CHAPTER 2

SIMPLE LINEAR REGRESSION

2.1 INTRODUCTION

We start with the simple case of studying the relationship between a response variable Y and a predictor variable X_1. Since we have only one predictor variable, we shall drop the subscript in X_1 and use X for simplicity. We discuss covariance and correlation coefficient as measures of the direction and strength of the linear relationship between the two variables. Simple linear regression model is then formulated and the key theoretical results are given without mathematical derivations, but illustrated by numerical examples. Readers interested in mathematical derivations are referred to the bibliographic notes at the end of the chapter, where books that contain a formal development of regression analysis are listed.

2.2 COVARIANCE AND CORRELATION COEFFICIENT

Suppose we have observations on n subjects consisting of a dependent or response variable Y and an explanatory variable X. The observations are usually recorded as in Table 2.1. We wish to measure both the *direction* and the *strength* of the relationship between Y and X. Two related measures, known as the *covariance* and the *correlation coefficient*, are developed below.

Regression Analysis by Example, Fourth Edition. By Samprit Chatterjee and Ali S. Hadi

Table 2.1 Notation for the Data Used in Simple Regression and Correlation

Observation Number	Response Y	Predictor X
1	y_1	x_1
2	y_2	x_2
$\vdots$	$\vdots$	$\vdots$
n	y_n	x_n

On the scatter plot of Y versus X, let us draw a vertical line at $\bar{x}$ and a horizontal line at $\bar{y}$, as shown in Figure 2.1, where

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \quad \text{and} \quad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \tag{2.1}$$

are the sample mean of Y and X, respectively. The two lines divide the graph into four quadrants. For each point i in the graph, compute the following quantities:

- $y_i - \bar{y}$, the deviation of each observation y_i from the mean of the response variable,
- $x_i - \bar{x}$, the deviation of each observation x_i from the mean of the predictor variable, and
- the product of the above two quantities, $(y_i - \bar{y})(x_i - \bar{x})$.

It is clear from the graph that the quantity $(y_i - \bar{y})$ is positive for every point in the first and second quadrants, and is negative for every point in the third and fourth

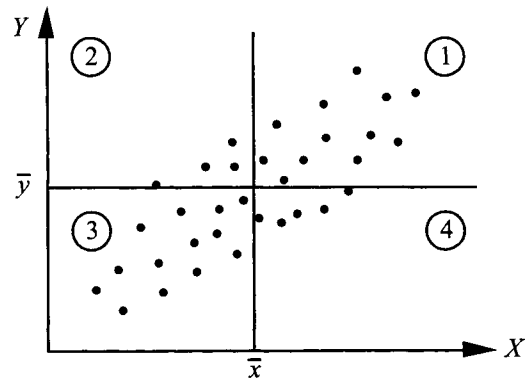

Figure 2.1 A graphical illustration of the correlation coefficient.

quadrants. Similarly, the quantity $(x_i - \bar{x})$ is positive for every point in the first and fourth quadrants, and is negative for every point in the second and third quadrants. These facts are summarized in Table 2.2.

Table 2.2 Algebraic Signs of the Quantities $(y_i - \bar{y})$ and $(x_i - \bar{x})$

Quadrant	$y_i - \bar{y}$	$x_i - \bar{x}$	$(y_i - \bar{y})(x_i - \bar{x})$
1	+	+	+
2	+	−	−
3	−	−	+
4	−	+	−

If the linear relationship between Y and X is positive (as X increases Y also increases), then there are more points in the first and third quadrants than in the second and fourth quadrants. In this case, the sum of the last column in Table 2.2 is likely to be positive because there are more positive than negative quantities. Conversely, if the relationship between Y and X is negative (as X increases Y decreases), then there are more points in the second and fourth quadrants than in the first and third quadrants. Hence the sum of the last column in Table 2.2 is likely to be negative. Therefore, the sign of the quantity

$$\text{Cov}(Y, X) = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{n - 1}, \tag{2.2}$$

which is known as the *covariance* between Y and X, indicates the direction of the linear relationship between Y and X. If $\text{Cov}(Y, X) > 0$, then there is a positive relationship between Y and X, but if $\text{Cov}(Y, X) < 0$, then the relationship is negative. Unfortunately, $\text{Cov}(Y, X)$ does not tell us much about the strength of such a relationship because it is affected by changes in the units of measurement. For example, we would get two different values for the $\text{Cov}(Y, X)$ if we report Y and/or X in terms of thousands of dollars instead of dollars. To avoid this disadvantage of the covariance, we *standardize* the data before computing the covariance. To standardize the Y data, we first subtract the mean from each observation then divide by the standard deviation, that is, we compute

