CHAPTER 1

Introduction

The ability to predict the future would certainly be a highly desirable skill. Soothsayers have claimed to have this ability for centuries, but history has shown them to be quite fallible. No statistical technique can be used to eliminate or explain all of the uncertainty in the world, but statistics can be used to quantify that uncertainty. Unlike the soothsayer, the user of *regression analysis* can make predictions and place error bounds on those predictions. Regression can be used to predict the value of a *dependent variable* from knowledge of the values of one or more *independent variables*. The technical details of regression are given starting in Section 1.2.

The parents of a young man who has applied to a prestigious private university would certainly welcome information on what his grade-point average (GPA) might be after four years. If his predicted average was not high enough to enable him to graduate, his parents would obviously be more willing to pay his tuition at a different school at which he would be expected to graduate.

Similarly, it would be nice to know how long a certain make of automobile would be expected to last before needing major repairs, or to know the expected yield of a manufacturing process given a certain combination of inputs, or to know what a company's sales should be for a given set of expenses.

We see predictions being made all around us. Colleges and universities do use regression analysis to predict what a student's four-year college GPA would be if the student were admitted, and then make a decision regarding admittance from that predicted value and from other information. There is a vast amount of literature on this subject, including Paolillo (1982), Graham (1991), and Wright and Palmer (1994, 1997, 1999).

Regression analysis is one of the two most widely used statistical techniques (analysis of variance is the other), and it is used in almost every field of application. The following sample of titles of relatively recent research papers provides some evidence of the breadth of range of possible applications:

Modern Regression Methods, Second Edition By Thomas P. Ryan

Regression analysis as an aid in making oboe reeds (Ceasar-Spall and Spall, 1997).

Using segmented *regression* models to fit soil nutrient and soybean grain yield changes due to liming (Shuai, Zhou, and Yost, 2003).

Human animation using nonparametric *regression* (Faraway, 2004).

Tiered polychotomous *regression*: Ranking NFL quarterbacks (White and Berry, 2002).

Semiparametric *regression* for periodic longitudinal hormone data from multiple menstrual cycles (Zhang, Lin, and Sowers, 2000).

Bayesian variable selection in logistic *regression*: Predicting company earnings direction (Gerlach, Bird, and Hall, 2002).

Calibration using a piecewise simple linear *regression* model (Ndlovu and Preater, 2001).

and an especially contemporary application paper

How to make millions on eBay, or using multiple *regression* to estimate prices (Robinson, Kamischke, and Tabor, 2005).

The first of these papers describes an interesting and seemingly novel use of regression. As stated in the abstract, professional oboe players make their own reeds, which is a time-consuming process. Regression was used to help predict the ultimate quality of a finished reed based on data available at the initial tryout, with the inputs being several different characteristics of the cane used in making the reeds. Work would cease on a reed when the prediction quality was acceptable. It was found that the predicted quality was close to the actual quality and thus the overall effort was successful.

It is worth noting that the strength of a regression model will vary over the various application fields, which raises the question of whether something other than regression should be used in fields where strong models generally cannot be developed from data. In a very interesting paper, Dana and Dawes (2004) concluded that regression will generally not be the best approach in social science applications. What they recommend is discussed in Chapter 4.

A complete listing and description of all of the research papers in which regression analysis has been applied in analyzing data from subject-matter fields would require a separate book. It has even been claimed that lifespan can be predicted using regression, and we examine this issue in Section 2.4.7.

In a nonstatistical context, the word *regression* means "to return to an earlier place or state," and reversion is listed in dictionaries as a synonym. We might then wonder how regression can be used to predict the future when the word literally means to go backward in some sense.

Regression analysis can be traced to Sir Francis Galton (1822–1911), who observed that children's heights tended to "revert" to the average height of the population rather than diverting from it. In other words, the future generations of

offspring who are taller than average are not progressively taller than their respective parents, and parents who are shorter than average do not beget successively smaller children. (This has been called "regression to the mean" or *regression toward the mean*, which Galton (1886) termed *regression towards mediocrity* in his famous paper: "Regression towards mediocrity in hereditary stature." See also Bland and Altman (1994).)

Galton originally used the word *reversion* to describe this tendency, and some years later used the word *regression* instead. This early use of the word regression in data analysis is unrelated, however, to what has become known as regression analysis. (For additional information on the invention of regression analysis, as well as some aids to understanding simple linear regression, see Stanton (2001). See also Stigler (1997) and Finney (1996).)

The user of regression analysis attempts to discern the relationship between a dependent variable and one or more independent variables. That relationship will not be a functional relationship, however, nor can a cause-and-effect relationship necessarily be inferred. (Regression can be used when there is a cause-and-effect relationship, however.) The equation $\mathrm{F} = \frac{9}{5}\mathrm{C} + 32$ expresses temperature in Fahrenheit as a function of temperature measured on the Celsius scale. This represents an exact functional relationship; one in which temperature in degrees Celsius could just as easily have been expressed as a function of temperature in degrees Fahrenheit. Thus there is no clear choice for the dependent variable.

Exact relationships do not exist in regression analysis, and in regression the variable that should be designated as the dependent variable is usually readily apparent. As a simple example, let's assume that we want to predict college GPA using high school GPA. Obviously, college GPA should be related to high school GPA and should depend on high school GPA to some extent. Thus, college GPA would logically be the dependent variable and high school GPA the independent variable.

Throughout this book the independent variables will be referred to as regressors, predictors, or regression variables, and the dependent variable will occasionally be referred to as the response variable.

1.1 SIMPLE LINEAR REGRESSION MODEL

The word *simple* means that there is a single independent variable, but the word *linear* does not have the meaning that would seem to be self-evident. Specifically, it does not mean that the relationship between the two variables can be displayed graphically as a straight line. Rather, it means that the model is linear in the parameters.

The basic model is

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1.1}$$

in which Y and X denote the dependent and independent variable, respectively, and β_0 and β_1 are parameters that must be estimated. The symbol ϵ represents the error term. This does not mean that a mistake is being made; it is simply a symbol used to indicate the absence of an exact relationship between X and Y.

The reader will recognize that, except for ϵ, Eq. (1.1) is in the general form of the equation for a straight line. That is, β_1 is the slope and β_0 is the Y-intercept.

1.2 USES OF REGRESSION MODELS

Once β_0 and β_1 have been estimated (estimation is covered in Section 1.4), the following *prediction equation* results:

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X \tag{1.2}$$

The "hats" (as they are called) above β_0 and β_1 signify that those parameters are being estimated, but the hat above Y means that the dependent variable is being predicted. (The reader will observe that there is no error term in Eq. (1.2), as we do not estimate error terms explicitly in regression models.) The equation would be used for values of X within the range of the sample data, or perhaps only slightly outside the range.

The prediction equation implies that prediction is one of the uses of regression analysis. Simply stated, the objective is to see if past data indicate a strong enough relationship between X and Y to enable future values of Y to be well predicted. The rationale is that if past values of Y can be closely fit by the prediction equation, then future values should similarly be closely predicted. As mentioned previously, a prediction equation can be used to predict what a student's college GPA would be after four years if he or she were admitted to a particular college or university.

Regression can also be used for the related purposes of estimation and description. Specifically, once $\widehat{\beta}_0$ and $\widehat{\beta}_1$ have been obtained, these parameter estimates can be used to describe the relationship between Y and X. For example, $\widehat{\beta}_1$ is the amount, positive or negative, by which we would predict Y to change per unit change in X, *for the range of* X in the sample that was used to determine the prediction equation. (*Note*: Whereas such an interpretation is possible when the prediction equation contains a single X, it will often not be possible to do so when there are multiple Xs, as is discussed in Section 4.3.) Similarly, $\widehat{\beta}_0$ is the value that we would predict for Y when $X = 0$.

A seldom mentioned but important use of regression analysis is for control. For example, Y might represent a measure of the pollutant contamination of a river, with the regressor(s) representing the (controllable) input pollutant(s) being varied (e.g., reduced) in such a way as to control Y at a tolerable level. The use of regression for control purposes is discussed by Hahn (1974) and Draper and Smith (1998) and is discussed in Section 1.9.1.

Table 1.1 Four Data Sets Given by Anscombe (1973) which Illustrate the Need to Plot Regression Data

Data Set:	1–3	1	2	3	4	4
Variable:	*X*	*Y*	*Y*	*Y*	*X*	*Y*
	10	8.04	9.14	7.46	8	6.58
	8	6.95	8.14	6.77	8	5.76
	13	7.58	8.74	12.74	8	7.71
	9	8.81	8.77	7.11	8	8.84
	11	8.33	9.26	7.81	8	8.47
	14	9.96	8.10	8.84	8	7.04
	6	7.24	6.13	6.08	8	5.25
	4	4.26	3.10	5.39	19	12.50
	12	10.84	9.13	8.15	8	5.56
	7	4.82	7.26	6.42	8	7.91
	5	5.68	4.74	5.73	8	6.89

1.3 GRAPH THE DATA!

The use of the model given in Eq. (1.1) assumes that we have already selected the model. As George E.P. Box has often stated, "all models are wrong, but some are useful." Any statistical analysis should begin with graphic displays of the data, and these displays can help us select a useful model.

The importance of graphing regression data is perhaps best illustrated by Anscombe (1973), who showed that completely different graphs can correspond to the same regression equation and summary statistics. In particular, Anscombe showed that the graph of a data set that clearly shows the need for a quadratic term in X can correspond to the same regression equation as a graph that clearly indicates that a linear term is sufficient. The four data sets are given in Table 1.1, and the reader is asked to compare the four (almost identical) regression equations against the corresponding scatter plots in Exercise 1.1.

■ EXAMPLE 1.1

Consider the data given in Table 1.2. Assume that the data have come from an industrial experiment in which temperature is varied from 375 to 420 °F in 5-degree increments, with the temperatures run in random order and the process yield recorded for each temperature setting. (It is assumed that all other factors are being held constant.)

The scatter plot for these data is given in Figure 1.1. The plot shows a strong linear relationship between X and Y, although there is some hint of curvature at the extreme values of X. Therefore, the model given by Eq. (1.1) is a good starting point, possibly to be modified later.

Table 1.2 Process Yield Data

Y (Yield)	X (°F)
26.15	400
28.45	410
25.20	395
29.30	415
24.35	390
23.10	385
27.40	405
29.60	420
22.05	380
21.30	375

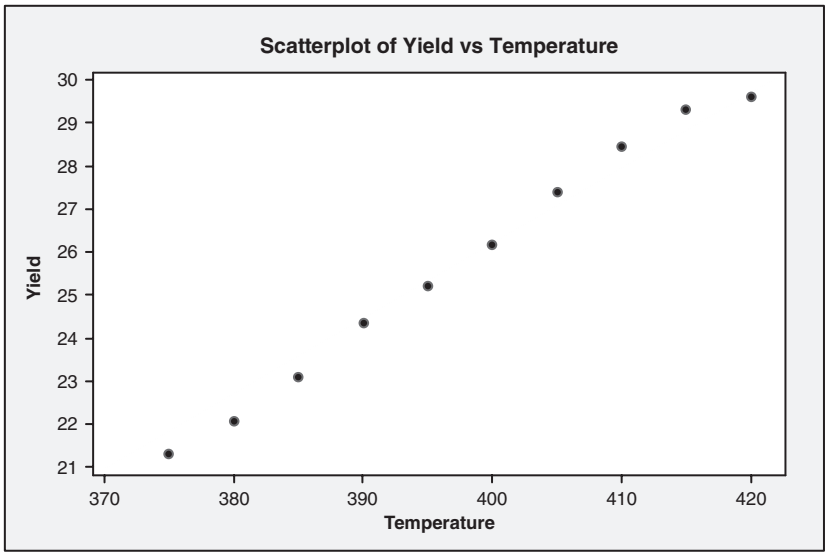

Figure 1.1 Scatter plot of yield versus temperature for the data in Table 1.2. ■

1.4 ESTIMATION OF β_0 AND β_1

Point estimates of β_0 and β_1 are needed to obtain the prediction equation given in Eq. (1.2). A crude approach would be to draw a line through the center of the points and then use the slope and Y-intercept of the line as the estimates of β_1 and β_0, respectively.

Before one could even attempt to do so, however, it would be necessary to define what is meant by "center." We could attempt to minimize the sum of the slant distances (with each distance measured from the point to the line), but since we will be using X to predict Y, it would make more sense to try to minimize the sum of the vertical distances. If we use the signs of those distances we have a

problem, however, because then the line is not uniquely defined (as the reader is asked to show in Exercise 1.2).

One way to eliminate this problem would be to minimize the sum of the absolute values of the distances (see, e.g., Birkes and Dodge (1993)). This has been suggested as a way to deal with extreme X-values and is discussed briefly in Chapter 11.

The standard approach, however, is to minimize the sum of the squares of the vertical distances, and this is accomplished by using the *method of least squares*. (The following is a presentation of ordinary listed squares (OLS), with the first word used to distinguish the method from other least squares methods, such as weighted least squares, which is discussed in Section 2.1.3.1.) For the purpose of illustration we must assume that Eq. (1.1) is the correct model, although as stated previously the correct model will generally be unknown.

The starting point is to write the model as

$$\epsilon = Y - (\beta_0 + \beta_1 X) \tag{1.3}$$

Since ϵ represents the vertical distance from the observed value of Y to the line represented by $Y = \beta_0 + \beta_1 X$ that we would have if β_0 and β_1 were known, we want to minimize $\sum \epsilon^2$.

For convenience we define L as

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \tag{1.4}$$

where n denotes the number of data points in a sample that has been obtained. To minimize L we take the partial derivative of L with respect to each of the two parameters that we are estimating and set the resulting expressions equal to zero. Thus,

$$\frac{\partial L}{\partial \beta_0} = 2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)(-1) = 0 \tag{1.5a}$$

and

$$\frac{\partial L}{\partial \beta_1} = 2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)(-X_i) = 0 \tag{1.5b}$$

Dropping the 2 and the -1 from Eqs. (1.5a) and (1.5b), the solutions for β_0 and β_1 would be obtained by solving the equations (which are generally called

normal equations):

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) = 0$$

and

$$\sum_{i=1}^{n} (X_i Y_i - X_i \beta_0 - \beta_1 X_i^2) = 0$$

which become

$$n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i$$

and

$$\beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i$$

The solution of these two equations (for the estimators of β_0 and β_1) produces the least squares estimators

$$\widehat{\beta}_1 = \frac{\sum X_i Y_i - \left(\sum X_i\right)\left(\sum Y_i\right)/n}{\sum X_i^2 - \left(\sum X_i\right)^2/n} \tag{1.6a}$$

and

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X} \tag{1.6b}$$

(The astute reader will recognize that L in Eq. (1.4) is not automatically minimized just by setting the first derivatives equal to zero and solving the resultant equations. It can be shown, however, that the determinant of the matrix of second-order partial derivatives of L is positive, thus ensuring that a minimum, rather than a maximum, has been attained. Matrix algebra for regression is covered in Chapter 3.)

Notice that in Eq. (1.6) we have dropped the beginning and ending points of the summation and have simply used $\sum$. We shall do so frequently in this and subsequent chapters, as all summations hereinafter will be assumed to start with 1 and end with n, unless indicated otherwise. We will also drop the subscript i in most summation expressions.

If computations were performed by hand calculator, $\widehat{\beta}_1$ would obviously have to be computed before $\widehat{\beta}_0$, since the latter is obtained using the former.

It is advantageous (and space saving, in particular) to use shorthand notation to represent the right side of Eq. (1.6). The symbol S_{xy} can be used to represent the numerator, and S_{xx} to represent the denominator. The S can be thought of as representing "sum," while xx represents X times X, which, of course, equals X^2. In statistical jargon, S_{xx} represents the "corrected" sum of squares for X, and S_{xy} denotes the corrected sum of products of X and Y. That is, $S_{xx} \neq \sum X^2$ and $S_{xy} \neq \sum XY$ Rather, $S_{xx} = \sum (X - \overline{X})^2$ and $S_{xy} = \sum (X - \overline{X})(Y - \overline{Y})$, where these two expressions can be shown to be equal to the denominator and numerator, respectively, of Eq. (1.6). Thus, the "correction" factors are $\overline{X}$ and $\overline{Y}$. Also, $S_{yy} = \sum (Y - \overline{Y})^2$.

When the calculations are performed using the data in Table 1.2, the following results are obtained:

$$\begin{aligned}\widehat{\beta}_1 &= \frac{102{,}523.5 - (3975)(256.9)/10}{1{,}582{,}125 - (3975)^2/10} \\ &= \frac{405.75}{2062.5} \\ &= 0.19673\end{aligned}$$

and

$$\begin{aligned}\widehat{\beta}_0 &= 25.69 - 0.19673(397.5) \\ &= -52.51\end{aligned}$$

The regression equation is thus $\widehat{Y} = -52.51 + 0.19673X$.

■ EXAMPLE 1.2

Background

An interesting and important application of simple linear regression was described by Current (1998). It is a frequent practice in anesthesiology and critical care medicine to estimate the body surface area (BSA) of infants and children, in particular. This is necessary for determining proper drug dosages and for other reasons. The most accurate formula for accomplishing this is $\text{BSA} = \text{Weight}^{(0.7285 - 0.0188 \log(\text{weight}))} \times \text{Height}^{(0.3)} \times 0.0003207$, as given by Boyd (1935), which was determined using data from the literature on 231 subjects, with the "known" BSA values determined using triangulation, surface integration, and various coating methods.

