Chapter 11

Regression

CHAPTER OUTLINE

Sir Francis Galton (1822–1911) had earned a Cambridge mathematics degree and completed two years of medical school when his father died, leaving him with a substantial inheritance. Free to travel, he became an explorer of some note, but when The Origin of Species *was published in 1859, his interests began to shift from geography to statistics and anthropology (Charles Darwin was his cousin). It was Galton's work on fingerprints that made possible their use in human identification. He was knighted in 1909.*

11.1 INTRODUCTION

High on the list of problems that experimenters most frequently need to deal with is the determination of the relationships that exist among the various components of a complex system. If those relationships are sufficiently understood, there is a good possibility that the system's output can be effectively modeled, maybe even controlled.

Consider, for example, the formidable problem of relating the incidence of cancer to its many contributing causes—diet, genetic makeup, pollution, and cigarette smoking, to name only a few. Or think of the Wall Street financier trying to anticipate trends in stock prices by tracking market indices and corporate performances, as well as the overall economic climate. In those situations, a host of variables are involved, and the analysis becomes very intricate. Fortunately, many of the fundamental ideas associated with the study of relationships can be nicely illustrated when only *two* variables are involved. This two-variable model will be the focus of Chapter 11.

Section 11.2 gives a computational technique for determining the "best" equation describing a set of points $(x_1, y_1), (x_2, y_2), \ldots$, and (x_n, y_n), where *best* is defined geometrically. Section 11.3 adds a probability distribution to the y-variable, which allows for a variety of inference procedures to be developed. The consequences of both measurements being random variables is the topic of Section 11.4. Then Section 11.5 takes up a special case of Section 11.4, where the variability in X and Y is described by the *bivariate normal pdf*.

11.2 The Method of Least Squares

We begin our study of the relationship between two variables by asking a simple geometry question. Given a set of n points—$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$—and a positive integer m, which polynomial of degree m is "closest" to the given points?

Suppose that the desired polynomial, $p(x)$, is written

$$p(x) = a + \sum_{i=1}^{m} b_i x^i$$

where $a, b_1, \ldots, b_m$ are to be determined. The *method of least squares* answers the question by finding the coefficient values that minimize the sum of the squares of the vertical distances from the data points to the presumed polynomial. That is, the polynomial $p(x)$ that we will call "best" is the one whose coefficients minimize the function L, where

$$L = \sum_{i=1}^{n} [y_i - p(x_i)]^2$$

Theorem 11.2.1 summarizes the method of least squares as it applies to the important special case where $p(x)$ is a *linear* polynomial. (*Note:* To simplify notation, the linear polynomial $y = a + b_1 x^1$ will be written $y = a + bx$.)

Theorem 11.2.1 *Given n points* $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, *the straight line* $y = a + bx$ *minimizing*

$$L = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

has slope

$$b = \frac{n \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i \right) \left(\sum_{i=1}^{n} y_i \right)}{n \left(\sum_{i=1}^{n} x_i^2 \right) - \left(\sum_{i=1}^{n} x_i \right)^2}$$

and y-intercept

$$a = \frac{\sum_{i=1}^{n} y_i - b \sum_{i=1}^{n} x_i}{n} = \bar{y} - b\bar{x}$$

Proof The proof is accomplished by the familiar calculus technique of taking the partial derivatives of L with respect to a and b, setting the resulting expressions equal to 0, and solving. By the first step we get

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{n} (-2) x_i [y_i - (a + bx_i)]$$

and

$$\frac{\partial L}{\partial a} = \sum_{i=1}^{n} (-2) [y_i - (a + bx_i)]$$

Setting the right-hand sides of $\partial L / \partial a$ and $\partial L / \partial b$ equal to 0 and simplifying yields the two equations

$$na + \left(\sum_{i=1}^{n} x_i \right) b = \sum_{i=1}^{n} y_i$$

(Continued on next page)

(Theorem 11.2.1 continued)

and

$$\left(\sum_{i=1}^{n} x_i\right) a + \left(\sum_{i=1}^{n} x_i^2\right) b = \sum_{i=1}^{n} x_i y_i$$

An application of Cramer's rule gives the solution for b stated in the theorem. The expression for a follows immediately.

CASE STUDY 11.2.1

A manufacturer of air conditioning units is having assembly problems due to the failure of a connecting rod to meet finished-weight specifications. Too many rods are being completely tooled, then rejected as overweight. To reduce that cost, the company's quality-control department wants to quantify the relationship between the weight of the finished rod, y, and that of the rough casting, x (149). Castings likely to produce rods that are too heavy can then be discarded before undergoing the final (and costly) tooling process.

As a first step in examining the xy-relationship, twenty-five (x_i, y_i) pairs are measured (see Table 11.2.1). Graphed, the points suggest that the weight (in ounces) of the finished rod is linearly related to the weight of the rough casting (see Figure 11.2.1). Use Theorem 11.2.1 to find the best straight line approximating the xy-relationship.

Table 11.2.1

Rod Number	Rough Weight, x	Finished Weight, y	Rod Number	Rough Weight, x	Finished Weight, y
1	2.745	2.080	14	2.635	1.990
2	2.700	2.045	15	2.630	1.990
3	2.690	2.050	16	2.625	1.995
4	2.680	2.005	17	2.625	1.985
5	2.675	2.035	18	2.620	1.970
6	2.670	2.035	19	2.615	1.985
7	2.665	2.020	20	2.615	1.990
8	2.660	2.005	21	2.615	1.995
9	2.655	2.010	22	2.610	1.990
10	2.655	2.000	23	2.590	1.975
11	2.650	2.000	24	2.590	1.995
12	2.650	2.005	25	2.565	1.955
13	2.645	2.015			

From Table 11.2.1, we find that

$$\sum_{i=1}^{25} x_i = 66.075 \qquad \sum_{i=1}^{25} x_i^2 = 174.672925$$

$$\sum_{i=1}^{25} y_i = 50.12 \qquad \sum_{i=1}^{25} y_i^2 = 100.49865$$

$$\sum_{i=1}^{25} x_i y_i = 132.490725$$

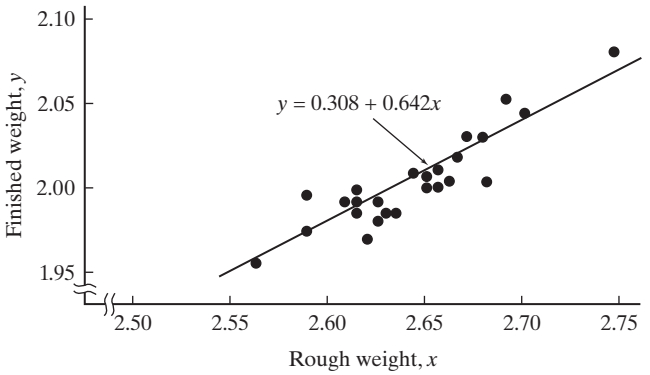


Figure 11.2.1

Therefore,

$$b = \frac{25(132.490725) - (66.075)(50.12)}{25(174.672925) - (66.075)^2} = 0.642$$

and

$$a = \frac{50.12 - 0.642(66.075)}{25} = 0.308$$

making the least squares line

$$y = 0.308 + 0.642x$$

The manufacturer is now in a position to make some informed policy decisions. If the weight of a rough casting is, say, 2.71 oz., the least squares line predicts that its finished weight will be 2.05 oz.:

$$\text{estimated weight} = a + b(2.71) = 0.308 + 0.642(2.71) = 2.05$$

In the event that finished weights of 2.05 oz. are considered to be too heavy, rough castings weighing 2.71 oz. (or more) should be discarded.

RESIDUALS

The difference between an observed y_i and the value of the least squares line when $x = x_i$ is called the ith *residual*. Its magnitude reflects the failure of the least squares line to "model" that particular point.

Definition 11.2.1
Let a and b be the least squares coefficients associated with the sample (x_1, y_1), $(x_2, y_2), \ldots, (x_n, y_n)$. For any value of x, the quantity $\hat{y} = a + bx$ is known as the *predicted value* of y. For any given i, $i = 1, 2, \ldots, n$, the difference $y_i - \hat{y}_i = y_i - (a + bx_i)$ is called the ith *residual*. A graph of $y_i - \hat{y}_i$ versus x_i, for all i, is called a *residual plot*.

INTERPRETING RESIDUAL PLOTS

Applied statisticians find residual plots to be very helpful in assessing the appropriateness of fitting a straight line through a given set of n points. If the relationship between x and y *is* linear, the corresponding residual plot typically shows no patterns, cycles, trends, or outliers. For nonlinear relationships, though, residual plots often take on dramatically nonrandom appearances that can very effectively highlight and illuminate the underlying association between x and y.

Example 11.2.1 Make the residual plot for the data in Case Study 11.2.1. What does its appearance imply about the suitability of fitting those points with a straight line?

We begin by calculating the residuals for each of the twenty-five data points. The first observation recorded, for example, was $(x_1, y_1) = (2.745, 2.080)$. The corresponding predicted value, $\hat{y}_1$, is 2.070:

$$\begin{aligned}\hat{y}_1 &= 0.308 + 0.642(2.745)\\ &= 2.070\end{aligned}$$

The first residual, then, is $y_1 - \hat{y}_1 = 2.080 - 2.070$, or 0.010. The complete set of residuals appears in the last column of Table 11.2.2.

Table 11.2.2

x_i	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$
2.745	2.080	2.070	0.010
2.700	2.045	2.041	0.004
2.690	2.050	2.035	0.015
2.680	2.005	2.029	−0.024
2.675	2.035	2.025	0.010
2.670	2.035	2.022	0.013
2.665	2.020	2.019	0.001
2.660	2.005	2.016	−0.011
2.655	2.010	2.013	−0.003
2.655	2.000	2.013	−0.013
2.650	2.000	2.009	−0.009
2.650	2.005	2.009	−0.004
2.645	2.015	2.006	0.009
2.635	1.990	2.000	−0.010
2.630	1.990	1.996	−0.006
2.625	1.995	1.993	0.002
2.625	1.985	1.993	−0.008
2.620	1.970	1.990	−0.020
2.615	1.985	1.987	−0.002
2.615	1.990	1.987	0.003
2.615	1.995	1.987	0.008
2.610	1.990	1.984	0.006
2.590	1.975	1.971	0.004
2.590	1.995	1.971	0.024
2.565	1.955	1.955	0.000

Figure 11.2.2

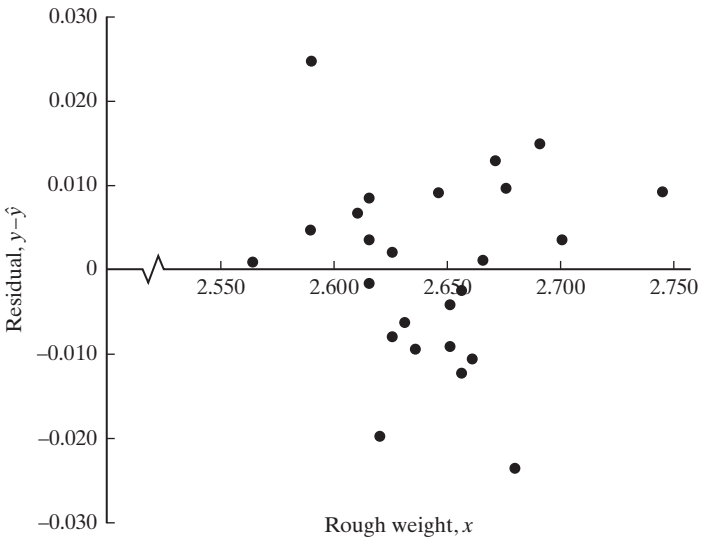


Figure 11.2.2 shows the residual plot generated by fitting the least squares straight line, $y = 0.308 + 0.642x$, to the twenty-five (x_i, y_i)'s. To an applied statistician, there is nothing here that would raise any serious doubts about using a straight line to describe the xy-relationship—the points appear to be randomly scattered and exhibit no obvious anomalies or patterns. ■

CASE STUDY 11.2.2

Table 11.2.3 lists Social Security expenditures for five-year intervals from 1965 to 2005. During that period, payouts rose from \$19.2 billion to \$529.9 billion. Substituting these nine (x_i, y_i)'s into the formulas in Theorem 11.2.1 gives

$$y = -38.0 + 12.9x$$

as the least squares straight line describing the xy-relationship. Based on the data from 1965 to 2005, was it reasonable to predict that Social Security costs in the year 2010 (when $x = 45$) would be \$543 billion $[= -38.0 + 12.9(45)]$?

Table 11.2.3

Year	Years after 1965, x	Social Security Expenditures (\$ billions), y
1965	0	19.2
1970	5	33.1
1975	10	69.2
1980	15	123.6
1985	20	190.6
1990	25	253.1
1995	30	339.8
2000	35	415.1
2005	40	529.9

Data from: www.socialsecurity.gov/history/trustfunds.html.

(Continued on next page)

(Case Study 11.2.2 continued)

Not at all. At first glance, the least squares line does appear to fit the data quite well (see Figure 11.2.3). A closer look, though, suggests that the underlying *xy*-relationship may be curvilinear rather than linear. The residual plot (Figure 11.2.4) confirms that suspicion—there we see a distinctly nonrandom pattern.

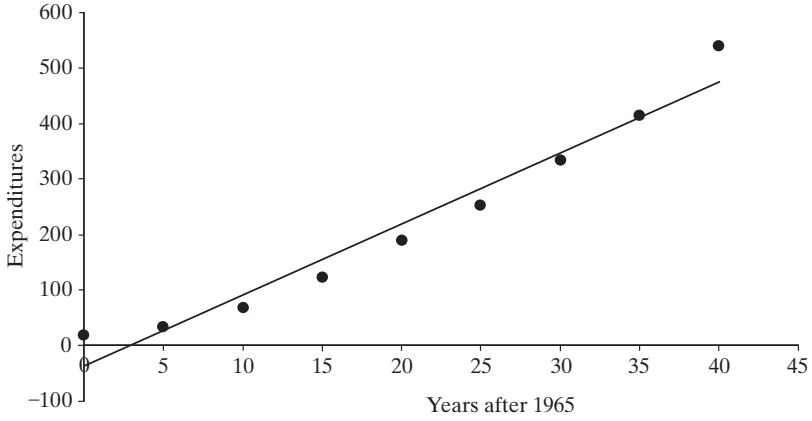


Figure 11.2.3

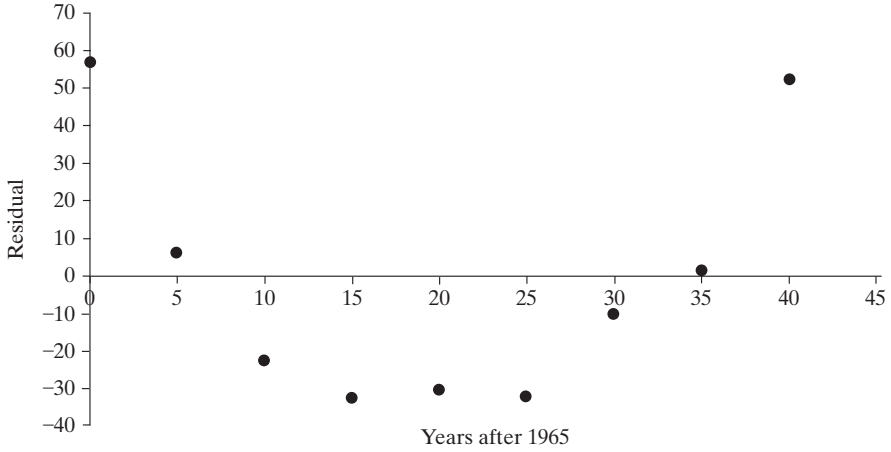


Figure 11.2.4

Clearly, extrapolating these data would be foolish. The figure for the next fifth year period, 2010, of $713 billion already exceeded the linear projection of $543 billion, leading economists to predict rapidly accelerating expenditures in the future.

Comment For the data in Table 11.2.3, the suggestion that the *xy*-relationship may be curvilinear is certainly present in Figure 11.2.3, but the residual plot makes the case much more emphatically. In point of fact, that will often be the case, which is why residual plots are such a valuable diagnostic tool—departures from randomness that may be only hinted at in an *xy*-plot will be illustrated more clearly in the corresponding residual plot.

CASE STUDY 11.2.3

A new, presumably simpler laboratory procedure has been proposed for recovering calcium oxide (CaO) from solutions that contain magnesium. Critics of the method argue that the results are too dependent on the person who performs the analysis. To demonstrate their concern, they arrange for the procedure to be run on ten samples, each containing a known amount of CaO. Nine of the ten tests are done by Chemist A; the other is run by Chemist B. Based on the results summarized in Table 11.2.4, does their criticism seem justified?

Table 11.2.4

Chemist	CaO Present (in mg), x	CaO Recovered (in mg), y
A	4.0	3.7
A	8.0	7.8
A	12.5	12.1
A	16.0	15.6
A	20.0	19.8
A	25.0	24.5
B	31.0	31.1
A	36.0	35.5
A	40.0	39.4
A	40.0	39.5

Figure 11.2.5 shows the scatterplot of y versus x. The linear function appears to fit all ten points exceptionally well, which would suggest that the critics' concerns are unwarranted. But look at the residual plot (Figure 11.2.6). The latter shows one point located noticeably farther away from zero than any of the others, and that point corresponds to the one measurement attributed to

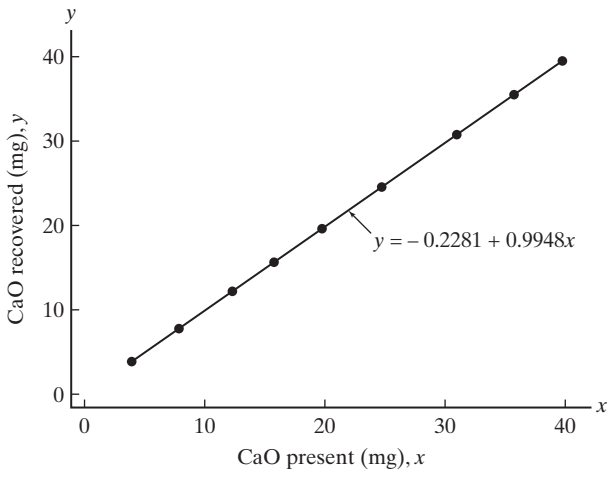


Figure 11.2.5

(Continued on next page)

(Case Study 11.2.3 continued)

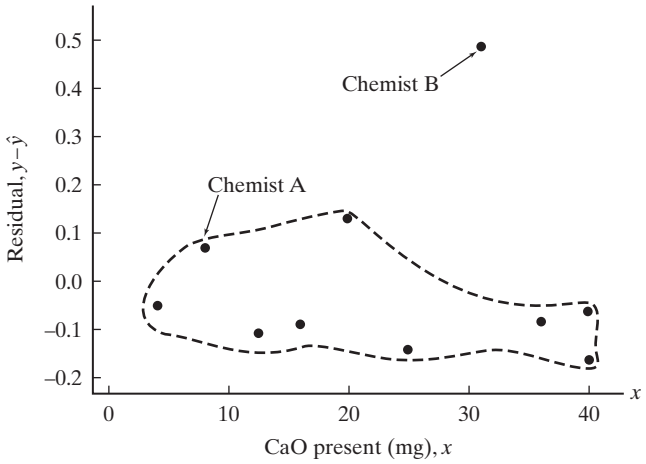


Figure 11.2.6

Chemist B. So, while the scatterplot has failed to identify anything unusual about the data, the residual plot has focused on precisely the question the data set out to answer.

Does the appearance of the residual plot—specifically, the separation between the Chemist B data point and the nine Chemist A data points—"prove" that the output from the new procedure is dependent on the analyst? No, but it does speak to the magnitude of the disparity and, in so doing, provides the critics with at least a partial answer to their original question.

Questions

11.2.1. Crickets make their chirping sound by sliding one wing cover very rapidly back and forth over the other. Biologists have long been aware that there is a linear relationship between *temperature* and the *frequency* with which a cricket chirps, although the slope and y-intercept of the relationship vary from species to species. The following table lists fifteen frequency-temperature observations recorded for the striped ground cricket, *Nemobius fasciatus fasciatus* (145). Plot these data and find the equation of the least squares line, $y = a + bx$. Suppose a cricket of this species is observed to chirp eighteen times per second. What would be the estimated temperature?