$$z_i = \frac{y_i - \bar{y}}{s_y}, \tag{2.3}$$

where

$$s_y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n - 1}}, \tag{2.4}$$

is the sample *standard deviation* of Y. It can be shown that the standardized variable Z in (2.3) has mean zero and standard deviation one. We standardize X in

a similar way by subtracting the mean $\bar{x}$ from each observation x_i then divide by the standard deviation s_x. The covariance between the standardized X and Y data is known as the *correlation coefficient* between Y and X and is given by

$$\text{Cor}(Y, X) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{y_i - \bar{y}}{s_y} \right) \left(\frac{x_i - \bar{x}}{s_x} \right). \tag{2.5}$$

Equivalent formulas for the correlation coefficient are

$$\text{Cor}(Y, X) = \frac{\text{Cov}(Y, X)}{s_y s_x} \tag{2.6}$$

$$= \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (x_i - \bar{x})^2}}. \tag{2.7}$$

Thus, $\text{Cor}(Y, X)$ can be interpreted either as the covariance between the standardized variables or the ratio of the covariance to the standard deviations of the two variables. From (2.5), it can be seen that the correlation coefficient is symmetric, that is, $\text{Cor}(Y, X) = \text{Cor}(X, Y)$.

Unlike $\text{Cov}(Y, X)$, $\text{Cor}(Y, X)$ is scale invariant, that is, it does not change if we change the units of measurements. Furthermore, $\text{Cor}(Y, X)$ satisfies

$$-1 \leq \text{Cor}(Y, X) \leq 1. \tag{2.8}$$

These properties make the $\text{Cor}(Y, X)$ a useful quantity for measuring both the direction and the strength of the relationship between Y and X. The magnitude of $\text{Cor}(Y, X)$ measures the strength of the linear relationship between Y and X. The closer $\text{Cor}(Y, X)$ is to 1 or -1, the stronger is the relationship between Y and X. The sign of $\text{Cor}(Y, X)$ indicates the direction of the relationship between Y and X. That is, $\text{Cor}(Y, X) > 0$ implies that Y and X are positively related. Conversely, $\text{Cor}(Y, X) < 0$, implies that Y and X are negatively related.

Note, however, that $\text{Cor}(Y, X) = 0$ does not necessarily mean that Y and X are not related. It only implies that they are not linearly related because the correlation coefficient measures only *linear* relationships. In other words, the $\text{Cor}(Y, X)$ can still be zero when Y and X are nonlinearly related. For example, Y and X in Table 2.3 have the perfect nonlinear relationship $Y = 50 - X^2$ (graphed in Figure 2.2), yet $\text{Cor}(Y, X) = 0$.

Furthermore, like many other summary statistics, the $\text{Cor}(Y, X)$ can be substantially influenced by one or few outliers in the data. To emphasize this point, Anscombe (1973) has constructed four data sets, known as Anscombe's quartet, each with a distinct pattern, but each having the same set of summary statistics (e.g., the same value of the correlation coefficient). The data and graphs are reproduced in Table 2.4 and Figure 2.3. The data can be found in the book's Web site.[1] An analysis based exclusively on an examination of summary statistics, such as the correlation coefficient, would have been unable to detect the differences in patterns.

[1] http://www.ilr.cornell.edu/~hadi/RABE4

Table 2.3 A Data Set With a Perfect Nonlinear Relationship Between Y and X, Yet $\text{Cor}(X, Y) = 0$

Y	X	Y	X	Y	X
1	−7	46	−2	41	3
14	−6	49	−1	34	4
25	−5	50	0	25	5
34	−4	49	1	14	6
41	−3	46	2	1	7

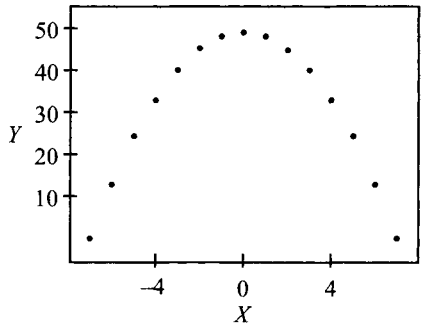

Figure 2.2 A scatter plot of Y versus X in Table 2.3.

