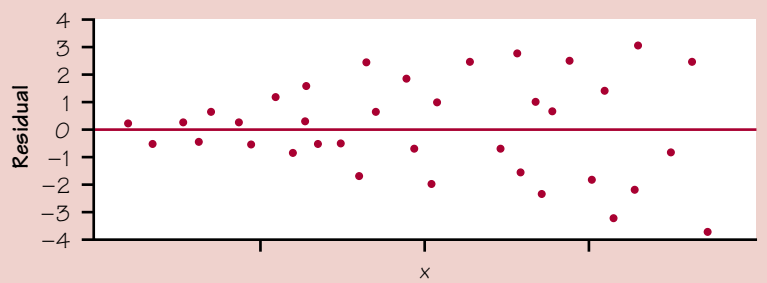


FIGURE 3.19(c) The response variable y has more spread for larger values of the explanatory variable x, so prediction will be less accurate when x is large.

• **Individual points with large residuals,** like Child 19 in Figures 3.17 and 3.18 are outliers in the vertical (y) direction because they lie far from the line that describes the overall pattern.

• **Individual points that are extreme in the x direction,** like Child 18 in Figures 3.17 and 3.18, may not have large residuals, but they can be very important. We address such points next.

Influential observations

Children 18 and 19 are both unusual in the Gesell example. They are unusual in different ways. Child 19 lies far from the regression line. This child's Gesell score is so high that we should check for a mistake in recording it. In fact, the score is correct. Child 18 is close to the line but far out in the x direction. He or she began to speak much later than any of the other children. *Because of its extreme position on the age scale, this point has a strong influence on the position of the regression line.* Figure 3.20 adds a second regression line, calculated after leaving out Child 18. You can see that this one point moves the line quite a bit. We call such points *influential*.

OUTLIERS AND INFLUENTIAL OBSERVATIONS IN REGRESSION

An **outlier** is an observation that lies outside the overall pattern of the other observations.

An observation is **influential** for a statistical calculation if removing it would markedly change the result of the calculation. Points that are outliers in the x direction of a scatterplot are often influential for the least-squares regression line.

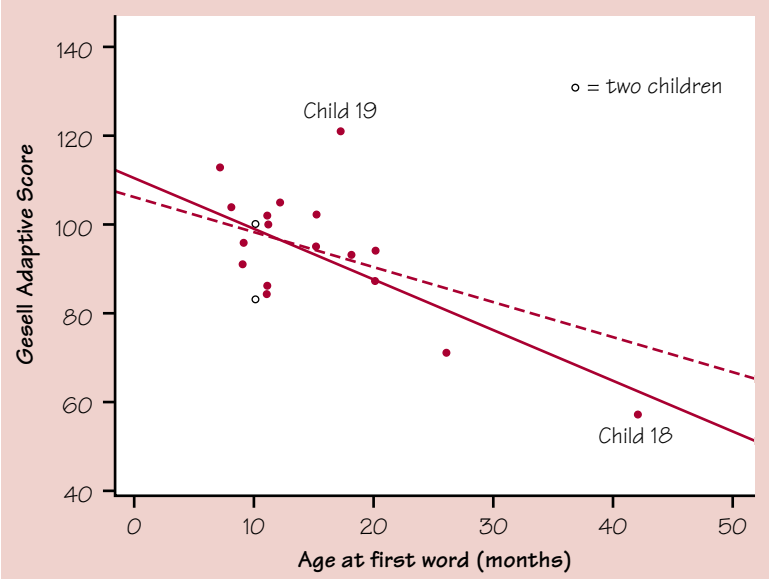


FIGURE 3.20 Two least-squares regression lines of Gesell score on age at first word. The solid line is calculated from all the data. The dashed line is calculated leaving out Child 18. Child 18 is an influential observation because leaving out this point moves the regression line quite a bit.

Children 18 and 19 are both outliers in Figure 3.20. Child 18 is an outlier in the x direction and influences the least-squares line. Child 19 is an outlier in the y direction. It has less influence on the regression line because the many other points with similar values of x anchor the line well below the outlying point. Influential points often have small residuals, because they pull the regression line toward themselves. If you just look at residuals, you will miss influential points. Influential observations can greatly change the interpretation of data.

EXAMPLE 3.15 AN INFLUENTIAL OBSERVATION

The strong influence of Child 18 makes the original regression of Gesell score on age at first word misleading. The original data have $r^2 = 0.41$. That is, the age at which a child begins to talk explains 41% of the variation on a later test of mental ability. This relationship is strong enough to be interesting to parents. If we leave out Child 18, r^2 drops to only 11%. The apparent strength of the association was largely due to a single influential observation.

What should the child development researcher do? She must decide whether Child 18 is so slow to speak that this individual should not be allowed to influence the analysis. If she excludes Child 18, much of the evidence for a connection between the age at which a child begins to talk and later ability score vanishes. If she keeps Child 18, she needs data on other children who were also slow to begin talking, so that the analysis no longer depends so heavily on just one child.

EXERCISES

3.46 DRIVING SPEED AND FUEL CONSUMPTION Exercise 3.11 (page 129) gives data on the fuel consumption y of a car at various speeds x. Fuel consumption is measured in liters of gasoline per 100 kilometers driven and speed is measured in kilometers per hour. A statistical software package gives the least-squares regression line and also the residuals. The regression line is

$$\hat{y} = 11.058 - 0.01466x$$

The residuals, in the same order as the observations, are

10.09	2.24	–0.62	–2.47	–3.33	–4.28	–3.73	–2.94
–2.17	–1.32	–0.42	0.57	1.64	2.76	3.97	

(a) Make a scatterplot of the observations and draw the regression line on your plot.

(b) Would you use the regression line to predict y from x? Explain your answer.

(c) Check that the residuals have sum zero (up to roundoff error).

(d) Make a plot of residuals against the values of x. Draw a horizontal line at height zero on your plot. Notice that the residuals show the same pattern about this line as the data points show about the regression line in the scatterplot in **(a)**. What do you conclude about the residual plot?

3.47 HOW MANY CALORIES? Exercise 3.20 (page 138) gives data on the true calories in ten foods and the average guesses made by a large group of people. Exercise 3.31 (page 147) explored the influence of two outlying observations on the correlation.

(a) Make a scatterplot suitable for predicting guessed calories from true calories. Circle the points for spaghetti and snack cake on your plot. These points lie outside the linear pattern of the other eight points.