For the data in the table, the sums needed are:

$$\sum_{i=1}^{15} x_i = 249.8 \qquad \sum_{i=1}^{15} x_i^2 = 4{,}200.56$$

$$\sum_{i=1}^{15} y_i = 1{,}200.6 \qquad \sum_{i=1}^{15} x_i y_i = 20{,}127.47$$

Observation Number	Chirps per Second, x	Temperature, y (°F)
1	20.0	88.6
2	16.0	71.6
3	19.8	93.3
4	18.4	84.3
5	17.1	80.6
6	15.5	75.2
7	14.7	69.7
8	17.1	82.0
9	15.4	69.4
10	16.2	83.3
11	15.0	79.6
12	17.2	82.6
13	16.0	80.6
14	17.0	83.5
15	14.4	76.3

11.2.2. The aging of whisky in charred oak barrels brings about a number of chemical changes that enhance its taste and darken its color. The following table shows the change in a whisky's proof as a function of the number of years it is stored (168).

Age, x (years)	Proof, y
0	104.6
0.5	104.1
1	104.4
2	105.0
3	106.0
4	106.8
5	107.7
6	108.7
7	110.6
8	112.1

(*Note:* The proof initially decreases because of dilution from moisture in the staves of the barrels.) Graph these data and draw in the least squares line.

11.2.3. As water temperature increases, sodium nitrate ($NaNO_3$) becomes more soluble. The following table (110) gives the number of parts of sodium nitrate that dissolve in one hundred parts of water.

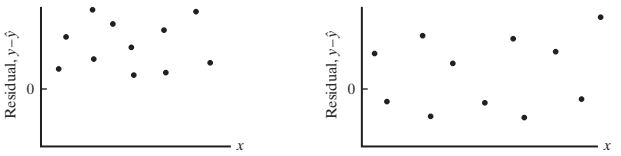
Temperature (degrees Celsius), x	Parts Dissolved, y
0	66.7
4	71.0
10	76.3
15	80.6
21	85.7
29	92.9
36	99.4
51	113.6
68	125.1

Calculate the residuals, $y_1-\hat{y}_1, \ldots, y_9-\hat{y}_9$, and draw the residual plot. Does it suggest that fitting a straight line through these data would be appropriate? Use the following sums:

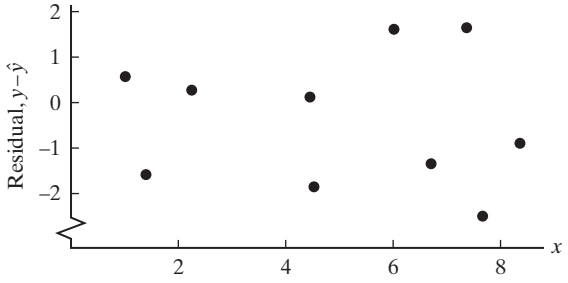
$$\sum_{i=1}^{9} x_i = 234 \qquad \sum_{i=1}^{9} y_i = 811.3$$

$$\sum_{i=1}^{9} x_i^2 = 10{,}144 \qquad \sum_{i=1}^{9} x_i y_i = 24{,}628.6$$

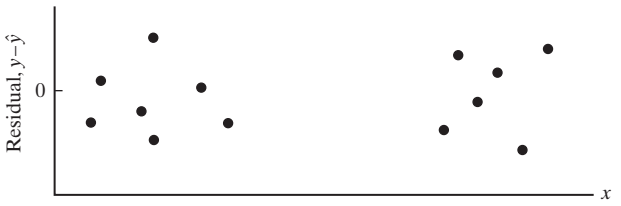
11.2.4. What, if anything, is unusual about the following residual plots?



11.2.5. The following is the residual plot that results from fitting the equation $y = 6.0+2.0x$ to a set of $n = 10$ points. What, if anything, would be wrong with predicting that y will equal 30.0 when $x = 12$?



11.2.6. Would the following residual plot produced by fitting a least squares straight line to a set of $n = 13$ points cause you to doubt that the underlying xy-relationship is linear? Explain.



11.2.7. The relationship between school funding and student performance continues to be a hotly debated political and philosophical issue. Typical of the data available are the following figures, showing the per-pupil expenditures and graduation rate for twenty-six randomly chosen districts in Massachusetts.

Graph the data and superimpose the least squares line, $y = a + bx$. What would you conclude about the xy-relationship? Use the following sums:

$$\sum_{i=1}^{26} x_i = 360 \qquad \sum_{i=1}^{26} y_i = 2{,}256.6$$

$$\sum_{i=1}^{26} x_i^2 = 5{,}365.08 \qquad \sum_{i=1}^{26} x_i y_i = 31{,}402$$

District	Spending per Pupil (in 1000s), x	Graduation Rate, y
Dighton-Rehoboth	$10.0	88.7
Duxbury	$10.2	93.2
Tyngsborough	$10.2	95.1
Lynnfield	$10.3	94.0
Southwick-Tolland	$10.3	88.3
Clinton	$10.8	89.9
Athol-Royalston	$11.0	67.7
Tantasqua	$11.0	90.2
Ayer	$11.2	95.5
Adams-Cheshire	$11.6	75.2
Danvers	$12.1	84.6
Lee	$12.3	85.0
Needham	$12.6	94.8
New Bedford	$12.7	56.1
Springfield	$12.9	54.4
Manchester Essex	$13.0	97.9
Dedham	$13.9	83.0
Lexington	$14.5	94.0
Chatham	$14.7	91.4
Newton	$15.5	94.2
Blackstone Valley	$16.4	97.2
Concord Carlisle	$17.5	94.4
Pathfinder	$18.1	78.6
Nantucket	$20.8	87.6
Essex	$22.4	93.3
Provincetown	$24.0	92.3

Data from: profiles.doe.mass.edu/state–report/ppx.aspx.

11.2.8. (a) Find the equation of the least squares straight line for the plant cover diversity/bird species diversity data given in Question 8.2.11.
(b) Make the residual plot associated with the least squares fit asked for in part (a). Based on the appearance of the residual plot, would you conclude that fitting a straight line to these data is appropriate? Explain.

11.2.9. An Atomic Energy Commission nuclear facility was established in Hanford, Washington, in 1943. Over the years, a significant amount of strontium 90 and cesium 137 leaked into the Columbia River. In a study to determine how much this radioactivity caused serious medical problems for those who lived along the river, public health officials created an index of radioactive exposure for nine Oregon counties in the vicinity of the river. As a covariate, cancer mortality was determined for each of the counties (45). The results are given in the table in the next column. For the nine (x_i, y_i)'s in the table,

$$\sum_{i=1}^{9} x_i = 41.56 \qquad \sum_{i=1}^{9} x_i^2 = 289.4222$$

$$\sum_{i=1}^{9} y_i = 1{,}416.1 \qquad \sum_{i=1}^{9} x_i y_i = 7{,}439.37$$

County	Index of Exposure, x	Cancer Mortality per 100,000, y
Umatilla	2.49	147.1
Morrow	2.57	130.1
Gilliam	3.41	129.9
Sherman	1.25	113.5
Wasco	1.62	137.5
Hood River	3.83	162.3
Portland	11.64	207.5
Columbia	6.41	177.9
Clatsop	8.34	210.3

Find the least squares straight line for these points. Also, construct the corresponding residual plot. Does it seem reasonable to conclude that x and y are linearly related?

11.2.10. Would you have any reservations about fitting the following data with a straight line? Explain.

x	y
3	20
7	37
5	29
1	10
10	59
12	69
6	39
11	58
8	47
9	48
2	18
4	29

11.2.11. When two closely related species are crossed, the progeny will tend to have physical traits that lie somewhere between those of the two parents. Whether a similar mixing occurs with behavioral traits was the focus of an experiment where the subjects were mallard and pintail ducks (173). A total of eleven males were studied; all were second-generation crosses. A rating scale was devised that measured the extent to which the plumage of each of the ducks resembled the plumage of the first generation's parents. A score of 0 indicated that the hybrid had the same appearance (phenotype) as a pure mallard; a score of 20 meant that the hybrid looked like a pintail. Similarly, certain behavioral traits were quantified and a second scale was constructed that ranged from 0 (completely mallard-like) to 15 (completely pintail-like). Use Theorem 11.2.1 and the following data to summarize the relationship between the plumage and behavioral indices. Does a linear model seem adequate?

Male	Plumage Index, x	Behavioral Index, y
R	7	3
S	13	10
D	14	11
F	6	5
W	14	15
K	15	15
U	4	7
O	8	10
V	7	4
J	9	9
L	14	11

11.2.12. Verify that the coefficients a and b of the least squares straight line are solutions of the matrix equation

$$\begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix}$$

11.2.13. Prove that a least squares straight line must necessarily pass through the point $(\bar{x}, \bar{y})$.

11.2.14. In some regression situations, there are *a priori* reasons for assuming that the xy-relationship being approximated passes through the origin. If so, the equation to be fit to the (x_i, y_i)'s has the form $y = bx$. Use the least squares criterion to show that the "best" slope in that case is given by

$$b = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

11.2.15. Case Study 8.2.6 discusses the expansion of the universe, as astronomer Edwin Hubble announced in 1929. Hubble's law states that $v = Hd$, where v is a galaxy's recession velocity relative to that of any other galaxy and d is its distance from that same galaxy. Table 8.2.6 gives distance and velocity measurements made on eleven galactic clusters (26). Use the formula cited in Question 11.2.14 and these data to estimate H, Hubble's constant.

11.2.16. Given a set of n linearly related points, $(x_1, y_1), (x_2, y_2), \ldots,$ and (x_n, y_n), use the least squares criterion to find formulas for

(a) a if the slope of the xy-relationship is known to be b^*.

(b) b if the y-intercept of the xy-relationship is known to be a^*.

11.2.17. Among the problems faced by job seekers wanting to reenter the workforce, eroded skills and outdated backgrounds are two of the most difficult to overcome. Knowing that, employers are often wary of hiring individuals who have spent lengthy periods of time away from the job. The following table shows the percentages of hospitals willing to rehire medical technicians who have been away from that career for x years (154). It can be argued that the fitted line should necessarily have a y-intercept of 100 because no employer would refuse to hire someone (due to outdated skills) whose career had not been interrupted at all—that is, applicants for whom $x = 0$. Under that assumption, use the result from Question 11.2.16 to fit these data with the model $y = 100 + bx$.

Years of Inactivity, x	Percent of Hospitals Willing to Hire, y
0.5	100
1.5	94
4	75
8	44
13	28
18	17

11.2.18. A graph of the luxury suite data in Question 8.2.5 suggests that the xy-relationship is linear. Moreover, it makes sense to constrain the fitted line to go through the origin, since $x = 0$ suites will necessarily produce $y = 0$ revenue.

(a) Find the equation of the least squares line, $y = bx$. (*Hint:* Recall Question 11.2.14.)

(b) How much revenue would 120 suites be expected to generate?

11.2.19. Set up (but do not solve) the equations necessary to determine the least squares estimates for the trigonometric model,

$$y = a + bx + c \sin x$$

Assume that the data consist of the random sample $(x_1, y_1), (x_2, y_2), \ldots,$ and (x_n, y_n).

NONLINEAR MODELS

In Chapter 3 it was acknowledged that an infinite number of functions qualify as being either discrete or continuous random variables, but at the end of the day only a handful or so are important in the sense that they accurately model the probabilistic

behavior of real-world measurements. A similar disclaimer holds for regression functions. Graphed, a set of regression data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ could be configured in any of an infinite number of fundamentally different patterns, but the reality is just the opposite—only a small number of functions are needed to describe a large proportion of the regression data likely to be encountered.

Concluding this section is a discussion of three of the most widely-used nonlinear regression models—*exponential regression, logarithmic regression,* and *logistic regression*. Each differs from the others in having its own unique "growth rate" for y as a function of x. And those different growth rates are the starting points for deriving the model equations themselves. The simplest example of that linkage can be seen in the *linear regression* described earlier. In that case, the change in y as a function of x is equal to some fixed number, b. That is,

$$dy/dx = b \quad \text{or, equivalently,} \quad \int dy = \int b\,dx$$

which implies that

$$y = a + bx, \quad \text{where } a \text{ is the constant of integration.}$$

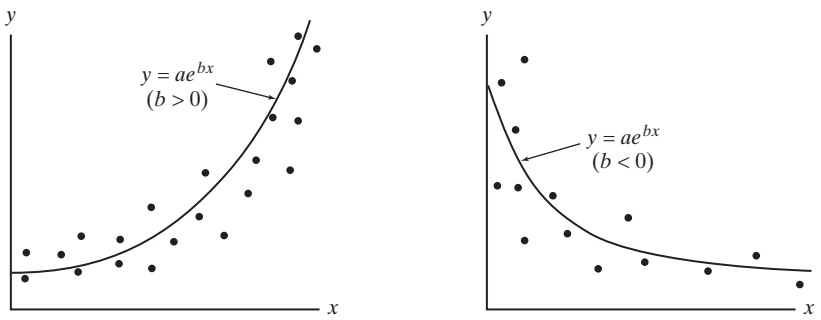
What these nonlinear models have in common is that each can be "linearized" by applying an appropriate transformation to either y and/or x. Doing so means that Theorem 11.2.1—which ostensibly dealt with straight lines—can be used to fit each of these nonlinear models as well.

Exponential Regression Suppose y depends on x, where changes in y are proportional to y, that is, $\frac{dy}{dx} = by$, where b is a constant. Then $\int \frac{dy}{y} = \int b\,dx$ or $\ln y = bx + c$ where c is a constant of integration. Exponentiation of both sides gives $y = e^{bx} \cdot e^{c}$. Since e^c is an arbitrary positive constant, for simplicity call it a. Thus, the growth relationship between the two variables implies that their relationship is of the form

$$y = ae^{bx} \tag{11.2.1}$$

Depending on the value of b, Equation 11.2.1 will look like one of the graphs pictured in Figure 11.2.7. Those curvilinear shapes notwithstanding, though, there is a *linear* model also related to Equation 11.2.1.

Figure 11.2.7



If $y = ae^{bx}$, it is necessarily true that

$$\ln y = \ln a + bx \tag{11.2.2}$$

which implies that *ln y and x have a linear relationship*. That being the case, the formulas of Theorem 11.2.1 *applied to x and ln y* should yield the slope and y-intercept of Equation 11.2.2.

Specifically,

$$b = \frac{n\sum_{i=1}^{n} x_i \ln y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} \ln y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

and

$$\ln a = \frac{\sum_{i=1}^{n} \ln y_i - b\sum_{i=1}^{n} x_i}{n}$$

Comment Transformations that induce linearity often require that the slope and/or y-intercept of the transformed model be transformed "back" to the original model. Here, for example, Theorem 11.2.1 leads to a formula for $\ln a$, which means that the constant a appearing in the original exponential model is evaluated by calculating $e^{\ln a}$.

CASE STUDY 11.2.4

Beginning in the 1970s, computers have steadily decreased in size as they have grown in power. The ability to have more computing potential in a four-pound laptop than in a mainframe of the 1970s is a result of engineers squeezing more and more transistors onto silicon chips. The rate at which this miniaturization occurs is known as Moore's law, after Gordon Moore, one of the founders of Intel Corporation. His prediction, first articulated in 1965, was that the number of transistors per chip would double every eighteen months.

Table 11.2.5 lists some of the growth benchmarks—namely, the number of transistors per chip—associated with the Intel chips marketed over the twenty-year period from 1975 through 1995. Based on these figures, is it believable that chip capacity is, in fact, doubling at a fixed rate (meaning that Equation 11.2.1 applies)? And if so, how close is the actual doubling time to Moore's prediction of eighteen months?

A plot of y versus x shows that their relationship is certainly not linear (see Figure 11.2.8). The scatterplot more closely resembles the graph of $y = ae^{bx}$ when $b > 0$, as shown in Figure 11.2.7.

Table 11.2.5

Chip	Year	Years after 1975, x	Transistors per Chip, y
8080	1975	0	4,500
8086	1978	3	29,000
80286	1982	7	90,000
80386	1985	10	229,000
80486	1989	14	1,200,000
Pentium	1993	18	3,100,000
Pentium Pro	1995	20	5,500,000

(Continued on next page)

(Case Study 11.2.4 continued)

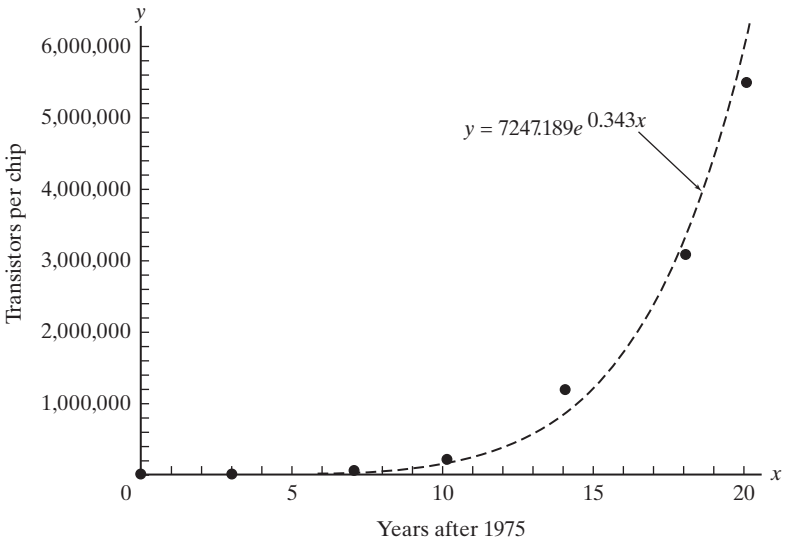


Figure 11.2.8

Table 11.2.6

Years after 1975, x_i	x_i^2	Transistors per Chip, y_i	$\ln y_i$	$x_i \cdot \ln y_i$
0	0	4,500	8.41183	0
3	9	29,000	10.27505	30.82515
7	49	90,000	11.40756	79.85292
10	100	229,000	12.34148	123.41480
14	196	1,200,000	13.99783	195.96962
18	324	3,100,000	14.94691	269.04438
20	400	5,500,000	15.52026	310.40520
72	1078		86.90093	1009.51207

Table 11.2.6 shows the calculation of the sums required to evaluate the formulas for b and $\ln a$. Here the slope and the y-intercept of the linearized model (Equation 11.2.2) are 0.342810 and 8.888369, respectively:

$$b = \frac{7(1009.51207) - 72(86.90093)}{7(1078) - (72)^2}$$

$$= 0.342810$$

and

$$\ln a = \frac{86.9093 - (0.342810)(72)}{7}$$

$$= 8.888369$$

Therefore,

$$a = e^{\ln a} = e^{8.888369} = 7247.189$$

which implies that the best-fitting exponential model describing Intel's technological advances in chip design has the equation

$$y = 7247.189e^{0.343x}$$

(see Figure 11.2.8).

To compare Equation 11.2.1 to Moore's "eighteen-month doubling time" prediction requires that we write $y = 7247.189e^{0.343x}$ in the form $y = 7247.189(2)^x$. But

$$e^{0.343} = 2^{0.495}$$

so another way to express the fitted curve would be

$$y = 7247.189(2^{0.495x}) \tag{11.2.3}$$

In Equation 11.2.3, though, y doubles when $2^{0.495x} = 2$, or, equivalently, when $0.495x = 1$, which implies that *2.0 years* is the empirically determined technology doubling time, a pace not too much slower than Moore's prediction of eighteen months.

About the Data In April of 2005, Gordon Moore pronounced his law dead. He said, "It can't continue forever. The nature of exponentials is that you push them out and eventually disaster happens." If by "disaster" he meant that technology often makes a quantum leap, moving well beyond what an extrapolated law could predict, he was quite correct. Indeed, he could have made this declaration in 2003. By that year, the Itanium 2 featured 220,000,000 transistors on a chip, whereas the model of the case study predicts the number to be only

$$y = 7247.189e^{0.343(28)} = 107,432,032$$

(In the equation, $x = 2003 - 1975 = 28$.)

Logarithmic Regression Another curvilinear model that can be linearized stems from the assumption that changes in y are proportional to the ratio of y to x, that is, $\frac{dy}{dx} = b\frac{y}{x}$, where b is a constant. Then $\int \frac{dy}{y} = \int \frac{b}{x}\, dx$ or $\log y = b \log x + \log a$, where $\log x$ is the logarithm base 10 and $\log a$ is a constant of integration. Therefore, $y = ax^b$ and *$\log y$ is linear with $\log x$*. Therefore,

$$b = \frac{n\sum_{i=1}^{n} \log x_i \cdot \log y_i - \left(\sum_{i=1}^{n} \log x_i\right)\left(\sum_{i=1}^{n} \log y_i\right)}{n\sum_{i=1}^{n} (\log x_i)^2 - \left(\sum_{i=1}^{n} \log x_i\right)^2}$$

and

$$\log a = \frac{\sum_{i=1}^{n} \log y_i - b\sum_{i=1}^{n} \log x_i}{n}$$

Regressions of this type have slower growth rates than exponential models and are particularly useful in describing biological and engineering phenomena.