Table 2.4 Anscombe's Quartet: Four Data Sets Having Same Values of Summary Statistics

Y_1	X_1	Y_2	X_2	Y_3	X_3	Y_4	X_4
8.04	10	9.14	10	7.46	10	6.58	8
6.95	8	8.14	8	6.77	8	5.76	8
7.58	13	8.74	13	12.74	13	7.71	8
8.81	9	8.77	9	7.11	9	8.84	8
8.33	11	9.26	11	7.81	11	8.47	8
9.96	14	8.10	14	8.84	14	7.04	8
7.24	6	6.13	6	6.08	6	5.25	8
4.26	4	3.10	4	5.39	4	12.50	19
10.84	12	9.13	12	8.15	12	5.56	8
4.82	7	7.26	7	6.42	7	7.91	8
5.68	5	4.74	5	5.73	5	6.89	8

Source: Anscombe (1973).

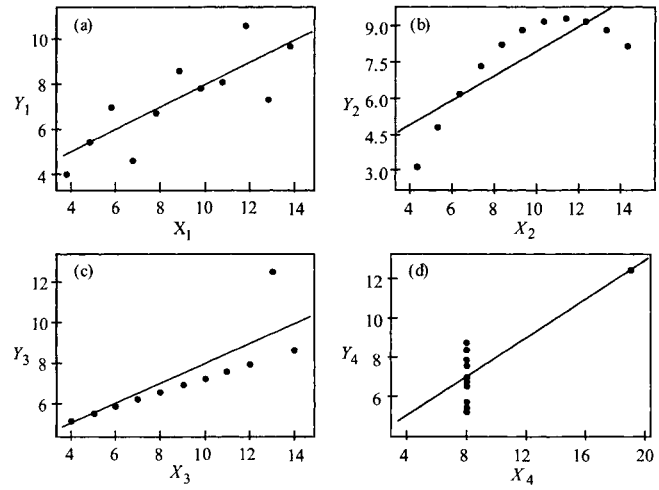

Figure 2.3 Scatter plots of the data in Table 2.4 with the fitted lines.

An examination of Figure 2.3 shows that only the first set, whose plot is given in (a), can be described by a linear model. The plot in (b) shows the second data set is distinctly nonlinear and would be better fitted by a quadratic function. The plot in (c) shows that the third data set has one point that distorts the slope and the intercept of the fitted line. The plot in (d) shows that the fourth data set is unsuitable for linear fitting, the fitted line being determined essentially by one extreme observation. Therefore, it is important to examine the scatter plot of Y versus X before interpreting the numerical value of $\text{Cor}(Y, X)$.

2.3 EXAMPLE: COMPUTER REPAIR DATA

As an illustrative example, consider a case of a company that markets and repairs small computers. To study the relationship between the length of a service call and the number of electronic components in the computer that must be repaired or replaced, a sample of records on service calls was taken. The data consist of the length of service calls in minutes (the response variable) and the number of components repaired (the predictor variable). The data are presented in Table 2.5. The Computer Repair data can also be found in the book's Web site. We use this data set throughout this chapter as an illustrative example. The quantities needed to compute $\bar{y}$, $\bar{x}$, $\text{Cov}(Y, X)$, and $\text{Cor}(Y, X)$ are shown in Table 2.6. We have

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{1361}{14} = 97.21 \text{ and } \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{84}{14} = 6,$$

Table 2.5 Length of Service Calls (in Minutes) and Number of Units Repaired

Row	Minutes	Units	Row	Minutes	Units
1	23	1	8	97	6
2	29	2	9	109	7
3	49	3	10	119	8
4	64	4	11	149	9
5	74	4	12	145	9
6	87	5	13	154	10
7	96	6	14	166	10

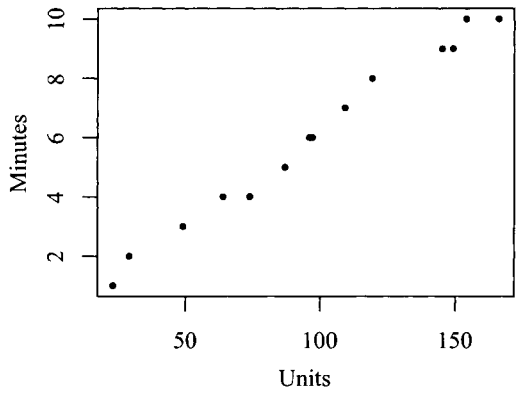

Figure 2.4 Computer Repair data: Scatter plot of Minutes versus Units.

$$\text{Cov}(Y, X) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{n-1} = \frac{1768}{13} = 136,$$

and

$$\text{Cor}(Y, X) = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum(y_i - \bar{y})^2 \sum(x_i - \bar{x})^2}} = \frac{1768}{\sqrt{27768.36 \times 114}} = 0.996.$$

Before drawing conclusions from this value of $\text{Cor}(Y, X)$, we should examine the corresponding scatter plot of Y versus X. This plot is given in Figure 2.4. The high value of $\text{Cor}(Y, X) = 0.996$ is consistent with the strong linear relationship between Y and X exhibited in Figure 2.4. We therefore conclude that there is a strong positive relationship between repair time and units repaired.