(b) Use your calculator to find the least-squares regression line of guessed calories on true calories. Do this twice, first for all ten data points and then leaving out spaghetti and snack cake.

(c) Plot both lines on your graph. (Make one dashed so that you can tell them apart.) Are spaghetti and snack cake, taken together, influential observations? Explain your answer.

3.48 INFLUENTIAL OR NOT? The discussion of Example 3.15 shows that Child 18 in the Gesell data in Table 3.4 is an influential observation. Now we will examine the effect of Child 19, who is also an outlier in Figure 3.20.

(a) Find the least-squares regression line of Gesell score on age at first word, leaving out Child 19. Example 3.14 gives the regression line from all the children. Plot both lines on the same graph. (You do not have to make a scatterplot of all the points—just plot the two lines.) Would you call Child 19 very influential? Why?

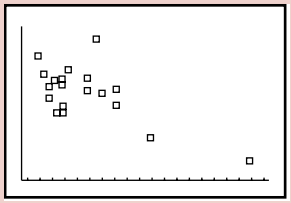
(b) How does removing Child 19 change the r^2 for this regression? Explain why r^2 changes in this direction when you drop Child 19.

TECHNOLOGY TOOLBOX *Residual plots by calculator*

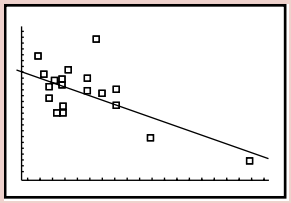
Here is a procedure for calculating residuals on your TI-83/89 and then displaying a residual plot.

• Enter the ages and Gesell scores from Table 3.4 (page 168). Plot the scatterplot and perform the linear regression. Store the regression equation in Y_1 (y1(x) on the TI-89) and superimpose the LSRL on the scatterplot.

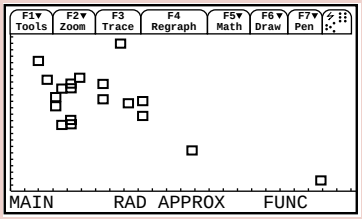
TI-83

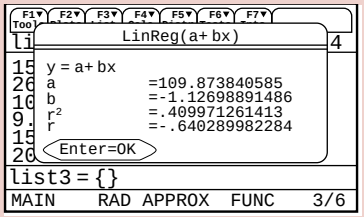


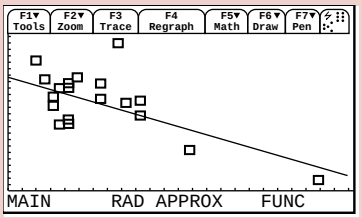
```
LinReg
 y=a+bx
 a=109.8738406
 b=-1.126988915
 r2=.4099712614
 r=-.6402899823
```



TI-89







This sets the stage. To graph the residual plot:

• Restore the six default lists using the SetUpEditor command.

• Press STAT, choose `5:SetUpEditor,` and press ENTER.

• Press CATALOG, choose `SetUpEd`(, type), and press ENTER.

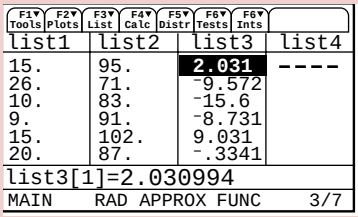
• Define L_3/list3 as the observed value minus the predicted value.

• With L_3 highlighted, enter the command `L2-Y1(L1)`. Press ENTER to show the residuals.

• With list3 highlighted, enter the command `list2 -y1(list1)`. Press ENTER to show the residuals.

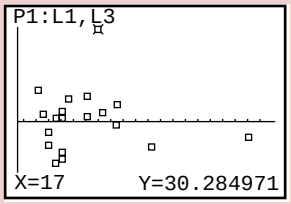
L1	L2	L3 3
15	95	2.031
26	71	-9.572
10	83	-15.6
9	91	-8.731
15	102	9.031
20	87	-.3341
18	93	3.412

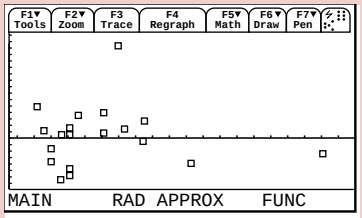
L3(1)=2.030993137…



TECHNOLOGY TOOLBOX *Residual plots by calculator (continued)*

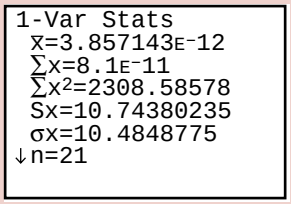
• Turn off Plot1 and deselect the regression equation. Specify Plot2 with L_1/list1 as the x variable and L_3/list3 as the y variable. Use ZoomStat (ZoomData) to see the residual plot.

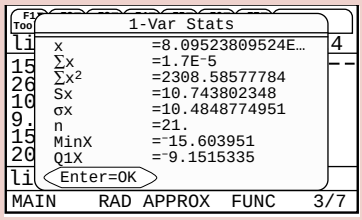




The x axis in the residual plot serves as a reference line, with points above this line corresponding to positive residuals and points below the line corresponding to negative residuals. We used TRACE to see the regression outlier at $x = 17$.

• Finally, we have previously noted that an important property of residuals is that their sum is zero. Calculate one-variable statistics on the residuals list to verify that $\sum$(residuals) = 0 and that, consequently, the mean of the residuals is also 0.





Note that the calculator is showing some roundoff error. You should recognize these peculiar looking numbers as equivalent to 0.

3.49 LEAN BODY MASS AS A PREDICTOR OF METABOLIC RATE Exercise 3.12 (page 132) provides data from a study of dieting for 12 women and 7 men subjects. We will explore the women's data further.

(a) Define two lists on your calculator, MASSF for female mass and METF for female metabolic rate. Then transfer the data to lists 1 and 2. Define Plot1 using the □ plotting symbol, and plot the scatterplot.

(b) Perform least-squares regression on your calculator and record the equation and the correlation. Lean body mass explains what percent of the variation in metabolic rate for the women?

(c) Does the least-squares line provide an adequate model for the data? Define Plot2 to be a residual plot on your calculator with residuals on the vertical axis and lean body mass (x-values) on the horizontal axis. Use the □ plotting symbol. Use ZoomStat/ZoomData to see the plot. Copy the plot onto your paper. Label both axes appropriately.