CASE STUDY 11.2.5

Entomologists who study the behavior of ants know that a certain number are assigned foraging duties, which require them to come and go from their colonies on a regular basis. Moreover, studies have shown that if y is the total number of ants in a colony and x is the number that forage, the relationship between x and y can be effectively modeled by a *logarithmic regression*,

$$y = ax^b \tag{11.2.4}$$

where a and b vary from species to species.

Finding good estimates for a and b is a worthwhile endeavor because knowing those values allows investigators to avoid the difficult problem of counting the very large number of ants in a colony; instead, they need simply count the much smaller number of foraging ants (x) and then use ax^b as an estimate for the colony size (y).

To that end, Table 11.2.7 shows the results of a "calibration" study done on the red wood ant (*Formica polyctena*). Listed are the colony sizes, y, and the foraging sizes, x, observed for fifteen of their colonies (102).

Table 11.2.7

Foraging Size, x	Colony Size, y	y/x
45	280	6.2
70	601	8.6
74	222	3.0
118	288	2.4
220	1205	5.5
338	7551	22.3
446	3229	7.2
611	8834	14.4
647	2828	4.4
765	3762	4.9
823	2769	3.4
850	12,605	14.8
4119	12,584	3.1
11,600	34,661	3.0
64,512	139,043	2.2

The first step in any regression analysis is to graph the data. Do the (x_i, y_i)'s show a pattern similar to what is expected? Data consistent with the model $y = ax^b$, for example, will necessarily have one of the two basic configurations pictured in Figure 11.2.9.

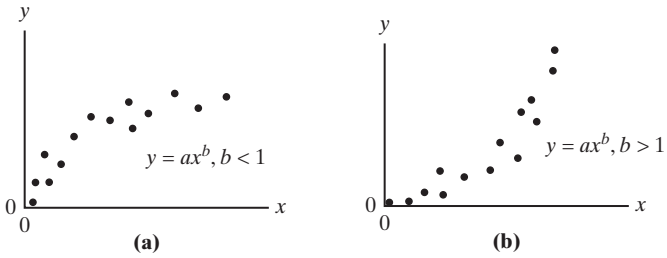


Figure 11.2.9

Here, plotting y versus x for the data in Table 11.2.7 is problematic because of the huge ranges for both variables (x goes from 45 to 64,512; y goes from 222 to 139,043). But, if the model $y = ax^b$ *is* appropriate, then graphing $\log y$ versus $\log x$ should produce a pattern that looks entirely consistent with data having a *linear* relationship. Based on Figure 11.2.10, it does.

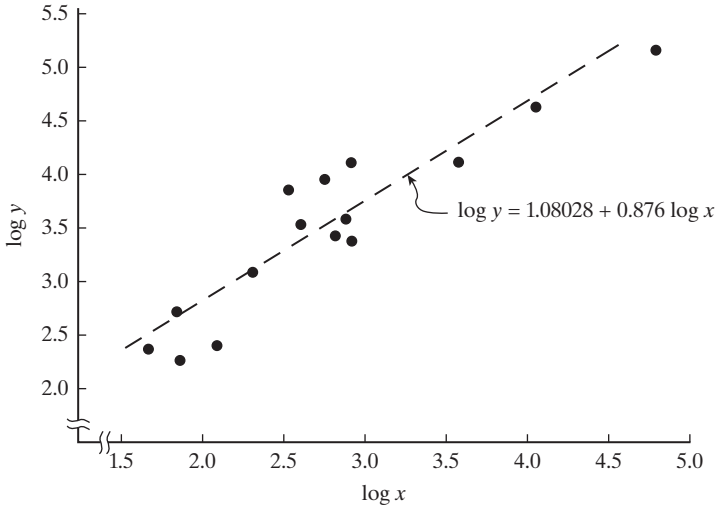


Figure 11.2.10

Suppose a sixteenth *Formica polyctena* colony is found and twenty-five hundred of its ants appear to be foraging. What would be a reasonable estimate for the size of that colony?

First we need to find a and b using the foraging size/colony size database summarized in Table 11.2.7. The necessary sums and sums of squares are

$$\sum_{i=1}^{15} \log x_i = 41.77441 \qquad \sum_{i=1}^{15} \log y_i = 52.79857$$

$$\sum_{i=1}^{15} (\log x_i)^2 = 126.60450 \qquad \sum_{i=1}^{15} \log x_i \cdot \log y_i = 156.03811$$

According to the formulas given in Example 11.2.1,

$$b = \frac{15(156.03811) - (41.77441)(52.79857)}{15(126.60450) - (41.77441)^2} = 0.876$$

$$\log a = \frac{52.79857 - 0.876(41.77441)}{15} = 1.08028$$

$$a = 10^{1.08028} = 12.03$$

Therefore, $y = 12.03x^{0.876}$ and the predicted colony size based on $x = 2500$ foraging ants would be

$$y = 12.03(2500)^{0.876} = 11,400$$

Comment The fact that $y = ax^b$ models these data so nicely should not have come as a complete surprise. The growth rate that defines logarithmic regression,

$$dy/dx = b \cdot (y/x)$$

(Continued on next page)

(Case Study 11.2.5 continued)

mimics the way many animal populations evolve, where dy refers to the population's future growth and x is some measure of its present size. For the *Formica polyctena* data described in Case Study 11.2.5, the number of ants foraging (x) is a proxy variable representing the current size of the colony.

Having calculated that $b < 1$ for the entries in Table 11.2.7, the general configuration of these data if y is plotted against x will look like Figure 11.2.9(a). That is, there will be an initial growth spurt as the colony gets established and quickly multiplies (and values of y/x will often be quite large because x is still small). But as time passes, increases in the denominator x make it more and more difficult for dy to increase enough to make the ratio y/x return to its earlier levels, and the colony's growth begins to level off. That particular scenario—with a few exceptions—is clearly playing out in the entries listed in the third column of Table 11.2.7. Four of the lowest six values of y/x, for example, occur among the five largest colonies.

Logistic Regression Growth is a fundamental characteristic of organisms, institutions, and ideas. In biology, it might refer to the change in size of a *Drosophila* population; in economics, to the proliferation of global markets; in political science, to the gradual acceptance of tax reform. Prominent among the many growth models capable of describing situations of this sort is one where changes in y are proportional to their present magnitudes and to their distance from an upper limit L. In that case, $\frac{dy}{dx} = k\frac{y}{L-y}$, where k and L are constants. Then $\int \frac{dy}{y(L-y)} = \int k\,dx$ or $\frac{1}{L} \cdot \ln\left(\frac{y}{L-y}\right) = kx + c$, where c is a constant of integration. Exponentiating and solving for y yields the *logistic* equation

$$y = \frac{L}{1 + e^{a+bx}} \tag{11.2.5}$$

where a, b, and L are constants. For different values of a and b, Equation 11.2.5 generates a variety of S-shaped curves.

To linearize Equation 11.2.5, we start with its reciprocal:

$$\frac{1}{y} = \frac{1 + e^{a+bx}}{L}$$

Therefore,

$$\frac{L}{y} = 1 + e^{a+bx}$$

and

$$\frac{L-y}{y} = e^{a+bx}$$

Equivalently,

$$\ln\left(\frac{L-y}{y}\right) = a + bx$$

which implies that $\ln\left(\frac{L-y}{y}\right)$ *is linear with* x.

Comment The parameter L is interpreted as the limit to which y is converging as x increases. In practice, L is often estimated simply by plotting the data and "eyeballing" the y-asymptote.

CASE STUDY 11.2.6

The scene is familiar; hundreds of birds are perched wing-to-wing along a stretch of power lines next to a busy highway (Are they hoping to see a five car pile-up?) when they suddenly fly away *all at the same time*. How do they do that? The short answer is they don't. Using high-speed photography, a team of researchers (86) documented the way birds take flight. Their pictures revealed that flocks get airborne in a very precise way, but the pattern plays out much too quickly for the naked eye ever to see.

The targets of their investigation were flocks of redshanks. These are mid-sized, migratory shorebirds that like to winter in the salt marshes of Scotland. A major predator of red shanks are sparrow hawks, and thirty-eight of their attacks were photographed, every 4/100-ths of a second, after the first redshank took flight. Table 11.2.8 shows the average proportions of the flock that were airborne as a function of time.

Table 11.2.8

Time (seconds)	Proportion
0.04	0.08
0.08	0.14
0.12	0.21
0.16	0.30
0.20	0.38
0.24	0.61
0.28	0.70
0.32	0.78
0.36	0.86
0.40	0.90
0.44	0.94
0.48	0.95
0.52	0.96
0.56	0.97
0.60	0.98

The scatterplot for these fifteen data points has a definite *S*-shaped appearance (see Figure 11.2.11), which makes Equation 11.2.5 a good candidate for modeling the *xy*-relationship. The limit to which the population is converging is the full proportion, that is, 1. Quantify the population/time relationship by fitting a logistic equation to these data. Let $L = 1$.

The form of the linearized version of Equation 11.2.5 requires that we find the following sums:

$$\sum_{i=1}^{51} x_i = 4.80, \sum_{i=1}^{15} \ln\left(\frac{1-y_i}{y_i}\right) = -15.91, \sum_{i=1}^{15} x_i^2 = 1.984, \sum_{i=1}^{15} x_i \cdot \ln\left(\frac{1-y_i}{y_i}\right) = -10.2096$$

Substituting $\ln\left(\frac{1-y_i}{y_i}\right)$ for y_i into the formulas for a and b in Theorem 11.2.1 gives

$$b = \frac{15(-10.2096) - (4.80)(-15.91)}{15(1.984) - (4.80)^2} = -11.425$$

and

$$a = \frac{-15.910 - (-11.425)(4.80)}{15} = 2.595$$

so the best-fitting logistic curve has equation $\frac{1}{1+e^{2.595-11.425x}}$.

(Continued on next page)

(Case Study 11.2.6 continued)

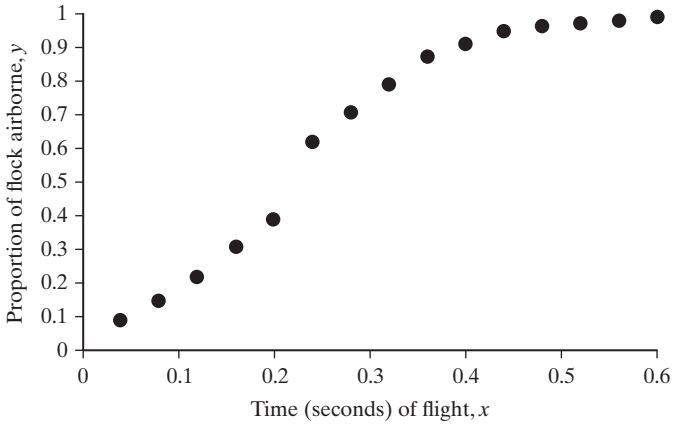


Figure 11.2.11

About the Data For many curve-fitting problems—maybe most—the particular differential equation that defines dy/dx is not known in advance. In its absence, the regression model chosen to describe the data is simply the one that fits the (x_i, y_i)'s the best. This is not one of those situations. Without seeing any data, it would have been reasonable to hypothesize that a graph showing "% of flock airborne" versus "time after first warning" would have the *S*-shape pictured in Figure 11.2.12.

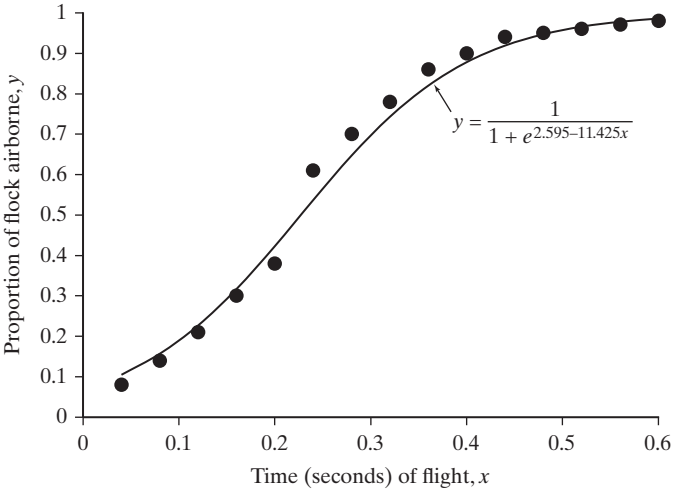


Figure 11.2.12

Why? Because the key factors in the differential equation that gives rise to the logistic curve—namely, $y(L - y)$—describe a growth pattern entirely consistent with the way flocks of birds might be expected to take flight. When the first bird that spots the approach of a predator begins to take off, only a few of the nearby birds would immediately become aware of the threat, so the values of dy/dx would be small for values of x close to 0. As additional birds take flight, though, more birds on the ground would respond to the imminent danger and dy/dx would increase sharply.

As soon as the number of birds remaining on the ground is smaller than the number of birds in the air, dy/dx would begin to decrease, and the values would get very small when virtually the entire flock was airborne. The changes that we would expect to see in dy/dx, then, are precisely what we do see in Figure 11.2.12.

Other Curvilinear Models While the exponential, logarithmic, and logistic equations are three of the most common curvilinear models, there are several others that deserve mention as well. Table 11.2.9 lists a total of six nonlinear equations, including the three already described. Along with each is the particular transformation that reduces the equation to a linear form. Proofs for parts (d), (e), and (f) will be left as exercises.

Table 11.2.9

a. If $y = ae^{bx}$, then $\ln y$ is linear with x.

b. If $y = ax^b$, then $\log y$ is linear with $\log x$.

c. If $y = L/(1 + e^{a+bx})$, then $\ln\left(\dfrac{L-y}{y}\right)$ is linear with x.

d. If $y = \dfrac{1}{a+bx}$, then $\dfrac{1}{y}$ is linear with x.

e. If $y = \dfrac{x}{a+bx}$, then $\dfrac{1}{y}$ is linear with $\dfrac{1}{x}$.

f. If $y = 1 - e^{-x^b/a}$, then $\ln\ln\left(\dfrac{1}{1-y}\right)$ is linear with $\ln x$.

Questions

11.2.20. Radioactive gold (^{195}Au-aurothiomalate) has an affinity for inflamed tissues and is sometimes used as a tracer to diagnose arthritis. The data in the following table (67) come from an experiment investigating the length of time and the concentrations that ^{195}Au-aurothiomalate is retained in a person's blood. Listed are the serum gold concentrations found in ten blood samples taken from patients given an initial dose of 50 mg. Follow-up readings were made at various times, ranging from one to seven days after injection. In each case, the retention is expressed as a percentage of the patient's day-zero serum gold concentration.

Days after Injection, x	Serum Gold % Concentration, y
1	94.5
1	86.4
2	71.0
2	80.5
2	81.4
3	67.4
5	49.3
6	46.8
6	42.3
7	36.6

(a) Fit an exponential curve to these data.

(b) Estimate the half-life of ^{195}Au-aurothiomalate; that is, how long does it take for half the gold to disappear from a person's blood?

If x denotes days after injection and y denotes serum gold % concentration, then $\sum_{i=1}^{10} x_i = 35$, $\sum_{i=1}^{10} x_i^2 = 169$, $\sum_{i=1}^{10} \ln\ y_i = 41.35720$, and $\sum_{i=1}^{10} x_i \ln y_i = 137.97415$.

11.2.21. The growth of federal expenditures is one of the characteristic features of the U.S. economy. The rapidity of the increases from 2000 to 2015, as shown in the table below, suggests an exponential model.

Year	Years after 2000, x	Gross Federal Debt (in \$ trillions), y
2000	0	5.629
2001	1	5.770
2002	2	6.198
2003	3	6.760
2004	4	7.355
2005	5	7.905
2006	6	8.451
2007	7	8.951
2008	8	9.986
2009	9	11.876
2010	10	13.529
2011	11	14.764
2012	12	16.051
2013	13	16.719
2014	14	17.794
2015	15	18.120

Source: https://www.whitehouse.gov/omb/budget/Historicals

(a) Find the best-fitting exponential curve, using the method of least squares together with an appropriate

linearizing transformation. Use the sums: $\sum_{i=0}^{15} x_i = 120$, $\sum_{i=0}^{15} \ln y_i = 37.04571$, and $\sum_{i=0}^{15} x_i \cdot \ln y_i = 307.6275$

(b) Calculate the residuals for the years 2009 through 2015. What does this say about the exponential model?

11.2.22. Used cars are often sold wholesale at auctions, and from these sales, retail sales prices are recommended. The following table gives the recommended prices in 2009 for a four-door manual transmission Toyota Corolla based on the age of the car.

Age (in years), x	Suggested Retail Price, y
1	$14,680
2	12,150
3	11,215
4	10,180
5	9,230
6	8,455
7	7,730
8	6,825
9	6,135
10	5,620

Data from: www.bb.com.

(a) Fit these data with a model of the form $y = ae^{bx}$. Graph the (x_i, y_i)'s and superimpose the least squares exponential curve.

(b) What would you predict the retail price of an eleven-year-old Toyota Corolla to be?

(c) The price of a new Corolla in 2009 was $16,200. Is that figure consistent with the widely held belief that a new car depreciates substantially the moment it is purchased? Explain.

11.2.23. The Super Bowl showed steady and significant growth in popularity from its beginning in 1967. This growth was reflected in ticket prices. The table gives the ticket prices in four-year intervals from 1967 to 2011.

Years after 1967, x	Ticket Cost ($), y
0	10
4	15
8	20
12	30
16	40
20	75
24	150
28	200
32	325
36	500
40	600
44	900

Data from: http://www.jsonline.com/sports/rank–file-7p40vnb-138455269.html

Use the fact that $\sum_{i=1}^{12} \ln y_i = 54.92066$ and $\sum_{i=1}^{12} x_i \cdot \ln y_i = 1453.58352$ to fit the data with an exponential model.

11.2.24. Suppose a set of $n(x_i, y_i)$'s are measured on a phenomenon whose theoretical xy-relationship is of the form $y = ae^{bx}$.

(a) Show that $\dfrac{dy}{dx} = by$ implies that $y = ae^{bx}$.

(b) On what kind of graph paper would the (x_i, y_i)'s show a linear relationship?

11.2.25. In 1959, the Ise Bay typhoon devastated parts of Japan. For seven metropolitan areas in the storm's path, the following table gives the number of homes damaged as a function of peak wind gust (126). Show that a function of the form $y = ax^b$ provides a good model for the data.

City	Peak Wind Gust (hundred mph), x	Numbers of Damaged Homes (in thousands), y
A	0.98	25.000
B	0.74	0.950
C	1.12	200.000
D	1.34	150.000
E	0.87	0.940
F	0.65	0.090
G	1.39	260.000

Use the following sums:

$$\sum_{i=1}^{7} \log\ x_i = -0.067772 \qquad \sum_{i=1}^{7} \log\ y_i = 7.1951$$

$$\sum_{i=1}^{7} (\log\ x_i)^2 = 0.0948679 \qquad \sum_{i=1}^{7} (\log\ x_i)(\log\ y_i) = 0.92314$$

11.2.26. Among mammals, the relationship between the age at which an animal develops locomotion and the age at which it first begins to play has been widely studied. The table below lists "onset" times for locomotion and for play in eleven different species (46). Fit the data to the $y = ax^b$ model.

Species	Locomotion Begins, x (days)	Play Begins, y (days)
Homo sapiens	360	90
Gorilla gorilla	165	105
Felis catus	21	21
Canis familiaris	23	26
Rattus norvegicus	11	14
Turdus merula	18	28
Macaca mulatta	18	21
Pan troglodytes	150	105
Saimiri sciurens	45	68
Cercocebus alb.	45	75
Tamiasciureus hud.	18	46

11.2.27. Over the years, many efforts have been made to demonstrate that the human brain is appreciably different

in structure from the brains of lower-order primates. In point of fact, such differences in gross anatomy are disconcertingly difficult to discern. The following are the average areas of the striate cortex (x) and the prestriate cortex (y) found for humans and for three species of chimpanzees (137).

	Area	
Primate	Striate Cortex, x (mm^2)	Prestriate Cortex, y (mm^2)
Homo	2613	7838
Pongo	1876	2864
Cercopithecus	933	1334
Galago	78.9	40.8

Plot the data and superimpose the least squares curve, $y = ax^b$.