Boyd's formula is obviously not a simple expression, however, and requires the use of a calculator. Consequently, Current (1998) sought something simpler and

used data for 112 patients ranging in weight from 3 to 30 kilograms given in the literature and reported by Boyd (1935). Current (1998) showed a graph of BSA (using Boyd's formula) plotted against Weight (in grams), with the simple linear regression line displayed on the graph (see `http://www.ispub.com/ostia/index.php?xmlFilePath=journals/ija/vol2n2/bsa.xml`). The fit is obviously very good at low weights but breaks down at high weights.

Regression Equation

The regression equation is BSA $= 1321 + 0.3433$Weight, and the R^2 value (see Section 1.5.2) was .9715. Remember, however, that BSA is from Boyd's formula; it is not the "known" BSA. Virtually all formulas used for any type of estimation will break down somewhere, so the question must be asked: Is the disagreement between the fitted BSA and Boyd's BSA at higher weights due to Boyd's formula not working well at those weights, or is it due to the simple linear regression model not being very applicable at high weights, since there should obviously be greater variability in BSA values for high weights than for low weights? Current (1998) stated that the regression equation should be used because of the high correlation (.9857) between the fitted values that it produces and the values obtained using Boyd's formula. (Correlation is discussed in Section 1.8.) High correlation by itself is not necessarily sufficient, however, because two sets of values can differ by a large constant and still be perfectly correlated. It is also worth noting that Current (1998) states that the regression equation applies only to well-proportioned infants and children who range from 3 to 30 kilograms and, furthermore, should be used only in noncritical applications, with Boyd's formula used in critical applications. Current (an M.D.) stated that he has found the regression equation to be quite useful in operating room management of infants and children for the application of pump flows and fresh gas flows and believes it to be adequate for most noncritical cases. We should also note that although the regression equation is simpler than Boyd's formula, medical personnel will still probably need to use a calculator and would also probably need a calculator if a simplified version, given by Current (1998), of the regression equation is used.

This is an interesting application of simple linear regression because a less complicated, easier to use model was sought. In essence, this is what linear regression is all about because the true (unknown) model in practically any application will be unknown, as was emphasized in Section 1.3 with the quote from George Box. It is also worth noting that Current (1998) emphasized that not only should the regression equation be used only for the weight range of 3–30 kilograms, but that the child must be "well proportioned." Thus, this is somewhat of a multidimensional problem relative to the subjects, even though the regression equation contains only a single predictor. So extrapolation would occur if the model were applied to a child who was not well proportioned. ■

1.4.1 Orthogonal Regression

It is reasonable to ask why $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are not obtained by minimizing the sum of the squares of the perpendicular (slant) distances from the points to the regression line rather than the sum of the squares of the vertical distances, especially since those distances are shorter. Indeed, this is sometimes done and that is called *orthogonal regression*. (For more technical details and a computer program for orthogonal regression, see, for example, `http://www.nlreg.com/orthogonal.htm`.)

If the regressor values are fixed, vertical distances seem to be most appropriate. Nevertheless, orthogonal regression, which corrects for measurement error in the predictors, has been found to be useful in various applications, but as Carroll and Ruppert (1996) pointed out, it is frequently misused as equation error is not considered.

Orthogonal regression assumes that there is an exact relationship between X and Y (i.e., no equation error), with measurement error, assumed for both X and Y, preventing the points from plotting on a straight line. It is often used when two methods measure the same quantity, or when Y is related to X through a physical law. We will not pursue further discussion of orthogonal regression. The interested reader is advised to read the corresponding material in Fuller (1987) and is especially advised to read Carroll and Ruppert (1996) carefully.

1.5 INFERENCES FROM REGRESSION EQUATIONS

What can we infer from the regression equation for the Table 1.2 data? Since the data have come from a designed experiment, we can state that, *within the range of temperatures used in the experiment*, process yield should increase by approximately 0.2 ($\times 100$) units per single degree increase in temperature. This assumes that all controllable factors that might influence process yield have been held constant during the experiment, and that the randomization of temperatures has prevented any "lurking variables" from undermining the inferences made from the data. (Lurking variables are factors that can have a systematic effect on the response variable; see Section 5.7 for further discussion of lurking variables.) See Chapter 14 for a discussion of inferences that can be made from controlled experiments versus the conclusions that can be drawn from observational studies. The latter refers to obtaining data without any intervention on the part of the data collector, such as taking a sample of data from a college registrar's records. From Table 1.2 we can see that if the temperatures are ordered from 375 °F to 420 °F, the average change in process yield for each degree increase in temperature is 0.184, which differs only slightly from $\widehat{\beta}_1$ (as we might expect).

In a study of this type, the primary benefit to be derived from this experiment would be a better understanding of how temperature affects process yield, and to gain a better understanding of what might seem to be the optimal temperature setting, as is done in *Evolutionary Operation* (see, e.g., Box and Draper (1969)

or Ryan (2000)). Notice, however, that this cannot be determined from the data in Table 1.2, as the yield is strictly increasing as temperature increases. It would appear though that the optimal temperature setting may be just slightly greater than 420 °F, as yield does level off somewhat as temperature is increased from 415 to 420 °F.

1.5.1 Predicting *Y*

To illustrate the other uses of regression analysis and the inferences that we can draw from such an analysis, let's assume that we want to use the regression equation to predict Y. That is, we are interested in predicting process yield for different temperature settings.

Let's assume that we would like to be able to predict what yield should be when the temperature is 400 °F. A very simple approach would be to use 26.2 as the predicted value since that was the process yield that resulted when temperature was set at 400 °F in the experiment. To do so, however, would be to ignore the question "Is there a relationship between X and Y?" and if a relationship does exist, to not utilize the extent of that relationship.

Consider this. Assume that a scatter plot of Y versus X resulted in a random configuration of points such that the line represented by the regression equation would have a slope of zero (which would be a horizontal line). It would then be foolish to use any particular value of Y in the data set as the predicted value of Y for the corresponding value of X, because the regression equation would be telling us that there is no relationship between X and Y, and that we might as well use the average value of Y in the data set in predicting Y. (Note that this is what happens algebraically when $\widehat{\beta}_1 = 0$, since $\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1\overline{X}$ and $\widehat{Y}$ then equals $\widehat{\beta}_0$, which in this case equals $\overline{Y}$.)

Thus, there is no point in trying to predict Y from X unless there is evidence of some type of a relationship (linear or nonlinear) between the two, and the apparent strength of the relationship suggests that Y will be well predicted from X.

How do we discern the strength of the relationship between X and Y? Figure 1.1 shows that there is a strong linear relationship between X and Y, and when this is the case, the observed values in Table 1.2 should be very close to the fitted values. In particular, when $X = 400$, $\widehat{Y} = -52.51 + 0.19673(400) = 26.18$, which hardly differs from $Y = 26.15$ in Table 1.2. (We shall use the term *fitted values* to denote the $\widehat{Y}$ values for the sample and the term *predicted value* to represent the prediction of a future (i.e., unobservable) value of Y.)

We would expect the other fitted values to be close to the observed values, and Table 1.3 shows that this is true. We can see that all of the fitted values are close to the observed values, with the largest difference occurring at $X = 420$. The latter is to be expected since the difference in the Y values for $X = 415$ and $X = 420$ is noticeably smaller than any of the other differences for successive X values.

Table 1.3 Y and $\widehat{Y}$ Values for Table 1.1 Data

X	Y	$\widehat{Y}$	$Y - \widehat{Y}$	$(Y - \widehat{Y})^2$
400	26.15	26.18	−0.03	0.0009
410	28.45	28.15	0.30	0.0900
395	25.20	25.20	0.00	0.0000
415	29.30	29.13	0.17	0.0289
390	24.35	24.21	0.14	0.0196
385	23.10	23.23	−0.13	0.0169
405	27.40	27.17	0.23	0.0529
420	29.60	30.12	−0.52	0.2704
380	22.05	22.25	−0.20	0.0400
375	21.30	21.26	0.04	0.0016
			0.00	0.5212

(It should be noted that the values for $(Y - \widehat{Y})^2$ differ slightly from the values that would be obtained if a computer (or more decimal places) were used, as the given values are computed directly from the two-decimal-place values for $Y - \widehat{Y}$. This also means that subsequent calculations in this chapter that utilize $\sum (Y - \widehat{Y})^2$ will also differ slightly from the results produced by a computer.)

It can also be observed that $\sum (Y - \widehat{Y}) = 0$. This is due to Eq. (1.5a), since that equation can be written, with the parameters replaced by their respective estimators, as $\sum (Y - (\widehat{\beta}_0 + \widehat{\beta}_1)) = 0 = \sum (Y - \widehat{Y})$. Thus, since the second of the two equations that were solved is not involved, having $\sum (Y - \widehat{Y}) = 0$ does not ensure that the correct values for $\widehat{\beta}_0$ and $\widehat{\beta}_1$ were obtained, as the reader is asked to investigate in Exercise 1.2.

For the correct $\widehat{\beta}_0$ and $\widehat{\beta}_1$ values, $\sum (Y - \widehat{Y})^2$ is minimized. This is because the method of least squares was used with the initial intent being to minimize $\sum \epsilon^2 = \sum (Y - \beta_0 - \beta_1 X)^2$. We cannot observe the values for ϵ, however, since β_0 and β_1 are unknown. What we do observe are the e values, where $e = Y - \widehat{Y}$ is called a *residual*, and $\sum e^2 = \sum (Y - \widehat{\beta}_0 - \widehat{\beta}_1 X)^2$ is minimized for a given set of data. (Obviously, e is defined only when $\widehat{Y}$ denotes a fitted value rather than a predicted value yet to be observed.)

1.5.2 Worth of the Regression Equation

For this example it is clear that the regression equation has value for prediction (or for control or descriptive purposes).

A scatter plot will not always show as obvious a linear relationship as was seen in Figure 1.1, however, and the size of the $Y - \widehat{Y}$ values will generally depend on the magnitude of Y. Consequently, it would be helpful to have a numerical measure that expresses the strength of the linear relationship between X and Y.

A measure of the variability in Y is $\sum (Y - \overline{Y})^2$. Since

$$\sum (Y - \overline{Y})^2 = \sum (Y - \widehat{Y})^2 + \sum (\widehat{Y} - \overline{Y})^2 \tag{1.7}$$

as is shown in the chapter Appendix, it would be logical to use either $\sum (Y - \widehat{Y})^2$ or $\sum (\widehat{Y} - \overline{Y})^2$ as a measure of the worth of the prediction equation, and divide the one that is used by $\sum (Y - \overline{Y})^2$ so as to produce a unit-free number.

It is a question of whether we want the measure to be large or small. If we define

$$R^2 = \frac{\sum (\widehat{Y} - \overline{Y})^2}{\sum (Y - \overline{Y})^2} \tag{1.8}$$

then R^2 represents the percentage of the variability in Y (as represented by $\sum (Y - \overline{Y})^2$) that is explained by using X to predict Y. Note that if $\widehat{\beta}_1 = 0$, then $\widehat{Y} = \overline{Y}$ for every value of X and $R^2 = 0$. At the other extreme, R^2 would equal 1 if $\widehat{Y} = Y$ for each value of Y in the data set. Thus, $0 \leq R^2 \leq 1$ and we would want R^2 to be as close to 1 as possible.

From the data in Table 1.3,

$$\begin{aligned} R^2 &= \frac{\sum (\widehat{Y} - \overline{Y})^2}{\sum (Y - \overline{Y})^2} \\ &= 1 - \frac{\sum (Y - \widehat{Y})^2}{S_{yy}} \quad \text{(from Eq. (1.7))} \\ &= 1 - \frac{0.5212}{80.3390} \\ &= .9935 \end{aligned} \tag{1.9}$$

Thus, R^2 is very close to 1 in this example.

How large must R^2 be for the regression equation to be useful? That depends on the area of application. If we could develop a regression equation to predict the stock market (which unfortunately we cannot), we would be ecstatic if $R^2 = .50$. On the other hand, if we were predicting college GPA, we would want the prediction equation to have strong predictive ability, since the consequences of a poor prediction could be quite serious.

Although R^2 is a well-accepted measure, a few criticisms can be made both for one X and for more than one X. For example, with a single X the slope of the plotted points will affect R^2, as is discussed by Barrett (1974). The argument is as follows. Consider the expression for R^2 given in Eq. (1.9). If we let the closeness of $\widehat{Y}_i$ to Y_i be fixed so that $\sum (Y_i - \widehat{Y}_i)^2$ remains constant, but then rotate the configuration of points so that the slope of the regression line is increased, R^2

must increase because $\sum (Y - \overline{Y})^2$ will increase. Under the assumption that the values of X are preselected, Ranney and Thigpen (1981) showed that the expected value of R^2 can also be increased by increasing the range of X. The value of R^2 may also be artificially large if the sample size is small relative to the number of parameters, and as pointed out by Draper and Smith (1998), when there are repeated X-values, it is the number of *distinct* X-values relative to the number of parameters that is important, not the sample size. Another disturbing feature of R^2 is that we can expect it to increase for each variable that is added to a known (true) model. (This result is discussed in the chapter Appendix.)

Despite these shortcomings, R^2 has value as a rough indicator of the worth of a regression model. Another form of R^2, R^2_{adjusted}, is sometimes used. Since

$$R^2_{\text{adjusted}} = 1 - \frac{SSE_p/(n-p)}{SST/(n-1)}$$

with p denoting the number of parameters in the model, R^2_{adjusted} can decrease as p is increased if $n - p$ declines at a faster rate than SSE_p. Thus, R^2_{adjusted} might be used to determine the "point of diminishing returns." R^2_{adjusted} is discussed further in Section 7.4.1. (Note that R^2_{adjusted} would be negative if SSE_p is close to SST but that is of no practical concern since this isn't likely to occur with a fitted model.)

1.5.3 Regression Assumptions

What has been presented to this point in the chapter has hardly involved statistics. Rather, calculus was used to minimize a function, and the estimators that resulted from this minimization, $\widehat{\beta}_0$ and $\widehat{\beta}_1$, were used to form the prediction equation. Thus, there are very few assumptions that need to be made at this point.

One assumption that obviously must be made is that the model given in Eq. (1.1) is a suitable proxy for the "correct" (but unknown) model. We also need to assume that the variance of ϵ_i (i.e., $\text{Var}(\epsilon_i)$) is the same for each value of i ($i = 1, 2, \ldots, n$). If this requirement is not met, then *weighted least squares* (see Section 2.1.3.1) should be used.

A typical analysis of regression data entails much more than the development of a prediction equation, however, and additional inferences require additional assumptions. These assumptions are stated in terms of the error term, ϵ, in the regression model. One very important assumption is that the error terms are uncorrelated. Specifically, any pair of errors (ϵ_i, ϵ_j) should be uncorrelated, and the errors must also be uncorrelated with X. This means that if we knew the value of ϵ_i, that value would not tell us anything about the value of ϵ_j. Similarly, the value of ϵ_i should not depend on the value of X_i.

The assumption of uncorrelated errors is frequently violated when data are collected over time, and the consequences can be serious (see Chapter 2). Another important assumption is that ϵ should have approximately a normal distribution.

The method of least squares should not be used when the distribution of ϵ is markedly nonnormal. For example, positive and negative residuals with equal absolute values should clearly not be weighted the same if the distribution of the errors is not symmetric. As discussed by Rousseeuw and Leroy (1987, p. 2) and others, Gauss introduced the normal distribution as the optimal error distribution for least squares, so using least squares when the error distribution is known (or believed) to be nonnormal is inappropriate. Although some theoretical statisticians would argue that asymptotic results support the use of least squares for nonnormal error distributions when certain conditions are met, such a stance will often lead to very bad results. It is important to realize that the methods of Chapter 11 (for nonnormal error distributions and other problems) will often be needed. With any statistical analysis (using regression or some other technique), it is a good idea to analyze the data first assuming that the assumptions are met, and then analyze the data not assuming that the assumptions are met. If the results differ considerably, then the results of the second analysis will generally be the more reliable.

This last point cannot be overemphasized, as the confidence intervals, prediction interval, and hypothesis tests that are presented in subsequent sections are sensitive (i.e., not robust) to more than a slight-to-moderate departure from normality (see, e.g., Rousseeuw and Leroy (1987, p. 41) or Hamilton (1992, p. 113)), with the prediction interval being the most sensitive. Nonnormality, bad data, and good data that are far removed from the rest of the data can create serious problems. Consequently, the regression user and serious student of regression are urged to study Chapter 11 carefully. Methods for checking the assumptions of normality, independence, and a constant error variance are given in Chapter 2.

Another assumption on ϵ is that the mean of ϵ is zero. This assumption is never checked; it simply states that the "true" regression line goes through the center of a set of data. These assumptions on ϵ can be represented by $\epsilon \sim NID(0, \sigma^2)$, with the additional assumption of normality meaning that the ϵ_i, ϵ_j are not only uncorrelated but are also independent (i.e., normal and independent; *NID*).

Another assumption that is necessary for the theory that immediately follows, but is not necessary for regression analysis in general, is for the values of X to be selected by the experimenter rather than being allowed to occur at random. (The use of regression analysis when X is random is discussed in Section 1.9.)

When X is fixed, the assumptions on ϵ translate into similar assumptions on Y. This is because $\beta_0 + \beta_1 X$ is then an (unknown) constant, and adding a constant to ϵ causes $Y = \beta_0 + \beta_1 X + \epsilon$ to have the same distribution and variance as ϵ. Only the mean is different. Specifically, for a given value of X, $Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$, and the assumption of uncorrelated errors means that the Y_i, Y_j are also uncorrelated. As with the errors, the assumption of normality allows us to state further that the Y_i, Y_j are also *independent*. These assumptions of normality and independence for Y are necessary for the inferential procedures

that are presented in subsequent sections. (In subsequent sections and chapters we will write $\text{Var}(Y)$ to represent the conditional variance, $\text{Var}(Y|X)$, as the variance of Y is assumed to be the same for each value of X.)