(d) Define list3 to be the predicted y-values: $Y_1(L_1)$ on the TI-83 or Y_1(list1) on the TI-89. Define Plot3 to be a residual plot on your calculator with residuals on the vertical axis and predicted metabolic rate on the horizontal axis. Use the + plotting symbol. Use ZoomStat/ZoomData to see the plot. Copy the plot onto your paper. Label both axes. Compare the two residual plots.

SUMMARY

A **regression line** is a straight line that describes how a response variable y changes as an explanatory variable x changes.

The most common method of fitting a line to a scatterplot is least squares. The **least-squares regression line** is the straight line $\hat{y} = a + bx$ that minimizes the sum of the squares of the vertical distances of the observed points from the line.

You can use a regression line to **predict** the value of y for any value of x by substituting this x into the equation of the line.

The **slope** b of a regression line $\hat{y} = a + bx$ is the rate at which the predicted response $\hat{y}$ changes along the line as the explanatory variable x changes. Specifically, b is the change in $\hat{y}$ when x increases by 1.

The **intercept** a of a regression line $\hat{y} = a + bx$ is the predicted response $\hat{y}$ when the explanatory variable $x = 0$. This prediction is of no statistical use unless x can actually take values near 0.

The least-squares regression line of y on x is the line with slope rs_y/s_x and intercept $a = \bar{y} - b\bar{x}$. This line always passes through the point $(\bar{x}, \bar{y})$.

Correlation and regression are closely connected. The correlation r is the slope of the least-squares regression line when we measure both x and y in standardized units. The square of the correlation r^2 is the fraction of the variance of one variable that is explained by least-squares regression on the other variable.

You can examine the fit of a regression line by studying the **residuals,** which are the differences between the observed and predicted values of y. Be on the lookout for outlying points with unusually large residuals and also for nonlinear patterns and uneven variation about the line.

Also look for **influential observations,** individual points that substantially change the regression line. Influential observations are often outliers in the x direction, but they need not have large residuals.

SECTION 3.3 EXERCISES

3.50 REVIEW OF STRAIGHT LINES Fred keeps his savings under his mattress. He began with \$500 from his mother and adds \$100 each year. His total savings y after x years are given by the equation

$$y = 500 + 100x$$

(a) Draw a graph of this equation. (Choose two values of x, such as 0 and 10. Compute the corresponding values of y from the equation. Plot these two points on graph paper and draw the straight line joining them.)

(b) After 20 years, how much will Fred have under his mattress?

(c) If Fred had added $200 instead of $100 each year to his initial $500, what is the equation that describes his savings after x years?

3.51 REVIEW OF STRAIGHT LINES During the period after birth, a male white rat gains exactly 40 grams (g) per week. (This rat is unusually regular in his growth, but 40 g per week is a realistic rate.)

(a) If the rat weighed 100 g at birth, give an equation for his weight after x weeks. What is the slope of this line?

(b) Draw a graph of this line between birth and 10 weeks of age.

(c) Would you be willing to use this line to predict the rat's weight at age 2 years? Do the prediction and think about the reasonableness of the result. (There are 454 grams in a pound. To help you assess the result, note that a large cat weighs about 10 pounds.)

3.52 IQ AND SCHOOL GPA Figure 3.5 (page 135) plots school grade point average (GPA) against IQ test score for 78 seventh-grade students. Calculation shows that the mean and standard deviation of the IQ scores are

$$\bar{x} = 108.9 \quad s_x = 13.17$$

For the grade point averages,

$$\bar{y} = 7.447 \quad s_y = 2.10$$

The correlation between IQ and GPA is $r = 0.6337$.

(a) Find the equation of the least-squares line for predicting GPA from IQ.

(b) What percent of the observed variation in these students' GPAs can be explained by the linear relationship between GPA and IQ?

(c) One student has an IQ of 103 but a very low GPA of 0.53. What is the predicted GPA for a student with IQ = 103? What is the residual for this particular student?

3.53 TAKE ME OUT TO THE BALL GAME What is the relationship between the price charged for a hot dog and the price charged for a 16-ounce soda in major league baseball stadiums? Here are some data:[16]

Team	Hot dog	Soda	Team	Hot dog	Soda	Team	Hot dog	Soda
Angels	2.50	1.75	Giants	2.75	2.17	Rangers	2.00	2.00
Astros	2.00	2.00	Indians	2.00	2.00	Red Sox	2.25	2.29
Braves	2.50	1.79	Marlins	2.25	1.80	Rockies	2.25	2.25
Brewers	2.00	2.00	Mets	2.50	2.50	Royals	1.75	1.99
Cardinals	3.50	2.00	Padres	1.75	2.25	Tigers	2.00	2.00
Dodgers	2.75	2.00	Phillies	2.75	2.20	Twins	2.50	2.22
Expos	1.75	2.00	Pirates	1.75	1.75	White Sox	2.00	2.00

(a) Make a scatterplot appropriate for predicting soda price from hot dog price. Describe the relationship that you see. Are there any outliers?

(b) Find the correlation between hot dog price and soda price. What percent of the variation in soda price does a linear relationship account for?

(c) Find the equation of the least-squares line for predicting soda price from hot dog price. Draw the line on your scatterplot. Based on your findings in **(b)**, explain why it is not surprising that the line is nearly horizontal (slope near zero).

(d) Circle the observation that is potentially the most influential. What team is this? Find the least-squares line without this one observation and draw it on your scatterplot. Was the observation in fact influential?

3.54 KEEPING WATER CLEAN Keeping water supplies clean requires regular measurement of levels of pollutants. The measurements are indirect—a typical analysis involves forming a dye by a chemical reaction with the dissolved pollutant, then passing light through the solution and measuring its "absorbence." To calibrate such measurements, the laboratory measures known standard solutions and uses regression to relate absorbence to pollutant concentration. This is usually done every day. Here is one series of data on the absorbence for different levels of nitrates. Nitrates are measured in milligrams per liter of water.[17]

Nitrates:	50	50	100	200	400	800	1200	1600	2000	2000
Absorbence:	7.0	7.5	12.8	24.0	47.0	93.0	138.0	183.0	230.0	226.0

(a) Chemical theory says that these data should lie on a straight line. If the correlation is not at least 0.997, something went wrong and the calibration procedure is repeated. Plot the data and find the correlation. Must the calibration be done again?