11.2.28. Years of experience buying and selling commercial real estate have convinced many investors that the value of land zoned for business (y) is inversely related to its distance (x) from the center of town—that is, $y = a + b \cdot \frac{1}{x}$. If that suspicion is correct, what should be the appraised value of a piece of property located $\frac{1}{4}$ mile from the town square, based on the sales listed below?

Land Parcel	Distance from Center of City (in thousand feet), x	Value (in thousands), y
H1	1.00	$20.5
B6	0.50	42.7
Q4	0.25	80.4
L4	2.00	10.5
T7	4.00	6.1
D9	6.00	6.0
E4	10.00	3.5

11.2.29. Verify the claims made in parts (d), (e), and (f) of Table 11.2.9—that is, prove that the transformations cited will linearize the original models.

11.2.30. Biological organisms, such as yeast, often exhibit exponential growth. However, in some cases, that rapid rate of growth cannot be sustained. Such factors as lack of nutrition to support a large population or the buildup of toxins limit the rate of growth. In such cases the curve begins concave up, inflects at some point, and becomes concave down and asymptotic to a limit. A now-classical experiment by Carlson (23) measured the amount of biomass of brewer's yeast (*Saccharomyces Cerevisiae*) at one-hour intervals. The table shows the results.

Hour	Yeast Count	Hour	Yeast Count
0	9.6	9	441.00
1	18.3	10	513.3
2	29.0	11	559.7
3	47.2	12	594.8
4	71.1	13	629.4
5	119.1	14	640.8
6	174.6	15	651.1
7	257.3	16	655.9
8	350.7	17	659.6

Quantify the population/time relationship by fitting a logistic equation to these data. Let $L = 700$.

11.2.31. The following table shows a portion of the results from a clinical trial investigating the effectiveness of a monoamine oxidase inhibitor as a treatment for depression (219). The relationship between y, the percentage of subjects showing improvement, and x, the patient's age, appears to be S-shaped. Graph the data and superimpose a graph of the least squares curve $y = \frac{L}{1 + e^{a+bx}}$. Take L to be 60.

Age Group	Age Mid-Point, x	% Improved, y	$\ln\left(\frac{60-y}{y}\right)$
[28, 32)	30	11	1.49393
[32, 36)	34	14	1.18958
[36, 40)	38	19	0.76913
[40, 44)	42	32	−0.13353
[44, 48)	46	42	−0.84730
[48, 52)	50	48	−1.38629
[52, 56)	54	50	−1.60944
[56, 60)	58	52	−1.87180

11.3 The Linear Model

Section 11.2 views the problem of "curve fitting" from a purely geometrical perspective. The observed (x_i, y_i)'s are assumed to be nothing more than points in the xy-plane, devoid of any statistical properties. It is more realistic, though, to think of each y as the value recorded for a random variable Y, meaning that a *distribution* of possible y-values is associated with every value of x.

Consider, for example, the connecting rod weights analyzed in Case Study 11.2.1. The first rod listed in Table 11.2.1 had an initial weight of $x = 2.745$ oz. and, after the tooling process was completed, a finished weight of $y = 2.080$ oz. It does not follow from that one observation, of course, that an initial weight of 2.745 oz. necessarily

leads to a finished weight of 2.080 oz. Common sense tells us that the tooling process will not always have exactly the same effect, even on rods having the same initial weight. Associated with each x, then, there *will* be a range of possible y-values. The symbol $f_{Y|x}(y)$ is used to denote the pdfs of these "conditional" distributions.

Definition 11.3.1
Let $f_{Y|x}(y)$ denote the pdf of the random variable Y for a given value x, and let $E(Y \mid x)$ denote the expected value associated with $f_{Y|x}(y)$. The function

$$y = E(Y \mid x)$$

is called the *regression curve of Y on x*.

Example 11.3.1

Suppose that corresponding to each value of x in the interval $0 \le x \le 1$ is a distribution of y-values having the pdf

$$f_{Y|x}(y) = \frac{x+y}{x+\frac{1}{2}}, \quad 0 \le y \le 1; \quad 0 \le x \le 1$$

Find and graph the regression curve of Y on x.

Notice, first of all, that for any x between 0 and 1, $f_{Y|x}(y)$ does qualify as a pdf:

1. $f_{Y|x}(y) \ge 0$, for $0 \le y \le 1$ and any $0 \le x \le 1$

2. $$\int_0^1 f_{Y|x}(y)\,dy = \int_0^1 \frac{x+y}{x+1/2}dy = \int_0^1 \frac{x}{x+1/2}dy + \int_0^1 \frac{y}{x+1/2}dy = \frac{x+1/2}{x+1/2} = 1$$

Moreover,

$$E(Y \mid x) = \int_0^1 y \cdot f_{Y|x}(y)\,dy = \int_0^1 y \cdot \frac{x+y}{x+\frac{1}{2}}\,dy$$

$$= \left[\frac{xy^2}{2\left(x+\frac{1}{2}\right)} + \frac{y^3}{3\left(x+\frac{1}{2}\right)}\right]\Bigg|_0^1$$

$$= \frac{3x+2}{6x+3}, \quad 0 \le x \le 1$$

Figure 11.3.1 shows the regression curve, $y = E(Y|x) = \dfrac{3x+2}{6x+3}$, together with three of the conditional distributions—$f_{Y|0}(y) = 2y$, $f_{Y|\frac{1}{2}}(y) = y + \frac{1}{2}$, and

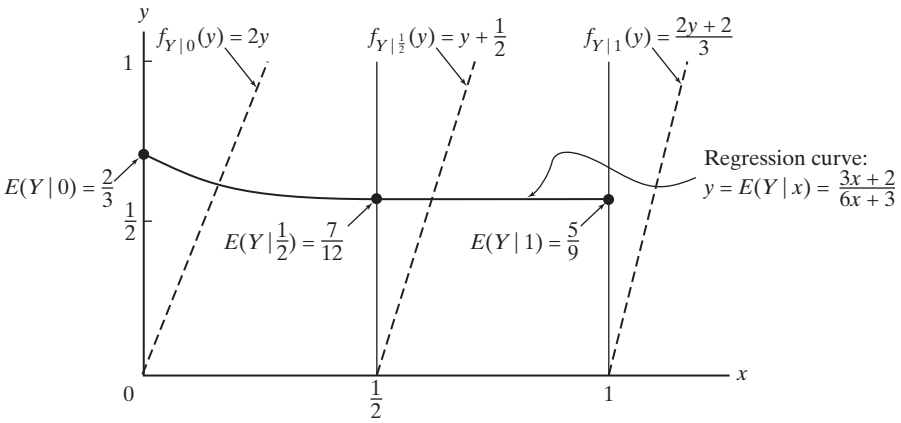


Figure 11.3.1

$f_{y|1}(y) = \dfrac{2y+2}{3}$. The $f_{Y|x}(y)$'s, of course, should be visualized as coming out of the plane of the paper. ■

A SPECIAL CASE

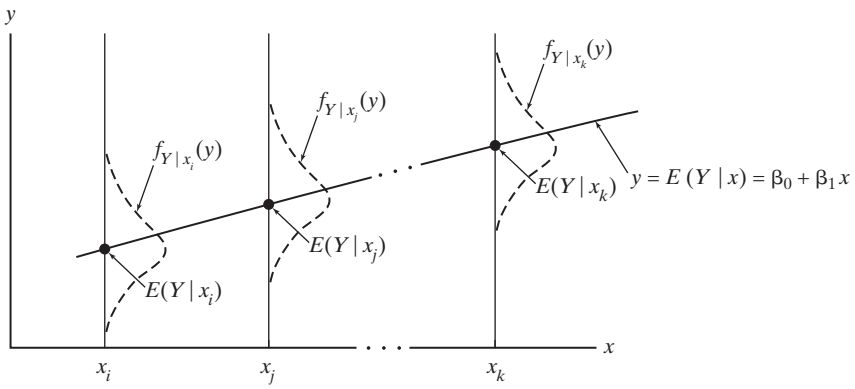
Definition 11.3.1 introduces the notion of a regression curve in the most general of contexts. In practice, there is one special case of the function $y = E(Y|x)$ that is particularly important. Known as the *simple linear model,* it makes four assumptions:

1. $f_{Y|x}(y)$ is a normal pdf for all x.
2. The standard deviation, σ, associated with $f_{Y|x}(y)$ is the same for all x.
3. The means of all the conditional Y distributions are collinear—that is,

$$y = E(Y|x) = \beta_0 + \beta_1 x$$

4. All of the conditional distributions represent independent random variables. (See Figure 11.3.2.)

Figure 11.3.2



ESTIMATING THE LINEAR MODEL PARAMETERS

Implicit in the simple linear model are three parameters—β_0, β_1, and σ^2. Typically, all three will be unknown and need to be estimated. Since the model assumes a probability structure for the Y-variable, estimates can be obtained using the method of maximum likelihood, as opposed to the method of least squares that we saw in Section 11.2. (Maximum likelihood estimates are preferable to least squares estimates because the former have probability distributions that can be used to set up hypothesis tests and confidence intervals.)

Comment It would be entirely consistent with the notation used previously to denote the sample in Theorem 11.3.1 as (x_1, y_1), (x_2, y_2), . . . , and (x_n, y_n). To emphasize the important distinction, though, between the (lack of) assumptions on the y_i's made in Section 11.2 and the conditional pdfs $f_{Y|x}(y)$ introduced in Definition 11.3.1, we will use random variable notation to write linear model data as (x_1, Y_1), (x_2, Y_2), . . . , and (x_n, Y_n).

Theorem 11.3.1 *Let $(x_1, Y_1), (x_2, Y_2), \ldots,$ and (x_n, Y_n) be a set of points satisfying the simple linear model, $E(Y|x) = \beta_0 + \beta_1 x$. The maximum likelihood estimators for β_0, β_1, and σ^2 are given by*

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i Y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

where $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, $i = 1, \ldots, n$.

Proof Since each Y_i is assumed to be normally distributed with mean equal to $\beta_0 + \beta_1 x_i$ and variance equal to σ^2, the sample's likelihood function, L, is the product

$$L = \prod_{i=1}^{n} f_{Y_i|x_i}(y_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma}\right)^2} = (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma}\right)^2}$$

The maximum of L for this case occurs when the partial derivatives with respect to β_0, β_1, and σ^2 all vanish. It will be easier, computationally, to differentiate $-2 \ln L$, and the latter will be minimized for the same parameter values that maximize L. Here,

$$\ln L = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma}\right)^2$$

and

$$-2 \ln L = n \ln 2\pi + n \ln \sigma^2 + \frac{1}{\sigma^2}\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Setting the three partial derivatives equal to 0 gives

$$\frac{\partial(-2 \ln L)}{\partial \beta_0} = \frac{2}{\sigma^2}\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)(-1) = 0$$

$$\frac{\partial(-2 \ln L)}{\partial \beta_1} = \frac{2}{\sigma^2}\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)(-x_i) = 0$$

$$\frac{\partial(-2 \ln L)}{\partial \sigma^2} = \frac{n}{\sigma^2} - \frac{1}{(\sigma^2)^2}\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = 0$$

The first two equations depend only on β_0 and β_1, and the resulting solutions for $\hat{\beta}_0$ and $\hat{\beta}_1$ have the same forms that are given in the statement of the theorem. Substituting the solutions from the first two equations into the third gives the expression for $\hat{\sigma}^2$.

Comment Note the similarity in the formulas for the maximum likelihood estimators and the least squares estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$. The least squares estimates, of course, are numbers, while the maximum likelihood estimators are random variables.

Up to this point, random variables have been denoted with uppercase letters and their values with lowercase letters. In this section, boldface $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}_1$ will represent

the maximum likelihood *random variables,* and plain-text $\hat{\beta}_0$ and $\hat{\beta}_1$ will refer to specific values taken on by those random variables.

PROPERTIES OF LINEAR MODEL ESTIMATORS

By virtue of the assumptions that define the simple linear model, we know that the estimators $\hat{\boldsymbol{\beta}}_0$, $\hat{\boldsymbol{\beta}}_1$, and $\hat{\boldsymbol{\sigma}}^2$ are random variables. Before those estimators can be used to set up inference procedures, though, we need to establish their basic statistical properties—specifically, their means, variances, and pdfs.

Theorem 11.3.2 *Let $(x_1, Y_1), (x_2, Y_2), \ldots,$ and (x_n, Y_n) be a set of points satisfying the simple linear model, $E(Y|x) = \beta_0 + \beta_1 x$. Let $\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_1,$ and $\hat{\boldsymbol{\sigma}}^2$ be the maximum likelihood estimators for $\beta_0, \beta_1,$ and σ^2, respectively. Then*

a. *$\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}_1$ are both normally distributed.*

b. *$\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}_1$ are both unbiased: $E(\hat{\boldsymbol{\beta}}_0) = \beta_0$ and $E(\hat{\boldsymbol{\beta}}_1) = \beta_1$.*

c. $\text{Var}(\hat{\boldsymbol{\beta}}_1) = \dfrac{\sigma^2}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}$

d. $\text{Var}(\hat{\boldsymbol{\beta}}_0) = \dfrac{\sigma^2 \sum\limits_{i=1}^{n} x_i^2}{n\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} = \sigma^2\left[\dfrac{1}{n} + \dfrac{\bar{x}^2}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}\right]$

Proof We will prove the statements for $\hat{\boldsymbol{\beta}}_1$; the results for $\hat{\boldsymbol{\beta}}_0$ follow similarly.

The equation for the estimator $\hat{\boldsymbol{\beta}}_1$ given in Theorem 11.3.1 is the simplest form that solves the likelihood equations (and the least squares equations as well). It is also convenient for computation. However, two other expressions for $\hat{\beta}_1$ are useful for theoretical results.

To begin, take the version of $\hat{\boldsymbol{\beta}}_1$ from Theorem 11.3.1:

$$\hat{\boldsymbol{\beta}}_1 = \frac{n\sum_{i=1}^{n} x_i Y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Dividing numerator and denominator by n gives

$$\begin{aligned}\hat{\boldsymbol{\beta}}_1 &= \frac{\sum_{i=1}^{n} x_i Y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} \\ &= \frac{\sum_{i=1}^{n} x_i Y_i - \bar{x}\left(\sum_{i=1}^{n} Y_i\right)}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \qquad (11.3.1)\end{aligned}$$

(Continued on next page)

(Theorem 11.3.2 continued)

Equation 11.3.1 expresses $\hat{\beta}_1$ as a linear combination of independent normal variables, so by the second corollary to Theorem 4.3.3, it is itself normal, proving part (a).

To see that $\hat{\beta}_1$ is unbiased, note that

$$E(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})E(Y_i)}{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2} = \frac{\beta_0\sum_{i=1}^{n}(x_i - \bar{x}) + \beta_1\sum_{i=1}^{n}(x_i - \bar{x})x_i}{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2}$$

$$= \frac{0 + \beta_1\sum_{i=1}^{n}(x_i - \bar{x})x_i}{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2} = \frac{\beta_1\left(\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\right)}{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2} = \beta_1$$

To find $\text{Var}(\hat{\beta}_1)$, rewrite the denominator of Equation 11.3.1 in the form

$$\sum_{i=1}^{n}x_i^2 - n\bar{x}^2 = \sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

which makes

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{11.3.2}$$

Using Equation 11.3.2, Theorem 3.6.2, and the second corollary to Theorem 3.9.5 gives

$$\text{Var}(\hat{\beta}_1) = \text{Var}\left[\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sum_{i=1}^{n}(x_i - \bar{x})Y_i\right]$$

$$= \frac{1}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\sigma^2$$

$$= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Theorem 11.3.3 *Let $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$ satisfy the assumptions of the simple linear model. Then*

a. *$\hat{\beta}_1, \bar{Y}$, and $\hat{\sigma}^2$ are mutually independent.*

b. *$\dfrac{n\hat{\sigma}^2}{\sigma^2}$ has a chi square distribution with $n - 2$ degrees of freedom.*

Proof See Appendix 11.A.1.

Corollary 11.3.1 *Let $\hat{\sigma}^2$ be the maximum likelihood estimator for σ^2 in a simple linear model. Then $\frac{n}{n-2} \cdot \hat{\sigma}^2$ is an unbiased estimator for σ^2.*

Proof Recall that the expected value of a χ^2_k distribution is k (see Theorems 4.6.3 and 7.3.1). Therefore,

$$E\left(\frac{n}{n-2} \cdot \hat{\sigma}^2\right) = \frac{\sigma^2}{n-2} E\left(\frac{n\hat{\sigma}^2}{\sigma^2}\right)$$

$$= \frac{\sigma^2}{n-2} \cdot (n-2) \qquad \text{[by part (b) of Theorem 11.3.3]}$$

$$= \sigma^2$$

Corollary 11.3.2 *The random variables $\hat{Y}$ and $\hat{\sigma}^2$ are independent.*

ESTIMATING σ^2

We know that the (biased) maximum likelihood estimator for σ^2 in a simple linear model is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

The unbiased estimator for σ^2 based on $\hat{\sigma}^2$ is denoted S^2, where

$$S^2 = \frac{n}{n-2} \hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Statistical software packages—including Minitab—typically print out s, rather than $\hat{\sigma}$, in summarizing the calculations associated with linear model data. To accommodate that convention, we will use s^2 rather than $\hat{\sigma}^2$ in writing the formulas for the test statistics and confidence intervals that arise in connection with the simple linear model.

Comment Calculating $\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ can be cumbersome. Three (algebraically equivalent) computing formulas are available that may be easier to use, depending on the data:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{11.3.3}$$

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n} \sum_{i=1}^{n} y_i - \frac{\left[\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)\right]^2}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\sum_{i=1}^{n} x_i} \tag{11.3.4}$$

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i \tag{11.3.5}$$

DRAWING INFERENCES ABOUT β_1

Hypothesis tests and confidence intervals for β_1 can be carried out by defining a t statistic based on the properties that appear in Theorems 11.3.2 and 11.3.3.

Theorem 11.3.4 *Let $(x_1, Y_1), (x_2, Y_2), \ldots,$ and (x_n, Y_n) be a set of points that satisfy the assumptions of the simple linear model, and let $S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$. Then*

$$T_{n-2} = \frac{\hat{\beta}_1 - \beta_1}{S \Big/ \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

has a Student t distribution with $n - 2$ degrees of freedom.

Proof We know from Theorem 11.3.2 that

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma \Big/ \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

has a standard normal pdf. Furthermore, $\frac{n\hat{\sigma}^2}{\sigma^2} = \frac{(n-2)S^2}{\sigma^2}$ has a χ^2 pdf with $n - 2$ degrees of freedom, and, by Theorem 11.3.3, Z and $\frac{(n-2)S^2}{\sigma^2}$ are independent. From Definition 7.3.3, then, it follows that

$$Z \Big/ \sqrt{\frac{(n-2)S^2}{\sigma^2} \Big/ (n-2)} = \frac{\hat{\beta}_1 - \beta_1}{S \Big/ \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

has a Student t distribution with $n - 2$ degrees of freedom.

Theorem 11.3.5 *Let $(x_1, Y_1), (x_2, Y_2), \ldots,$ and (x_n, Y_n) be a set of points that satisfy the assumptions of the simple linear model. Let*

$$t = \frac{\hat{\beta}_1 - \beta_1'}{s \Big/ \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

a. *To test $H_0: \beta_1 = \beta_1'$ versus $H_1: \beta_1 > \beta_1'$ at the α level of significance, reject H_0 if $t \geq t_{\alpha,n-2}$.*

b. *To test $H_0: \beta_1 = \beta_1'$ versus $H_1: \beta_1 < \beta_1'$ at the α level of significance, reject H_0 if $t \leq -t_{\alpha,n-2}$.*

c. *To test $H_0: \beta_1 = \beta_1'$ versus $H_1: \beta_1 \neq \beta_1'$ at the α level of significance, reject H_0 if t is either (1) $\leq -t_{\alpha/2,n-2}$ or (2) $\geq t_{\alpha/2,n-2}$.*

Proof The decision rule given here is, in fact, a GLRT. A formal proof proceeds along the lines followed in Appendix 7.A.4. We will omit the details.

Comment A particularly common application of Theorem 11.3.5 is to test $H_0: \beta_1 = 0$. If the null hypothesis that the slope is zero is rejected, it can be concluded (at the α level of significance) that $E(Y)$ changes with x. Conversely, if $H_0: \beta_1 = 0$ is *not* rejected, the data have not ruled out the possibility that variation in Y is unaffected by x.