Although $\text{Cor}(Y, X)$ is a useful quantity for measuring the direction and the strength of linear relationships, it cannot be used for prediction purposes, that is, we cannot use $\text{Cor}(Y, X)$ to predict the value of one variable given the value of the other. Furthermore, $\text{Cor}(Y, X)$ measures only pairwise relationships. Regression analysis, however, can be used to relate one or more response variable to one or more predictor variables. It can also be used in prediction. Regression analysis

Table 2.6 Quantities Needed for the Computation of the Correlation Coefficient Between the Length of Service Calls, Y, and Number of Units Repaired, X

i	y_i	x_i	$(y_i - \bar{y})$	$(x_i - \bar{x})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})(x_i - \bar{x})$
1	23	1	−74.21	−5	5507.76	25	371.07
2	29	2	−68.21	−4	4653.19	16	272.86
3	49	3	−48.21	−3	2324.62	9	144.64
4	64	4	−33.21	−2	1103.19	4	66.43
5	74	4	−23.21	−2	538.90	4	46.43
6	87	5	−10.21	−1	104.33	1	10.21
7	96	6	−1.21	0	1.47	0	0.00
8	97	6	−0.21	0	0.05	0	0.00
9	109	7	11.79	1	138.90	1	11.79
10	119	8	21.79	2	474.62	4	43.57
11	149	9	51.79	3	2681.76	9	155.36
12	145	9	47.79	3	2283.47	9	143.36
13	154	10	56.79	4	3224.62	16	227.14
14	166	10	68.79	4	4731.47	16	275.14
Total	1361	84	0.00	0	27768.36	114	1768.00

is an attractive extension to correlation analysis because it postulates a model that can be used not only to measure the direction and the strength of a relationship between the response and predictor variables, but also to numerically describe that relationship. We discuss simple linear regression models in the rest of this chapter. Chapter 3 is devoted to multiple regression models.

2.4 THE SIMPLE LINEAR REGRESSION MODEL

The relationship between a response variable Y and a predictor variable X is postulated as a linear[2] model

$$Y = \beta_0 + \beta_1 X + \varepsilon, \tag{2.9}$$

where β_0 and β_1, are constants called the *model regression coefficients* or *parameters*, and ε is a random disturbance or error. It is assumed that in the range of the observations studied, the linear equation (2.9) provides an acceptable approximation to the true relation between Y and X. In other words, Y is approximately a linear function of X, and ε measures the discrepancy in that approximation.

[2]The adjective *linear* has a dual role here. It may be taken to describe the fact that the relationship between Y and X is linear. More generally, the word *linear* refers to the fact that the regression parameters, β_0 and β_1, enter (2.9) in a linear fashion. Thus, for example, $Y = \beta_0 + \beta_1 X^2 + \varepsilon$ is also a linear model even though the relationship between Y and X is quadratic.

In particular ε contains no systematic information for determining Y that is not already captured in X. The coefficient β_1, called the *slope*, may be interpreted as the change in Y for unit change in X. The coefficient β_0, called the *constant* coefficient or *intercept*, is the predicted value of Y when $X = 0$.

According to (2.9), each observation in Table 2.1 can be written as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{2.10}$$

where y_i represents the ith value of the response variable Y, x_i represents the ith value of the predictor variable X, and ε_i represents the error in the approximation of y_i.

Regression analysis differs in an important way from correlation analysis. The correlation coefficient is symmetric in the sense that $\text{Cor}(Y, X)$ is the same as $\text{Cor}(X, Y)$. The variables X and Y are of equal importance. In regression analysis the response variable Y is of primary importance. The importance of the predictor X lies on its ability to account for the variability of the response variable Y and not in itself per se. Hence Y is of primary importance.

Returning to the Computer Repair Data example, suppose that the company wants to forecast the number of service engineers that will be required over the next few years. A linear model,

$$\text{Minutes} = \beta_0 + \beta_1 \cdot \text{Units} + \varepsilon, \tag{2.11}$$

is assumed to represent the relationship between the length of service calls and the number of electronic components in the computer that must be repaired or replaced. To validate this assumption, we examine the graph of the response variable versus the explanatory variable. This graph, shown in Figure 2.4, suggests that the straight line relationship in (2.11) is a reasonable assumption.