(b) What is the equation of the least-squares line for predicting absorbence from concentration? If the lab analyzed a specimen with 500 milligrams of nitrates per liter, what do you expect the absorbence to be? Based on your plot and the correlation, do you expect your predicted absorbence to be very accurate?

3.55 A GROWING CHILD Sarah's parents are concerned that she seems short for her age. Their doctor has the following record of Sarah's height:

Age (months):	36	48	51	54	57	60
Height (cm):	86	90	91	93	94	95

(a) Make a scatterplot of these data. Note the strong linear pattern.

(b) Using your calculator, find the equation of the least-squares regression line of height on age.

(c) Predict Sarah's height at 40 months and at 60 months. Use your results to draw the regression line on your scatterplot.

(d) What is Sarah's rate of growth, in centimeters per month? Normally growing girls gain about 6 cm in height between ages 4 (48 months) and 5 (60 months). What rate of growth is this in centimeters per month? Is Sarah growing more slowly than normal?

3.56 INVESTING AT HOME AND OVERSEAS Investors ask about the relationship between returns on investments in the United States and on investments overseas. Table 3.5 gives the total returns on U.S. and overseas common stocks over a 26-year period. (The total return is change in price plus any dividends paid, converted into U.S. dollars. Both returns are averages over many individual stocks.)

TABLE 3.5 Annual total return on overseas and U.S. stocks

Year	Overseas % return	U.S. % return	Year	Overseas % return	U.S. % return	Year	Overseas % return	U.S. % return
1971	29.6	14.6	1980	22.6	32.3	1989	10.6	31.5
1972	36.3	18.9	1981	–2.3	–5.0	1990	–23.0	–3.1
1973	–14.9	–14.8	1982	–1.9	21.5	1991	12.8	30.4
1974	–23.2	–26.4	1983	23.7	22.4	1992	–12.1	7.6
1975	35.4	37.2	1984	7.4	6.1	1993	32.9	10.1
1976	2.5	23.6	1985	56.2	31.6	1994	6.2	1.3
1977	18.1	–7.4	1986	69.4	18.6	1995	11.2	37.6
1978	32.6	6.4	1987	24.6	5.1	1996	6.4	23.0
1979	4.8	18.2	1988	28.5	16.8	1997	2.1	33.4

Source: The U.S. returns are for the Standard & Poor's 500 Index. The overseas returns are for the Morgan Stanley Europe, Australasia, Far East (EAFE) index.

(a) Make a scatterplot suitable for predicting overseas returns from U.S. returns.

(b) Find the correlation and r^2. Describe the relationship between U.S. and overseas returns in words, using r and r^2 to make your description more precise.

(c) Find the least-squares regression line of overseas returns on U.S. returns. Draw the line on the scatterplot.

(d) In 1997, the return on U.S. stocks was 33.4%. Use the regression line to predict the return on overseas stocks. The actual overseas return was 2.1%. Are you confident that predictions using the regression line will be quite accurate? Why?

(e) Circle the point that has the largest residual (either positive or negative). What year is this? Are there any points that seem likely to be very influential?

3.57 WHAT'S MY GRADE? In Professor Friedman's economics course the correlation between the students' total scores prior to the final examination and their final examination scores is $r = 0.6$. The pre-exam totals for all students in the course have mean 280 and standard deviation 30. The final exam scores have mean 75 and standard deviation 8. Professor Friedman has lost Julie's final exam but knows that her total before the exam was 300. He decides to predict her final exam score from her pre-exam total.

(a) What is the slope of the least-squares regression line of final exam scores on pre-exam total scores in this course? What is the intercept?

(b) Use the regression line to predict Julie's final exam score.

(c) Julie doesn't think this method accurately predicts how well she did on the final exam. Calculate r^2 and use the value you get to argue that her actual score could have been much higher (or much lower) than the predicted value.

3.58 A NONSENSE PREDICTION Use the least-squares regression line for the data in Exercise 3.55 to predict Sarah's height at age 40 years (480 months). Your prediction is in centimeters. Convert it to inches using the fact that a centimeter is 0.3937 inch.

The prediction is impossibly large. It is not reasonable to use data for 36 to 60 months to predict height at 480 months.

3.59 INVESTING AT HOME AND OVERSEAS Exercise 3.56 examined the relationship between returns on U.S. and overseas stocks. Investors also want to know what typical returns are and how much year-to-year variability (called *volatility* in finance) there is. Regression and correlation do not answer these questions.

(a) Find the five-number summaries for both U.S. and overseas returns, and make side-by-side boxplots to compare the two distributions.

(b) Were returns generally higher in the United States or overseas during this period? Explain your answer.

(c) Were returns more volatile (more variable) in the United States or overseas during this period? Explain your answer.

3.60 WILL I BOMB THE FINAL? We expect that students who do well on the midterm exam in a course will usually also do well on the final exam. Gary Smith of Pomona College looked at the exam scores of all 346 students who took his statistics class over a 10-year period.[18] The least-squares line for predicting final exam score from midterm exam score was $\hat{y} = 46.6 + 0.41x$.

Octavio scores 10 points above the class mean on the midterm. How many points above the class mean do you predict that he will score on the final? (*Hint:* Use the fact that the least-squares line passes through the point $(\bar{x}, \bar{y})$ and the fact that Octavio's midterm score is $\bar{x} + 10$. This is an example of the phenomenon that gave "regression" its name: students who do well on the midterm will on the average do less well, but still above average, on the final.)

3.61 NAHYA INFANT WEIGHTS A study of nutrition in developing countries collected data from the Egyptian village of Nahya. Here are the mean weights (in kilograms) for 170 infants in Nahya who were weighed each month during their first year of life.

Age (months):	1	2	3	4	5	6	7	8	9	10	11	12
Weight (kg):	4.3	5.1	5.7	6.3	6.8	7.1	7.2	7.2	7.2	7.2	7.5	7.8

(a) Plot the weight against time.

(b) A hasty user of statistics enters the data into software and computes the least-squares line without plotting the data. The result is

```
THE REGRESSION EQUATION IS
WEIGHT = 4.88 + 0.267 AGE
```

Plot this line on your graph. Is it an acceptable summary of the overall pattern of growth? Remember that you can calculate the least-squares line for *any* set of two-variable data. It's up to you to decide if it makes sense to fit a line.