CASE STUDY 11.3.1

By late 1971, all cigarette packs had to be labeled with the words, "Warning: The Surgeon General Has Determined That Smoking Is Dangerous To Your Health." The case against smoking rested heavily on statistical, rather than laboratory, evidence. Extensive surveys of smokers and nonsmokers had revealed the former to have much higher risks of dying from a variety of causes, including heart disease.

Typical of that research are the data in Table 11.3.1, showing the annual cigarette consumption, x, and the corresponding mortality rate, Y, due to coronary heart disease (CHD) for twenty-one countries (124). Do these data support the suspicion that smoking contributes to CHD mortality? Test $H_0: \beta_1 = 0$ versus $H_1: \beta_1 > 0$ at the $\alpha = 0.05$ level of significance.

Table 11.3.1

Country	Cigarette Consumption per Adult per Year, x	CHD Mortality per 100,000 (ages 35–64), y
United States	3900	256.9
Canada	3350	211.6
Australia	3220	238.1
New Zealand	3220	211.8
United Kingdom	2790	194.1
Switzerland	2780	124.5
Ireland	2770	187.3
Iceland	2290	110.5
Finland	2160	233.1
West Germany	1890	150.3
Netherlands	1810	124.7
Greece	1800	41.2
Austria	1770	182.1
Belgium	1700	118.1
Mexico	1680	31.9
Italy	1510	114.3
Denmark	1500	144.9
France	1410	59.7
Sweden	1270	126.9
Spain	1200	43.9
Norway	1090	136.3

From Table 11.3.1,

$$\sum_{i=1}^{21} x_i = 45{,}110 \qquad \sum_{i=1}^{21} y_i = 3{,}042.2$$

$$\sum_{i=1}^{21} x_i^2 = 109{,}957{,}100 \qquad \sum_{i=1}^{21} y_i^2 = 529{,}321.58$$

$$\sum_{i=1}^{21} x_i y_i = 7{,}319{,}602$$

(Continued on next page)

(Case Study 11.3.1 continued)

and it follows that

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$= \frac{21(7{,}319{,}602) - (45{,}110)(3{,}042.2)}{21(109{,}957{,}100) - (45{,}110)^2} = 0.0601$$

and

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i}{n}$$

$$= \frac{3{,}042.2 - 0.0601(45{,}110)}{21} = 15.771$$

The two other quantities needed for the test statistic are

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\right)\left(\sum_{i=1}^{n} x_i\right)^2$$

$$= 109{,}957{,}100 - \left(\frac{1}{21}\right)(45{,}100)^2 = 13{,}056{,}523.81$$

so $\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \sqrt{13{,}056{,}523.81} = 3{,}613.38.$

From Equation 11.3.5,

$$s^2 = \frac{1}{21-2}\left(\sum_{i=1}^{21} y_i^2 - \hat{\beta}_0 \sum_{i=1}^{21} y_i - \hat{\beta}_1 \sum_{i=1}^{21} x_i y_i\right)$$

$$= \frac{1}{19}[529{,}321.58 - (15.766)(3{,}042.2) - (0.0601)(7{,}319{,}602)] = 2{,}181.588$$

and $s = \sqrt{2{,}181.588} = 46.707$

To test

$$H_0: \beta_1 = 0$$
versus
$$H_0: \beta_1 > 0$$

at the $\alpha = 0.05$ level of significance, we should reject the null hypothesis if $t \geq t_{.05,19} = 1.7291$. But

$$t = \frac{\hat{\beta}_1 - \beta_1'}{s \Big/ \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}} = \frac{0.0601 - 0}{46.707/3{,}613.38}$$

$$= 4.65$$

so our conclusion is clear-cut—reject H_0. It would appear that the level of CHD mortality in a country *is* affected by its citizens' smoking habits. More specifically, as the number of people who smoke increases, so will the number who die of coronary heart disease.

Theorem 11.3.6 *Let* $(x_1, Y_1), (x_2, Y_2), \ldots,$ *and* (x_n, Y_n) *be a set of points that satisfy the assumptions of the simple linear model, and let* $s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$. *Then*

$$\left[\hat{\beta}_1 - t_{\alpha/2,n-2} \cdot \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \hat{\beta}_1 + t_{\alpha/2,n-2} \cdot \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right]$$

is a $100(1-\alpha)\%$ *confidence interval for* β_1.

Proof Let T_{n-2} denote a Student t random variable with $n-2$ degrees of freedom, in which case

$$P(-t_{\alpha/2,n-2} \le T_{n-2} \le t_{\alpha/2,n-2}) = 1 - \alpha$$

Substitute the expression for T_{n-2} given in Theorem 11.3.4 and isolate β_1 in the center of the inequalities. The resulting endpoints will be the expressions appearing in the statement of the theorem.

CASE STUDY 11.3.2

Not surprisingly, the older a car is, the less its value as a used car. But, in some cases the price of a car can show a predictable increase by year. This can occur even though a model of a given year may be "improved" over the previous year. Table 11.3.2 gives the suggested retail price in 2016 of each year's model of a four-door Toyota Camry sedan. Graphed, the xy-relationship is described very well by the line $y = 8188.67 + 748.41x$, where 8188.67 and 748.41 are the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ calculated from the formulas of Theorem 11.3.1 (see Figure 11.3.3). To simplify the calculations, x is the number of years after 2005, as in the second column of Table 11.3.2.

Table 11.3.2

Year	Year after 2005	Suggested Retail Price ($)
2005	0	7,935
2006	1	8,495
2007	2	10,160
2008	3	10,817
2009	4	11,078
2010	5	11,967
2011	6	12,658
2012	7	13,844
2013	8	13,982
2014	9	14,629

Data from: kbb.com

(Continued on next page)

(Case Study 11.3.2 continued)

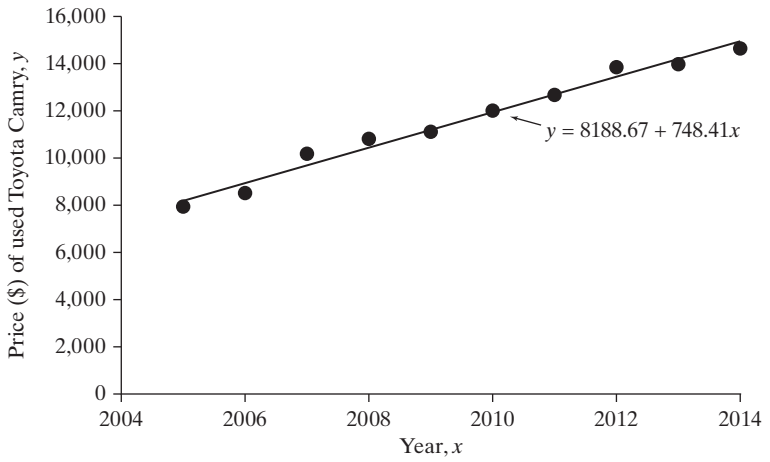


Figure 11.3.3

The slope of the line, $\hat{\beta}_1$, represents the amount of increase year-by-year in the cost of an older model. Often a range of values is better than a single estimate, so a good way to provide this is using a confidence interval for the true value β_1.

Here, $$\sqrt{\sum_{i=0}^{9}(x_i - \bar{x})^2} = \sqrt{82.5} = 9.083$$

and from Equation 11.3.5, $s^2 = \frac{1}{10-2}\left(\sum_{i=0}^{9} y_i^2 - \hat{\beta}_0 \sum_{i=0}^{9} y_i - \hat{\beta}_1 \sum_{i=0}^{9} x_i y_i\right)$

$$\frac{1}{8}[1{,}382{,}678{,}777 - (8188.67)(115{,}565) - (748.41)(581{,}786)] = 117{,}727.98$$

so $s = \sqrt{117{,}727.98} = 343.11$.

Using $t_{.025,8} = 2.3060$, the expression given in Theorem 11.3.6 reduces to $\left(748.41 - 2.3060\frac{343.11}{9.083}, 748.41 + 2.3060\frac{343.11}{9.083}\right) = (\$661.30, \$835.52)$

About the Data The suggested retail value of a used car may not be the actual sales price, which depends in part on the consumer's value of such a car. The predictive value of the regression equation in Case Study 11.3.2 depends on a continuing buyers' sense of value. Perhaps for a variety of reasons, the value of a one-year-old car does not fit the model. The regression equation, even using the upper value in the confidence interval for $\hat{\beta}_1$, 835.52, gives a predicted value of $\hat{y} = 8188.67 + 835.52(10) = 16,543.87$. This value is well below the suggested retail value for the 2015 model of $20,879.

DRAWING INFERENCES ABOUT β_0

In practice, the value of β_0 is not likely to be as important as the value of β_1. Slopes often quantify particularly important aspects of *xy*-relationships, which was true, for example, in Case Study 11.3.2. Nevertheless, hypothesis tests and confidence intervals for β_0 can be easily derived from the results given in Theorems 11.3.2 and 11.3.3.

The GLRT procedure for assessing the credibility of $H_0 : \beta_0 = \beta_0'$ is based on a Student t random variable with $n - 2$ degrees of freedom:

$$T_{n-2} = \frac{(\hat{\beta}_0 - \beta_0')\sqrt{n}\sqrt{\sum_{i=i}^{n}(x_i - \bar{x})^2}}{S\sqrt{\sum_{i=1}^{n} x_i^2}} = \frac{\hat{\beta}_0 - \beta_0'}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_0)}} \tag{11.3.6}$$

"Inverting" Equation 11.3.6 (recall the proof of Theorem 11.3.6) yields

$$\left[\hat{\beta}_0 - t_{\alpha/2,n-2} \cdot \frac{s\sqrt{\sum_{i=1}^{n} x_i^2}}{\sqrt{n}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \hat{\beta}_0 + t_{\alpha/2,n-2} \cdot \frac{s\sqrt{\sum_{i=1}^{n} x_i^2}}{\sqrt{n}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right]$$

as the formula for a $100(1 - \alpha)\%$ confidence interval for β_0.

DRAWING INFERENCES ABOUT σ^2

Since $(n - 2)S^2/\sigma^2$ has a χ^2 pdf with $n - 2$ df (if the n observations satisfy the stipulations implicit in the simple linear model), it follows that

$$P\left[\chi^2_{\alpha/2,n-2} \le \frac{(n-2)S^2}{\sigma^2} \le \chi^2_{1-\alpha/2,n-2}\right] = 1 - \alpha$$

Equivalently,

$$P\left[\frac{(n-2)S^2}{\chi^2_{1-a/2,n-2}} \le \sigma^2 \le \frac{(n-2)S^2}{\chi^2_{\alpha/2,n-2}}\right] = 1 - \alpha$$

in which case

$$\left[\frac{(n-2)s^2}{\chi^2_{1-\alpha/2,n-2}}, \frac{(n-2)s^2}{\chi^2_{\alpha/2,n-2}}\right]$$

becomes the *100(1 − α)% confidence interval for* σ^2 (recall Theorem 7.5.1). Testing $H_0 : \sigma^2 = \sigma_0^2$ is done by calculating the ratio

$$\chi^2 = \frac{(n-2)s^2}{\sigma_0^2}$$

which has a χ^2 distribution with $n - 2$ df when the null hypothesis is true. Except for the degrees of freedom ($n - 2$ rather than $n - 1$), the appropriate decision rules for one-sided and two-sided H_1's are similar to those given in Theorem 7.5.2.

Questions

11.3.1. Insect flight ability can be measured in a laboratory by attaching the insect to a nearly frictionless rotating arm with a thin wire. The "tethered" insect then flies in circles until exhausted. The nonstop distance flown can easily be calculated from the number of revolutions made by the arm. The following are measurements of this sort made on *Culex tarsalis* mosquitos of four different ages. The response variable is the average distance flown until exhaustion for forty females of the species (159).

Age, x (weeks)	Distance Flown, y (thousand meters)
1	12.6
2	11.6
3	6.8
4	9.2

Fit a straight line to these data and test that the slope is zero. Use a two-sided alternative and the 0.05 level of significance.

11.3.2. The best straight line through the Massachusetts funding/graduation rate data described in Question 11.2.7 has the equation $y = 81.088 + 0.412x$, where $s = 11.78848$.

(a) Construct a 95% confidence interval for β_1.

(b) What does your answer to part (a) imply about the outcome of testing H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$ at the $\alpha = 0.05$ level of significance?

(c) Graph the data and superimpose the regression line. How would you summarize these data, and their implications, to a meeting of the state School Board?

11.3.3. Based on the data in Question 11.2.1, the relationship between y, the ambient temperature, and x, the frequency of a cricket's chirping, is given by $y = 25.2 + 3.29x$, where $s = 3.83$. At the $\alpha = 0.01$ level of significance, can the hypothesis that chirping frequency is not related to temperature be rejected?

11.3.4. Suppose an experimenter intends to do a regression analysis by taking a total of $2n$ data points, where the x_i's are restricted to the interval [0, 5]. If the xy-relationship is assumed to be linear and if the objective is to estimate the slope with the greatest possible precision, what values should be assigned to the x_i's?

11.3.5. Suppose a total of $n = 9$ measurements are to be taken on a simple linear model, where the x_i's will be set equal to $1, 2, \ldots,$ and 9. If the variance associated with the xy-relationship is known to be 45.0, what is the probability that the estimated slope will be within 1.5 units of the true slope?

11.3.6. Prove the useful computing formula (Equation 11.3.5) that

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^{n} y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i$$

11.3.7. The sodium nitrate ($NaNO_3$) solubility data in Question 11.2.3 is described nicely by the regression line $y = 67.508 + 0.871x$, where $s = 0.959$. Construct a 90% confidence interval for the y-intercept, β_0.

11.3.8. Set up and carry out an appropriate hypothesis test for the Hanford radioactive contamination data given in Question 11.2.9. Let $\alpha = 0.05$. Justify your choice of H_0 and H_1. What do you conclude?

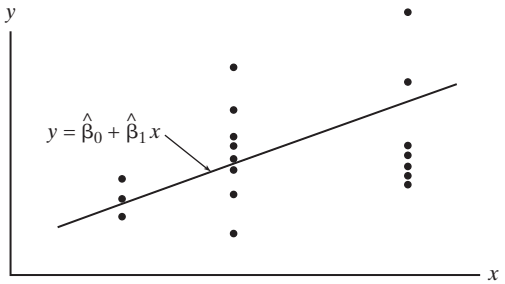
11.3.9. Test H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$ for the plumage index/behavioral index data given in Question 11.2.11. Let $\alpha = 0.05$. Use the fact that $y = 0.61 + 0.84x$ is the best straight line describing the xy-relationship.

11.3.10. Let (x_1, Y_1), (x_2, Y_2), $\ldots$, and (x_n, Y_n) be a set of points satisfying the assumptions of the simple linear model. Prove that

$$E(\bar{Y}) = \beta_0 + \beta_1 \bar{x}$$

11.3.11. Derive a formula for a 95% confidence interval for β_0 if $n(x_i, Y_i)$'s are taken on a simple linear model where σ is known.

11.3.12. Which, if any, of the assumptions of the simple linear model appear to be violated in the following scatterplot? Which, if any, appear to be satisfied? Which, if any, cannot be assessed by looking at the scatterplot?



11.3.13. State the decision rule and the conclusion if H_0: $\sigma^2 = 12.6$ is to be tested against H_1: $\sigma^2 \neq 12.6$ where $n = 24$, $s^2 = 18.2$, and $\alpha = 0.05$.

11.3.14. Construct a 90% confidence interval for σ^2 in the cigarette-consumption/CHD mortality data given in Case Study 11.3.1.

11.3.15. Recall Kepler's Third Law data given in Question 8.2.1. The estimated regression line describing the xy-relationship has the equation $y = 1.795 + 0.181x$, where $s = 1.8$. Construct a 90% confidence interval for σ^2.

DRAWING INFERENCES ABOUT $E(Y \mid x)$

In Case Study 11.3.1, the random variable Y represents the CHD mortality resulting from cigarette consumption, x. A public health official might want to have some idea of the range of mortality likely to be encountered in a country where x is, say, 4200.

Intuition tells us that a reasonable point estimator for $E(Y \mid x)$ is the height of the regression line at x, that is, $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$. By Theorem 11.3.2, the latter is unbiased:

$$E(\hat{Y}) = E(\hat{\beta}_0 + \hat{\beta}_1 x) = E(\hat{\beta}_0) + xE(\hat{\beta}_1) = \beta_0 + \beta_1 x$$

Of course, to use $\hat{Y}$ in any inference procedure requires that we know its variance. But

$$\begin{aligned}\text{Var}(\hat{Y}) &= \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x) = \text{Var}(\bar{Y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x) \\ &= \text{Var}[\bar{Y} + \hat{\beta}_1 (x - \bar{x})] \\ &= \text{Var}(\bar{Y}) + (x - \bar{x})^2 \text{Var}(\hat{\beta}_1) \quad \text{(why?)} \\ &= \frac{1}{n}\sigma^2 + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\sigma^2 \\ &= \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]\end{aligned}$$

An application of Definition 7.3.3, then, allows us to construct a Student t random variable based on $\hat{Y}$. Specifically,

$$T_{n-2} = \frac{\hat{Y} - (\beta_0 + \beta_1 x)}{\sigma \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}} \Bigg/ \sqrt{\frac{(n-2)S^2}{\sigma^2 (n-2)}} = \frac{\hat{Y} - (\beta_0 + \beta_1 x)}{S \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}}$$

has a Student t distribution with $n-2$ degrees of freedom. Isolating $\beta_0 + \beta_1 x = E(Y \mid x)$ in the center of the inequalities $P(-t_{\alpha/2,n-2} \le T_{n-2} \le t_{\alpha/2,n-2}) = 1 - \alpha$ produces a $100(1-\alpha)\%$ confidence interval for $E(Y \mid x)$.

Theorem 11.3.7

Let $(x_1, Y_1), (x_2, Y_2), \ldots,$ and (x_n, Y_n) be a set of points that satisfy the assumptions of the simple linear model. A $100(1-\alpha)\%$ confidence interval for $E(Y \mid x) = \beta_0 + \beta_1 x$ is given by $(\hat{y} - w, \hat{y} + w)$, where

$$w = t_{\alpha/2,n-2} \cdot s \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

and $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

Example 11.3.2

Look again at Case Study 11.3.1. Suppose a country's public health officials estimate cigarette consumption to be 4200 per adult per year. If that were the case, what CHD mortality would they expect? Answer the question by constructing a 95% confidence interval for $E(Y|4200)$.

Here, $n = 21$, $t_{.025,19} = 2.0930$, $\sum_{i=1}^{21} (x_i - \bar{x})^2 = 13{,}056{,}523.81$, $s = 46.707$, $\hat{\beta}_0 = 15.7661$, $\hat{\beta}_1 = 0.0601$, and $\bar{x} = 2148.095$. From Theorem 11.3.7, then,

$$\hat{y} = 15.7661 + 0.0601(4200) = 268.1861$$

$$w = 2.0930(46.707)\sqrt{\frac{1}{21} + \frac{(4200 - 2148.095)^2}{13{,}056{,}523.81}} = 59.4714$$

and the 95% confidence interval for $E(Y|4200)$ is

$$(268.1861 - 59.4714,\ 268.1861 + 59.4714)$$

which rounded to two decimal places is

$$(208.71, 327.66)$$

Comment Notice from the formula in Theorem 11.3.7 that the width of a confidence interval for $E(Y \mid x)$ increases as the value of x becomes more extreme. That is, we are better able to predict the location of the regression line for an x-value close to $\bar{x}$ than we are for x-values that are either very small or very large.

Figure 11.3.4 shows the dependence of w on x for the data from Case Study 11.3.1. The lower and upper limits for the 95% confidence interval for $E(Y \mid x)$ have been calculated for all x. Pictured is the dotted curve (or *95% confidence band*) connecting those endpoints. The width of the band is smallest when $x = 2148.1\ (= \bar{x})$.

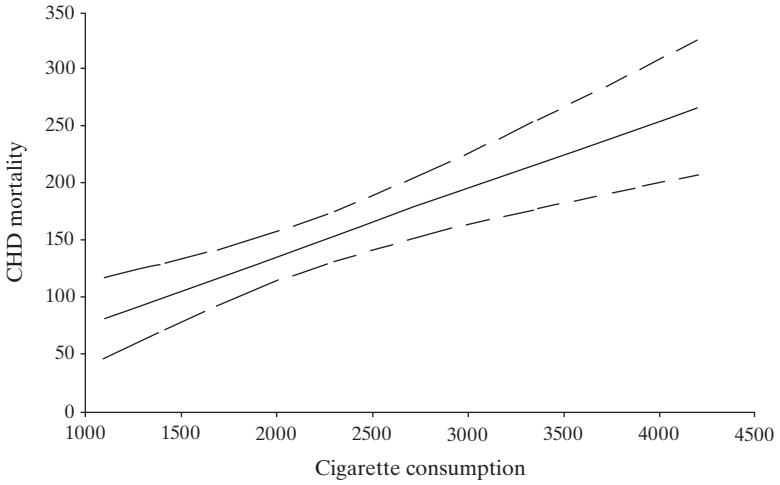


Figure 11.3.4

■

DRAWING INFERENCES ABOUT FUTURE OBSERVATIONS

A variation on Theorem 11.3.7 is the determination of a range of numbers that would have a high probability of including the value Y of a *single future observation* to be recorded at some given level of x. In Case Study 11.3.1, public health officials might want to predict the *actual* (not the *average*) CHD mortality that would occur if cigarette consumption is x.