2.5 PARAMETER ESTIMATION

Based on the available data, we wish to estimate the parameters β_0 and β_1. This is equivalent to finding the straight line that gives the *best fit* (representation) of the points in the scatter plot of the response versus the predictor variable (see Figure 2.4). We estimate the parameters using the popular *least squares method*, which gives the line that minimizes the sum of squares of the *vertical distances*[3] from each point to the line. The vertical distances represent the errors in the response variable. These errors can be obtained by rewriting (2.10) as

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i, \quad i = 1, 2, \ldots, n. \tag{2.12}$$

The sum of squares of these distances can then be written as

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2. \tag{2.13}$$

[3] An alternative to the vertical distance is the *perpendicular* (shortest) distance from each point to the line. The resultant line is called the *orthogonal regression* line.

The values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize $S(\beta_0, \beta_1)$ are given by

$$\hat{\beta}_1 = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} \tag{2.14}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{2.15}$$

Note that we give the formula for $\hat{\beta}_1$ before the formula for $\hat{\beta}_0$ because $\hat{\beta}_0$ uses $\hat{\beta}_1$. The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are called the least squares estimates of β_0 and β_1 because they are the solution to the *least squares method*, the intercept and the slope of the line that has the smallest possible sum of squares of the vertical distances from each point to the line. For this reason, the line is called the *least squares regression line*. The least squares regression line is given by

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X. \tag{2.16}$$

Note that a least squares line always exists because we can always find a line that gives the minimum sum of squares of the vertical distances. In fact, as we shall see later, in some cases a least squares line may not be unique. These cases are not common in practice.

For each observation in our data we can compute

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, 2, \ldots, n. \tag{2.17}$$

These are called the *fitted* values. Thus, the ith fitted value, $\hat{y}_i$, is the point on the least squares regression line (2.16) corresponding to x_i. The vertical distance corresponding to the ith observation is

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \ldots, n. \tag{2.18}$$

These vertical distances are called the *ordinary*[4] *least squares residuals*. One properties of the residuals in (2.18) is that their sum is zero (see Exercise 2.5(a)). This means that the sum of the distances above the line is equal to the sum of the distances below the line.

Using the Computer Repair data and the quantities in Table 2.6, we have

$$\hat{\beta}_1 = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} = \frac{1768}{114} = 15.509,$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 97.21 - 15.509 \times 6 = 4.162.$$

Then the equation of the least squares regression line is

$$\text{Minutes} = 4.162 + 15.509 \cdot \text{Units}. \tag{2.19}$$

[4]To be distinguished from other types of residuals to be presented later.

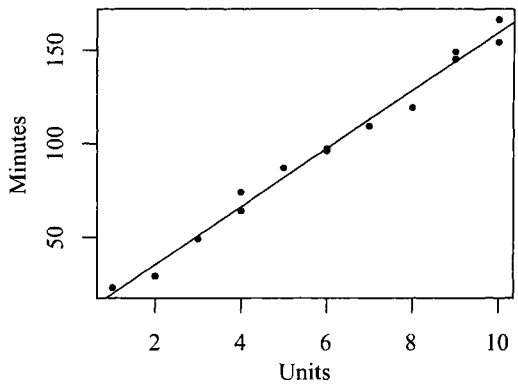

Figure 2.5 Plot of Minutes versus Units with the fitted least squares regression line.

This least squares line is shown together with the scatter plot of Minutes versus Units in Figure 2.5. The fitted values in (2.17) and the residuals in (2.18) are shown in Table 2.7.

The coefficients in (2.19) can be interpreted in physical terms. The constant term represents the setup or startup time for each repair and is approximately 4 minutes. The coefficient of Units represents the increase in the length of a service call for each additional component that has to be repaired. From the data given, we estimate that it takes about 16 minutes (15.509) for each additional component that has to be repaired. For example, the length of a service call in which four components had to be repaired is obtained by substituting Units = 4 in the equation of the regression line (2.19) and obtaining $\hat{y} = 4.162 + 15.509 \times 4 = 66.20$. Since Units = 4, corresponds to two observations in our data set (observations 4 and 5), the value 66.198 is the fitted value for both observations 4 and 5, as can be seen from Table 2.7. Note, however, that since observations 4 and 5 have different values for the response variable Minutes, they have different residuals.