1.5.4 Inferences on β_1

Confidence intervals and hypothesis tests are covered in introductory statistics courses, and the relationship between them is often discussed. Although the information provided by a hypothesis test is also provided by a confidence interval, the confidence interval provides additional information. Obviously, it would be helpful to have a confidence interval on the true rate of change of Y per unit change in X, acting as if Eq. (1.1) is the true model.

To obtain a confidence interval for β_1, we need an estimate of the standard deviation of $\widehat{\beta}_1$, after first deriving the expression for the standard deviation. The latter can be obtained as follows. We first write $\widehat{\beta}_1$ as a linear combination of the Y values. Specifically,

$$\begin{aligned}\widehat{\beta}_1 &= \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} \\ &= \frac{\sum (X_i - \overline{X})Y_i}{S_{xx}}\end{aligned} \tag{1.10}$$

with the second line resulting from the fact that $\sum (X_i - \overline{X})\overline{Y} = 0$. (Notice that the expression for $\widehat{\beta}_1$ in Eq. (1.10) differs from the expression in Eq. (1.6). The latter is used for hand calculation; the reader is asked to show the equivalence of the two expressions in Exercise 1.3.)

Using the expression in the second line, we may write $\widehat{\beta}_1 = \sum k_i Y_i$, with $k_i = (X_i - \overline{X})/S_{xx}$. Recall that Y is assumed to have a normal distribution, so the fact that $\widehat{\beta}_1$ can be written as a linear combination of normally distributed random variables means that $\widehat{\beta}_1$ also has a normal distribution. Furthermore, writing $\widehat{\beta}_1$ in this manner makes it easy to obtain its standard deviation.

We may also use this expression for $\widehat{\beta}_1$ to show that it is an unbiased estimator of β_1, as is shown in the chapter Appendix. In general, an *unbiased estimator* is one for which the expected value of the estimator is equal to the parameter, the unknown value of which the value of the estimator serves to estimate. It will be seen later in this section that $\widehat{\beta}_1$ is in the center of the confidence interval for β_1. This would be illogical if $\widehat{\beta}_1$ were a biased estimator, and the amount of the bias were known. It should be noted that the unbiasedness property of $\widehat{\beta}_1$ is based on the assumption that the fitted model is the true model. This is a rather strong assumption that generally will be false. But since the true model is unknown, we do not know the extent to which $\widehat{\beta}_1$ is biased. Therefore, the usual textbook approach is to regard the $\widehat{\beta}_i$ as being unbiased when the method of least squares is used.

We initially obtain the variance of $\widehat{\beta}_1$ as

$$\begin{aligned}\text{Var}(\widehat{\beta}_1) &= \text{Var}\left(\sum k_i Y_i\right) \\ &= \sum \text{Var}(k_i Y_i) \\ &= \sum k_i^2 \text{Var}(Y_i) \\ &= \sigma^2 \sum k_i^2 \\ &= \sigma^2 / S_{xx}\end{aligned}$$

Thus, the standard deviation of $\widehat{\beta}_1$ is $\sigma/\sqrt{S_{xx}}$, so the estimated standard deviation is $\widehat{\sigma}/\sqrt{S_{xx}}$. Therefore, we need to estimate σ. (The estimated standard deviation of an estimator is frequently called the *standard error* of the estimator.)

Before doing so, however, we need to clearly understand what σ and σ^2 represent. To this point in the chapter (and in Section 1.5.3, in particular), σ^2 has been used to represent $\text{Var}(\epsilon_i)$ and $\text{Var}(Y_i|X_i)$. Can we state that σ^2 is also the variance of Y? Not quite.

Our interest is in the variance of Y given X, not the variance of Y ignoring X. Whereas we could estimate σ_y^2 as $s_y^2 = \sum (Y - \overline{Y})^2/(n-1)$, this would ignore the fact that X is being used to predict Y. Consequently, the estimate of σ_ϵ^2 should be less than s_y^2, and the variance of Y that we should speak of is, in statistical parlance, the conditional variance of Y given X, as was stated previously, remembering that the variance is assumed to be the same for each value of X. We need not be concerned about the latter, however, since it is equal to σ_ϵ^2 when the postulated model is the true model. As stated previously, we cannot expect to know the true model, but we act as if these two variances are the same in the absence of information that would suggest that the model given by Eq. (1.1) is not an appropriate model.

In estimating σ_ϵ^2 we can think about how the variance of a random variable is estimated in an introductory statistics course. For a sample of observations on some random variable, W, σ_w^2 is estimated by $s_w^2 = \sum (W - \overline{W})^2/(n-1)$, with the divisor making s_w^2 an unbiased estimator of σ_w^2.

If we proceed to estimate σ_ϵ^2 in an analogous manner, we would compute $s_e^2 = \sum (e - \overline{e})^2/(n-2)$. Note that e is substituted for ϵ since ϵ is not observable. Notice also that the divisor is $n-2$ instead of $n-1$. A divisor of $n-2$ is needed to make s_e^2 an unbiased estimator of σ_ϵ^2, as is shown in the chapter Appendix. (Another reason why the divisor must be $n-2$ is given later in this section.)

Since $e = Y - \widehat{Y}$ and $\overline{e} = 0$, s_e^2 can be written as $s_e^2 = \sum (Y - \widehat{Y})^2/(n-2)$. Putting these results together, we obtain

$$\begin{aligned}s_{\widehat{\beta}_1} &= \frac{s_e}{\sqrt{S_{xx}}} \\ &= \sqrt{\frac{\sum (Y - \widehat{Y})^2}{(n-2)S_{xx}}}\end{aligned}$$

If we write

$$t = \frac{\widehat{\beta}_1 - \beta_1}{s_{\widehat{\beta}_1}}$$

it can be shown that t has a Student's-t distribution with $n - 2$ degrees of freedom (see the chapter Appendix).

It follows that

$$P\left(-t_{\alpha/2,n-2} \leq \frac{\widehat{\beta}_1 - \beta_1}{s_{\widehat{\beta}_1}} \leq t_{\alpha/2,n-2}\right) = 1 - \alpha \tag{1.11}$$

provides the starting point for obtaining a $100(1 - \alpha)\%$ confidence interval for β_1, with the value of $t_{\alpha/2,n-2}$ obtained from a t table with a tail area of $\alpha/2$ and $n - 2$ degrees of freedom. Rearranging Eq. (1.11) so as to give the end points of the interval produces

$$P(\widehat{\beta} - t_{\alpha/2,n-2}s_{\widehat{\beta}_1} \leq \beta_1 \leq \widehat{\beta}_1 + t_{\alpha/2,n-2}s_{\widehat{\beta}_1}) = 1 - \alpha$$

The lower limit (*L.L.*) is thus $\widehat{\beta}_1 - t_{\alpha/2,n-2}s_{\widehat{\beta}_1}$ and the upper limit (*U.L.*) is $\widehat{\beta}_1 + t_{\alpha/2,n-2}s_{\widehat{\beta}_1}$.

For the data in Table 1.2, the values of $\sum \left(Y - \widehat{Y}\right)^2$ and S_{xx} were given previously. We thus have

$$\begin{aligned} s_{\widehat{\beta}_1} &= \sqrt{\frac{\sum (Y - \widehat{Y})^2}{(n-2)S_{xx}}} \\ &= \sqrt{\frac{0.5212}{(8)2062.5}} \\ &= 0.0056 \end{aligned}$$

A 95% confidence interval for β_1 would then have $L.L. = 0.19673 - 2.306(0.0056) = 0.1838$, and $U.L. = 0.19673 + 2.306(0.0056) = 0.2096$.

We would certainly want our confidence interval to not include zero because if $\beta_1 = 0$, there would be no regression equation. Since R^2 was quite large (.9935), we would expect the interval not to include zero. The converse is not true, however. That is, if the interval does not include zero, R^2 could be much less than one and could even be less than .5. This will be discussed further in the context of hypothesis testing.

Statistics books that cover regression analysis generally present a hypothesis test of $\beta_1 = 0$ in which the null and alternative hypotheses are H_o: $\beta_1 = 0$ and H_a: $\beta_1 \neq 0$, respectively. The rejection of H_0, using a (typical) significance level

of $\alpha = .05$ or $\alpha = .01$, will not ensure that the regression equation will have much value, however.

This can be demonstrated as follows, continuing with the current example. For testing H_o: $\beta_1 = 0$,

$$\begin{aligned} t &= \frac{\widehat{\beta}_1 - 0}{s_{\widehat{\beta}_1}} \\ &= \frac{0.19673}{0.0056} \\ &= 35.13 \end{aligned}$$

Since $t_{8,.025} = 2.306$ and $t_{8,.005} = 3.355$, we would reject H_o: $\beta_1 = 0$ at either significance level. As inferred earlier, the hypothesis test and confidence interval for β_1 are *not* robust (i.e., not insensitive) to nonnormality, or to extreme data points that can result from nonnormality. Therefore, the hypothesis test and confidence interval should be used with caution, with methods such as those given in Chapter 2 used for detecting a clear departure from normality.

With some algebra we can show (see chapter Appendix) that

$$R^2 = \frac{t^2}{n - 2 + t^2} \tag{1.12}$$

so $R^2 < 0.50$ if $t^2 < n - 2$. If the calculated value of t had equaled $t_{8,.025}$, then R^2 would have equaled .3993. Using a significance level of $\alpha = .01$ would not help much, because if $t = t_{8,.005}$ then $R^2 = .5845$.

It should thus be clear that for the regression equation to be of value (for not only prediction, but also for the other uses of regression), the calculated value of t should exceed some multiple of the tabular value. If we adapt to a t-test the recommendation of Draper and Smith (1998), which is based on the work of Wetz (1964), we obtain the result that the calculated t-value should be *at least* twice the tabular value.

For the current example, if $t \geq 2t_{8,.025}$ then $R^2 \geq .7267$, and if $t \geq 2t_{8,.005}$ then $R^2 \geq .8491$. Obviously, these numbers look better than the pair of R^2 values given previously. As stated earlier, there is no clear dividing line between high and low R^2 values, but we can see that the "double t" rule could result in some rather small R^2 values being declared acceptable when n is much larger than 10. For example, for $n = 20$, $t \geq 2t_{18,.025}$ implies $R^2 \geq .4952$, whereas $t \geq 3t_{18,.025}$ implies $R^2 \geq .6882$. Therefore, three might be a more suitable multiplier if n is much larger than 10.

How large should n be? This decision is also somewhat arbitrary, just as are the choices for α and the multiplier of t. Draper and Smith (1998) suggest that n should be at least equal to ten times the number of regressors. Thus, we might consider having $n \geq 10$ in simple regression. This should be considered as only a very rough rule-of-thumb, however, as data are expensive in certain fields of

application, such as the physical sciences. As stated by Frank and Friedman (1992), "In chemometrics applications the number of predictor variables often (greatly) exceeds the number of observations."

Because of the relationship between confidence intervals and hypothesis tests for β_1, the results that were stated in terms of hypothesis tests also apply to confidence intervals. Specifically, if $t = t_{8,.025}$ the lower limit of the 95% confidence interval would be zero, and if $t > t_{8,.025}$ the lower limit would exceed zero (and similarly the upper limit would be less than zero if $t < -t_{8,.025}$). Thus, a 95% (or 99%) confidence interval for β_1 that does not cover zero does not guarantee a reasonable value for R^2.

1.5.5 Inferences on β_0

We are usually not interested in constructing a confidence interval for β_0, and rarely would we want to test the hypothesis that $\beta_0 = 0$. There are, however, a few situations in which a confidence interval for β_0 would be useful, and the form of the confidence interval can be obtained as follows.

Analogous to the form of the confidence interval for β_1, the confidence interval for β_0 is given by

$$\widehat{\beta}_0 \pm t_{\alpha/2,n-2} s_{\widehat{\beta}_0}$$

It is shown in the chapter Appendix that $\text{Cov}(\overline{Y}, \widehat{\beta}_1) = 0$, with Cov denoting covariance. It then follows that $\text{Var}(\widehat{\beta}_0) = \text{Var}(\overline{Y} - \widehat{\beta}_1 \overline{X}) = \text{Var}(\overline{Y}) + \overline{X}^2 \text{Var}(\widehat{\beta}_1) = \sigma^2/n + \overline{X}^2(\sigma^2/S_{xx})$, which can be written more concisely as $\sigma^2 \left(\sum X^2/nS_{xx}\right)$. It then follows that

$$s_{\widehat{\beta}_0} = \widehat{\sigma} \left(\frac{\sum X^2}{nS_{xx}} \right)^{1/2}$$

The use of a hypothesis test of H_o: $\beta_0 = 0$ against H_a: $\beta_0 \neq 0$ is another matter, however, as this relates to choosing between the models $Y = \beta_0 + \beta_1 X + \epsilon$ and $Y = \beta_1' X + \epsilon'$. A choice between the two models should be made before any data are collected, and rarely would we use the no-intercept model. We would generally use the latter when (1) we know that $Y = 0$ when $X = 0$, (2) we are using a data set (for parameter estimation) where the points are at least very close to the point (0, 0), and preferably cover that point, and (3) we are interested in using the regression equation (for prediction, say) when X is close to zero.

Certainly, there are many potential applications of regression analysis for which the first condition is met. One such application is described by Casella (1983), in which the independent variable is the weight of a car and the dependent variable is gallons per mile. The author contends that a no-intercept model (which is also termed *regression through the origin*) is appropriate in view of the physical considerations. Obviously, no car weighs close to zero pounds, however, so the

second and third conditions are not met, and to say that the first condition is met in this example is to state the obvious: if we do not drive a car then we will not realize any mileage from a car, so no gas will be used. Thus, since the first condition is satisfied only trivially, one might argue that a hypothesis test of H_o: $\beta_0 = 0$ is needed to determine if a no-intercept model should be used, and in this example Casella (1983) found that H_o was not rejected. It was also shown, however, that the intercept model had a slightly higher R^2 value, so one could argue that the intercept model should be used. (*Note*: Different forms of R^2 have been recommended for different types of regression models; this is discussed in Section 1.6 and in subsequent chapters.)

Recall a potential application of regression for control that was mentioned briefly in Section 1.2, in which the objective is to control the level of a certain type of river pollutant (and, ideally, to eliminate the pollutant). If the dependent variable were a measure of pollution that could have a nonzero value only if the pollutant were nonzero, then regression through the origin would be appropriate because we would probably be interested in predicting Y when X is close to zero.

Even if we know that Y must equal zero when $X = 0$, that is not a sufficient reason for using the no-intercept model, especially when this condition is trivially satisfied. (The latter will certainly be true in many regression applications, because if X and Y are related and X is "absent" in the sense that $X = 0$ is an implausible value, then Y is also likely to be absent.) For the "absent–absent" scenario we are not likely to have data that are close to the point (0, 0), so if we force the regression line to go through the origin the line may not fit the data as well as a regression line with an intercept. An example of this is given by Draper and Smith (1998), who indicate that the intercept model provides a better fit to the data even though it was known that $Y = 0$ when $X = 0$. (Y was the height of soap suds in a dishpan, and X was grams of the detergent.)

Thus, there are limited conditions for which linear regression through the origin is appropriate. Accordingly, the topic is treated only briefly in Section 1.6.

1.5.6 Inferences for *Y*

Fitted values of Y have been illustrated previously, but we would generally like to have an interval about the predicted values. Such an interval is termed a *prediction interval* rather than a confidence interval because the latter is constructed only for a parameter, and Y is not a parameter.

1.5.6.1 Prediction Interval for Y

Confidence intervals for a parameter θ often have the general form $\widehat{\theta} \pm ts_{\widehat{\theta}}$. The confidence interval presented in the next section will thus be of this form, but the prediction interval for Y, which is considerably more useful, is not of this form. Specifically, the prediction interval is given by

$$\widehat{Y} \pm t_{\alpha/2,n-2}\sqrt{\widehat{\sigma}_y^2 + \widehat{\sigma}_\epsilon^2} \tag{1.13}$$

Notice that we would have the general form of a confidence interval (for the parameter that $\widehat{Y}$ estimates) if it were not for the second term under the radical in Eq. (1.13). The presence of that term can be explained as follows. Since the mean of ϵ is zero, it follows that the conditional mean of Y given X, $\mu_{y|x}$, is equal to $\beta_0 + \beta_1 X$. Because we are predicting an individual Y value, we must account for the variability of Y about its mean. Since $\epsilon = Y - (\beta_0 + \beta_1 X) = Y - \mu_{y|x}$, the additional variance component must be σ_ϵ^2. Thus, since $Y = \mu_{y|x} + \epsilon$ and $\widehat{Y} = \widehat{\mu}_{y|x}$, we must add $\sigma_{\widehat{y}}^2$ to σ_ϵ^2 and then use the appropriate estimators for each of the two variance components. (Note that $\text{Var}(\widehat{Y} + \epsilon)$ is equal to the sum of the individual variances because ϵ is assumed to be independent of $\widehat{Y}$, which follows from the assumption stated earlier that ϵ must be uncorrelated, and hence independent under normality, of X.)

It is shown in the chapter Appendix that

$$\text{Var}(\widehat{Y}) = \sigma_\epsilon^2 \left(\frac{1}{n} + \frac{(x - \overline{x})^2}{S_{xx}} \right) \tag{1.14}$$

Thus, $\widehat{\sigma}_\epsilon^2 + \widehat{\sigma}_{\widehat{y}}^2 = \widehat{\sigma}_\epsilon^2(1 + 1/n + (x - \overline{x})^2/S_{xx})$, and the square root of this last expression would be used in obtaining the prediction interval given in Eq. (1.13).