Let $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$ be a set of n points that satisfy the assumptions of the simple linear model, and let (x, Y) be a hypothetical future observation, where Y is independent of the n Y_i's. A *prediction interval* is a range of numbers that contains Y with a specified probability.

Consider the difference $\hat{Y} - Y$. Clearly,

$$E(\hat{Y} - Y) = E(\hat{Y}) - E(Y) = (\beta_0 + \beta_1 x) - (\beta_0 + \beta_1 x) = 0$$

and

$$\begin{aligned}\text{Var}(\hat{Y} - Y) &= \text{Var}(\hat{Y}) + \text{Var}(Y) \\ &= \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] + \sigma^2 \\ &= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]\end{aligned}$$

Following exactly the same steps that were taken in the derivation of Theorem 11.3.7, a Student t random variable with $n - 2$ degrees of freedom can be constructed from $\hat{Y} - Y$ (using Definition 7.3.3). Inverting the equation $P(-t_{\alpha/2,n-2} \le T_{n-2} \le t_{\alpha/2,n-2}) = 1 - \alpha$ will then yield the prediction interval $(\hat{y} - w, \hat{y} + w)$ given in Theorem 11.3.8.

Theorem 11.3.8

Let $(x_1, Y_1), (x_2, Y_2), \ldots,$ and (x_n, Y_n) be a set of n points that satisfy the assumptions of the simple linear model. A $100(1-\alpha)\%$ prediction interval for Y *at the fixed value x is given by $(\hat{y} - w, \hat{y} + w)$, where*

$$w = t_{\alpha/2,n-2} \cdot s \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

and $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

Example 11.3.3

Based on the data in Case Study 11.3.1, we calculated in Example 11.3.2 that a 95% confidence interval for $E(Y|4200)$ is (208.71, 327.66). How does that compare to the corresponding 95% prediction interval for Y?

When $x = 4200$, $\hat{y} = 268.1861$ for both intervals. From Theorem 11.3.8, the width of the 95% prediction interval for Y is:

$$w = 2.0930(46.707)\sqrt{1 + \frac{1}{21} + \frac{(4200 - 2148.095)^2}{13{,}056{,}523.81}} = 114.4725$$

The 95% prediction interval, then, is

$$(268.1861 - 114.4725, 268.1861 + 114.4725)$$

which rounded to two decimal places is

$$(153.76, 382.61)$$

which makes it 92% wider than the 95% confidence interval for $E(Y|4200)$. ■

TESTING THE EQUALITY OF TWO SLOPES

We saw in Chapter 9 that the comparison of two treatments or two conditions often leads to a hypothesis test that the mean of one is equal to the mean of the other. Similarly, the comparison of two linear xy-relationships often requires that we test H_0: $\beta_1 = \beta_1^*$, where β_1 and β_1^* are the true slopes associated with the two regressions.

If the data points taken on the two regressions are all independent, a two-sample t test can be set up based on the properties in Theorems 11.3.2 and 11.3.3. Theorem 11.3.9 identifies the appropriate test statistic and summarizes the GLRT decision rule. Details of the proof will be omitted.

Theorem 11.3.9 *Let $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$ and $(x_1^*, Y_1^*), (x_2^*, Y_2^*), \ldots, (x_m^*, Y_m^*)$ be two independent sets of points, each satisfying the assumptions of the simple linear model, that is, $E(Y \mid x) = \beta_0 + \beta_1 x$ and $E(Y^* \mid x^*) = \beta_0^* + \beta_1^* x^*$.*

a. *Let*

$$T = \frac{\hat{\beta}_1 - \hat{\beta}_1^* - (\beta_1 - \beta_1^*)}{S\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + \frac{1}{\sum_{i=1}^{m}(x_i^* - \bar{x}^*)^2}}}$$

where

$$S = \sqrt{\frac{\sum_{i=1}^{n}[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 + \sum_{i=1}^{m}[Y_i^* - (\hat{\beta}_0^* + \hat{\beta}_1^* x_i^*)]^2}{n + m - 4}}$$

Then T has a Student t distribution with $n + m - 4$ degrees of freedom.

b. *To test $H_0 : \beta_1 = \beta_1^*$ versus $H_1 : \beta_1 \neq \beta_1^*$ at the α level of significance, reject H_0 if t is either (1) $\leq -t_{\alpha/2,n+m-4}$ or (2) $\geq t_{\alpha/2,n+m-4}$, where*

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_1^*}{s\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + \frac{1}{\sum_{i=1}^{m}(x_i^* - \bar{x}^*)^2}}}$$

(One-sided tests are defined in the usual way by replacing $\pm t_{\alpha/2,n+m-4}$ with either $t_{\alpha,n+m-4}$ or $-t_{\alpha,n+m-4}$.)

Example 11.3.4 Genetic variability is thought to be a key factor in the survival of a species, the idea being that "diverse" populations should have a better chance of coping with changing environments. Table 11.3.3 summarizes the results of a study designed to test that hypothesis experimentally [data slightly modified from (4)]. Two populations of fruit

Table 11.3.3

Date	Day no., $x(= x^*)$	Strain A pop^n, y	Strain B pop^n, y^*
Feb 2	0	100	100
May 13	100	250	203
Aug 21	200	304	214
Nov 29	300	403	295
Mar 8	400	446	330
Jun 16	500	482	324

flies (*Drosophila serrata*)—one that was cross-bred (Strain A) and the other, in-bred (Strain B)—were put into sealed containers where food and space were kept to a minimum. Recorded every hundred days were the numbers of *Drosophila* alive in each population.

Figure 11.3.5 shows a graph of the two sets of population figures. For both strains, growth was approximately linear over the period covered. Strain A, though, with an estimated slope of 0.74, increased at a faster rate than did Strain B, where the estimated slope was 0.45. The question is, do we have enough evidence here to reject the null hypothesis that the two true slopes are equal? Is the difference between 0.74 and 0.45, in other words, statistically significant?

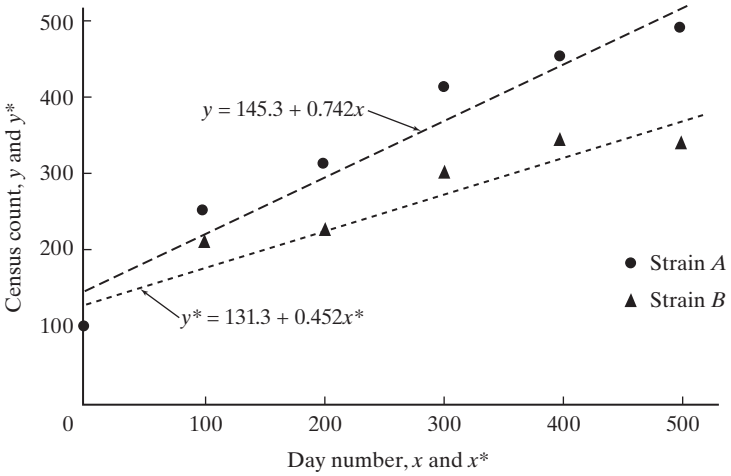


Figure 11.3.5

Let $\alpha = 0.05$ and let (x_i, y_i), $i = 1, 2, \ldots, 6$, and (x_i^*, y_i^*), $i = 1, 2, \ldots, 6$, denote the times and population sizes for Strain A and Strain B, respectively. Our objective is to test $H_0 : \beta_1 = \beta_1^*$ versus $H_1 : \beta_1 > \beta_1^*$. Rejecting H_0, of course, would support the contention that genetic variability benefits a species' chances of survival.

From Table 11.3.3, $\bar{x} = \bar{x}^* = 250$ and

$$\sum_{i=1}^{6} (x_i - \bar{x})^2 = \sum_{i=1}^{6} (x_i^* - \bar{x}^*)^2 = 175{,}000$$

Also,

$$\sum_{i=1}^{6} [y_i - (145.3 + 0.742x_i)]^2 = 5512.14$$

and

$$\sum_{i=1}^{6} [y_i^* - (131.3 + 0.452x_i^*)]^2 = 3960.14$$

so

$$s = \sqrt{\frac{5512.14 + 3960.14}{6 + 6 - 4}} = 34.41$$

Since H_1 is one-sided to the right, we should reject H_0 if $t \geq t_{.05,8} = 1.8595$. But

$$t = \frac{0.742 - 0.452}{34.41\sqrt{\frac{1}{175{,}000} + \frac{1}{175{,}000}}}$$

$$= 2.50$$

These data, then, *do* support the theory that genetically mixed populations have a better chance of survival in hostile environments. ■

Questions

11.3.16. Regression techniques can be very useful in situations where one variable—say, y—is difficult to measure but x is not. Once such an xy-relationship has been "calibrated," based on a set of (x_i, y_i)'s, future values of Y can be easily estimated using $\hat{\beta}_0 + \hat{\beta}_1 x$. Determining the volume of an irregularly shaped object, for example, is often difficult, but weighing that object is likely to be easy. The following table shows the weights (in kilograms) and the volumes (in cubic decimeters) of eighteen children between the ages of five and eight (15). The estimated regression line has the equation $y = -0.104 + 0.988x$, where $s = 0.202$.

(a) Construct a 95% confidence interval for $E(Y|14.0)$.

(b) Construct a 95% prediction interval for the volume of a child weighing 14.0 kilograms.

Weight, x	Volume, y	Weight, x	Volume, y
17.1	16.7	15.8	15.2
10.5	10.4	15.1	14.8
13.8	13.5	12.1	11.9
15.7	15.7	18.4	18.3
11.9	11.6	17.1	16.7
10.4	10.2	16.7	16.6
15.0	14.5	16.5	15.9
16.0	15.8	15.1	15.1
17.8	17.6	15.1	14.5

11.3.17. Construct a 95% confidence interval for $E(Y \mid 2.750)$ using the connecting rod data given in Case Study 11.2.1.

11.3.18. For the CHD mortality data of Case Study 11.3.1, construct a 99% confidence interval for the expected death rate in a country where the cigarette consumption is 2500 per adult per year. Is a public health official more likely to be interested in a 99% confidence interval for $E(Y \mid 2500)$ or a 99% prediction interval for Y when $x = 2500$?

11.3.19. The fuel economy (in miles per gallon) of an automobile can depend on a number of factors, but the table below suggests that the weight of vehicle is a very good predictor.

Model	Weight, x (lbs)	Fuel usage, y, (mpg)
Toyota Yaris Liftback (manual)	2370	34
Toyota Yaris Sedan	2430	33
Scion xB (manual)	2450	32
Honda Fit Sport (manual)	2495	34
Honda Fit	2535	32
Chevrolet Malibu (4-cyl.)	3135	24
Honda Accord (4-cyl.)	3195	24
Nissan Altima 2.5 S	3215	25
Toyota Camry LE (4-cyl.)	3280	24
BMW 325i	3460	24
Volkswagen Passat 2.0T	3465	24
Lexus IS250	3510	24

Find the 95% confidence interval for $E(Y|2890)$, where 2890 is the weight of the Honda Civic Hybrid. Does the interval include 37, the miles per gallon of the Civic?

11.3.20. In the radioactive exposure example in Question 11.2.9, find the 95% confidence interval for $E(Y|9.00)$ and the prediction interval for the value 9.00.

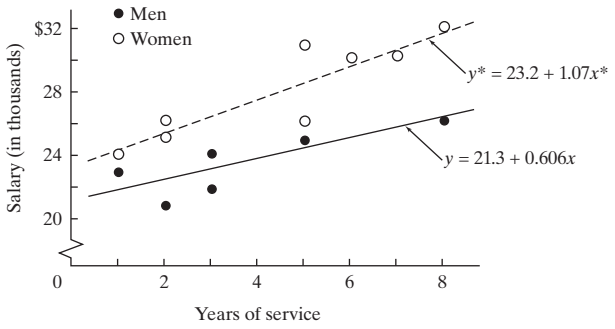
11.3.21. Attorneys representing a group of male buyers employed by Flirty Fashions are filing a reverse discrimination suit against the female-owned company. Central to their case are the following data, showing the relationship between years of service and annual salary for the firm's fourteen buyers, six of whom are men. The plaintiffs claim that the difference in slopes (0.606 for men versus 1.07 for women) is prima facie evidence that the company's salary policies discriminate against men. As the lawyer for Flirty Fashions, how would you respond? Use the following sums:

$$\sum_{i=1}^{6} (y_i - 21.3 - 0.606x_i)^2 = 5.983$$

and

$$\sum_{i=1}^{8} (y_i^* - 23.2 - 1.07x_i^*)^2 = 13.804$$

Also, $\sum_{i=1}^{6} (x_i - \bar{x})^2 = 31.33$ and $\sum_{i=1}^{8} (x_i^* - \bar{x}^*)^2 = 46$.



11.3.22. Polls taken during a city's last two administrations (one Democratic, one Republican) suggested that public support of the two mayors fell off linearly with years in office. Can it be concluded from the following data that the rates at which the two administrations lost favor were significantly different? Let $\alpha = 0.05$. (*Note:* $y = 69.3077 - 3.4615x$ with an estimated standard deviation of 0.9058 and $y^* = 59.9407 - 2.7373x^*$ with an estimated standard deviation of 1.2368.)

Democratic Mayor		Republican Mayor	
Years after Taking Office, x	Percent in Support, y	Years after Taking Office, x^*	Percent in Support, y^*
2	63	1	58
3	58	2	55
5	52	4	47
7	46	6	43
8	41	7	41
		8	39

11.3.23. Prove that the variance of $\hat{Y}$ can also be written

$$\text{Var}(\hat{Y}) = \frac{\sigma^2 \sum_{i=1}^{n} (x_i - x)^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

11.3.24. Show that

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

for any set of points (x_i, Y_i), $i = 1, 2, \ldots, n$.

11.4 Covariance and Correlation

Our discussion of xy-relationships in Chapter 11 began with the simplest possible setup from a statistical standpoint—the case where the (x_i, y_i)'s are just numbers and have no probabilistic structure whatsoever. Then we examined the more complicated (and more "inference-friendly") scenario where x_i is a constant but Y_i is a random variable. Introduced in this section is the next level of complexity—problems where both X_i and Y_i are assumed to be random variables. (Measurements of the form (x_i, y_i) or (x_i, Y_i) are typically referred to as *regression data*; observations satisfying the assumptions made in this section—that is, measurements of the form (X_i, Y_i)—are more commonly referred to as *correlation data*.)

MEASURING THE DEPENDENCE BETWEEN TWO RANDOM VARIABLES

Given a pair of random variables, it makes sense to inquire how one varies *with respect to the other*. If X increases, for example, does Y also tend to increase? And if so, how strong is the dependence between the two?

The first step in addressing such questions was taken in Section 3.9 with the definition of covariance. In that section, its role was primarily as a tool for finding the variance of a sum of random variables. Here, it will serve as the basis for measuring the relationship between X and Y.

THE CORRELATION COEFFICIENT

The covariance of X and Y necessarily reflects the *units* of both random variables, which can make it difficult to interpret. In applied settings, it helps to have a *dimensionless* measure of dependency so that one xy-relationship can be compared to another. Dividing $\text{Cov}(X, Y)$ by $\sigma_X \sigma_Y$ accomplishes not only that objective but also scales the quotient to be a number between -1 and $+1$.

Definition 11.4.1
Let X and Y be any two random variables. The *correlation coefficient of X and Y,* denoted $\rho(X, Y)$, is given by

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \text{Cov}(X^*, Y^*)$$

where $X^* = (X - \mu_X)/\sigma_X$ and $Y^* = (Y - \mu_Y)/\sigma_Y$.

Theorem 11.4.1 *For any two random variables X and Y,*

a. $|\rho(X, Y)| \leq 1$.

b. $|\rho(X, Y)| = 1$ *if and only if* $Y = aX + b$ *for some constants a and b (except possibly on a set of probability zero).*

Proof Following the notation of Definition 11.4.1, let X^* and Y^* denote the standardized transformations of X and Y. Then

$$\begin{aligned} 0 \leq \text{Var}(X^* \pm Y^*) &= \text{Var}(X^*) \pm 2\,\text{Cov}(X^*, Y^*) + \text{Var}(Y^*) \\ &= 1 \pm 2\rho(X, Y) + 1 \\ &= 2\,[1 \pm \rho(X, Y)] \end{aligned}$$

But $1 \pm \rho(X, Y) \geq 0$ implies that $|\rho(X, Y)| \leq 1$, and part (a) of the theorem is proved.

Next, suppose that $\rho(X, Y) = 1$. Then $\text{Var}(X^* - Y^*) = 0$; however, a random variable with zero variance is constant, except possibly on a set of probability zero. From the constancy of $X^* - Y^*$, it readily follows that Y is a linear function of X. The case for $\rho(X, Y) = -1$ is similar.

The converse of part (b) is left as an exercise.

Questions

11.4.1. Let X and Y have the joint pdf

$$f_{X,Y}(x, y) = \begin{cases} \frac{x+2y}{22}, & \text{for } (x, y) = (1, 1), (1, 3), (2, 1), (2, 3) \\ 0, & \text{elsewhere} \end{cases}$$

Find $\text{Cov}(X, Y)$ and $\rho(X, Y)$.

11.4.2. Suppose that X and Y have the joint pdf

$$f_{X,Y}(x, y) = x + y, \quad 0 < x < 1, 0 < y < 1$$

Find $\rho(X, Y)$.

11.4.3. If the random variables X and Y have the joint pdf

$$f_{X,Y}(x, y) = \begin{cases} 8xy, & 0 \leq y \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

show that $\text{Cov}(X, Y) = \frac{8}{450}$. Calculate $\rho(X, Y)$.

11.4.4. Suppose that X and Y are discrete random variables with the joint pdf

(x, y)	$f_{X,Y}(x, y)$
$(1, 2)$	$\frac{1}{2}$
$(1, 3)$	$\frac{1}{4}$
$(2, 1)$	$\frac{1}{8}$
$(2, 4)$	$\frac{1}{8}$

Find the correlation coefficient between X and Y.

11.4.5. Prove that $\rho(a + bX, c + dY) = \rho(X, Y)$ for constants a, b, c, and d where b and d are positive. Note that this result allows for a change of scale to one convenient for computation.

11.4.6. Let the random variable X take on the values $1, 2, \ldots, n$, each with probability $1/n$. Define Y to be X^2. Find $\rho(X, Y)$ and $\lim_{n\to\infty} \rho(X, Y)$.

11.4.7. (a) For random variables X and Y, show that

$$\text{Cov}(X + Y, X - Y) = \text{Var}(X) - \text{Var}(Y)$$

(b) Suppose that $\text{Cov}(X, Y) = 0$. Prove that

$$\rho(X + Y, X - Y) = \frac{\text{Var}(X) - \text{Var}(Y)}{\text{Var}(X) + \text{Var}(Y)}$$

ESTIMATING $\rho(X, Y)$: THE SAMPLE CORRELATION COEFFICIENT

We conclude this section with an estimation problem. Suppose the correlation coefficient between X and Y is unknown, but we have some relevant information about its value in the form of n measurements $(X_1, Y_1), (X_2, Y_2), \ldots,$ and (X_n, Y_n). How can we use those data to estimate $\rho(X, Y)$?

Since the correlation coefficient can be written in terms of various theoretical moments,

$$\rho(X, Y) = \frac{E(XY) - E(X)E(Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$$

it would seem reasonable to estimate each component of $\rho(X, Y)$ with its corresponding *sample* moment. That is, let $\bar{X}$ and $\bar{Y}$ approximate $E(X)$ and $E(Y)$, replace $E(XY)$ with

$$\frac{1}{n}\sum_{i=1}^{n} X_i Y_i$$

and substitute

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

for $\text{Var}(X)$ and $\text{Var}(Y)$, respectively.