We should note here that by comparing (2.2), (2.7), and (2.14), an alternative formula for $\hat{\beta}_1$ can be expressed as

$$\hat{\beta}_1 = \frac{\text{Cov}(Y, X)}{Var(X)} = \text{Cor}(Y, X)\frac{s_y}{s_x}, \tag{2.20}$$

from which it can be seen that $\hat{\beta}_1$, $\text{Cov}(Y, X)$, and $\text{Cor}(Y, X)$ have the same sign. This makes intuitive sense because positive (negative) slope means positive (negative) correlation.

So far in our analysis we have made only one assumption, namely, that Y and X are linearly related. This assumption is referred to as the *linearity* assumption. This is merely an assumption or a hypothesis about the relationship between the response and predictor variables. An early step in the analysis should always be the validation of this assumption. We wish to determine if the data at hand support

Table 2.7 The Fitted Values, $\hat{y}_i$, and the Ordinary Least Squares Residuals, e_i, for the Computer Repair Data

i	x_i	y_i	$\hat{y}_i$	e_i	i	x_i	y_i	$\hat{y}_i$	e_i
1	1	23	19.67	3.33	8	6	97	97.21	−0.21
2	2	29	35.18	−6.18	9	7	109	112.72	−3.72
3	3	49	50.69	−1.69	10	8	119	128.23	−9.23
4	4	64	66.20	−2.20	11	9	149	143.74	5.26
5	4	74	66.20	7.80	12	9	145	143.74	1.26
6	5	87	81.71	5.29	13	10	154	159.25	−5.25
7	6	96	97.21	−1.21	14	10	166	159.25	6.75

the assumption that Y and X are linearly related. An informal way to check this assumption is to examine the scatter plot of the response versus the predictor variable, preferably drawn with the least squares line superimposed on the graph (see Figure 2.5). If we observe a nonlinear pattern, we will have to take corrective action. For example, we may *re-express* or *transform* the data before we continue the analysis. *Data transformation* is discussed in Chapter 6.

If the scatter of points resemble a straight line, then we conclude that the linearity assumption is reasonable and continue with our analysis. The least squares estimators have several desirable properties when some additional assumptions hold. The required assumptions are stated in Chapter 4. The validity of these assumptions must be checked before meaningful conclusions can be reached from the analysis. Chapter 4 also presents methods for the validation of these assumptions. Using the properties of least squares estimators, one can develop statistical inference procedures (e.g., confidence interval estimation, tests of hypothesis, and goodness-of-fit tests). These are presented in Sections 2.6 to 2.9.

2.6 TESTS OF HYPOTHESES

As stated earlier, the usefulness of X as a predictor of Y can be measured informally by examining the correlation coefficient and the corresponding scatter plot of Y versus X. A more formal way of measuring the usefulness of X as a predictor of Y is to conduct a test of hypothesis about the regression parameter β_1. Note that the hypothesis $\beta_1 = 0$ means that there is no linear relationship between Y and X. A test of this hypothesis requires the following assumption. For every fixed value of X, the ε's are assumed to be independent random quantities normally distributed with mean zero and a common variance σ^2. With these assumptions, the quantities,

$\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased[5] estimates of β_0 and β_1, respectively. Their variances are

$$Var(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right], \tag{2.21}$$

and

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} . \tag{2.22}$$

Furthermore, the *sampling distributions* of the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are normal with means β_0 and β_1 and variance as given in (2.21) and (2.22), respectively.

The variances of $\hat{\beta}_0$ and $\hat{\beta}_1$ depend on the unknown parameter σ^2. So, we need to estimate σ^2 from the data. An unbiased estimate of σ^2 is given by

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{\text{SSE}}{n-2} , \tag{2.23}$$

where SSE is the sum of squares of the residuals (errors). The number $n - 2$ in the denominator of (2.23) is called the *degrees of freedom* (*df*). It is equal to the number of observations minus the number of estimated regression coefficients.

Replacing σ^2 in (2.21) and (2.22) by $\hat{\sigma}^2$ in (2.23), we get unbiased estimates of the variances of $\hat{\beta}_0$ and $\hat{\beta}_1$. An *estimate* of the *standard deviation* is called the *standard error* (s.e.) of the estimate. Thus, the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$ are

$$\text{s.e.}(\hat{\beta}_0) = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}} \tag{2.24}$$

and

$$\text{s.e.}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum (x_i - \bar{x})^2}} , \tag{2.25}$$

respectively, where $\hat{\sigma}$ is the square root of $\hat{\sigma}^2$ in (2.23). The standard errors of $\hat{\beta}_1$ is a measure of how precisely the slope has been estimated. The smaller the standard error the more precise the estimator.