It was stated previously that σ_ϵ^2 could be estimated as $\widehat{\sigma}_\epsilon^2 = \sum (Y - \widehat{Y})^2/(n-2)$. This is not the best form of $\widehat{\sigma}_\epsilon^2$ for computational purposes, however, as each of the n $\widehat{Y}$ values would have to be computed. It is shown in the chapter Appendix that $\sum (Y - \widehat{Y})^2 = S_{yy} - \widehat{\beta}_1 S_{xy} - \widehat{\beta}_1^2 S_{xx}$. Thus, $\widehat{\sigma}_\epsilon^2 = (S_{yy} - \widehat{\beta}_1 S_{xy})/(n-2)$, which is frequently denoted as simply s^2.

Since $\widehat{\sigma}_\epsilon^2$ and $\widehat{\sigma}_{\widehat{\beta}_1}^2$ both contain s^2, Eq. (1.13) may be written in the form

$$\widehat{Y}_o \pm t_{\alpha/2,n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}} \tag{1.15}$$

where $\widehat{Y}_o$ is the predicted value of Y using a particular value of X denoted by x_o. The latter may be one of the values used in developing the prediction equation, or it may be some other value as long as that value is within the range of X values in the data set used in constructing the interval (or at least very close to that range). (Note that here we are using a lowercase letter to denote a specific value of a random variable, as is customary.)

To obtain predicted values and prediction intervals for values of X outside that range would be to engage in *extrapolation*, as we would be extrapolating from a region where we can approximate the relationship between Y and X to a region where we have no information regarding that relationship.

For the data in Table 1.2, a 95% prediction interval for Y given that $x_o = 380$ can be obtained as follows. For $x_o = 380$, $\widehat{y}_o = 22.05$, as was given in Table 1.3.

Using $s^2 = (S_{yy} - \widehat{\beta}_1 S_{xy})/(n-2)$ we obtain

$$\begin{aligned} s^2 &= \frac{80.339 - 0.19673(405.75)}{8} \\ &= 0.0645 \end{aligned}$$

so that $s = 0.2540$. The prediction interval would then be obtained as

$$\widehat{Y}_o \pm t_{\alpha/2,n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}} \tag{1.16}$$

$$= 22.05 \pm 2.306(0.2540)\sqrt{1 + \frac{1}{10} + \frac{(380 - 397.5)^2}{2062.5}}$$

$$= 22.05 \pm 0.65$$

Thus, the prediction interval is (21.4, 22.7) when $x = 380$. Prediction intervals are interpreted analogous to the way that confidence intervals are interpreted. That is, if the experiment described in Section 1.2 were repeated 100 times, $x = 380$ was one of the temperature readings used, the normality assumption was met, and the true model was given by Eq. (1.1), the expected number of times that the interval (21.4, 22.7) covers the observed value of Y when $x = 380$ is 95.

One might ask why we are interested in obtaining a prediction interval for Y for $x = 380$ when we have an observed value of Y corresponding to that value of X. That value of Y, 22.05 in this example, is of course a random variable, and we would much prefer to have an interval estimate for a future value of Y than a point estimate obtained by using an observed sample value. Therefore, even if we wish to estimate future process yield for an x_0 that is one of the sample values, it is desirable to have an interval estimate in addition to the predicted value.

It should be noted that the width of the prediction interval given by Eq. (1.16) will depend on the distance that x_0 is from $\overline{x}$. This should be intuitively apparent, as we should certainly be able to do a better job of predicting Y when x_0 is in the middle of a data set than when it is at either extreme.

There are many areas of application in which prediction intervals are extremely important. Consider again the scenario where Y = college GPA, and consider two possible prediction intervals given by (1.8, 2.6) and (2.1, 2.3). The value of $\widehat{Y}$ is 2.2 in each case (since $\widehat{Y}$ lies in the center of the interval), but an admissions officer would undoubtedly look more favorably upon the second interval than the first interval. In other areas of application such as fisheries management, having a well-estimated prediction interval is thought to be probably as important as having good estimates of the regression parameters (Ruppert and Aldershof, 1989).

The assumption of normality is quite important for a prediction interval (unlike the confidence interval presented in the next section), so there should be evidence of approximate normality before a prediction interval is constructed.

If the normality assumption appears not to be valid, methods of constructing prediction intervals given by Olive (2006) should be considered as they are not based on any error distribution assumption.

1.5.6.2 Confidence Interval for $\mu_{Y|X}$

A confidence interval for $\mu_{Y|X}$ is similar to the prediction interval in terms of the computational form, but quite different in terms of interpretation. In particular, a prediction interval is for the future, whereas a confidence interval is for the present. The only difference computationally is that the "1" under the radical in Eq. (1.16) is not used for the confidence interval. That is, the latter is of the form $\widehat{Y}_o \pm t_{\alpha/2,n-2} s_{\widehat{Y}}$. The confidence interval would be an interval on the average value of Y in a population when $X = x_o$. This will generally be less useful than the prediction interval, however. For example, using $Y =$ college GPA and $X =$ high school GPA, of what value would it be to have a confidence interval for the average GPA of all students who have entered a certain college with a particular value of X and subsequently graduated, when we are trying to reach a decision regarding the admittance of a particular student?

1.5.7 ANOVA Tables

An *Analysis of Variance* (ANOVA) table is usually constructed in analyzing regression data. Loosely speaking, the information provided by such tables is essentially the square of the information used in conjunction with the t-distribution that was presented in Section 1.5.4. The hypothesis test and confidence and prediction intervals presented in the preceding sections have utilized the standard deviation of the appropriate estimators, whereas ANOVA tables utilize variance-type information in showing variation due to different sources.

An ANOVA table for the example used in this chapter will have components that correspond to Eq. (1.7). Those three components are all *sums of squares*. Since $\sum (Y - \overline{Y})^2$ represents the total variability in Y, disregarding X, it would be logical to call it the *total sum of squares* (SS_{total}). The calculation of $\sum (Y - \widehat{Y})^2$ produces the residual *sum of squares* (SS_{residual}); $\sum (\widehat{Y} - \overline{Y})^2$ is the reduction in SS_{total} that results from using X to predict Y, and this is labeled $SS_{\text{regression}}$. Thus, $SS_{\text{total}} = SS_{\text{regression}} + SS_{\text{residual}}$.

Mean squares in ANOVA tables are always obtained by dividing each sum of squares by the corresponding *degrees of freedom* (*d.f.*). When SS_{total} is defined as $\sum (Y - \overline{Y})^2$, which is the usual way, "total" will always have $n - 1$ *d.f.* This can be explained as follows. First, there are $n - 1$ of the n components of $\sum (Y - \overline{Y})$ that are free to vary since $\sum (Y - \overline{Y}) = 0$. We then obtain $n - 1$ as the number of degrees of freedom by applying the rule which states that degrees of freedom are given by the number of observations, n, minus the number of linear restrictions. Another way to view *d.f.*(*total*) is to recognize that a degree of freedom is used, in general, whenever a parameter is estimated, and it can be shown that $SS(\widehat{\beta}_0) = \left(\sum Y\right)^2 / n$, which is part of SS_{total} since

Table 1.4 ANOVA Table for Data in Table 1.2

Source	df	Sum of Squares	Mean Square	F	p
Regression	1	79.822	79.822	1235.38	0.000
Residual error	8	0.517	0.065		
Total	9	80.339			

$\sum (Y - \overline{Y})^2 = \sum Y^2 - \left(\sum Y\right)^2 / n$. Thus, this "corrected" sum of squares that corrects for the mean also incorporates the sum of squares for one of the parameters that has been estimated. In order to compute $\sum (Y - \widehat{Y})^2$, both parameters must have been estimated, so this sum of squares has $n - 2$ *d.f.* Since degrees of freedom are generally additive, $\sum (\widehat{Y} - \overline{Y})^2$ must have one *d.f.*, and this one *d.f.* corresponds to the estimation of β_1.

Using the data in Table 1.2, we obtain the ANOVA table given in Table 1.4. The number in the column labeled F is obtained by computing the ratio of the two mean squares, and the p-value is the probability of obtaining a value of F that is at least this large when $\beta_1 = 0$.

In introductory courses, the relationship between the t- and F-distributions is usually discussed. Specifically, $t^2_{\nu_1,\alpha/2} = F_{1,\nu_1,\alpha}$, where ν_1 is the degrees of freedom of the t-statistic, and 1 and ν_1 are the degrees of freedom for the numerator and denominator, respectively, of the two components whose ratio comprises the F-statistic. (The upper tail areas for the t- and F-distributions are here denoted by $\alpha/2$ and α, respectively.)

It is easy to show that $t^2 = F$, where $t = \widehat{\beta}_1 / s_{\widehat{\beta}_1}$. We proceed by writing

$$t = \frac{\widehat{\beta}_1}{(s_e/\sqrt{S_{xx}})}$$

so that $t^2 = \widehat{\beta}_1^2 S_{xx}/s_e^2$. It was mentioned in Section 1.5.6.1 that $\sum (Y - \widehat{Y})^2 = S_{yy} - \widehat{\beta}_1^2 S_{xx}$, where $\sum (Y - \widehat{Y})^2 = SS_{\text{residual}}$ and $S_{yy} = SS_{\text{total}}$. It follows that $\widehat{\beta}_1^2 S_{xx} = SS_{\text{regression}}$. Since $s_e^2 = MS_{\text{residual}}$, $t^2 = SS_{\text{regression}}/MS_{\text{residual}} = MS_{\text{regression}}/MS_{\text{residual}} = F$. (It is shown in the chapter Appendix that the ratio of these two mean squares has an F-distribution, starting from the assumption of a normal distribution for the error term.)

Whether we look at $t = 35.13$ or $F = 1235.38$, the magnitude of each of these two numbers coupled with the fact that $R^2 = .9935$ indicates that there is a strong relationship between yield and temperature for the range of temperatures covered in the experiment.

1.5.8 Lack of Fit

Frequently, a regression model can be improved by using nonlinear terms in addition to the linear terms. In simple regression the need for nonlinear terms

is generally indicated by the scatter plot of the data, but with more than one regressor the need for nonlinear terms will be more difficult to detect.

A *lack-of-fit* test can be performed to see if there is a need for one or more nonlinear terms. We should note that we are still concerned here with linear regression models, which are linear in the parameters; we are simply trying to determine the possible need for terms such as polynomial terms or trigonometric terms that would enter the model in a linear manner.

The general idea is to separate SS_{residual} into two components: $SS_{\text{pure error}}$ and $SS_{\text{lack of fit}}$. Simply stated, $SS_{\text{pure error}}$ is the portion of SS_{residual} that cannot be reduced by improving the model. This can be seen from the formula for $SS_{\text{pure error}}$, which is

$$SS_{\text{pure error}} = \sum_{j=1}^{n_i} \sum_{i=1}^{k} (Y_{ij} - \overline{Y}_i)^2 \tag{1.17}$$

where $\overline{Y}_i$ is the average of the n_i Y values corresponding to X_i, k is the number of different X-values, and Y_{ij} is the jth value of Y corresponding to the ith distinct value of X.

■ EXAMPLE 1.3

Consider the data in Table 1.5 for which $k = 6$ and, for example, $Y_{42} = 33$. Thus, $SS_{\text{pure error}} = (15 - 15.5)^2 + (16 - 15.5)^2 + (20 - 21)^2 + (22 - 21)^2 + (31 - 32)^2 + (33 - 32)^2 + (46 - 47.5)^2 + (49 - 47.5)^2 = 9.0$. For a given value of X, different values of Y will plot vertically on a scatter plot, and there is no way that a regression model can be modified to accommodate a vertical line segment, as the slope would be undefined.

Table 1.5 Data for Illustrating Lack of Fit

Y	X
15	10
16	10
14	15
20	20
22	20
31	25
33	25
46	30
49	30
60	35

Using the formulas given in the preceding sections, $SS_{\text{residual}} = 261.85$. Thus, $SS_{\text{lack of fit}} = 261.85 - 9.00 = 252.85$, so almost all of the residual is due to lack of fit, thereby indicating that the wrong model is being used. When this is the case, σ_ϵ^2 should *not* be estimated by MS_{residual} as the latter contains more than just experimental error. Some authors use *error* instead of *residual* in ANOVA tables such as Table 1.4. The two terms are conceptually different as the former is the *experimental error* (in Y) that results when an experiment is repeated, whereas the latter represents the collection of factors that result in the model not providing an exact fit to the data. The latter might consist of not having the right functional form of the regressor(s) or not using all relevant regressors, in addition to experimental error (i.e., variation) resulting from factors that cannot be identified and/or controlled during the experiment.

By extracting the experimental error component from the residual, we can see whether we should attempt to improve the model. A formal lack-of-fit test is performed by computing $F = MS_{\text{lack of fit}}/MS_{\text{pure error}}$ and rejecting the hypothesis of no lack of fit if F exceeds the value from the F-table.

Each mean square is computed in the usual manner by dividing the sum of squares by the corresponding degrees of freedom. The *d.f.* for pure error is best viewed as the sum of the degrees of freedom for each X_i that is repeated, where each such *d.f.* is $n_i - 1$. When $SS_{\text{pure error}}$ is computed using Eq. (1.17) and then divided by the *d.f.*, this is equivalent to taking a weighted average of the s_i^2 values, where s_i^2 is the sample variance of the Y-values corresponding to X_i, with the weights being $n_i - 1$. Thus, $\sigma^2_{\text{pure error}}$ is estimated in a logical manner. The *d.f.* for lack of fit is obtained from *d.f.*(residual) − *d.f.*(pure error).

For the Table 1.5 data,

$$F = \frac{252.85/4}{9.00/4} = 28.09$$

Since $F_{4,4,.05} = 6.39$ and $F_{4,4,.01} = 15.98$, we would conclude that there is evidence of lack of fit.

Since the degree of lack of fit is so extreme, we would expect that the nonlinearity would be evident in the scatter plot. The latter is shown in Figure 1.2, and we can see the obvious curvature.

It is important to recognize that this lack-of-fit test does not indicate how the model should be modified; it simply shows that some modification is necessary. Other methods must be employed to determine how the model should be modified. Some of these methods are discussed in detail in Chapter 5.

When the regressor values are not preselected, we may not have any repeated values. When this happens, there are other methods that can be used. Daniel and Wood (1980, p. 133) discuss a *method of nearest neighbors* that is based on forming pseudo-replicates from regressor values that are close together. Another method is the variagraph approach as described by Robinson and Weisberg

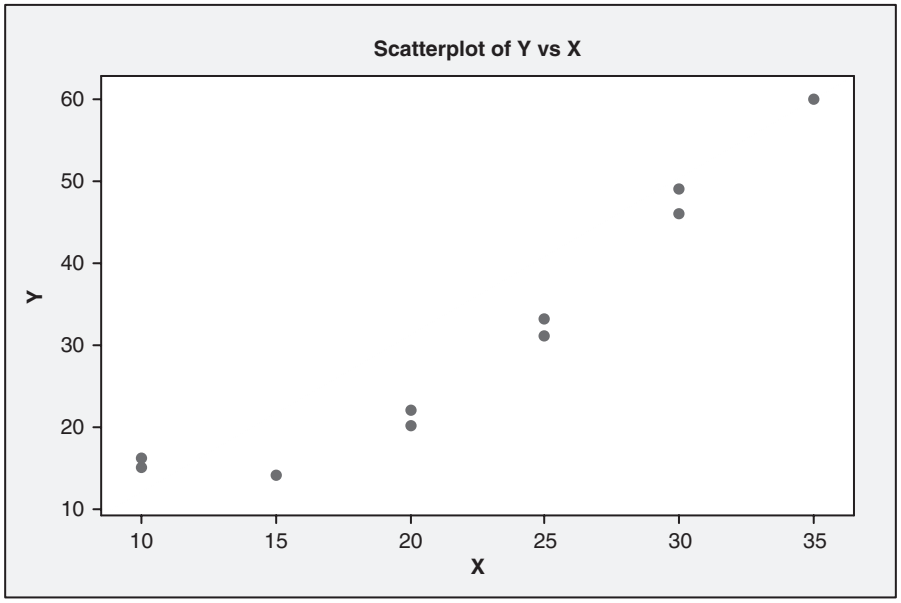

Figure 1.2 Scatter plot of Y versus X for the data in Table 1.5.

(2003). The performance of each of three lack-of-fit tests were compared by Wang and Conerly (2003). ■

1.6 REGRESSION THROUGH THE ORIGIN

This topic was mentioned briefly in Section 1.5.5, in which it was stated that linear regression through the origin will rarely be applicable. The model is

$$Y = \beta X + \epsilon \tag{1.18}$$

and to derive the least squares estimator of β we could use the same approach as was used in Section 1.4. Doing so would produce the estimator

$$\widehat{\beta} = \frac{\sum XY}{\sum X^2}$$

Inferences for the model given by Eq. (1.18) are similar to the inferences that were developed in Section 1.5 for the model given by Eq. (1.1). There are, however, a few notable differences. In particular, the residuals will not sum to zero. For the intercept model, the solution of Eq. (1.5a) is what causes the residuals to sum to zero. There is no counterpart to that equation for the no-intercept model, however, so the residuals are not forced to sum to zero.