(c) Fortunately, the software also prints out the residuals from the least-squares line. In order of age along the rows, they are

–0.85	–0.31	0.02	0.35	0.58	0.62
0.45	0.18	–0.08	–0.35	–0.32	–0.28

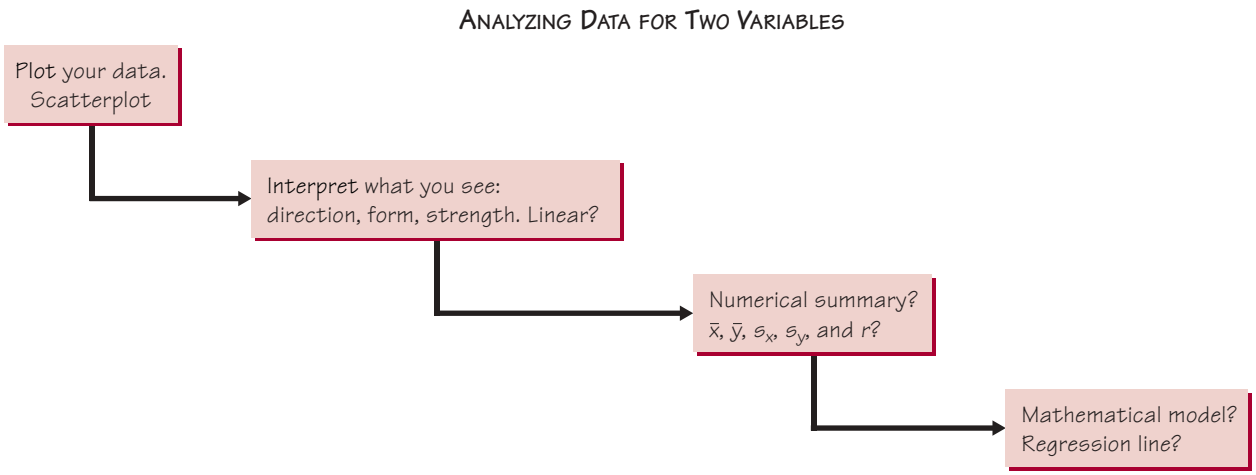
Verify that the residuals have sum 0 (except for roundoff error). Plot the residuals against age and add a horizontal line at 0. Describe carefully the pattern that you see.

CHAPTER REVIEW

Chapters 1 and 2 dealt with data analysis for a single variable. In this chapter, we have studied analysis of data for two or more variables. The proper analysis depends on whether the variables are categorical or quantitative and on whether one is an explanatory variable and the other a response variable.

Data analysis begins with graphs and then adds numerical summaries of specific aspects of the data.

This chapter concentrates on relations between two quantitative variables. Scatterplots show the relationship, whether or not there is an explanatory-response distinction. Correlation describes the strength of a linear relationship, and least-squares regression fits a line to data that have an explanatory-response relation.



Here is a review list of the most important skills you should have gained from studying this chapter.

A. DATA

1. Recognize whether each variable is quantitative or categorical.

2. Identify the explanatory and response variables in situations where one variable explains or influences another.

B. SCATTERPLOTS

1. Make a scatterplot to display the relationship between two quantitative variables. Place the explanatory variable (if any) on the horizontal scale of the plot.

2. Add a categorical variable to a scatterplot by using a different plotting symbol or color.

3. Describe the form, direction, and strength of the overall pattern of a scatterplot. In particular, recognize positive or negative association and linear (straight-line) patterns. Recognize outliers in a scatterplot.

C. CORRELATION

1. Using a calculator, find the correlation r between two quantitative variables.

2. Know the basic properties of correlation: r measures the strength and direction of only linear relationships; $-1 \le r \le 1$ always; $r = \pm 1$ only for perfect straight-line relations; r moves away from 0 toward ± 1 as the linear relation gets stronger.

D. STRAIGHT LINES

1. Explain what the slope b and the intercept a mean in the equation $y = a + bx$ of a straight line.

2. Draw a graph of the straight line when you are given its equation.

E. REGRESSION

1. Using a calculator, find the least-squares regression line of a response variable y on an explanatory variable x from data.

2. Find the slope and intercept of the least-squares regression line from the means and standard deviations of x and y and their correlation.

3. Use the regression line to predict y for a given x. Recognize extrapolation and be aware of its dangers.

4. Use r^2 to describe how much of the variation in one variable can be accounted for by a straight-line relationship with another variable.

5. Recognize outliers and potentially influential observations from a scatterplot with the regression line drawn on it.

6. Calculate the residuals and plot them against the explanatory variable x or against other variables. Recognize unusual patterns.

CHAPTER 3 REVIEW EXERCISES

3.62 Figure 3.21 is a scatterplot that displays the heights of 53 pairs of parents. The mother's height is plotted on the vertical axis and the father's height on the horizontal axis.[20]

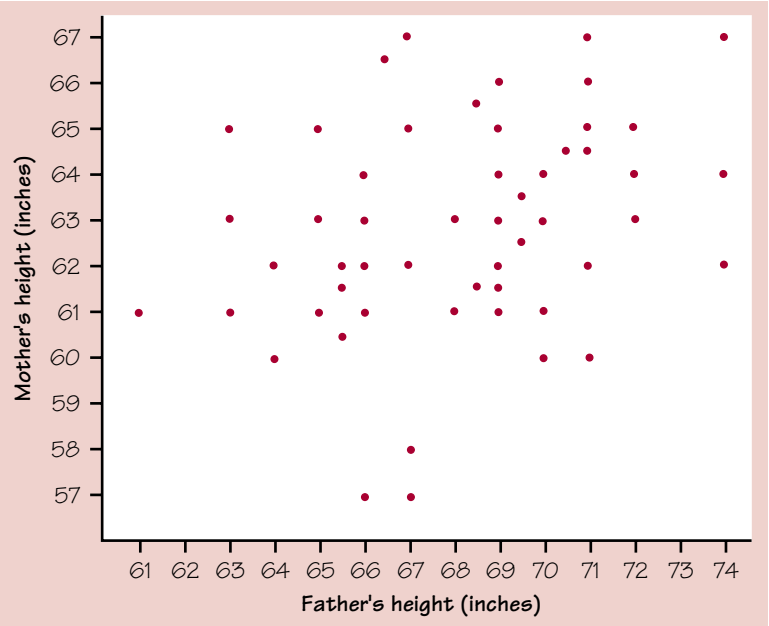


FIGURE 3.21 Scatterplot of the heights of the mother and father in 53 pairs of parents.