With the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$, we are now in position to perform statistical analysis concerning the usefulness of X as a predictor of Y. Under the normality assumption, an appropriate test statistic for testing the null hypothesis $H_0 : \beta_1 = 0$ against the alternative $H_1 : \beta_1 \neq 0$ is the t-test,

$$t_1 = \frac{\hat{\beta}_1}{\text{s.e.}(\hat{\beta}_1)} . \tag{2.26}$$

The statistic t_1 is distributed as a Student's t with $(n - 2)$ degrees of freedom. The test is carried out by comparing this observed value with the appropriate critical

[5] An estimate $\hat{\theta}$ is said to be an unbiased estimate of a parameter θ if the expected value of $\hat{\theta}$ is equal to θ.

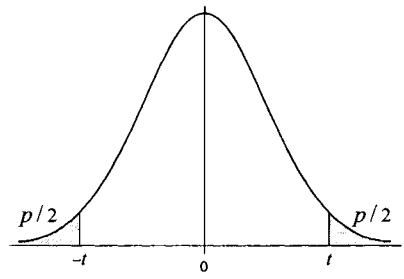

Figure 2.6 A graph of the probability density function of a t-distribution. The p-value for the t-test is the shaded areas under the curve.

value obtained from the t-table given in the Appendix to this book (see Table A.2), which is $t_{(n-2,\alpha/2)}$, where α is a specified significance level. Note that we divide α by 2 because we have a two-sided alternative hypothesis. Accordingly, H_0 is to be rejected at the significance level α if

$$|t_1| \geq t_{(n-2,\alpha/2)}, \tag{2.27}$$

where $|t_1|$ denotes the absolute value of t_1. A criterion equivalent to that in (2.27) is to compare the p-value for the t-test with α and reject H_0 if

$$p(|t_1|) \leq \alpha, \tag{2.28}$$

where $p(|t_1|)$, called the *p-value*, is the probability that a random variable having a Student t distribution with $(n-2)$ is greater than $|t_1|$ (the absolute value of the observed value of the t-test). Figure 2.6 is a graph of the density function of a t-distribution. The p-value is the sum of the two shaded areas under the curve. The p-value is usually computed and supplied as part of the regression output by statistical packages. Note that the rejection of $H_0 : \beta_1 = 0$ would mean that β_1 is likely to be different from 0, and hence the predictor variable X is a statistically significant predictor of the response variable Y.

To complete the picture of hypotheses testing regarding regression parameters, we give here tests for three other hypotheses that may arise in practice.

Testing H_0: $\beta_1 = \beta_1^0$

The above t-test can be generalized to test the more general hypothesis $H_0 : \beta_1 = \beta_1^0$, where β_1^0 is a constant chosen by the investigator, against the two-sided alternative $H_1 : \beta_1 \neq \beta_1^0$. The appropriate test statistic in this case is the t-test,

$$t_1 = \frac{\hat{\beta}_1 - \beta_1^0}{\text{s.e.}(\hat{\beta}_1)}. \tag{2.29}$$

Note that when $\beta_1^0 = 0$, the t-test in (2.29) reduces to the t-test in (2.26). The statistic t_1 in (2.29) is also distributed as a Student's t with $(n-2)$ degrees of

freedom. Thus, $H_0 : \beta_1 = \beta_1^0$ is rejected if (2.27) holds (or, equivalently, if (2.28) holds).

For illustration, using the Computer Repair data, let us suppose that the management expected the increase in service time for each additional unit to be repaired to be 12 minutes. Do the data support this conjecture? The answer may be obtained by testing $H_0 : \beta_1 = 12$ against $H_1 : \beta_1 \neq 12$. The appropriate statistic is

$$t_1 = \frac{\hat{\beta}_1 - 12}{\text{s.e.}(\hat{\beta}_1)} = \frac{15.509 - 12}{0.505} = 6.948,$$

with 12 degrees of freedom. The critical value for this test is $t_{(n-2,\alpha/2)} = t_{(12,0.025)} = 2.18$. Since $t_1 = 6.948 > 2.18$, the result is highly significant, leading to the rejection of the null hypothesis. The management's estimate of the increase in time for each additional component to be repaired is not supported by the data. Their estimate is too low.