Other differences include the fact that the estimate of σ^2 is based on $n - 1$ *d.f.* instead of $n - 2$ since there is only one parameter in the model. A major difference is the way in which R^2 is defined. As discussed in Section 1.5.2, the definition of R^2 for the intercept model is quite intuitive in that it measures the (scaled) improvement in the prediction of Y that results for $\widehat{\beta}_1 \neq 0$ over $\widehat{\beta}_1 = 0$. For the intercept model, $\widehat{Y} = \overline{Y}$ if $\widehat{\beta}_1 = 0$, but for the no-intercept model $\widehat{Y} = 0$ if $\widehat{\beta} = 0$. Therefore, an analogous definition of R^2 for the no-intercept model would be one that measures improvement over $\widehat{Y} = 0$. This suggests that R^2 *might* be defined as $R^2 = \sum \widehat{Y}^2 / \sum (Y - \overline{Y})^2$.

Some reflection should indicate that this is not logical, however, because if we knew that there was no relationship between X and Y, the most reasonable choice for $\widehat{Y}$ is still $\widehat{Y} = \overline{Y}$. Also, it should be apparent that R^2 would exceed one if we had a perfect fit (i.e., $\widehat{Y}_i = Y_i$, $i = 1, 2, \ldots, n$). Therefore, something must be subtracted from the numerator.

The use of $\widehat{Y} = \overline{Y}$ when there is no regression relationship might suggest that we also use R^2 as given in Eq. (1.8) for the no-intercept model, but whereas we know from Eq. (1.7) that R^2 when defined in this manner has an upper bound of one, we have no such assurance for the no-intercept model. Kvålseth (1985) gave an example in which R^2 exceeds one when defined by Eq. (1.8) and recommends that R^2 be defined as in Eq. (1.9) for both the intercept and no-intercept models. Although Eq. (1.9) is equivalent to Eq. (1.8) for an intercept model, they are not equivalent for a no-intercept model. Simple algebra reveals that using Eq. (1.9) for a no-intercept model produces

$$R_0^2 = \frac{\sum \widehat{Y}^2 - \left(\sum Y\right)^2 / n}{\sum \left(Y - \overline{Y}\right)^2} \tag{1.19}$$

Using Eq. (1.9) avoids the issue of how $\widehat{Y}$ would be computed if a regression relationship did not exist, but the use of either (1.9) or (1.19) is not totally devoid of deficiencies, however, as R_o^2 could be negative. This would simply indicate the total inappropriateness of the model given in Eq. (1.18). We also note that Eq. (1.19) is equivalent to the form of R^2 recommended by Gordon (1981).

It is worthwhile to reiterate a point made in Section 1.5.5 regarding a no-intercept model. In discussing such models, Kvålseth (1985) stated "Occasionally, an analyst may force such a model to be fitted by empirical data by relying purely on theoretical reasoning, whereas a careful analysis of the data may reveal that an intercept model is preferable."

1.7 ADDITIONAL EXAMPLES

■ EXAMPLE 1.4

An interesting application of regression, to which we can all relate, was given by Armstrong and Cuzan (2006). They refer to a "Keys model" due to Lichtman

(2005) for predicting the winner of the presidential election. There were 13 keys identified by Lichtman as being crucial in determining whether or not an incumbent would be reelected. Each key was a statement which, if true, favored the incumbent, and if false, favored the opponent. True statements are scored a zero and false statements are scored a one. If fewer than six statements are false (i.e., at least 8 are true), the incumbent is forecast to win, and is forecast to lose if six or more are false (at most 7 are true). Armstrong and Cuzan (2006) get tangled up a bit when they subsequently state "The model challenges credibility because to win, . . . the incumbent needs 7 of the 13 keys in his or her favor." This contradicts their earlier statement, which implies that at least 8 keys must favor the incumbent.

One might question this simple approach because (1) each key is treated equally, and (2) the assignment of a 0 or 1 for the outcome is done subjectively. Accordingly, Armstrong and Cuzan (2006) applied simple linear regression to all historical data (1860 through 2004) and obtained the following regression equation:

$$\widehat{Y} = 37.3 + 1.8X$$

with Y denoting the percentage of the two-party vote that the incumbent received, and X denoting the number of keys that are favorable to the incumbent.

They found that this simple regression equation correctly predicted the winner of each presidential election, going backward in time. Note that at least 7 keys must be favorable to the incumbent in order for the latter to be predicted to win the majority vote. Armstrong and Cuzan (2006) stated that the incumbent is considered to have an advantage and seemingly wouldn't have to win a majority of the keys, but the regression equation indicates that they do need that majority. It is hard to argue against perfect prediction, however. ■

To this point, we have assumed that a single linear regression is adequate for all of the sample data. There will be many applications in which this will not be true, however, as the relationship between Y and X may not be constant over the range of values of X in the sample. Values of Y at each extreme might also suggest that a different model would be needed at those extremes, depending on the application. Indeed, Wright and Palmer (1977, 1999) found that such an approach was necessary in predicting the performance of graduate business students.

We explore methods for fitting something other than a single model to a set of a data in Chapter 10, as well as doing fitting without specifying any model.

1.8 CORRELATION

If X is a *random* variable, we may properly speak of the *correlation* between X and Y. This refers to the extent that the two random variables are related, with the strength of the relationship measured by the (sample) correlation coefficient,

r_{xy}. The latter is computed as

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \tag{1.20}$$

with the components of Eq. (1.20) as previously defined. This is sometimes called the Pearson correlation coefficient to distinguish it from the other correlation coefficients that are used. The relationship between X and Y is assumed to be linear when Eq. (1.20) is used, and X and Y are assumed to have a bivariate normal distribution, although the second assumption is generally not checked. Formally, the population correlation coefficient, ρ_{xy}, is defined as the covariance between X and Y divided by the standard deviation of X times the standard deviation of Y. When those parameters are estimated by the sample statistics, the result reduces to Eq. (1.20).

The possible values of r_{xy} range from -1 to $+1$. The former represents perfect negative correlation, as would happen if all the points fell on a line with a negative slope, and the latter represents perfect positive correlation. In regression there is nothing really "negative" about negative correlation, as a strong negative correlation is just as valuable as a strong positive correlation as far as estimation and prediction are concerned. A zero correlation between X and Y would signify that it would not be meaningful to construct a linear regression equation using that regressor, but when there is more than one regressor, we would prefer that the regressors be "uncorrelated" (orthogonal) with each other. This is discussed and illustrated in Chapter 14.

When there is only a single regressor, $r_{xy}^2 = R^2$. Thus, for the one-regressor random-X case, R^2 is the square of the correlation between X and Y. When X is fixed, R^2 must then be viewed (and labeled) somewhat differently, and it is customary to refer to R^2 as the *coefficient of determination* (and the coefficient of multiple determination when there is more than one regressor).

Another connection between the square of a correlation coefficient and R^2 is $r_{y\widehat{y}}^2 = R^2$, as the reader is asked to show in Exercise 1.10. (Note that this last result holds whether X is random or not, since both Y and $\widehat{Y}$ are random variables.) Clearly, Y and $\widehat{Y}$ must be highly correlated for the regression model to have value.

1.9 MISCELLANEOUS USES OF REGRESSION

Three uses of simple linear regression are discussed in the next three sections.

1.9.1 Regression for Control

As mentioned in Section 1.2, an important but infrequently discussed application of regression is to attempt to control Y at a desired level through the manipulation of X.

Doing so may be difficult, however, for a number of reasons. First, since a cause-and-effect relationship is being inferred, the prediction equation must have been produced from preselected X values, and there must not be any other independent variables that are related to Y. The latter is not likely to be true when only a single X is being used, however, and the consequences when it is not true may be great.

Box (1966) gives a good example of this in regard to a hypothetical chemical process in which Y = process yield, X_1 = pressure, and X_2 = an unsuspected impurity. The scenario is that undesirable frothing can be reduced by increasing pressure, but a high value of X_2 is what actually produces frothing and also lowers yield, with the latter unrelated to pressure. Assume that the regression equation $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1$ is developed with the intention of increasing Y through manipulation of X_1. Even if an experiment consisted of systematically varying X_1 and recording the corresponding Y value, the regression equation would have no value for control because Y depends solely on X_2. Therefore, an experimenter who attempted to rely on the value of $\widehat{\beta}_1$ to effect a desired change in Y through manipulation of X_1 would not be successful.

The equation could still have value for prediction, however, because if R^2 is large, X_1 is essentially serving as a suitable proxy for X_2. So even though the (high) correlation between Y and X_1 is a spurious correlation, X_1 can still be used to predict Y. Thus, a spurious correlation can produce suitable predictions, but $\widehat{\beta}_1$ has no meaning by itself because the correct model does not include X_2. If the true relationship between X_1 and X_2 were known, then that relationship could be incorporated into the prediction equation, but the appropriate coefficient would be different from $\widehat{\beta}_1$.

Following Box (1966), let's assume that $X_2 = \beta_0^* + \beta_1^* X_1$ and the true model in terms of Y is $Y = \beta_0' + \beta_2 X_2$. Then with $\widehat{X}_2 = \widehat{\beta}_0^* + \widehat{\beta}_1^* X_1$ we have

$$\begin{aligned}\widehat{Y} &= \widehat{\beta}_0' + \widehat{\beta}_2(\widehat{\beta}_0^* + \widehat{\beta}_1^* X_1) \\ &= (\widehat{\beta}_0' + \widehat{\beta}_0^* \widehat{\beta}_2) + \widehat{\beta}_2 \widehat{\beta}_1^* X_1\end{aligned}$$

Thus, the experimenter who attempts to manipulate Y through X_1 will do poorly if $\widehat{\beta}_2 \widehat{\beta}_1^*$ differs from $\widehat{\beta}_1$ by more than a small amount, and there should generally be no reason to expect them to be close.

Additional information on the use of regression for control can be found in Hahn (1974) and Draper and Smith (1998).

1.9.2 Inverse Regression

Two specialized uses of simple linear regression are discussed in this section and in the following section. The first of these, *inverse regression*, can be used effectively in calibration work, although its applications are not limited to calibration.

Assume that we have two measuring instruments; one is quite accurate but is both expensive to use and slow, and the other is fast and less expensive to use,

but is also less accurate. If the measurements obtained from the two devices are highly correlated, then the measurement that would have been made using the expensive measuring device could be predicted fairly well from the measurement that is actually obtained using the less expensive device.

In particular, if the less expensive device has an almost constant bias and we define X = measurement from the accurate instrument, and Y = measurement from the inaccurate device, and then regress X on Y to obtain $\widehat{X}^* = \widehat{\beta}_0^* + \widehat{\beta}_1^* Y$, we would expect $\widehat{\beta}_1$ to be close to 1.0 and $\widehat{\beta}_0$ to be approximately equal to the bias.

This is termed *inverse regression* because X is being regressed on Y instead of Y regressed on X.

Since X and Y might seem to be just arbitrary labels, why not reverse them so that we would then have just simple linear regression? Recall that Y must be a random variable, and classical regression theory holds that X is fixed. A measurement from an accurate device should, theoretically, have a zero variance, and the redefined X would be a random variable.

But we have exactly the same problem if we regress X on Y, with X and Y as originally defined. The fact that the independent variable Y is a random variable is not really a problem, as regression can still be used when the independent variable is random, as is discussed in Section 1.8.

If the *dependent* variable is not truly a random variable, however, then classical regression is being "bent" considerably, and this is one reason why inverse regression has been somewhat controversial. Krutchkoff (1967) reintroduced inverse regression, and it was subsequently criticized by other writers.

There is an alternative to inverse regression that avoids these problems, however, and which should frequently produce almost identical results. In the *classical theory of calibration*, Y is regressed against X and X is then solved for in terms of Y for the purpose of predicting X for a given value of Y. Specifically,

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X$$

so that

$$X = (\widehat{Y} - \widehat{\beta}_0)/\widehat{\beta}_1$$

For a given value of Y, say, Y_c, X is then predicted as

$$\widehat{X}_c = (Y_c - \widehat{\beta}_0)/\widehat{\beta}_1$$

As in Section 1.7, let r_{xy} denote the correlation between X and Y. If $r_{xy} \doteq 1$, then $\widehat{X}_c$ and $\widehat{X}_c^*$ (from inverse regression) will be almost identical. Recall that for inverse regression to be effective, the two measurements must be highly correlated, so under conditions for which we would want to use either inverse regression or the classical method of calibration, the two approaches will give

very similar results. Therefore, we need not be concerned about the controversy surrounding inverse regression.

■ EXAMPLE 1.5

Data and Analysis

The real data given in Table 1.6 will be used to illustrate the two approaches.

Regressing Y on X produces the equation

$$\widehat{Y} = 0.76 + 0.9873\ X$$

so for a given value of Y, X would be estimated as

$$\begin{aligned}\widehat{X} &= \frac{Y - 0.76}{0.9873} \\ &= -0.7698 + 1.0129\ Y\end{aligned}$$

Table 1.6 Calibration Data

Measured Amount of Molybdenum (Y)	Known Amount of Molybdenum (X)
1.8	1
1.6	1
3.1	2
2.6	2
3.6	3
3.4	3
4.9	4
4.2	4
6.0	5
5.9	5
6.8	6
6.9	6
8.2	7
7.3	7
8.8	8
8.5	8
9.5	9
9.5	9
10.6	10
10.6	10

Source: The data were originally provided by G. E. P. Box and are given (without X and Y designated) as Table 1 in Hunter and Lamboy (1981).

Table 1.7 Predicted Values for the Table 1.6 Data Using Inverse Regression ($\widehat{X}^*$) and Calibration ($\widehat{X}$)

X	$\widehat{X}^*$	$\widehat{X}$
1	1.08882	1.05341
1	0.88786	0.85083
2	2.39509	2.37017
2	1.89268	1.86373
3	2.89751	2.87662
3	2.69654	2.67404
4	4.20378	4.19338
4	3.50040	3.48436
5	5.30908	5.30757
5	5.20860	5.20628
6	6.11294	6.11788
6	6.21342	6.21917
7	7.51970	7.53593
7	6.61535	6.62433
8	8.12259	8.14367
8	7.82114	7.83980
9	8.82597	8.85269
9	8.82597	8.85269
10	9.93127	9.96688
10	9.93127	9.96688

Regressing X on Y produces the inverse regression estimator

$$\widehat{X}^* = -0.7199 + 1.0048\, Y$$

Discussion

Thus, the equations differ only slightly, and the $\widehat{X}$ and $\widehat{X}^*$ values are given for comparison in Table 1.7. We can see that the estimated values produced by the two procedures differ only slightly, and this is because X and Y are very highly correlated. (The value of the correlation coefficient is .996.)

It is also important to note that $\widehat{X}$ and $\widehat{X}^*$ are closer to X than is Y in all but two cases, and that the difference in most cases is a considerable amount.

For this example the average squared error for $\widehat{X}^*$ is 0.0657 compared to 0.0662 for $\widehat{X}$, but this small difference is hardly justification for using $\widehat{X}^*$. Furthermore, such a comparison would be considered inappropriate by some people since the classical estimator has an infinite mean (expected) square error, whereas the inverse regression estimator has a finite mean square error. (The mean square error of an estimator is defined as the variance plus the square of the bias. This is discussed further in the chapter Appendix.) ■

Various methods have been proposed for obtaining a confidence interval for X_0; for example, see Seber (1977), Draper and Smith (1998), and Cox (2005).

The reader is referred to Chow and Shao (1990) for a related discussion on a comparison of the classical and inverse methods, and an application for which

the two approaches were not equally useful. See also Brown (1993), Graybill and Iyer (1994), Hunter and Lamboy (1981), Mee and Eberhardt (1996), and Ndlovu and Preater (2001).

1.9.3 Regression Control Chart

A regression (quality) control chart was presented by Mandel (1969). The objective is similar to the use of regression for control, although with a regression control chart the intention is to see whether Y is "in control" rather than trying to produce a particular value of Y.

For example, let Y be some measure of tool wear, with $X =$ time. Since Y can be expected to change as X changes, it would be reasonable to determine a prediction interval for Y for a given X, and this is what is done. Specifically, Eq. (1.15) is used with $t_{\alpha/2,n-2}$ replaced by 2.

A regression control chart has the same potential weakness as the use of regression for control, however, in that Y could be affected by other variables in addition to X.

1.9.4 Monitoring Linear Profiles

Whereas a regression control chart is for monitoring Y, it is also desirable to monitor the regression relationship. Since nothing is ever static, the parameter estimates will very likely become inappropriate in time, and a simple linear regression model may even become inappropriate. Mahmoud and Woodall (2004) addressed the issue of monitoring a regression model, with applications to quality control/improvement, by constructing control charts for β_0, β_1, and σ^2. A test for the equality of several regression lines is performed, using indicator variables (see Section 4.4). Briefly, the test consists of fitting a model with a common slope and intercept and then fitting a model that additionally contains the slope and intercepts for each sample. The hypothesis of a common slope and intercept is rejected if SS_{residual} drops substantially when the full model is fit.

If the test results in rejection of a common slope and intercept, control charts are constructed for the slope and intercept in an effort to determine the sample(s) that caused the hypothesis test of a common slope and intercept to be rejected.

In statistical process control parlance, this approach is for Phase I (only), which is analysis of historical data. If an in-control state is suggested for Phase I, a model with a single slope and intercept would be used for process monitoring in Phase II.

Krieger, Pollak, and Yakir (2003) also considered the monitoring of a simple linear regression equation but used a different approach.

1.10 FIXED VERSUS RANDOM REGRESSORS

It is important to realize at the outset that even though classical regression theory is based on the assumption that the values of X are selected, regression can

still be used when X is a random variable. This is important because regression data frequently come from observational studies in which the values of X occur randomly, and we could probably go further and state that most regression data sets have random regressors.

The issue of random regressors has been discussed by many writers, and there is some disagreement. The problem is compounded somewhat by the fact that in the literature it sometimes isn't clear whether regressors "with error" are considered to have random error, or measurement error, or both. Here we consider only random error; measurement error is discussed in Section 2.6.