(a) What is the smallest height of any mother in the group? How many mothers have that height? What are the heights of the fathers in these pairs?

(b) What is the greatest height of any father in the group? How many fathers have that height? How tall are the mothers in these pairs?

(c) Are there clear explanatory and response variables, or could we freely choose which variable to plot horizontally?

(d) Say in words what a positive association between these variables means. The scatterplot shows a weak positive association. Why do we say the association is weak?

3.63 IS WINE GOOD FOR YOUR HEART? Table 3.6 below gives data on average per capita wine consumption and heart disease death rates in 19 countries.

TABLE 3.6 Wine consumption and heart disease

Country	Alcohol from wine (liters/year)	Heart disease death rate (per 100,000)	Country	Alcohol from wine (liters/year)	Heart disease death rate (per 100,000)
Australia	2.5	211	Netherlands	1.8	167
Austria	3.9	167	New Zealand	1.9	266
Belgium/Lux.	2.9	131	Norway	0.8	227
Canada	2.4	191	Spain	6.5	86
Denmark	2.9	220	Sweden	1.6	207
Finland	0.8	297	Switzerland	5.8	115
France	9.1	71	United Kingdom	1.3	285
Iceland	0.8	211	United States	1.2	199
Ireland	0.7	300	West Germany	2.7	172
Italy	7.9	107			

Source: M. H. Criqui, University of California, San Diego, reported in the *New York Times*, December 28, 1994.

(a) Construct a scatterplot for these data. Describe the relationship between the two variables.

(b) Determine the equation of the least-squares line for predicting heart disease death rate from wine consumption using the data in Table 3.6. Determine the correlation.

(c) Interpret the correlation. About what percent of the variation among countries in heart disease death rates is explained by the straight-line relationship with wine consumption?

(d) Predict the heart disease death rate in another country where adults average 4 liters of alcohol from wine each year.

(e) The correlation and the slope of the least-squares line in **(b)** are both negative. Is it possible for these two quantities to have opposite signs? Explain your answer.

3.64 AGE AND EDUCATION IN THE STATES Because older people as a group have less education than younger people, we might suspect a relationship between the percent of state residents aged 65 and over and the percent who are not high school graduates. Figure 3.22 is a scatterplot of these variables. The data appear in Tables 1.5 and 1.15 (pages 23 and 70).

(a) There are at least two and perhaps three outliers in the plot. Identify these states, and give plausible reasons for why they might be outliers.

(b) If we ignore the outliers, does the relationship have a clear form and direction? Explain your answer.

(c) If we calculate the correlation with and without the three outliers, we get $r = 0.067$ and $r = 0.267$. Which of these is the correlation without the outliers? Explain your answer.

3.65 ALWAYS PLOT YOUR DATA! Table 3.7 presents four sets of data prepared by the statistician Frank Anscombe to illustrate the dangers of calculating without first plotting the data.

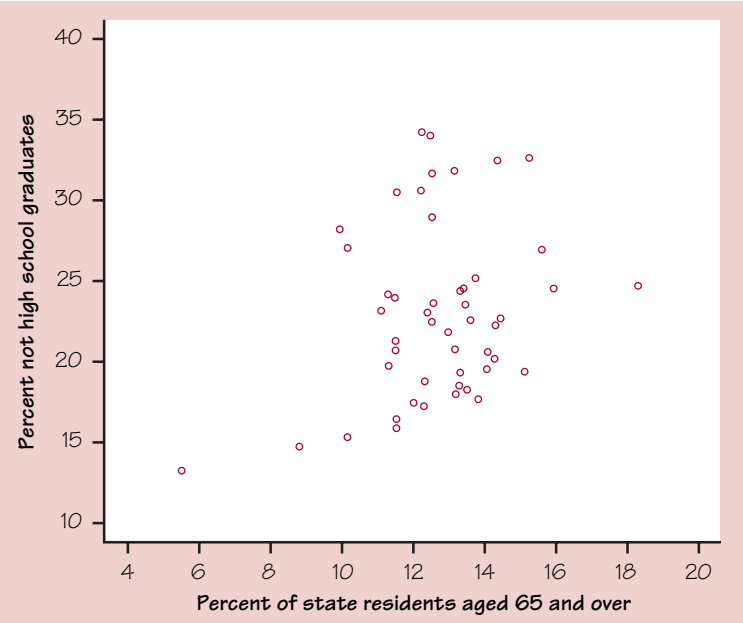


FIGURE 3.22 Scatterplot of the percent of residents who are not high school graduates against the percent of residents aged 65 and over in the 50 states, for Exercise 3.64.

TABLE 3.7 Four data sets for exploring correlation and regression

Data Set A											
x	10	8	13	9	11	14	6	4	12	7	5
y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68
Data Set B											
x	10	8	13	9	11	14	6	4	12	7	5
y	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74
Data Set C											
x	10	8	13	9	11	14	6	4	12	7	5
y	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73
Data Set D											
x	8	8	8	8	8	8	8	8	8	8	19
y	6.58	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.50

Source: Frank J. Anscombe, "Graphs in statistical analysis," *American Statistician*, 27 (1973), pp. 17–21.

(a) Without making scatterplots, find the correlation and the least-squares regression line for all four data sets. What do you notice? Use the regression line to predict y for $x = 10$.

(b) Make a scatterplot for each of the data sets and add the regression line to each plot.

(c) In which of the four cases would you be willing to use the regression line to describe the dependence of y on x? Explain your answer in each case.

3.66 FOOD POISONING Here are data on 18 people who fell ill from an incident of food poisoning.[21] The data give each person's age in years, the incubation period (the time in hours between eating the infected food and the first signs of illness), and whether the victim survived (S) or died (D).

Person:	1	2	3	4	5	6	7	8	9
Age:	29	39	44	37	42	17	38	43	51
Incubation:	13	46	43	34	20	20	18	72	19
Outcome:	D	S	S	D	D	S	D	S	D

Person:	10	11	12	13	14	15	16	17	18
Age:	30	32	59	33	31	32	32	36	50
Incubation:	36	48	44	21	32	86	48	28	16
Outcome:	D	D	S	D	D	S	D	S	D

(a) Make a scatterplot of incubation period against age, using different symbols for people who died and those who survived.