We define the *sample correlation coefficient,* then, to be the ratio

$$R = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}\,\bar{Y}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{11.4.1}$$

or, equivalently,

$$R = \frac{n\sum_{i=1}^{n} X_i Y_i - \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{\sqrt{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}\sqrt{n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2}} \tag{11.4.2}$$

(Sometimes R is referred to as the *Pearson product-moment correlation coefficient*, in honor of the eminent British statistician Karl Pearson.)

Questions

11.4.8. Derive Equation 11.4.2 from Equation 11.4.1.

11.4.9. Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be a set of measurements whose sample correlation coefficient is r. Show that

$$r = \hat{\beta}_1 \cdot \frac{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}}{\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

where $\hat{\beta}_1$ is the maximum likelihood estimate for the slope.

INTERPRETING R

The properties cited for $\rho(X, Y)$ in Theorem 11.4.1 are not sufficient to provide a useful interpretation of R. What does it mean, for example, to say that the sample correlation coefficient is 0.73, or 0.55, or -0.24? One way to answer such a question focuses on the *square* of R, rather than on R itself.

We know from Equation 11.3.3 that

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 - \hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$$

Using the relationship between $\hat{\beta}_1$ and r in Question 11.4.9—together with the fact that

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \Big/ n$$

we can write

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 - r^2 \cdot \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sum_{i=1}^{n}(x_i - \bar{x})^2$$

which reduces to

$$r^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{11.4.3}$$

Equation 11.4.3 has a nice, simple interpretation. Notice that

1. $\sum_{i=1}^{n}(y_i - \bar{y})^2$ represents the *total variability* in the dependent variable, that is, the extent to which the y_i's are not all the same.
2. $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$ represents the variation in the y_i's *not explained* (or accounted for) by the linear regression with x.
3. $\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$ represents the variation in the y_i's that *is explained* by the linear regression with x.

Therefore, r^2 is the *proportion of the total variation in the y_i's that can be attributed to the linear relationship with x*. So, if $r = 0.60$, we can say that 36% of the variation in Y is explained by the linear regression with X (and that 64% is associated with other factors).

Comment The quantity r^2 is sometimes called the *coefficient of determination*.

CASE STUDY 11.4.1

The Scholastic Aptitude Test (SAT) is widely used by colleges and universities to help choose their incoming classes. It was never designed to measure the quality of education provided by secondary schools, but critics and supporters alike often force it into that role. The problem is that average SAT scores associated with schools or districts or states reflect a variety of factors, some of which have little or nothing to do with the quality of instruction that students are receiving. One of these factors, as Section 3.13 pointed out is the states' participation rate, the percentage of eligible students who take the SAT.

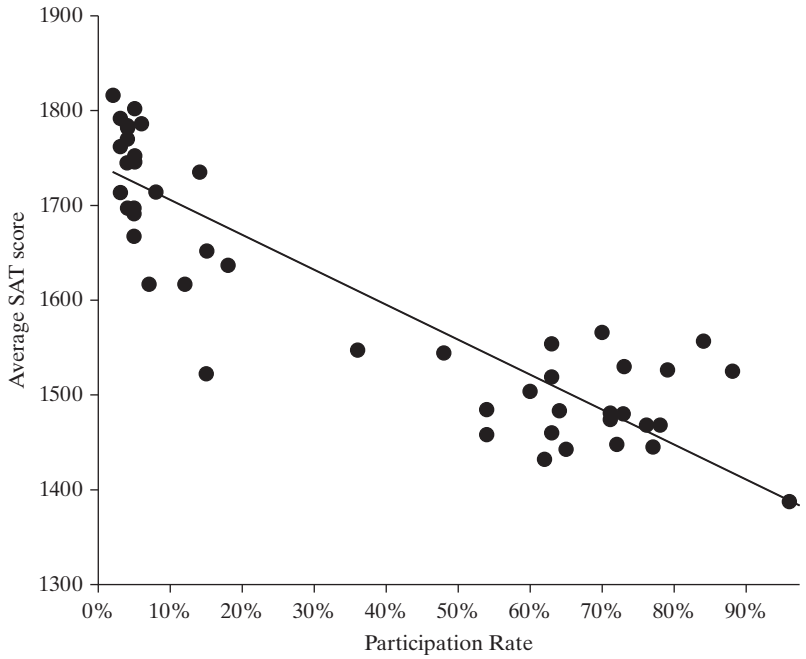


Figure 11.4.1

Table 11.4.1 shows one testing period's average SAT scores (y), by state, as a function of participation rate (x). Figure 11.4.1 suggests a strong dependency between the two measurements—as a state's participation rate goes down, its average SAT score goes up. In North Dakota, for example, only 2% of the students eligible to take the test actually did; in Maryland, the participation rate was a dramatically larger 78%. The average SAT score in Maryland was 1468; in North Dakota the average score of 1816 was 24% higher.

A good way to quantify the overall relationship between test scores and participation rates is to calculate the data's sample correlation coefficient, r. From Table 11.4.1, we can calculate the sums necessary to evaluate Equation 11.4.2

$$\sum_{i=1}^{51} x_i = 2{,}099 \qquad \sum_{i=1}^{51} y_i = 81{,}108$$

$$\sum_{i=1}^{51} x_i^2 = 147{,}507 \qquad \sum_{i=1}^{51} y_i^2 = 129{,}989{,}648$$

$$\sum_{i=1}^{51} x_i y_i = 3{,}112{,}824$$

Substituting the sums into the formula for r, then, shows that the sample correlation coefficient is -0.912:

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\left(\sum_{i=1}^{n} y_i^2\right) - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

$$= \frac{51(3,112,824) - (2099)(81,108)}{\sqrt{51(147,507) - 2099^2}\sqrt{51(129,909,648) - 81,108^2}}$$

(Continued on next page)

(Case Study 11.4.1 continued)

Table 11.4.1

State	Participation Rate	Average SAT Score	State	Participation Rate	Average SAT Score
Alabama	7%	1617	Montana	18%	1637
Alaska	54%	1485	Nebraska	4%	1745
Arizona	36%	1547	Nevada	54%	1458
Arkansas	4%	1698	New Hampshire	70%	1566
California	60%	1504	New Jersey	79%	1526
Colorado	14%	1735	New Mexico	12%	1617
Connecticut	88%	1525	New York	76%	1468
Delaware	100%	1359	North Carolina	64%	1483
Dist. Columbia	100%	1309	North Dakota	2%	1816
Florida	72%	1448	Ohio	15%	1652
Georgia	77%	1445	Oklahoma	5%	1697
Hawaii	63%	1460	Oregon	48%	1544
Idaho	100%	1364	Pennsylvania	71%	1481
Illinois	5%	1802	Rhode Island	73%	1480
Indiana	71%	1474	South Carolina	65%	1443
Iowa	3%	1794	South Dakota	3%	1792
Kansas	5%	1753	Tennessee	8%	1714
Kentucky	5%	1746	Texas	62%	1432
Louisiana	5%	1667	Utah	5%	1690
Maine	96%	1387	Vermont	63%	1554
Maryland	78%	1468	Virginia	73%	1530
Massachusetts	84%	1556	Washington	63%	1519
Michigan	4%	1784	West Virginia	15%	1522
Minnesota	6%	1786	Wisconsin	4%	1782
Mississippi	3%	1714	Wyoming	3%	1762
Missouri	4%	1771			

Based on data from: http://blog.prepscholar.com/average-sat-scores-by-state-most-recent

Since $r^2 = (0.912)^2 = 0.832$, we can say that 83.2% of the variability in SAT scores from state to state can be attributed to the linear relationship between test scores and participation rates.

About the Data The magnitude of r^2 for these data should be a clear warning that comparing average SATs at face value from state to state or school system to school system is largely meaningless. It would make more sense to examine the residuals associated with $y = \hat{\beta}_0 + \hat{\beta}_1 x$. States with particularly large positive values for $y - \hat{y}$ may be doing something that other states might be well advised to copy.

Questions

11.4.10. In Case Study 11.3.1, how much of the variability in CHD mortality is explained by cigarette consumption?

11.4.11. Some baseball fans believe that the number of home runs a team hits is markedly affected by the altitude of the club's home park. The rationale is that the air is thinner at the higher altitudes, and balls would be expected to travel farther. The following table shows the altitudes (X) of American League ballparks and the number of home runs (Y) that each team hit during a recent season (183). Calculate the sample correlation coefficient, r, using the sums below. What would you conclude?

$$\sum_{i=1}^{12} x_i = 4936 \qquad \sum_{i=1}^{12} y_i = 1175$$

$$\sum_{i=1}^{12} x_i^2 = 3{,}071{,}116 \qquad \sum_{i=1}^{12} y_i^2 = 123{,}349$$

$$\sum_{i=1}^{12} x_i y_i = 480{,}565$$

Club	Altitude, x	Number of Home Runs, y
Cleveland	660	138
Milwaukee	635	81
Detroit	585	135
New York	55	90
Boston	21	120
Baltimore	20	84
Minnesota	815	106
Kansas City	750	57
Chicago	595	109
Texas	435	74
California	340	61
Oakland	25	120

11.4.12. Many people believe that a salary bonus is a reward for good performance. The corporate world may have a different understanding. A random sample of thirty chief executive officers of large capitalization public companies recorded the cash bonus paid, x (in \$100,000), and the performance of the company, y, as measured by percentage change in company revenues. The following sums resulted.

$$\sum_{i=1}^{30} x_i = 1{,}300.69 \qquad \sum_{i=1}^{30} y_i = 323$$

$$\sum_{i=1}^{30} x_i^2 = 86{,}754.6939 \qquad \sum_{i=1}^{30} y_i^2 = 11,881$$

$$\sum_{i=1}^{30} x_i y_i = 7{,}807.36$$

Find the sample coefficient of correlation. What does this study say about the relationship between bonuses and performance?

11.4.13. The extent to which stress is a contributing factor to the severity of chronic illnesses was the focus of the study summarized in the following table (221). Seventeen conditions were compared on a Seriousness of Illness Rating Scale (SIRS). Patients with each of these conditions were asked to fill out a Schedule of Recent Experience (SRE) questionnaire. Higher scores on the SRE reflect presumably greater levels of stress. How much of the variation in the SIRS values can be attributed to the linear regression with SRE?

Admitting Diagnosis	Average SRE, x	SIRS, y
Dandruff	26	21
Varicose veins	130	173
Psoriasis	317	174
Eczema	231	204
Anemia	325	312
Hyperthyroidism	816	393
Gallstones	563	454
Arthritis	312	468
Peptic ulcer	603	500
High blood pressure	405	520
Diabetes	599	621
Emphysema	357	636
Alcoholism	688	688
Cirrhosis	443	733
Schizophrenia	609	776
Heart failure	772	824
Cancer	777	1020

Use the following sums:

$$\sum_{i=1}^{17} x_i = 7{,}973 \qquad \sum_{i=1}^{17} y_i = 8{,}517$$

$$\sum_{i=1}^{17} x_i^2 = 4{,}611{,}291 \qquad \sum_{i=1}^{17} y_i^2 = 5{,}421{,}917$$

$$\sum_{i=1}^{17} x_i y_i = 4{,}759{,}470$$

11.4.14. Burglary and larceny both involve the illegal taking of something of value. The difference, simply put, is that burglary involves unlawful entry to a structure, while larceny does not. While the two crimes might seem similar, the correlation between the two is quite low. A data set to be used for such an analysis is the annual rates of burglary, x, and larceny, y, from 1975 to 2010. Both variables give the number of crimes per 100,000 U.S. citizens. Calculate the xy correlation. Use the following sums:

$$\sum_{i=1}^{36} x_i = 994.7700, \quad \sum_{i=1}^{36} x_i^2 = 28462.1047,$$

$$\sum_{i=1}^{36} y_i = 254.6900, \quad \sum_{i=1}^{36} y_i^2 = 1816.1417,$$

$$\sum_{i=1}^{36} x_i y_i = 7051.2633$$

11.4.15. A common saying in golf is "You drive for show, but you putt for dough." To see if there is any truth in this assertion, data for ninety-six top money-winning golfers were examined. For each, their money earnings in 2014 (y, in \$ millions), their average yards per drive (v), and their average number of putts (x) were tallied.

(a) Show that the correlation coefficient between the putting average and earnings reveal a slightly stronger

relationship than that for driving and earnings. Use the following sums:

$$\sum_{i=1}^{96} v_i = 27{,}989, \qquad \sum_{i=1}^{96} v_i^2 = 8{,}167{,}723,$$
$$\sum_{i=1}^{96} y_i = 230.87, \qquad \sum_{i=1}^{96} y_i^2 = 734.32,$$
$$\sum_{i=1}^{96} v_i y_i = 67658.00,$$

and also $\sum_{i=1}^{96} x_i = 169.31, \sum_{i=1}^{96} x_i^2 = 298.64, \sum_{i=1}^{96} x_i y_i = 406.37$

(b) For each correlation r, compute r^2 to show that neither the v nor the x variable alone is a good predictor of earnings.

11.5 The Bivariate Normal Distribution

The singular importance of the normal distribution in univariate inference procedures should, by now, be abundantly clear. In dealing with problems that involve *two* random variables—for example, the calculation of $\rho(X, Y)$—it should come as no surprise that the most frequently encountered *joint* pdf, $f_{X,Y}(x, y)$, is a bivariate version of the normal curve. Our objectives in this section are twofold: (1) to deduce the form of the bivariate normal from basic principles and (2) to identify the particular properties of that pdf that pertain to the problem of assessing the nature of the dependence between X and Y.

GENERALIZING THE UNIVARIATE NORMAL PDF

At this point, we know many things about the *univariate* normal pdf,

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}, \quad -\infty < y < \infty$$

Sections upon sections have been devoted to estimating and testing its parameters, studying its transformations, and learning about its role as an approximation to the distribution of sums and averages. What has not been discussed is the generalization of $f_Y(y)$ itself, to a *bivariate, trivariate,* or *multivariate* pdf.

Given the mathematical complexities inherent in the univariate normal pdf, it should come as no surprise that its extension to higher dimensions is not a simple matter. In the bivariate case, for example, which is the only generalization we will consider, $f_{X,Y}(x, y)$ has *five* different parameters and its functional form is decidely unpleasant.

We will begin by "constructing" a bivariate normal pdf, $f_{X,Y}(x, y)$, using properties suggested by what we already know holds true for the univariate normal, $f_Y(y)$. As a first condition to impose, it seems reasonable to require that the marginal pdfs associated with $f_{X,Y}(x, y)$ be univariate normal densities. It will be sufficient to consider the case where the two marginals are *standard* normals.

If X and Y are *independent* standard normal random variables,

$$f_{X,Y}(x, y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)}, \quad \begin{matrix} -\infty < x < \infty \\ -\infty < y < \infty \end{matrix} \tag{11.5.1}$$

Notice that the simplest extension of $f_{X,Y}(x, y)$ in Equation 11.5.1 is to replace $-\frac{1}{2}(x^2 + y^2)$ with $-\frac{1}{2}c(x^2 + uxy + y^2)$, or, equivalently, with $-\frac{1}{2}c(x^2 - 2vxy + y^2)$, where c and v are constants. The desired joint pdf, then, would have the general form

$$f_{X,Y}(x, y) = Ke^{-\frac{1}{2}c(x^2-2vxy+y^2)} \tag{11.5.2}$$

where K is the constant that makes the double integral of $f_{X,Y}(x, y)$ from $-\infty$ to ∞ equal to 1.

Now, what must be true of K, c, and v if the marginal pdfs based on $f_{X,Y}(x, y)$ are to be standard normals? Note, first, that completing the square in the exponent makes

$$\begin{aligned} x^2 - 2vxy + y^2 &= x^2 - v^2x^2 + (y^2 - 2vxy + v^2x^2) \\ &= (1 - v^2)x^2 + (y - vx)^2 \end{aligned}$$

so

$$f_{X,Y}(x, y) = Ke^{-\frac{1}{2}c(1-v^2)x^2}e^{-\frac{1}{2}c(y-vx)^2}$$

The exponents, though, must be negative, which implies that $1 - v^2 > 0$, or, equivalently, $|v| < 1$.

To find K, we start by calculating

$$\begin{aligned} &\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(1/2)c(1-v^2)x^2} \cdot e^{-(1/2)c(y-vx)^2}\, dy\, dx \\ &\quad = \int_{-\infty}^{\infty} e^{-(1/2)c(1-v^2)x^2}\left[\int_{-\infty}^{\infty} e^{-(1/2)c(y-vx)^2} dy\right] dx \\ &\quad = \int_{-\infty}^{\infty} e^{-(1/2)c(1-v^2)x^2}\left(\frac{\sqrt{2\pi}}{\sqrt{c}}\right) dx \\ &\quad = \frac{\sqrt{2\pi}}{\sqrt{c}}\frac{\sqrt{2\pi}}{\sqrt{c}\sqrt{1-v^2}} \\ &\quad = \frac{2\pi}{c\sqrt{1-v^2}} \end{aligned}$$

It follows that

$$K = \frac{c\sqrt{1-v^2}}{2\pi}$$

The constant c can be any positive value, but a convenient choice proves to be $c = 1/(1 - v^2)$. Substituting K and c, then, into Equation 11.5.2 gives

$$\begin{aligned} f_{X,Y}(x, y) &= \frac{1}{2\pi\sqrt{1-v^2}} e^{-(1/2)[1/(1-v^2)](x^2-2vxy+y^2)} \\ &= \frac{1}{2\pi\sqrt{1-v^2}} e^{-x^2} \cdot e^{-(1/2)[1/(1-v^2)](y-vx)^2} \end{aligned} \tag{11.5.3}$$

Recall that our choice of the form of $f_{X,Y}(x, y)$ was predicated on a wish for the marginal pdfs to be normal. A simple integration shows that to be the case:

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy \\ &= \frac{1}{2\pi\sqrt{1-v^2}} e^{-(1/2)x^2} \int_{-\infty}^{\infty} e^{-(1/2)[1/(1-v^2)](y-vx)^2}\, dy \\ &= \frac{1}{2\pi\sqrt{1-v^2}} e^{-(1/2)x^2} \cdot \sqrt{2\pi}\sqrt{1-v^2} \\ &= \frac{1}{\sqrt{2\pi}} e^{-(1/2)x^2} \end{aligned}$$

Since $f_{X,Y}(x, y)$ is symmetric in x and y, $f_Y(y)$ is also the standard normal.

The constant v is actually the correlation coefficient between X and Y. Since $E(X) = E(Y) = 0$ and $\sigma_X = \sigma_Y = 1$,

$$\begin{aligned}\rho(X,Y) = E(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_{X,Y}(x,y)\,dx\,dy \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} xe^{-(1/2)x^2}\left[\frac{1}{\sqrt{2\pi}\sqrt{1-v^2}}\int_{-\infty}^{\infty} ye^{-(1/2)[1/(1-v^2)](y-vx)^2}dy\right]dx \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} xe^{-(1/2)x^2}\cdot vx\,dx \quad \text{(why?)} \\ &= v\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^2e^{-(1/2)x^2}dx = v\,\text{Var}(X) = v\end{aligned}$$

Finally, we can replace x with $(x-\mu_X)/\sigma_X$ and y with $(y-\mu_Y)/\sigma_Y$. Doing so requires that the original pdf be multiplied by the derivative of both the X-transformation and the Y-transformation, that is, by $\dfrac{1}{\sigma_X\sigma_Y}$ [see (109)].

Definition 11.5.1
Let X and Y be random variables with joint pdf

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \cdot \exp\left\{-\frac{1}{2}\left(\frac{1}{1-\rho^2}\right)\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\frac{x-\mu_X}{\sigma_X}\cdot\frac{y-\mu_Y}{\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right\}$$

for all x and y. Then X and Y are said to have the *bivariate normal distribution* (with parameters μ_X, σ_X^2, μ_Y, σ_Y^2, and ρ).

Comment For bivariate normal densities, $\rho(X,Y) = 0$ implies that X and Y are independent, a result not true in general.

PROPERTIES OF THE BIVARIATE NORMAL DISTRIBUTION

Francis Galton, the renowned British biologist and scientist, perhaps more than any other person was responsible for launching *regression analysis* as a worthwhile field of statistical inquiry. Galton was a redoubtable data analyst whose keen insight enabled him to intuit much of the basic mathematical structure that we now associate with correlation and regression.

One of his more famous endeavors (63) was an examination of the relationship between parents' heights (X) and their adult children's heights (Y). Those particular variables have a bivariate normal distribution, the mathematical properties of which Galton knew nothing. Just by looking at cross-tabulations of X and Y, though, Galton postulated that (1) the marginal distributions of X and Y are both normal, (2) $E(Y \mid x)$ is a linear function of x, and (3) $\text{Var}(Y \mid x)$ is constant with x. As Theorem 11.5.1 shows, all of his empirically based deductions proved to be true.