One popular position is that we may proceed as if X were fixed, provided that the conditional distribution of Y_i given X_i is normal and has a constant variance for each i, and the X_i are independent (as are the Y_i), and the distribution of X_i does not depend on β_0, β_1, or σ^2 (see Neter, Wasserman, and Kutner, 1989, p. 86). Other writers speak of random regressors, while simultaneously assuming true, unobservable values for each X_i (see Seber, 1977, pp. 210–211). If X is a random variable that is devoid of measurement error, there is clearly no such thing as a true value, so the assumption of a true value that is different from the observed value implies that there is measurement error. In compiling and elaborating on results given in earlier papers, Sampson (1974) concluded that in the random regressor case we may proceed the same as in the fixed regressor case. (A detailed discussion of that paper is beyond the scope of this chapter.) A somewhat different view can be found in Mandel (1984), however. Properties of the regression coefficients when X is random are discussed by Shaffer (1991).

1.11 MISSING DATA

Missing data is often a problem in linear regression work. If data are missing on Y or X (but not both), the most commonly used approach is to discard the data on the variable for which it is not missing (obviously there is nothing to discard if the data are missing on both Y_i or X_i for one or more values of i). This usually does not create a problem if the data are missing at random.

The user of simple linear regression has a problem if the sample size is small and there are several missing values. Imputation might be attempted, but this does have some drawbacks. See Bello (1995) for a discussion of imputation methods in regression.

1.12 SPURIOUS RELATIONSHIPS

Just because a significant relationship exists between X and Y doesn't necessarily mean that the relationship makes any sense, as there are many examples of nonsensical relationships that produce significant regression results. One such example is given in Section 2.1.1.1, which is also discussed in Box (1995). There are many unrelated variables that are related to a third variable and move

in the same direction as that variable. One classic classroom example is the "significant" regression relationship between the salaries of college professors and alcohol consumption. They can both be expected to change in accordance with the state of the economy and measures of that state.

Oftentimes spurious relationships are exposed by indicating that there is no causal relationship between X and Y. This isn't necessary, however, and no causal relationship could be inferred without fixing the values of X, as in a designed experiment. (Experimental designs for regression are discussed in Chapter 12.)

1.13 SOFTWARE

The use of statistical software is essential if regression data are to be thoroughly analyzed. All general-purpose statistical software packages have basic regression analysis capability, although capabilities across software vary considerably as one moves past the basics. This will become apparent with the detailed discussion that begins in Section 2.7. The SPSS Web Book by X. Chen, P. Ender, M. Mitchell, and C. Wells (`http://www.ats.ucla.edu/stat/spss/webbooks/reg/default.htm`) informs readers how to use SPSS to perform regression analyses. Similarly, *Regression with SAS* (`http://www.ats.ucla.edu/stat/sas/webbooks/reg/default.htm`) and *Regression with Stata* (`http://www.ats.ucla.edu/stat/stata/webbooks/reg/default.htm`) by the same authors show users how to employ SAS Software and STATA, respectively, for regression analyses. For STATA, see also Chapter 6 of Hamilton (2006).

R and *S*-Plus are popular with statisticians. The first is freeware and the second is commercial; they are both command line statistical packages. They are especially appropriate for a small part of this book (especially Chapter 11) but are not as user friendly as software like MINITAB and thus have a steeper learning curve. (Note that *R* should not be confused with the *R*-code (short for "regression code") developed by Dennis Cook and Sandy Weisberg and described in Cook and Weisberg (1994), which is essentially a user's manual for the software. The software will be discussed in later chapters.)

Students are best advised to use software such as MINITAB for learning the basic concepts of regression analysis. *R* and S-Plus enthusiasts are referred to Weisberg (2005a) for information on using *R* and S-Plus in regression, although that is really intended to be a companion to Weisberg's (2005b) book.

Of these software packages, MINITAB is the software chosen for statistics courses by most colleges and universities. STATA has always sought to provide statistical capabilities that are not found in many, if any, other statistical software. Some of its regression capabilities are unique and are utilized in this book. MINITAB and STATA are both used extensively and JMP (from SAS Institute, Inc.) is also used on a limited basis. SYSTAT is used to produce two scatter plots in Chapter 2 and two scatter plots in Chapter 16, although a much earlier version of the software has better features for those (influence) plots than does the current version (SYSTAT 12), so the plots were produced by the earlier version. This is not to suggest that other software such as SAS, SPSS, S-Plus, and *R* do

not have worthy regression capabilities. Software popularity, availability to the author, ease of use, and capabilities needed for this book were all determining factors in deciding which software to use and discuss.