(b) Is there an overall relationship between age and incubation period? If so, describe it.

(c) More important, is there a relationship between either age or incubation period and whether the victim survived? Describe any relations that seem important here.

(d) Are there any unusual cases that may require individual investigation?

3.67 NEMATODES AND TOMATOES Nematodes are microscopic worms. Here are data from an experiment to study the effect of nematodes in the soil on plant growth. The experimenter prepared 16 planting pots and introduced different numbers of nematodes. Then he placed a tomato seedling in each pot and measured its growth (in centimeters) after 16 days.[22]

Nematodes	Seedling growth (cm)			
0	10.8	9.1	13.5	9.2
1,000	11.1	11.1	8.2	11.3
5,000	5.4	4.6	7.4	5.0
10,000	5.8	5.3	3.2	7.5

Analyze these data and give your conclusions about the effects of nematodes on plant growth.

3.68 A HOT STOCK? It is usual in finance to describe the returns from investing in a single stock by regressing the stock's returns on the returns from the stock market as a whole. This helps us see how closely the stock follows the market. We analyzed the monthly percent total return y on Philip Morris common stock and the monthly return x on the Standard & Poor's 500 Index, which represents the market, for the period between July 1990 and May 1997. Here are the results:

$$\bar{x} = 1.304 \quad s_x = 3.392 \quad r = 0.5251$$
$$\bar{y} = 1.878 \quad s_y = 7.554$$

A scatterplot shows no very influential observations.

(a) Find the equation of the least-squares line from this information. What percent of the variation in Philip Morris stock is explained by the linear relationship with the market as a whole?

(b) Explain carefully what the slope of the line tells us about how Philip Morris stock responds to changes in the market. This slope is called "beta" in investment theory.

(c) Returns on most individual stocks have a positive correlation with returns on the entire market. That is, when the market goes up, an individual stock tends to also go up. Explain why an investor should prefer stocks with beta > 1 when the market is rising and stocks with beta < 1 when the market is falling.

3.69 HUSBANDS AND WIVES The mean height of American women in their early twenties is about 64.5 inches and the standard deviation is about 2.5 inches. The mean height of men the same age is about 68.5 inches, with standard deviation about 2.7 inches. If the correlation between the heights of husbands and wives is about $r = 0.5$, what is the slope of the regression line of the husband's height on the wife's height in young couples? Draw a graph of this regression line. Predict the height of the husband of a woman who is 67 inches tall.

3.70 MEASURING ROAD STRENGTH Concrete road pavement gains strength over time as it cures. Highway builders use regression lines to predict the strength after 28 days (when curing is complete) from measurements made after 7 days. Let x be strength after 7 days (in pounds per square inch) and y the strength after 28 days. One set of data gives this least-squares regression line:

$$\hat{y} = 1389 + 0.96x$$

(a) Draw a graph of this line, with x running from 3000 to 4000 pounds per square inch.

(b) Explain what the slope $b = 0.96$ in this equation says about how concrete gains strength as it cures.

(c) A test of some new pavement after 7 days shows that its strength is 3300 pounds per square inch. Use the equation of the regression line to predict the strength of this pavement after 28 days. Also draw the "up and over" lines from $x = 3300$ on your graph (as in Figure 3.10. page 150).

3.71 COMPETITIVE RUNNERS Good runners take more steps per second as they speed up. Here are the average numbers of steps per second for a group of top female runners at different speeds. The speeds are in feet per second.[23]

Speed (ft/s):	15.86	16.88	17.50	18.62	19.97	21.06	22.11
Steps per second:	3.05	3.12	3.17	3.25	3.36	3.46	3.55

(a) You want to predict steps per second from running speed. Make a scatterplot of the data with this goal in mind.

(b) Describe the pattern of the data and find the correlation.

(c) Find the least-squares regression line of steps per second on running speed. Draw this line on your scatterplot.

(d) Does running speed explain most of the variation in the number of steps a runner takes per second? Calculate r^2 and use it to answer this question.

(e) If you wanted to predict running speed from a runner's steps per second, would you use the same line? Explain your answer. Would r^2 stay the same?

3.72 RESISTANCE REVISITED

(a) Is correlation a resistant measure? Give an example to support your answer.

(b) Is the least-squares regression line resistant? Give an example to support your answer.

3.73 BANK FAILURES The Franklin National Bank failed in 1974. Franklin was one of the 20 largest banks in the nation, and the largest ever to fail. Could Franklin's weakened condition have been detected in advance by simple data analysis? The table below gives the total assets (in billions of dollars) and net income (in millions of dollars) for the 20 largest banks in 1973, the year before Franklin failed.[24] Franklin is bank number 19.

Bank:	1	2	3	4	5	6	7	8	9	10
Assets:	49.0	42.3	36.3	16.4	14.9	14.2	13.5	13.4	13.2	11.8
Income:	218.8	265.6	170.9	85.9	88.1	63.6	96.9	60.9	144.2	53.6
Bank:	11	12	13	14	15	16	17	18	19	20
Assets:	11.6	9.5	9.4	7.5	7.2	6.7	6.0	4.6	3.8	3.4
Income:	42.9	32.4	68.3	48.6	32.2	42.7	28.9	40.7	13.8	22.2

(a) We expect banks with more assets to earn higher income. Make a scatterplot of these data that displays the relation between assets and income. Mark Franklin (Bank 19) with a separate symbol.

(b) Describe the overall pattern of your plot. Are there any banks with unusually high or low income relative to their assets? Does Franklin stand out from other banks in your plot?

(c) Find the least-squares regression line for predicting a bank's income from its assets. Draw the regression line on your scatterplot.

(d) Use the regression line to predict Franklin's income. Was the actual income higher or lower than predicted? What is the residual?

3.74 CAN YOU THINK OF A SCATTERPLOT?

(a) Draw a scatterplot that has a positive correlation such that when one point is added, the correlation becomes negative. Circle the influential point.

(b) Draw a scatterplot that has a correlation close to 0 (say less than 0.1) such that when one point is added, the correlation is close to 1 (say greater than 0.9). Circle the influential point.