Theorem 11.5.1 *Suppose that X and Y are random variables having the bivariate normal distribution given in Definition 11.5.1. Then*

a. *$f_X(x)$ is a normal pdf with mean μ_X and variance σ_X^2; $f_Y(y)$ is a normal pdf with mean μ_Y and variance σ_Y^2.*

b. *ρ is the correlation coefficient between X and Y.*

c. $E(Y \mid x) = \mu_Y + \dfrac{\rho\sigma_Y}{\sigma_X}(x-\mu_X)$.

d. $\text{Var}(Y \mid x) = (1-\rho^2)\sigma_Y^2$.

Proof We have already established (a) and (b). Properties (c) and (d) will be examined for the special case $\mu_X = \mu_Y = 0$ and $\sigma_x = \sigma_y = 1$. The extension to arbitrary μ_X, μ_Y, σ_X, and σ_Y is straightforward.

First, note that

$$f_{Y|x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$
$$= \frac{\frac{1}{2\pi\sqrt{1-\rho^2}} e^{-(1/2)x^2} e^{-(1/2)[1/(1-\rho^2)](y-\rho x)^2}}{\frac{1}{\sqrt{2\pi}} e^{-(1/2)x^2}}$$
$$= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} e^{-(1/2)[1/(1-\rho^2)](y-\rho x)^2} \tag{11.5.4}$$

By inspection, we see that Equation 11.5.4 is the pdf of a normal random variable with mean ρx and variance $1-\rho^2$. Therefore, $E(Y \mid x) = \rho x$ and $\text{Var}(Y \mid x) = 1-\rho^2$. Replacing y with $(y-\mu_Y)/\sigma_Y$ and x with $(x-\mu_X)/\sigma_X$ gives the desired results.

Comment The term *regression line* derives from a consequence of part (c) of Theorem 11.5.1. Suppose we make the simplifying assumption that $\mu_X = \mu_Y = \mu$ and $\sigma_X = \sigma_Y$. Then part (c) reduces to

$$E(Y \mid x) - \mu = \rho(X,Y)(x-\mu)$$

But recall that $|\rho(X,Y)| \le 1$ and, in this case, $0 < \rho(X,Y) < 1$. Here, the positive sign of $\rho(X,Y)$ tells us that, on the average, tall parents have tall children. However, $\rho(X,Y) < 1$ means (again, *on the average*) that the children's heights are closer to the mean than are the parents'. Galton called this phenomenon "regression to mediocrity."

Questions

11.5.1. Suppose that X and Y have a bivariate normal pdf with $\mu_X = 3$, $\mu_Y = 6$, $\sigma_X^2 = 4$, $\sigma_Y^2 = 10$, and $\rho = \frac{1}{2}$. Find $P(5 < Y < 6\frac{1}{2})$ and $P(5 < Y < 6\frac{1}{2} \mid x = 2)$.

11.5.2. Suppose that X and Y have a bivariate normal distribution with $\text{Var}(X) = \text{Var}(Y)$.

(a) Show that X and $Y - \rho X$ are independent.

(b) Show that $X + Y$ and $X - Y$ are independent. [*Hint:* See Question 11.4.7(a).]

11.5.3. Suppose that X and Y have a bivariate normal distribution.

(a) Prove that $X + Y$ has a normal distribution when X and Y are standard normal random variables.

(b) Find $E(cX + dY)$ and $\text{Var}(cX + dY)$ in terms of μ_X, μ_Y, σ_X, σ_Y, and $\rho(X,Y)$, where X and Y are arbitrary normal random variables.

11.5.4. Suppose that the random variables X and Y have a bivariate normal pdf with $\mu_X = 56$, $\mu_Y = 11$, $\sigma_X^2 = 1.2$, $\sigma_Y^2 = 2.6$, and $\rho = 0.6$. Compute $P(10 < Y < 10.5 \mid x = 55)$. Suppose that $n = 4$ values were to be observed with x fixed at 55. Find $P(10.5 < \bar{Y} < 11 \mid x = 55)$.

11.5.5. If the joint pdf of the random variables X and Y is

$$f_{X,Y}(x,y) = ke^{-(2/3)[(1/4)x^2 - (1/2)xy + y^2]}$$

find $E(X)$, $E(Y)$, $\text{Var}(X)$, $\text{Var}(Y)$, $\rho(X,Y)$, and k.

11.5.6. Give conditions on $a > 0$, $b > 0$, and u so that

$$f_{X,Y}(x,y) = ke^{-(ax^2 - 2uxy + by^2)}$$

is the bivariate normal density of random variables X and Y each having expected value 0. Also, find $\text{Var}(X)$, $\text{Var}(Y)$, and $\rho(X,Y)$.

ESTIMATING PARAMETERS IN THE BIVARIATE NORMAL PDF

The five parameters in $f_{X,Y}(x,y)$ can be estimated in the usual way with the method of maximum likelihood. Given a random sample of size n from $f_{X,Y}(x,y)$—(x_1, y_1),

$(x_2, y_2), \ldots, (x_n, y_n)$—we define $L = \prod_{i=1}^{n} f_{X,Y}(x_i, y_i)$ and take the derivatives of $\ln L$ with respect to each of the parameters. Solved simultaneously, the resulting five equations (each derivative set equal to 0) yield the maximum likelihood estimators given in Theorem 11.5.2. Details of the derivation will be left as an exercise.

Theorem 11.5.2 *Given that $f_{X,Y}(x, y)$ is a bivariate normal pdf, the maximum likelihood estimators for μ_X, μ_Y, σ_X^2, σ_Y^2, and ρ, assuming that all five are unknown, are $\bar{X}$, $\bar{Y}$, $\left(\frac{1}{n}\right)\sum_{i=1}^{n}(X_i - \bar{X})^2$, $\left(\frac{1}{n}\right)\sum_{i=1}^{n}(Y_i - \bar{Y})^2$, and R, respectively.*

TESTING H_0: $\rho = 0$

If X and Y have a bivariate normal distribution, testing whether the two variables are independent is equivalent to testing whether their correlation coefficient, ρ, equals 0 (recall the Comment following Definition 11.5.1). Two different procedures are widely used for testing H_0: $\rho = 0$. One is an exact test based on the T_{n-2} random variable given in Theorem 11.5.3; the other is an approximate test based on the standard normal distribution.

Theorem 11.5.3 *Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be a random sample of size n drawn from a bivariate normal distribution, and let R be the sample correlation coefficient. Under the null hypothesis that $\rho = 0$, the statistic*

$$T_{n-2} = \frac{\sqrt{n-2}\,R}{\sqrt{1-R^2}}$$

has a Student t distribution with $n-2$ degrees of freedom.

Proof See (54).

Example 11.5.1 Table 11.5.1 gives the mean temperature for twenty successive days in April and the average daily butterfat content in the milk of ten cows (148). Can we conclude that temperature and butterfat content have a nonzero correlation?

Let ρ denote the true correlation coefficient between X and Y. The hypotheses to be tested are

$$H_0: \rho = 0$$
versus
$$H_1: \rho \neq 0$$

Let $\alpha = 0.05$. Given that $n = 20$, the statistic

$$t = \frac{\sqrt{n-2}\cdot r}{\sqrt{1-r^2}}$$

follows a Student t distribution with 18 df (if H_0: $\rho = 0$ is true). That being the case, the null hypothesis will be rejected if t is either (1) ≤ -2.1009 $(= -t_{0.025,18})$ or (2) $\geq +2.1009 (= t_{0.025,18})$.

Table 11.5.1

Date	Temperature, x	Percent Butterfat, y
April 3	64	4.65
4	65	4.58
5	65	4.67
6	64	4.60
7	61	4.83
8	55	4.55
9	39	5.14
10	41	4.71
11	46	4.69
12	59	4.65
13	56	4.36
14	56	4.82
15	62	4.65
16	37	4.66
17	37	4.95
18	45	4.60
19	57	4.68
20	58	4.65
21	60	4.60
22	55	4.46

For the data in Table 11.5.1,

$$\sum_{i=1}^{20} x_i = 1{,}082 \qquad \sum_{i=1}^{20} y_i = 93.5$$

$$\sum_{i=1}^{20} x_i^2 = 60{,}304 \qquad \sum_{i=1}^{20} y_i^2 = 437.6406$$

$$\sum_{i=1}^{20} x_i y_i = 5{,}044.5$$

so

$$r = \frac{20(5{,}044.5) - (1{,}082)(93.5)}{\sqrt{20(60{,}304) - (1{,}082)^2}\sqrt{20(437.6406) - (93.5)^2}}$$

$$= -0.453$$

Therefore,

$$t = \frac{\sqrt{n-2}\cdot r}{\sqrt{1-r^2}} = \frac{\sqrt{18}(-0.453)}{\sqrt{1-(-0.453)^2}} = -2.156$$

and our conclusion is *reject* H_0. It would appear that temperature and butterfat content are not independent.

Comment An alternate approach to testing H_0: $\rho = 0$ was given by Fisher (51). He showed that the statistic

$$\frac{1}{2}\ln\frac{1+R}{1-R}$$

is asymptotically normal with mean $\frac{1}{2}\ln[(1+\rho)/(1-\rho)]$ and variance approximately $1/(n-3)$. Fisher's formulation makes it relatively easy to determine the power of a correlation test—a computation that would be much more difficult if the inference had to be based on $\sqrt{n-2}\,R/\sqrt{1-R^2}$. ■

Questions

11.5.7. What would be the conclusion for the test of Example 11.5.1 if $\alpha = 0.01$?

11.5.8. In a study of heart disease (79), the weight (in pounds) and the blood cholesterol (in mg/dl) of fourteen men without a history of coronary incidents were recorded. At the $\alpha = 0.05$ level, can we conclude from these data that the two variables are independent?

Subject	Weight, x	Cholesterol, y
1	168	135
2	175	403
3	173	294
4	158	312
5	154	311
6	214	222
7	176	302
8	262	269
9	181	311
10	143	286
11	140	403
12	187	244
13	163	353
14	164	252

The data in the table give the following sums:

$$\sum_{i=1}^{14} x_i = 2{,}458 \qquad \sum_{i=1}^{14} y_i = 4{,}097$$

$$\sum_{i=1}^{14} x_i^2 = 444{,}118 \qquad \sum_{i=1}^{14} y_i^2 = 1{,}262{,}559$$

$$\sum_{i=1}^{14} x_i y_i = 710{,}499$$

11.5.9. Recall the baseball data in Question 11.4.11. Test whether home run frequency and home park altitude are independent. Let $\alpha = 0.05$.

11.5.10. Test H_0: $\rho = 0$ versus H_1: $\rho \neq 0$ for the SRE/SIRS data described in Question 11.4.13. Let 0.01 be the level of significance.

11.5.11. The National Collegiate Athletic Association has had a long-standing concern about the graduation rate of athletes. Under the urging of the Association, some prominent athletic programs increased the funds for tutoring athletes. The table below gives the amount spent (in millions of dollars) and the resulting percentage of athletes graduating in 2007. Does the money matter? Test H_0 : $\rho = 0$ versus H_1 : $\rho \neq 0$ at the 0.10 level of significance.

University	Money Spent on Athletes' Tutoring, x	Graduation Rate 2007, y
Minnesota	1.61	72
Kansas	1.61	70
Florida	1.67	87
LSU	1.74	69
Georgia	1.77	70
Tennessee	1.83	78
Kentucky	1.86	73
Ohio St.	1.89	78
Texas	1.90	72
Oklahoma	2.45	69

Based on data from: Pensacola News Journal (Florida), December 21, 2008.

11.6 Taking a Second Look at Statistics (How *Not* to Interpret the Sample Correlation Coefficient)

Of all the "numbers" that statisticians and experimenters routinely compute, the correlation coefficient is one of the most frequently *misinterpreted*. Two errors in particular are common. First, there is a tendency to assume, either implicitly or explicitly, that a high sample correlation coefficient implies causality. It does not. Even if the linear relationship between x and y is perfect—that is, even if $r = -1$ or $r = +1$—we cannot conclude that X *causes* Y (or that Y *causes* X). The sample correlation coefficient is simply a measure of the strength of a linear relationship. *Why* the xy-relationship exists in the first place is a different question altogether.

George Bernard Shaw (an unlikely contributor to a mathematics text!) described elegantly the fallacy of using statistical relationships to infer underlying causality. Commenting on the "correlations" that exist between lifestyle and health, he wrote in *The Doctor's Dilemma* (174):

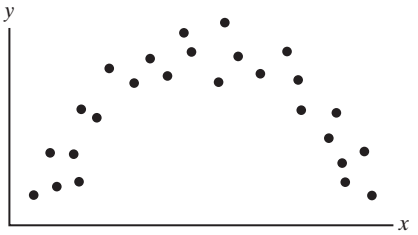
> It is easy to prove that the wearing of tall hats and the carrying of umbrellas enlarges the chest, prolongs life, and confers comparative immunity from disease; for the statistics show that the classes which use these articles are bigger, healthier, and live longer than the class which never dreams of possessing such things. It does not

> take much perspicacity to see that what really makes this difference is not the tall hat and the umbrella, but the wealth and nourishment of which they are evidence, and that a gold watch or membership of a club in Pall Mall might be proved in the same way to have the like sovereign virtues. A university degree, a daily bath, the owning of thirty pairs of trousers, a knowledge of Wagner's music, a pew in church, anything, in short, that implies more means and better nurture than the mass of laborers enjoy, can be statistically palmed off as a magic-spell conferring all sorts of privileges.

Examples of "spurious" correlations similar to those cited by Shaw are disturbingly commonplace. Between 1875 and 1920, for example, the correlation between the annual birthrate in Great Britain and the annual production of pig iron in the United States was an almost "perfect" −0.98. High correlations have also been found between salaries of Presbyterian ministers in Massachusetts and the price of rum in Havana and between the academic achievement of U.S. schoolchildren and the number of miles they live from the Canadian border. All too often, what looks like a cause is not a cause at all, but simply the effect of one or more factors that were not even measured. Researchers need to be very careful not to read more into the value of r than the number legitimately implies.

The second error frequently made when interpreting sample correlation coefficients is to forget that r measures the strength of a *linear* relationship. It says nothing about the strength of a *curvilinear* relationship. Computing r for the points shown in Figure 11.6.1, for example, is totally inappropriate. The (x_i, y_i) values in that scatterplot are clearly related but not in a linear way. Quoting the value of r would be misleading.

Figure 11.6.1



It is unfortunately true, though, that some xy-relationships having values of r close to either +1 or −1 are not as linear as their value of r might suggest. Recall the Social Security expenditures described in Case Study 11.2.2. The value of r for the 9 data points graphed in Figure 11.2.3 is 0.98, but the residual plot in Figure 11.2.4 makes it abundantly clear that the relationship is, in fact, curvilinear (as data for subsequent years confirmed).

The lesson to be learned from Figure 11.6.1 and Case Study 11.2.2 is clear—*always graph the data!* No correlation should ever, ever be calculated (much less interpreted) without first plotting the (x_i, y_i)'s to gain assurance that the relationship is linear. Digital cameras and photoshopping may have diminished the value of photographs as evidence in courts of law, but for statisticians, a picture is still worth a thousand words.

Appendix 11.A.1 A Proof of Theorem 11.3.3

The strategy for the proof is to express $n\hat{\sigma}^2$ in terms of the squares of normal random variables and then apply Fisher's Lemma (see Appendix 7.A.2). The random variables to be used are $\hat{\beta}_1 - \beta_1$, $W_i = Y_i - \beta_0 - \beta_1 x_i$, $i = 1, \ldots, n$, and $\bar{W} = \frac{1}{n}\sum_{i=1}^{n} W_i = \bar{Y} - \beta_0 - \beta_1 \bar{x}$. Note that

$$W_i - \bar{W} = (Y_i - \bar{Y}) - \beta_1 (x_i - \bar{x})$$

or, equivalently,

$$Y_i - \bar{Y} = (W_i - \bar{W}) + \beta_1(x_i - \bar{x})$$

Next, we express $\hat{\boldsymbol{\beta}}_1 - \beta_1$ as a linear combination of the W_i's. The argument begins by using Equation 11.3.1 to express $\hat{\boldsymbol{\beta}}_1$:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_1 - \beta_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} - \beta_1 \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}) - \beta_1 \sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})[(W_i - \bar{W}) + \beta_1(x_i - \bar{x})] - \beta_1 \sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(W_i - \bar{W})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (11.A.1.1)
\end{aligned}$$

Recall from Equation 11.3.3 that

$$n\hat{\boldsymbol{\sigma}}^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \hat{\boldsymbol{\beta}}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (11.A.1.2)$$

We need to express Equation 11.A.1.2 in terms of the W_i's—that is,

$$\begin{aligned}
n\hat{\boldsymbol{\sigma}}^2 &= \sum_{i=1}^{n}[(W_i - \bar{W}) + \beta_1(x_i - \bar{x})]^2 - \hat{\boldsymbol{\beta}}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \sum_{i=1}^{n}(W_i - \bar{W})^2 + 2\beta_1 \sum_{i=1}^{n}(x_i - \bar{x})(W_i - \bar{W}) + \beta_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&\quad - \hat{\boldsymbol{\beta}}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (11.A.1.3)
\end{aligned}$$

From Equation 11.A.1.1, we can write

$$\sum_{i=1}^{n}(x_i - \bar{x})(W_i - \bar{W}) = (\hat{\boldsymbol{\beta}}_1 - \beta_1)\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Substituting the right-hand side of the preceding expression for $\sum_{i=1}^{n}(x_i - \bar{x})(W_i - \bar{W})$ in Equation 11.A.1.3 gives

$$
\begin{aligned}
n\hat{\sigma}^2 &= \sum_{i=1}^{n}(W_i - \bar{W})^2 + 2\beta_1(\hat{\beta}_1 - \beta_1)\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&\quad + \beta_1^2\sum_{i=1}^{n}(x_i - \bar{x})^2 - \hat{\beta}_1^2\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \sum_{i=1}^{n}(W_i - \bar{W})^2 + \sum_{i=1}^{n}(x_i - \bar{x})^2\left[2\beta_1(\hat{\beta}_1 - \beta_1) + \beta_1^2 - \hat{\beta}_1^2\right] \\
&= \sum_{i=1}^{n}(W_i - \bar{W})^2 - \sum_{i=1}^{n}(x_i - \bar{x})^2\left[\hat{\beta}_1^2 - 2\hat{\beta}_1\beta_1 + \beta_1^2\right] \\
&= \sum_{i=1}^{n}(W_i - \bar{W})^2 - \sum_{i=1}^{n}(x_i - \bar{x})^2(\hat{\beta}_1 - \beta_1)^2 \\
&= \sum_{i=1}^{n}W_i^2 - n\bar{W}^2 - \sum_{i=1}^{n}(x_i - \bar{x})^2(\hat{\beta}_1 - \beta_1)^2
\end{aligned}
$$

Now, choose an orthogonal matrix, $\mathbf{M}$, whose first two rows are

$$\frac{x_1 - \bar{x}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \cdots \frac{x_n - \bar{x}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

and

$$\frac{1}{\sqrt{n}} \cdots \frac{1}{\sqrt{n}}$$

Define the random variables $Z_1, \ldots, Z_n$ through the transformation

$$\begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix} = \mathbf{M}\begin{pmatrix} W_1 \\ \vdots \\ W_n \end{pmatrix}$$

By Fisher's Lemma, the Z_i's are independent, normal random variables with mean zero and variance σ^2, and

$$\sum_{i=1}^{n}Z_i^2 = \sum_{i=1}^{n}W_i^2$$

Also, by Equation 11.A.1.1 and the choice of the first row of $\mathbf{M}$,

$$Z_1^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2(\hat{\beta}_1 - \beta_1)^2$$

and, by the selection of the second row of $\mathbf{M}$,

$$Z_2^2 = n\overline{W}^2$$

Thus,

$$n\hat{\sigma}^2 = \sum_{i=1}^{n}W_i^2 - Z_1^2 - Z_2^2 = \sum_{i=3}^{n}Z_i^2$$

From this follows the independence of $n\hat{\sigma}^2$, $\hat{\beta}_1$, and $\bar{Y}$.

Finally, notice that

$$\frac{n\hat{\sigma}^2}{\hat{\sigma}^2} = \sum_{i=3}^{n}\left(\frac{Z_i}{\sigma}\right)^2$$

The fact that the sum has a chi square distribution with $n - 2$ degrees of freedom proves the last part of the theorem.