Given below is the basic output obtained using MINITAB for the data in Table 1.2. A "complete" regression analysis entails going far beyond the production of such output, however, and includes investigation of "unusual observations" as designated in the last part of this output.

```
The regression equation is
Y = - 52.5 + 0.197 X

Predictor       Coef   SE Coef       T      P
Constant    -52.509     2.226  -23.59  0.000
X          0.196727  0.005597   35.15  0.000

S = 0.254192   R-Sq = 99.4%   R-Sq(adj) = 99.3%

Analysis of Variance

Source          DF      SS      MS        F      P
Regression       1  79.822  79.822  1235.38  0.000
Residual Error   8   0.517   0.065
Total            9  80.339

Unusual Observations

Obs   X         Y      Fit  SE Fit  Residual St Resid
8    420  29.6000  30.1164  0.1494   -0.5164   -2.51R

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

1.14 SUMMARY

The material presented in this chapter constitutes the first part of an introduction to the fundamentals of simple linear regression; the introduction is completed in the next chapter in which diagnostic procedures and additional plots are discussed.

When there is only a single regressor, a scatter plot of the data can be very informative and can frequently provide as much information as the combined information from all of the inferential procedures for determining model adequacy that were discussed in this chapter. The situation is quite different when there is more than one regressor, however, so it is desirable to master the techniques used in simple regression so as to have a foundation for understanding multiple regression, which is discussed starting in Chapter 4. (We will discover in subsequent chapters, however, that additional techniques are needed for multiple regression.)

We may use regression when X is either fixed or random, with X quite frequently being random because regression data are often obtained from observational studies.

Unthinking application of regression can produce poor results, so it is important for the reader to understand the array of available methods and when they should be used. This is emphasized in Chapter 16, in which the regression methods presented in this and subsequent chapters are reviewed in the context of a strategy for analyzing regression data. The reader should also study Watts (1981), in which the knowledge and skills necessary to analyze regression data are discussed, and a step-by-step approach to the analysis of linear and nonlinear regression data is given. A detailed discussion of the considerations that should be made in analyzing regression data is given in Section 16.6.

There are a large number of references cited in this text; many other references can be found in Draper (1998, 2002).

APPENDIX

1.A Analysis of Variance Identity

We wish to show that $\sum (Y - \overline{Y})^2 \equiv \sum (Y - \widehat{Y})^2 + \sum (\widehat{Y} - \overline{Y})^2$. We proceed by writing

$$\begin{aligned}\sum \left(Y - \overline{Y}\right)^2 &\equiv \sum \left(Y - \widehat{Y} + \widehat{Y} - \overline{Y}\right)^2 \\ &\equiv \sum \left(\left(Y - \widehat{Y}\right) + \left(\widehat{Y} - \overline{Y}\right)\right)^2 \\ &\equiv \sum \left(Y - \widehat{Y}\right)^2 + \sum \left(\widehat{Y} - \overline{Y}\right)^2 + 2\sum \left(Y - \widehat{Y}\right)\left(\widehat{Y} - \overline{Y}\right)\end{aligned}$$

The cross-product term can be shown to vanish as follows:

$$\begin{aligned}\sum \left(Y - \widehat{Y}\right)\left(\widehat{Y} - \overline{Y}\right) &= \sum \left(Y - \widehat{Y}\right)\widehat{Y} - \sum \left(Y - \widehat{Y}\right)\overline{Y} \\ &= \sum \left(Y - \widehat{Y}\right)\widehat{Y}\end{aligned}$$

since $\overline{Y} \sum (Y - \widehat{Y}) = 0$. Thus, we need only show that $\sum (Y - \widehat{Y})\widehat{Y} = 0$. Notice that this can be written as $\sum e\widehat{Y} = 0$, so as a by-product we will obtain the important result that the residuals are orthogonal to $\widehat{Y}$. We obtain

$$\begin{aligned}\sum Y\widehat{Y} &= \sum Y\left(\overline{Y} + \widehat{\beta}_1\left(X - \overline{X}\right)\right) \\ &= n\overline{Y}^2 + \widehat{\beta}_1 \sum \left(X - \overline{X}\right)Y \\ &= n\overline{Y}^2 + \widehat{\beta}_1 S_{xy}\end{aligned}$$

and

$$\begin{aligned}\sum \widehat{Y}^2 &= \sum \left(\overline{Y} + \widehat{\beta}_1 \left(X - \overline{X}\right)\right)^2 \\ &= \sum \overline{Y}^2 + \widehat{\beta}_1^2 \sum \left(X - \overline{X}\right)^2 + 2\overline{Y}\widehat{\beta}_1 \sum \left(X - \overline{X}\right) \\ &= n\overline{Y}^2 + \widehat{\beta}_1^2 S_{xx}\end{aligned}$$

Thus, $\sum Y\widehat{Y} = \sum \widehat{Y}^2$ since $\widehat{\beta}_1 S_{xy} = (S_{xy}/S_{xx})S_{xy} = S_{xy}^2/S_{xx} = (S_{xy}/S_{xx})^2 S_{xx} = \widehat{\beta}_1^2 S_{xx}$

1.B Unbiased Estimator of β_1

It was stated in Section 1.5.4 that $\widehat{\beta}_i$ is an unbiased estimator of β_i, under the assumption that the fitted model is the true model. We show this for β_1 as follows. Writing $\widehat{\beta}_1 = \sum (X_i - \overline{X})Y_i/S_{xx}$ and assuming that X is fixed, we obtain

$$\begin{aligned}E(\widehat{\beta}_1) &= \frac{1}{S_{xx}} \sum \left(X_i - \overline{X}\right) E(Y_i) \\ &= \frac{1}{S_{xx}} \sum \left(X_i - \overline{X}\right) (\beta_0 + \beta_1 X_i) \\ &= \frac{1}{S_{xx}} \left(0 + \beta_1 \sum \left(X_i - \overline{X}\right) X_i\right) \\ &= \frac{1}{S_{xx}} (\beta_1 S_{xx}) = \beta_1\end{aligned}$$

1.C Unbiased Estimator of σ_ϵ^2

The fact that s_e^2 is an unbiased estimator of σ_ϵ^2 can be shown as follows. Since Y is assumed to have a normal distribution with variance σ^2, it follows that $(Y - \mu_y)^2/\sigma^2$ will have a chi-square distribution with one degree of freedom, and $\sum (Y - \mu_y)^2/\sigma^2$ will have a chi-square distribution with n degrees of freedom. But μ_y is unknown, and two degrees of freedom are lost (corresponding to β_0 and β_1) when $\widehat{Y}$ is used in place of μ_y, so $\chi_{n-2}^2 = \sum (Y - \widehat{Y})^2/\sigma^2$ must be chi-square with $n - 2$ degrees of freedom. Since $E(\chi_{n-2}^2) = n - 2$, it follows that $E(S_e^2) = \sigma^2$ if s_e^2 is defined as $s_e^2 = \sum (Y - \widehat{Y})^2/(n - 2)$.

This proof depends on $E(\widehat{Y}) = \mu_y$, but it can also be shown (Myers, 1990) that s^2 is unbiased when the fitted model contains the true model in addition to extraneous variables.

1.D Mean Square Error of an Estimator

The term *mean square error* was used in Section 1.9.2. When an estimator is unbiased, the mean square error of an estimator is equal to the variance of

that estimator. For an arbitrary estimator $\widehat{\theta}_i$, the mean square error, $MSE(\widehat{\theta}_i)$, is defined as $MSE(\widehat{\theta}_i) = \text{Var}(\widehat{\theta}_i) + (E(\widehat{\theta}_i) - \theta)^2$. If $\widehat{\theta}_i$ is an estimator of a regression parameter, we must know the true model in order to determine $E(\widehat{\theta}_i)$. Thus, whereas we may speak conceptually of the mean square error of an estimator, we will rarely be able to derive it.

1.E How Extraneous Variables Can Affect R^2

Since a degree of freedom is lost whenever a variable is added to a regression model, the fact that s^2 is an unbiased estimator of σ^2 for an overfitted model implies that the expected value of SS_{error} must decrease, and so the expected value of $SS_{\text{regresion}}$, and hence R^2, must increase.

1.F The Distribution of $\left(\widehat{\beta}_1 - \beta_1\right)/s_{\widehat{\beta}_1}$

We will use the fact that a t random variable with ν degrees of freedom results from a standard normal random variable divided by the square root of a chi-square random variable that is first divided by its degrees of freedom, which is also ν.

Let

$$T = \frac{\left(\widehat{\beta}_1 - \beta_1\right)/\sigma_{\widehat{\beta}_1}}{s_{\widehat{\beta}_1}/\sigma_{\widehat{\beta}_1}}$$

The numerator of T is *normal*(0, 1), and the denominator reduces to the square root of s_e^2/σ^2. From the result given in 1.B, we know that $(n-2)s_e^2/\sigma^2$ has a chi-square distribution with $n-2$ degrees of freedom, so $s_{\widehat{\beta}_1}/\sigma_{\widehat{\beta}_1}$ is the square root of a chi-square random variable divided by its degrees of freedom, and, hence, T has a t-distribution with $n-2$ degrees of freedom.

1.G Relationship Between R^2 and t^2

If we expand the numerator of R^2 in Eq. (1.8) we obtain

$$\begin{aligned}\sum\left(\widehat{Y} - \overline{Y}\right)^2 &= \sum \widehat{Y}^2 - 2\overline{Y}\sum \widehat{Y} + \sum \overline{Y}^2 \\ &= \sum \widehat{Y}^2 - n\overline{Y}^2\end{aligned}$$

with the second line resulting from the fact that $\sum Y = \sum \widehat{Y}$ since $\sum (Y - \widehat{Y}) = 0$. From the derivations in 1.A. we have $\sum \widehat{Y}^2 - n\overline{Y}^2 = \widehat{\beta}_1^2 S_{xx}$. Therefore, we may write R^2 as

$$R^2 = \frac{\widehat{\beta}_1^2 S_{xx}}{S_{yy}}$$

with $S_{yy} = \sum (Y - \overline{Y})^2$. If we write t as

$$t = \frac{\widehat{\beta}_1 \sqrt{S_{xx}}}{s_e}$$

so that

$$t^2 = \frac{(n-2)\widehat{\beta}_1^2 S_{xx}}{S_{yy} - \widehat{\beta}_1^2 S_{xx}},$$

simple algebra then shows the result given in Eq. (1.12).

1.H Variance of $\widehat{Y}$

We first write $\widehat{Y}$ as $\widehat{Y} = \overline{Y} + \widehat{\beta}_1(X - \overline{X})$. Then

$$\begin{aligned}
\operatorname{Var}\left(\widehat{Y}\right) &= \operatorname{Var}\left(\overline{Y} + \widehat{\beta}_1\left(X - \overline{X}\right)\right) \\
&= \operatorname{Var}\left(\overline{Y}\right) + \operatorname{Var}\left(\widehat{\beta}_1\left(X - \overline{X}\right)\right) + 2\operatorname{Cov}\left(\overline{Y}, \widehat{\beta}_1\left(X - \overline{X}\right)\right) \\
&= \frac{\sigma^2}{n} + \left(X - \overline{X}\right)^2 \sigma^2_{\widehat{\beta}_1}
\end{aligned}$$

since $\operatorname{Cov}(\overline{Y}, \widehat{\beta}_1(X - \overline{X})) = 0$, where Cov represents covariance. The covariance result may be established as follows. Since X is assumed to be fixed (i.e., not a random variable), it follows that $\operatorname{Cov}(\overline{Y}, \widehat{\beta}_1(X - \overline{X})) = (X - \overline{X})\operatorname{Cov}(\overline{Y}, \widehat{\beta}_1)$, and

$$\begin{aligned}
\operatorname{Cov}\left(\overline{Y}, \widehat{\beta}_1\right) &= \operatorname{Cov}\left(\overline{Y}, \frac{\sum\left(X - \overline{X}\right)Y}{S_{xx}}\right) \\
&= \frac{1}{S_{xx}}\left\{\operatorname{Cov}\left(\frac{Y_1 + Y_2 + \cdots + Y_n}{n}, X_1Y_1 + X_2Y_2 + \cdots \right.\right. \\
&\quad \left.\left. + X_nY_n - \overline{X}\left(Y_1 + Y_2 + \cdots + Y_n\right)\right)\right\} \\
&= \frac{1}{S_{xx}}\left\{\sigma^2\left(\frac{\sum X_i}{n}\right) - \overline{X}\left(\frac{n\sigma^2}{n}\right)\right\} \\
&= 0
\end{aligned}$$

Since $\text{Var}(\widehat{\beta}_1) = \sigma^2/S_{xx}$, we may write the final result for $\text{Var}(\widehat{Y})$ as in Eq. (1.14).

1.I Distribution of $MS_{\text{regression}}/MS_{\text{residual}}$

When two independent chi-square random variables are each divided by their respective degrees of freedom, and a ratio of the two quotients is formed, that ratio will be a random variable that has an F-distribution. Therefore, we need to show that $SS_{\text{regression}}$ and SS_{residual} are independent chi-square random variables.

From the derivation in 1.B, we know that $SS_{\text{residual}}/\sigma^2$ has a chi-square distribution with $n - 2$ degrees of freedom. Applying the same approach to the distribution of $\sum (Y - \overline{Y})^2/\sigma^2$, it is apparent that this random variable has a chi-square distribution with $n - 1$ degrees of freedom (one *d.f.* is lost since $\overline{Y}$ is used to estimate μ_y). It then follows that $\sum (\widehat{Y} - \overline{Y})^2/\sigma^2$ must have a chi-square distribution with 1 *d.f.* because the difference of two chi-square random variables is also a chi-square random variable with degrees of freedom equal to the difference between the two degrees of freedom. Independence of the two chi-square random variables could be established by applying Cochran's (1934) theorem on the decomposition of squared functions of normal random variables.

REFERENCES

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, **27**, 17–21.

Armstrong, J. S. and A. G. Cuzan (2006). Index methods for forecasting: An application to presidential elections. *Foresight: The International Journal of Applied Forecasting*, Issue 3 (February), 10–13. (Paper available at `http://www.forecastingprinciples.com/Political/PDFs/13KeysbyArmstrong&Cuzan.pdf`)

Barrett, J. P. (1974). The coefficient of determination—some limitations. *The American Statistician*, **28**, 19–20.

Bello, A. L. (1995). Imputation techniques in regression analysis: Looking closely at their implementation. *Computational Statistics & Data Analysis*, **20**, 45–57.

Birkes, D. and Y. Dodge (1993). *Alternative Methods of Regression*. New York: Wiley.

Bland, J. M. and D. G. Altman (1994). Statistics Notes: Regression toward the mean. *British Medical Journal (BMJ)*, **308**, 1499. (Available at `http://www.bmj.com/cgi/content/full/308/6942/1499`.)

Box, G. E. P. (1966). Use and abuse of regression. *Technometrics*, **8**, 625–629.

Box, G. (1995). Regression analysis applied to happenstance data. *Quality Engineering*, **7**, 841–846. (Available as Report 149, Center for Quality and Productivity Improvement, University of Wisconsin, 1996, `http://www.engr.wisc.edu/centers/cqpi/reports.html`.)

Box, G. E. P. and N. R. Draper (1969). *Evolutionary Operation*. New York: Wiley.

Boyd, E. (1935). *The Growth of the Surface Area of the Human Body*. Minneapolis: University of Minnesota Press.

Brown, P. J. (1993). *Measurement, Regression, and Calibration*. Oxford, UK: Clarendon Press.

Carroll, R. J. and D. Ruppert (1988). *Transformation and Weighting in Regression*. New York: Chapman and Hall.

Carroll, R. J. and D. Ruppert (1996). The use and misuse of orthogonal regression in errors-in-variables models. *The American Statistician*, **50**, 1–6.

Casella, G. (1983). Leverage and regression through the origin. *The American Statistician*, **37**, 147–152.

Ceasar-Spall, K. and J. C. Spall (1997). Regression analysis as an aid in making oboe reeds. *Journal of Testing and Evaluation*, **25**, 439–444. (Abstract available at `http://cat.inist.fr/?aModele=afficheN&cpsidt=2780782.`) (Also published in *The Journal of the International Double Reed Society*; see p. 124 of `http://idrs.colorado.edu/Publications/Journal/JNL25.code1/JNL25.1997.pdf.`)

Chow, S.-C. and J. Shao (1990). On the difference between the classical and inverse methods of calibration. *Applied Statistics*, **39**, 219–228.

Chu, S. (1996). Diamond ring pricing using linear regression. *Journal of Statistics Education*, **4**, 6–6.

Cochran, W. G. (1934). The distribution of quadratic forms in a normal system with applications to the analysis of variance. *Proceedings of the Cambridge Philosophical Society*, **30**, 178–191.

Cook, R. D. and S. Weisberg (1994). *An Introduction to Regression Graphics*. New York: Wiley.

Cox, C. (2005). Limits of quantitation for laboratory assays. *Journal of the Royal Statistical Society, Series C*, **54**, 63–76.

Current, J. D. (1998). A linear equation for estimating the body surface area in infants and children. *Internet Journal of Anesthesiology*, **2** (2). (Available at `http://www.ispub.com/ostia/index.php?xmlFilePath=journals/ija/vol2n2/bsa.xml.`)

Dana, J. and R. M. Dawes (2004). The superiority of simple alternatives to regression for social science predictions. *Journal of Educational and Behavioral Statistics*, **29**, 317–331.

Daniel, C. and F. S. Wood (1980). *Fitting Equations to Data*, 2nd ed. New York: Wiley.

Draper, N. R. (1998). Applied regression analysis bibliography update 1994–97. *Communications in Statistics: Theory and Methods*, **27**, 2581–2623.

Draper, N. R. (2002). Applied regression bibliography update 2000–2001. *Communications in Statistics: Theory and Methods*, **31**, 2051–2075.

Draper, N. R. and H. Smith (1998). *Applied Regression Analysis*, 3rd ed. New York: Wiley.

Faraway, J. (2004). Human animation using nonparametric regression. *Journal of Computational and Graphical Statistics*, **13**, 537–553.

Finney, D. J. (1996). A note on the history of regression. *Journal of Applied Statistics*, **23**, 555–558.

Frank, I. E. and J. H. Friedman (1992). A statistical view of some chemometrics regression tools. *Technometrics*, **35**, 109–135 (discussion: 136–148).

Fuller, W. A. (1987). *Measurement Error Models*. New York: Wiley. (Paperback published in 2006.)

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute*, **15**, 246–263.

Gerlach, R., R. Bird, and A. Hall (2002). Bayesian variable selection in logistic regression: Predicting company earnings direction. *Australian and New Zealand Journal of Statistics*, **44**, 155–168.

Gordon, H. A. (1981). Errors in computer packages. Least squares regression through the origin. *The Statistician*, **30**, 23–29.

Graham, L. D. (1991). Predicting academic success of students in a master of business administration program. *Educational and Psychological Measurement*, **51**, 721–727.

Graybill, F.A. and H. K. Iyer (1994). *Regression Analysis: Concepts and Applications*. Belmont, CA: Duxbury.

Hahn, G. J. (1974). Regression for prediction versus regression for control. *Chemtech*, 574–576.

Hamilton, L. C. (1992). *Regression with Graphics*. Pacific Grove, CA: Brooks/Cole Publishing Company.

Hamilton, L. C. (2006). *Statistics with Stata*. Belmont, CA: Duxbury.

Hunter, W. G. and W. F. Lamboy (1981). A Bayesian analysis of the linear calibration problem. *Technometrics*, **23**, 323–328 (discussion: 329–350).

Krieger, A. M., M. Pollak, and B. Yakir (2003). Surveillance of a simple linear regression. *Journal of the American Statistical Association*, **98**, 456–469.

Kvålseth, T. O. (1985). Cautionary note about R^2. *The American Statistician*, **39**, 279–285.

Krutchkoff, R. G. (1967). Classical and inverse regression methods of calibration. *Technometrics*, **9**, 425–439.

Lichtman, A. J. (2005). The keys to the White House: Forecast for 2008. *Foresight: The International Journal of Applied Forecasting*, Issue 3 (February), 5–9.

Mahmoud, M. A. and W. H. Woodall (2004). Phase I analysis of linear profiles with calibration applications. *Technometrics*, **46**, 380–391.

Mandel, B. J. (1969). The regression control chart. *Journal of Quality Technology*, **1**, 1–9.

Mandel, J. (1984). Fitting straight lines when both variables are subject to error. *Journal of Quality Technology*, **16**, 1–14.

Mee, R. W. and K. R. Eberhardt (1996). A comparison of uncertainty criteria for calibration. *Technometrics*, **38**, 221–229.

Myers, R. H. (1990). *Classical and Modern Regression with Applications*, 2nd ed. North Scituate, MA: PWS-Kent.

Natrella, M. (1963). *Experimental Statistics*. Handbook 91, U.S. Department of Commerce, National Bureau of Standards.

Neter, J., W. Wasserman, and M. H. Kutner (1989). *Applied Linear Regression Models*, 2nd ed. Homewood, IL: Irwin. (The current edition is Kutner, Nachtsheim, and Wasserman (2004), 4th ed.)

Ndlovu, P. and J. Preater (2001). Calibration using a piecewise simple linear regression model. *Communications in Statistics: Theory and Methods*, **30**, 229–242.

Olive, D. J. (2006). Prediction intervals for regression models. Manuscript. (Available at `http://www.math.siu.edu/olive/ppspi.pdf`.)

Paolillo, J. (1982). The predictive validity of selected admissions variables relative to grade point average earned in a master of business administration program. *Educational and Psychological Measurement*, **42**, 1163–1167.

Ranney, G. B. and C. C. Thigpen (1981). The sample coefficient of determination in simple linear regression. *The American Statistician*, **35**, 152–153.

Robinson, A. P. and S. Weisberg (2003). Using the variagraph to test lack of fit of parametric regression model without replication. *Communications in Statistics: Simulation and Computation*, **32**, 733–745.

Robinson, K., E. Kamischke, and J. Tabor (2005). How to make millions on eBay, or using multiple regression to estimate prices. *Stats*, **6**, 8–11.

Rousseeuw, P. J. and A. Leroy (1987). *Robust Regression and Outlier Detection*. New York: Wiley.

Ruppert, D. and B. Aldershof (1989). Transformations to symmetry and homoscedasticity. *Journal of the American Statistical Association*, **84**, 437–446.

Ryan, T. P. (2000). *Statistical Methods for Quality Improvement*, 2nd ed. New York: Wiley.

Sampson, A. R. (1974). A tale of two regressions. *Journal of the American Statistical Association*, **69**, 682–689.

Seber, G. A. F. (1977). *Linear Regression Analysis*. New York: Wiley.

Shaffer, J. P. (1991). The Gauss–Markov theorem and random regressors. *The American Statistician*, **45**, 269-273.

Shuai, X., Z. Zhou, and R.S. Yost (2003). Using segmented regression models to fit soil nutrient and soybean grain yield changes due to liming. *Environmental Statistics*, **8**, 240–252.

Stanton, J. M. (2001). Galton, Pearson, and the peas: A brief history of linear regression for statistics instructors. *Journal of Statistics Education*, **9**, 3–3. (Available at `http://www.amstat.org/publications/jse/v9n3/stanton.html`.)

Stigler, S. M. (1997). Regression towards the mean, historically considered. *Statistical Methods in Medical Research*, **6**, 103–114.

Wang, D. X. and M. D. Conerly (2003). Evaluation of three lack of fit tests in linear regression models. *Journal of Applied Statistics*, **30**, 683–696.

Watts, D. G. (1981). A task-analysis approach to designing a regression course. *The American Statistician*, **35**, 77–84.

Weisberg, S. (2005a). *Computing Primer for Applied Linear Regression, Third Edition Using R and S-Plus*. (Online publication available at `http://math.fullerton.edu/mori/Math439/Links/RSprimer.pdf`.)

Weisberg, S. (2005b). *Applied Linear Regression*, 3rd ed. Hoboken, NJ: Wiley.

Wetz, J. (1964). Criteria for judging adequacy of estimation by an approximating response function. Unpublished Ph.D. thesis, University of Wisconsin.

White, C. and S. Berry (2002). Tiered polychotomous regression: ranking NFL quarterbacks. *American Statistician*, **56**, 10–21.

Wright, R. and J. Palmer (1994). GMAT scores and undergraduate GPAs as predictors of performance in graduate business programs. *Journal of Education for Business*, **69**, 344–348.

Wright, R. and J. Palmer (1997). Examining performance predictors for differentially successful MBA students. *College Student Journal*, **31** (2), 276–281.

Wright, R. and J. Palmer (1999). Predicting performance of above and below average performers in graduate business schools: A split sample regression analysis. *Educational Research Quarterly*, **22** (4), 35–44.

Zhang, D., X. Lin, and M. Sowers (2000). Semiparametric regression for periodic longitudinal hormone data from multiple menstrual cycles. *Biometrics*, **56**, 31–39.

EXERCISES

1.1. Plot each of the four data sets in Table 1.1 and then determine the prediction equation $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X$. Do your results suggest that regression data should be plotted before any computations are performed?

1.2. Given the following data,

X	1.6	3.2	3.8	4.2	4.4	5.8	6.0	6.7	7.1	7.8
Y	5.6	7.9	8.0	8.2	8.1	9.2	9.5	9.4	9.6	9.9

compute $\widehat{\beta}_0$ and $\widehat{\beta}_1$ using either computer software or a hand calculator. Now make up five different values for $\widehat{\beta}_{1\,(\text{wrong})}$ that are of the same order of magnitude as $\widehat{\beta}_1$, and for each of these compute $\widehat{\beta}_{0\,(\text{wrong})} = \overline{Y} - \widehat{\beta}_{1\,(\text{wrong})}$.

Then compute $\sum(Y - \widehat{Y})$ and $\sum(Y - \widehat{Y})^2$ for each of these six solutions, using four or five decimal places in the calculations. Explain your results relative to what was discussed in Sections 1.4 and 1.5.1.

1.3. Show that Eq. (1.10) is equivalent to Eq. (1.6).

1.4. Express R^2 as a function of F, the test statistic used in the ANOVA table.

1.5. Graph the data in Exercise 1.2, and then compute R^2 and construct the ANOVA table. Could the magnitude of the values of R^2 and F have been anticipated from your graph?

1.6. Assume that $\widehat{\beta}_1 = 2.54$. What does this number mean relative to X and Y?

1.7. What does $1 - R^2$ represent in words?

1.8. Given the following numbers,

Y	3	5	6	7	2	4	8
$\widehat{Y}$	2.53	___	5.30	6.68	3.22	5.30	8.06

fill in the blank. Could the prediction equation be determined from what is given? Explain.

1.9. Construct an example in which the correlation coefficient, r_{xy}, is equal to (plus) one, then multiply the numbers by an appropriate constant that will convert the correlation coefficient to minus one.

1.10. Prove that $r^2_{Y\widehat{Y}} = R^2$.

1.11. Compute the value of the correlation coefficient for the data in Exercise 1.2 and show that the sign of the correlation coefficient must be the same as the sign of $\widehat{\beta}_1$. Explain why this should be expected.

1.12. Using the data in Exercise 1.2, construct a 95% prediction interval for Y when $X = 5.7$.

1.13. Consider the following data:

X	1	2	3	4	5	6	7
Y	5	6	7	8	7	6	5

First obtain the prediction equation and compute R^2. *Then*, graph the data. Looking at your results, what does this suggest about which step should come first?

1.14. Explain why it would be inappropriate to use the regression equation obtained in Exercise 1.2 when $X = 9.5$.

1.15. Perform a lack-of-fit test for the following data (use $\alpha = .05$):

X	3.2	2.1	2.6	3.2	3.0	2.1	3.2	2.9	3.0	2.6	2.4
Y	3.6	3.7	3.0	3.7	3.3	3.8	3.4	2.7	2.8	3.1	3.4

Now construct a scatter plot of the data. Does the plot support the result of the lack-of-fit test?

1.16. Explain the conditions under which an experimenter should consider using regression through the origin. Then consider the following data:

X	0.6	0.7	0.4	1.1	1.3	0.9	1.6	1.2
Y	0.7	0.8	0.6	1.4	1.6	0.7	1.3	1.5

Assume that the (other) conditions are met for the proper use of regression through the origin, and obtain the prediction equation.

1.17. Extrapolation was mentioned in Section 1.5.6.1 as something that should be avoided, especially extrapolation well beyond the range of the data used in computing the prediction equation. Dallal (`http://www.tufts.edu/~gdallal/slr.htm`) gives an example of muscle strength regressed against lean body mass and states that the prediction equation is Strength $= -13.971 + 3.016$ LBM, with "LBM" denoting lean body mass in kilograms and Strength measured by "slow right extensor peak torque in the knee." Obviously, the predicted value for Strength will be small and perhaps even be negative if the LBM value is small. Explain this prediction equation to someone who knows nothing about simple linear regression but who knows that no reasonable measure of Strength can be negative.

1.18. Consider Table 1.2 and the graph of those data in Figure 1.1. We can see from the latter that there is obviously a strong linear relationship but it is frequently helpful to see a line fit to the data, which can help indicate points that deviate from linearity. Use appropriate software to construct a graph that shows the fitted line. (This can be accomplished using a Java applet for simple linear regression, some of which are Internet accessible, or by using standard statistical software such as MINITAB. For the latter, the FITLINE command will display the fitted line.) Does the fitted line suggest a departure from linearity? Explain.

1.19. A simple linear regression will often suffice when there is a single predictor variable, but it will often not suffice because, as is often said, "the world is nonlinear." The following is the well-known Skeena River Sockeye Salmon data from Carroll and Ruppert (1988).

Year	Spawners (X)	Recruits (Y)	Year	Spawners (X)	Recruits (Y)
1940	963	2215	1954	511	1393
1941	572	1334	1955	87	363
1942	305	800	1956	370	668
1943	272	438	1957	448	2067
1944	824	3071	1958	819	644
1945	940	957	1959	799	1747
1946	486	934	1960	273	744
1947	307	971	1961	936	1087
1948	1066	2257	1962	558	1335
1949	480	1451	1963	597	1981
1950	393	686	1964	848	627
1951	176	127	1965	619	1099
1952	237	700	1966	397	1532
1953	700	1381	1967	616	2086

Use appropriate software to graph the data, determine the simple linear regression equation and R^2. Would you recommend that this equation be used? Why or why not? We will return to these data in Exercise 13.20.

1.20. Radioimmunoassay data that has been used by R. J. Carroll (`http://www.stat.tamu.edu/~carroll/data.php`) is given below, Y denoting radioimmunoassay (RIA) counts and X denoting concentration. Use simple linear regression and perform a lack-of-fit test. What do you conclude? Now construct a scatter plot of Y against X. What does this suggest relative to the heading of Section 1.3?

X	Y	X	Y	X	Y
0.0000	1.70	0.550	5.68	6.00	16.740
0.0000	1.66	0.750	6.06	6.00	16.870
0.0000	1.95	0.750	5.07	8.25	18.980
0.0000	2.07	0.750	5.00	8.25	19.850
0.0750	1.91	0.750	5.98	8.25	18.750
0.0750	2.27	1.000	5.84	8.25	18.510
0.0750	2.11	1.000	5.79	11.25	21.666
0.0750	2.39	1.000	6.10	11.25	21.218
0.1025	2.22	1.000	7.81	11.25	19.790
0.1025	2.25	1.375	7.31	11.25	22.669
0.1025	3.26	1.375	7.08	15.00	23.206
0.1025	2.92	1.375	7.06	15.00	22.239
0.1350	2.80	1.375	6.87	15.00	22.436
0.1350	2.94	1.850	9.88	15.00	22.597
0.1350	2.38	1.850	10.12	20.25	23.922
0.1350	2.70	1.850	9.22	20.25	24.871
0.1850	2.78	1.850	9.96	20.25	23.815
0.1850	2.64	2.500	11.04	20.25	24.871
0.1850	2.71	2.500	10.46	27.50	25.478
0.1850	2.85	2.500	10.88	27.50	25.874
0.2500	3.54	2.500	11.65	27.50	24.907
0.2500	2.86	3.250	13.51	27.50	24.871
0.2500	3.15	3.250	15.47	37.00	24.441
0.2500	3.32	3.250	14.21	37.00	25.874
0.4000	3.91	3.250	13.92	37.00	25.748
0.4000	3.83	4.500	16.07	37.00	27.270
0.4000	4.88	4.500	14.67	50.00	29.580
0.4000	4.21	4.500	14.78	50.00	26.698
0.5500	4.54	4.500	15.21	50.00	26.536
0.5500	4.47	6.000	17.34	50.00	27.181
0.5500	4.79	6.000	16.85		

1.21. In her section entitled "Is the Assumption of Linear Regression Justified?" Natrella (1963) gave an illustrative example of Young's modulus (coded) of sapphire rods (Y) as a function of temperature (X). The data are as

X	Y	X	Y
500	328	750	175
550	296	750	154
600	266	800	152
603	260	800	146
603	244	800	124
650	240	850	117
650	232	850	94
650	213	900	97
700	204	900	61
700	203	950	38
700	184	1000	30
750	174	1000	5

follows. Answer the question that she posed using any methods that you prefer.

1.22. Natrella (1963) gave another example with Young's modulus and temperature but subtracted 4000 from the Young's modulus values before performing the computations. If later there is interest in expressing the regression equation in terms of the original units, what adjustment(s) to the computed equation would be necessary to accomplish this?

1.23. Croarkin and Varner (1982, *National Institute of Standards and Technology Technical Note*) illustrated the use of a linear calibration function to calibrate optical imaging systems. With $Y =$ linewidth and $X =$ NIST-certified linewidth, they obtained a regression equation of $\widehat{Y} = 0.282 + 0.977X$. With $n = 40$, $S_{xx} = 325.1$, and $s = 0.0683$, determine each of the following:

(a) What is the numerical value of the t-statistic for testing the hypothesis that $\beta_1 = 0$?

(b) Is the t-statistic large enough to suggest that the equation should be useful for calibration?

(c) What is the expression for $\widehat{X}_c$, assuming that the classical theory of calibration is used?

1.24. (Harder problem) J. A. Hoeting and A. R. Olsen gave a table in their article "Are the fish safe to eat? Assessing mercury levels in fish in Maine lakes" (in *Statistical Case Studies: A Collaboration Between Academe and Industry*, R. Peck, L. D. Haugh, and A. Goodman, eds. ASA-SIAM series) in which the p-value for a simple linear regression model is .0002 but the R^2 value is only .13. The messages are thus conflicting as to whether or not the model is an adequate fit to the data. Explain how this could happen and express the value of the t-statistic as a function of n.

1.25. To see the connection between the least squares regression equation and the equation for a straight line from algebra, consider a sample with only two points (x_1, y_1) and (x_2, y_2). Show that the least squares solution for $\widehat{\beta}_1$ is mathematically equivalent to $(y_2 - y_1)/(x_2 - x_1)$, which of course in algebra is the slope of the line. What is the numerical value of R^2 for this line?

1.26. Critique the following statement: "The least squares point estimates minimize the sum of the squares of the errors."

1.27. Explain what the word "least" means in the "method of least squares."

1.28. Consider a (small) sample of three observations on X_1 and X_2. If the values on X_1 are -2, 0, and 2, respectively, give corresponding values for X_2 such

that the correlation between X_1 and X_2 is zero. (*Hint:* There is more than one correct answer.)

1.29. What is a consequence, if any, of trying to fit a simple linear regression model to a data set that has a large amount of pure error?

1.30. Critique the following statement (question): "How can we expect the least squares regression line to go through the center of the data, such that the sum of the residuals is zero, if we don't know the true model?"

1.31. A housewife asks the following question: "I prepare popcorn for my family every Friday night and I need to know how thick a layer of kernels to use in the cooker in order to produce the desired amount of popcorn. Of course I know that some kernels won't pop and the number that won't pop is a direct function of how many are used. I found on the Internet the results of a group study performed in a statistics class, with a regression equation being given. I don't understand the use of the constant (i.e., intercept) in the equation, however, because obviously if I don't use any kernels, I won't have any popcorn. How can a regression equation with an intercept make any sense in my application?" Respond to the question.

1.32. For simple linear regression with the predictor being a random variable (so that r_{YX}^2 is a correlation coefficient), show that both $r_{Y\widehat{Y}}^2$ and r_{YX}^2 are equal to R^2. (*Hint:* What type of relationship exists between $\widehat{Y}$ and X?)

1.33. Assume that in simple linear regression a 95% prediction interval is to be constructed for Y_0, with $X_0 = \overline{x}$. If $n = 20$, $R^2 = .82$, and $S_{yy} = 40$, what number will be subtracted from $\widehat{Y}_0$ to produce the lower limit?

1.34. A simple linear regression equation is obtained with $\widehat{Y} = 52.4 - 0.805X$ and $R^2 = .652$. What is the numerical value of r_{YX}?

1.35. Assume that a predictor is measured in feet and its coefficient in a simple linear regression equation is 5.13. What will be the value of the coefficient if the predictor were changed from feet to inches? Would the value of the intercept change? If so, what would be the change?

1.36. Consider the following statement: "I know that $\text{Var}(\widehat{\beta}_1)$ is affected by the spread of the predictor values, so I am going to use a measurement unit of centimeters instead of inches for the model predictor so as to increase the spread." Will this work? Show what will happen algebraically when feet are initially used and then inches is used as the measurement unit. Comment.

1.37. A simple linear regression model is fit to a set of 60 data points, for which $\overline{x} = 23.8$. If $\sum y = 612$, what are the coordinates of one point through

which the regression line must pass? (*Hint:* Write the regression equation in a form that is equivalent to Eq. (1.6).)

1.38. Consider the information that is provided by the value of a correlation coefficient while considering the following statement: "It has been found that there is a positive correlation between the number of firefighters battling a blaze and the amount of damage due to the fire." That is, the more firefighters, the greater the dollar damage. Does this suggest that too many firefighters may be used at many fires with the result that more damage than expected may be resulting because the firefighters are getting in each other's way and not working efficiently?

1.39. Critique the following statement: "I am going to use a simple linear regression model without an intercept because I know that Y must be zero when X is zero."

1.40. Consider the following set of predictor values: 4, 8, 9, 7, 2, and 5. Now compute values for the independent variable as $Y = 5 + 7X$. What will be the numerical value of the correlation coefficient? What would be the numerical value if $Y = 7 - 0.5X$?

1.41. Consider the following sample of (y, x) data: (10,1), (5,2), (2,3), (1,4), (2,5), (5,6), and (10,7). Compute the value of the correlation coefficient. Does the value suggest that there is any relationship between X and Y? Then graph the data. What does this suggest about what should be done before computing the value of the correlation coefficient in Section 1.7?

1.42. Consider the following data for a simple linear regression problem. The sample size was $n = 25$ and when $H_o : \beta_1 = 0$ was tested against $H_a : \beta_1 \neq 0$, it was found that the value of the test statistic was positive and was equal to the critical (tabular) value.

(a) What is the numerical value of the correlation coefficient, assuming that X is a random variable?

(b) What does your answer to part (a) suggest about relying on the outcome of the hypothesis test given in this problem for determining whether or not the regression equation has predictive value?

1.43. It has been said by many that sometimes we discard good data when we should be discarding questionable models. Explain the logic in this.

1.44. Consider a simple linear regression prediction equation written in an alternative way. If $\widehat{\beta}_1 X$ is replaced by $\widehat{\beta}_1 \left(X - \overline{X}\right)$, what will be the numerical value of the intercept if $\sum X = 33$, $\sum Y = 56$, and $n = 20$? Can the numerical value of $\widehat{\beta}_1$ be determined from what is given here, combined with the value of $\widehat{\beta}_0$? Why, or why not? If possible, what is the value?

1.45. For which of the following computations is it necessary to assume that the error term has a normal distribution: regression equation, R^2, prediction interval for Y, or test that the slope parameter is zero.

1.46. In production flow–shop problems, performance is often evaluated by minimum makespan, this being the total elapsed time from starting the first job on the first machine until the last job is completed on the last machine. We might expect that minimum makespan would be linearly related, at least approximately, to the number of jobs. Consider the following data, with X denoting the number of jobs and Y denoting the minimum makespan in hours.

X	3	4	5	6	7	8	9	10	11	12	13
Y	6.50	7.25	8.00	8.50	9.50	10.25	11.50	12.25	13.00	13.75	14.50

(a) From the standpoint of engineering economics, what would a nonlinear relationship signify?
(b) What does a scatter plot of the data suggest about the relationship?
(c) If appropriate, fit a simple linear regression model to the data and estimate the increase in the minimum make span for each additional job. If doing this would be inappropriate, explain why.

1.47. A study was conducted at a large engineering firm to examine the relationship between the number of active projects (X) and the number of man-hours required per week for a graphics project (Y), using the firm's data for the preceding year. An approximately linear relationship was found and a simple linear regression model was fit to the data. The results obtained using $n = 75$ showed that $\widehat{\beta}_0 = 1.2$, $\widehat{\beta}_1 = 3.4$, $\overline{Y} = 9.6$, $S_{xx} = 74.2$, the residual mean square is 26.8, and a 95% prediction interval is desired for $X = 3$.
(a) Notice that the intercept is not zero. Should it be zero? Explain.
(b) Construct the prediction interval.

1.48. Fill in the blanks in the following regression output.