3.75 WILL WOMEN SOON OUTRUN MEN? Table 3.8 shows the men's and women's world records in the 800-meter run.

TABLE 3.8 Men's and women's world records in the 800-meter run

Year	Men's record	Women's record	Year	Men's record	Women's record
1905	113.4	—	1955	105.7	125.0
1915	111.9	—	1965	104.3	118.0
1925	111.9	144.0	1975	104.1	117.5
1935	109.7	135.6	1985	101.73	113.28
1945	106.6	132.0	1995	101.73	113.28

Source: This exercise was suggested in an article by Edward Wallace in *Mathematics Teacher*, 86, no. 9 (December 1993), p. 741.

(a) For each gender separately, do the following: Enter the data into your calculator or computer package and then plot a scatterplot. (Use the box plotting symbol for the men, and use the + plotting symbol for the women.) Describe the trend, if there is one. Perform least-squares regression and calculate the correlation. Comment on the suitability of the LSRL as a model for the data and interpret the correlation. Identify any regression outliers and influential observations.

(b) Brian Whipp and Susan Ward wrote an article based on the 800-meter run data entitled "Will Women Soon Outrun Men?" which appeared in the British journal *Nature* in 1992. They suggested in the article that women have made more progress in track events over the last half-century than men, hence the title of the article. Extend your calculator viewing window so that you can see both data sets and least-squares lines, and determine the intersection of the two LSRLs. Then comment on the premise of the *Nature* article.

3.76 MORE ON MANATEES Exercises 3.6 (page 125), 3.9 (page 129) and 3.41 (page 157) investigated the association between manatees killed and the number of powerboat registrations. For this exercise, you are to use the data for the years 1977 to 1994. Here is part of the output from the regression command in the Minitab statistical software:

```
The regression equation is
Killed = -35.2 + 0.113 Boats

Unusual Observations
Obs.  Boats  Killed    Fit  Stdev.Fit   Residual  St.Resid
  17    716   35.00  45.51       1.92     -10.51    -2.08R

R denotes an obs. with a large st. resid.
```

(a) Minitab checks for large residuals and influential observations. It calls attention to one observation that has a somewhat large residual. Circle this observation on your plot. We have no reason to remove it.

(b) Residuals from least-squares regression often have a distribution that is roughly normal. So Minitab reports the *standardized* residuals—that's what `St.Resid` means. Use the 68–95–99.7 rule for normal distributions to say how surprising a residual with standardized value –2.08 is.

3.77 JET SKI FATALITIES Exercise 3.7 (page 125) examined the association between the number of jet ski accidents and the number of jet skis in use during the period 1987 to 1996. The data also included the number of fatalities during those years.

(a) Use the methods of this chapter to investigate a possible association between the number of fatalities and the number of jet skis in use. Report your findings and support them with the appropriate numerical and graphical analyses.

(b) Use a search engine on the Internet to see which states have passed laws to regulate the use of jet skis in an attempt to reduce the number of accidents and fatalities. Are there any federal regulations for the operation of jet skis?

NOTES AND DATA SOURCES

1. Data from Personal Watercraft Industry Association, U.S. Coast Guard.
2. Based on T. N. Lam, "Estimating fuel consumption from engine size," *Journal of Transportation Engineering*, 111 (1985), pp. 339–357. The data for 10 to 50 km/h are measured; those for 60 and higher are calculated from a model given in the paper and are therefore smoothed.
3. A sophisticated treatment of improvements and additions to scatterplots is W. S. Cleveland and R. McGill, "The many faces of a scatterplot," *Journal of the American Statistical Association*, 79 (1984), pp. 807–822.
4. Data provided by Darlene Gordon, Purdue University.
5. Data from *Consumer Reports*, June 1986, pp. 366–367.
6. Data for 1995, from the 1997 *Statistical Abstract of the United States*.
7. The data are from M. A. Houck et al., "Allometric scaling in the earliest fossil bird, *Archaeopteryx lithographica*," *Science*, 247 (1990), pp. 195–198. The authors conclude from a variety of evidence that all specimens represent the same species.
8. From a survey by the Wheat Industry Council reported in *USA Today*, October 20, 1983.
9. The data are from W. L. Colville and D. P. McGill, "Effect of rate and method of planting on several plant characters and yield of irrigated corn," *Agronomy Journal*, 54 (1962), pp. 235–238.
10. Modified from M. C. Wilson and R. E. Shade, "Relative attractiveness of various luminescent colors to the cereal leaf beetle and the meadow spittlebug," *Journal of Economic Entomology*, 60 (1967), pp. 578–580.
11. A careful study of this phenomenon is W. S. Cleveland, P. Diaconis, and R. McGill, "Variables on scatterplots look more highly correlated when the scales are increased," *Science*, 216 (1982), pp. 1138–1141.
12. *T. Rowe Price Report*, winter 1997, p. 4.
13. From W. M. Lewis and M. C. Grant, "Acid precipitation in the western United States," *Science*, 207 (1980), pp. 176–177.
14. Data from E. P. Hubble, "A relation between distance and radial velocity among extra-galactic nebulae," *Proceedings of the National Academy of Sciences*, 15 (1929), pp. 168–173.
15. Based on a plot in G. D. Martinsen, E. M. Driebe, and T. G. Whitham, "Indirect interactions mediated by changing plant chemistry: beaver browsing benefits beetles," *Ecology*, 79 (1998), pp. 192–200.

16. From the *Philadelphia City Paper*, May 23–29, 1997. Because the sodas served vary in size, we have converted soda prices to the price of a 16-ounce soda at each price per ounce.
17. From a presentation by Charles Knauf, Monroe County (New York) Environmental Health Laboratory.
18. Gary Smith, "Do statistics test scores regress toward the mean?" *Chance*, 10, No. 4(1997), pp. 42–45.
19. Data provided by Peter Cook, Purdue University.
20. The data are a random sample of 53 from the 1079 pairs recorded by K. Pearson and A. Lee, "On the laws of inheritance in man," *Biometrika*, November 1903, p. 408.
21. Modified from data provided by Dana Quade, University of North Carolina.
22. Data provided by Matthew Moore.
23. Data from R.C. Nelson, C.M. Brooks, and N.L. Pike, "Biomechanical comparison of male and female distance runners," in P. Milvy (ed.), *The Marathon: Physiological, Medical, Epidemiological, and Psychological Studies*, New York Academy of Sciences, 1977, pp. 793–807.
24. Data from D.E. Booth, *Regression Methods and Problem Banks*, COMAP, Inc., 1986.