```
Regression Analysis: Volume versus Diameter
The regression equation is
Volume = - 36.9 + 5.07 Diameter

Predictor      Coef    SE Coef         T        P
Constant    -36.943     ______    -10.98    0.000
Diameter     5.0659     0.2474    ______    0.000

S = 4.25199    R-Sq = ______   R-Sq(adj) = 93.3%
```

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	7581.8	_____	419.36	0.000
Residual Error	29	_____	_____		
Total	30	_____			

1.49. Chu (1996) used data from an ad placed by a Singapore retailer of diamond jewelry in the February 29, 1992 issue of the *Straits Times* in illustrating simple linear regression. The objective is to examine the relationship between the prices of diamond rings and the weights of their diamond stones. The data are available at `http://www.amstat.org/publications/jse/datasets/diamond.dat` and the description of the data is given at `http://www.amstat.org/publications/jse/datasets/diamond.txt`.

(a) When the regression equation is fit to the data, the equation is $\widehat{Y} = -259.6 + 3271X$. Thus, as noted by Chu (1996), the intercept is negative but obviously the price can't be. Is this a problem? How would you respond to someone who says that the regression equation is useless because of the negative intercept?

(b) Graph the data and show the regression equation fit to the data. (This can be done in MINITAB by using the FITLINE command.) Does your graph suggest that a model of a different form should be used? Explain.

(c) The value of R^2 is .978. Does this value alone suggest that the regression equation is useful, despite the negative intercept? Explain.

(d) Compare your responses to the discussion in the article, which is available at `http://www.amstat.org/publications/jse/v4n3/datasets.chu.html`. Do you agree with the idea of using an alternative model? Explain.

1.50. Consider the data given in Exercise 1.21. Does the scatter plot of Y against X suggest that a lack-of-fit test would produce a significant result? Perform the test. Does the result conform to the visual impression from the scatter plot? Explain.

1.51. The data in the MINITAB datafile `BALLPARK.MTW` that comes with the software (in the Student14 folders) shows a moderate and significant correlation (.499) between a major league team's winning percentage for all games and its home attendance. Armed with this knowledge, a team's marketing promotion people decide to try to boost home attendance, reasoning that this will improve the team's record. What is wrong with that line of thinking?

1.52. An indication of how σ^2 causes variability in the regression equations can be seen with the following exercise. Consider the data in Table 4.2 and the

values given for $\widehat{\beta}_0$ and $\widehat{\beta}_1$, as well as the value for $\widehat{\sigma}^2$, which was given as 0.0645 in Section 1.5.6.1. Use these as parameter values and generate two sets of Y-values, with $Y = \widehat{\beta}_0 + \widehat{\beta}_1 X + \epsilon$, with $\epsilon \sim NID\,(0, \sqrt{0.0645})$. Plot the "true" equation on an X-Y scatter plot along with the two simulated regression lines and comment. (This can be done in MINITAB by using the OVERLAY subcommand of the PLOT command.)

1.53. The following data, given in the *World Almanac and Book of Facts* (1975) is the average weight in pounds for women aged 30–39 of the indicated height in inches.

Height	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72
Weight	115	117	120	123	126	129	132	135	139	142	146	150	154	159	164

Fit a simple linear regression model with weight as the dependent variable. Is the fit satisfactory? Would you expect it to be satisfactory? Explain. We will return to this dataset in Exercise 2.22.

1.54. The following data are the appraised property values for 2007 and the house square footage for a particular (short) street in northeast Atlanta:

Appraised (thousands)	333.8	308.6	330.6	433.5	317.6	332.9	348.9	329.0	290.5	324.6	318.0	355.0	342.1
Square footage (heated)	2924	2731	2976	3006	2752	2278	2686	2567	2432	2502	2440	2384	2691

Graph the data, using the appraised value as the dependent variable. Would you fit a simple linear regression model based on the graph? Why, or why not. Does the configuration of points in the graph conform to what you would expect considering the nature of the data? Explain.

1.55. There is a dataset for illustrating regression on the Internet that gives height data on a brother–sister pair for each of 11 families. This could hardly be called an example for illustrating regression, however, because R^2 is only .31 when either the brother's height or the sister's height is used as the dependent variable. Judicious use of regression analysis will frequently result only when subject-matter knowledge is employed. Why would we expect to have a poor model fit for this type of example?

1.56. One of the many Java applets for regression and correlation is the one given at `http://www.ruf.rice.edu/~lane/stat_sim/reg_by_eye/index.html`. The user is asked to guess the value of the correlation coefficient from five choices. Do this 10 times. What was your score?

1.57. Two students ("A" and "B") work a simple linear regression homework problem by hand (for practice) and obtain the following set of predictor values:

Y	$\widehat{Y}$(A)	$\widehat{Y}$(B)
1	1.33	1.18
3	2.38	2.27
4	5.55	5.53
6	6.60	6.62
8	6.60	6.62
9	8.71	8.80
11	10.83	10.98

(a) The course instructor decides to test the understanding of the other class members and gives these two solutions, without identifying the students. The instructor proclaims that one of the students has the correct answer and asks the class to determine which student it is and how the determination can be made. Answer these questions.

(b) Then the instructor asks if it can be determined that either student is correct without assuming that one of the students is correct. Answer this question.

1.58. Compute $\widehat{\sigma}^2$ for the correct set of $\widehat{Y}$ values in Exercise 1.57.

1.59. Assume that $R^2 = .75$ pertaining to the prediction equation $\widehat{Y} = 1.76 - 3.27X$. What is the numerical value of r_{xy}, the correlation between X and Y, assuming here that X is a random variable?

1.60. Write R^2_{adjusted} in terms of R^2 for simple linear regression and use that expression to show that R^2_{adjusted} must always be less than R^2.

1.61. Derive $\text{Var}(\widehat{\beta}')$ for the no-intercept model mentioned in Section 1.5.5: $Y = \beta' X + \epsilon$. Should the magnitude of this variance relative to $\text{Var}(\widehat{\beta})$ for the standard model with an intercept be a factor in choosing between the two models? Explain.

1.62. A regression user looks at the following output and declares that R^2_{adjusted} can't have any worth because "everybody knows that a squared quantity cannot be negative". Explain what happened.

```
regress var2 var6
  Source |       SS       df       MS
---------+------------------------------
   Model |  104.041893     1  104.041893
Residual |  35594.826     51  697.937765
---------+------------------------------
   Total | 35698.8679     52  686.516691

Number of obs   =        53
F (1, 51)       =      0.15
Prob > F        =    0.7010
R-squared       =    0.0029
Adj R-squarred  =   -0.0166
Root MSE        =    26.419

------------------------------------------------------------------------------
        var2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        var6 |   .2293197    .593944     0.39   0.701    -.9630727    1.421712
       _cons |    55.7987   35.45303     1.57   0.122    -15.37624    126.9736
------------------------------------------------------------------------------
```