Testing H_0: $\beta_0 = \beta_0^0$

The need for testing hypotheses regarding the regression parameter β_0 may also arise in practice. More specifically, suppose we wish to test $H_0 : \beta_0 = \beta_0^0$ against the alternative $H_1 : \beta_0 \neq \beta_0^0$, where β_0^0 is a constant chosen by the investigator. The appropriate test in this case is given by

$$t_0 = \frac{\hat{\beta}_0 - \beta_0^0}{\text{s.e.}(\hat{\beta}_0)} . \tag{2.30}$$

If we set $\beta_0^0 = 0$, a special case of this test is obtained as

$$t_0 = \frac{\hat{\beta}_0}{\text{s.e.}(\hat{\beta}_0)} , \tag{2.31}$$

which tests $H_0 : \beta_0 = 0$ against the alternative $H_1 : \beta_0 \neq 0$.

The least squares estimates of the regression coefficients, their standard errors, the t-tests for testing that the corresponding coefficient is zero, and the p-values are usually given as part of the regression output by statistical packages. These values are usually displayed in a table such as the one in Table 2.8. This table is known as the *coefficients table*. To facilitate the connection between a value in the table and the formula used to obtain it, the equation number of the formula is given in parentheses.

As an illustrative example, Table 2.9 shows a part of the regression output for the Computer Repair data in Table 2.5. Thus, for example, $\hat{\beta}_1 = 15.509$, the s.e.$(\hat{\beta}_1) = 0.505$, and hence $t_1 = 15.509/0.505 = 30.71$. The critical value for this test using $\alpha = 0.05$, for example, is $t_{(12,0.025)} = 2.18$. The $t_1 = 30.71$ is much larger than its critical value 2.18. Consequently, according to (2.27), $H_0 : \beta_1 = 0$ is

Table 2.8 A Standard Regression Output. The Equation Number of the Corresponding Formulas are Given in Parentheses

Variable	Coefficient (Formula)	s.e. (Formula)	t-test (Formula)	p-value
Constant	$\hat{\beta}_0$ (2.15)	s.e.$(\hat{\beta}_0)$ (2.24)	t_0 (2.31)	p_0
X	$\hat{\beta}_1$ (2.14)	s.e.$(\hat{\beta}_1)$ (2.25)	t_1 (2.26)	p_1

Table 2.9 Regression Output for the Computer Repair Data

Variable	Coefficient	s.e.	t-test	p-value
Constant	4.162	3.355	1.24	0.2385
Units	15.509	0.505	30.71	< 0.0001

rejected, which means that the predictor variable Units is a statistically significant predictor of the response variable Minutes. This conclusion can also be reached using (2.28) by observing that the p-value ($p_1 < 0.0001$) is much less than $\alpha = 0.05$ indicating very high significance.

A Test Using Correlation Coefficient

As mentioned above, a test of $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ can be thought of as a test for determining whether the response and the predictor variables are linearly related. We used the t-test in (2.26) to test this hypothesis. An alternative test, which involves the correlation coefficient between Y and X, can be developed. Suppose that the population correlation coefficient between Y and X is denoted by ρ. If $\rho \neq 0$, then Y and X are linearly related. An appropriate test for testing $H_0 : \rho = 0$ against $H_1 : \rho \neq 0$ is given by

$$t_1 = \frac{\text{Cor}(Y, X)\sqrt{n-2}}{\sqrt{1 - [\text{Cor}(Y, X)]^2}}, \tag{2.32}$$

where $\text{Cor}(Y, X)$ is the sample correlation coefficient between Y and X, defined in (2.6), which is considered here to be an estimate of ρ. The t-test in (2.32) is distributed as a Student's t with $(n-2)$ degrees of freedom. Thus, $H_0 : \rho = 0$ is rejected if (2.27) holds (or, equivalently, if (2.28) holds). Again if $H_0 : \rho = 0$ is rejected, it means that there is a statistically significant linear relationship between Y and X.

It is clear that if no linear relationship exists between Y and X, then $\beta_1 = 0$. Consequently, the statistical tests for $H_0 : \beta_1 = 0$ and $H_0 : \rho = 0$ should be identical. Although the statistics for testing these hypotheses given in (2.26) and (2.32) look different, it can be demonstrated that they are indeed algebraically equivalent.

2.7 CONFIDENCE INTERVALS

To construct confidence intervals for the regression parameters, we also need to assume that the ε's have a normal distribution, which will enable us to conclude that the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ are normal, as discussed in Section 2.6. Consequently, the $(1-\alpha)\times 100\%$ confidence interval for β_0 is given by

$$\hat{\beta}_0 \pm t_{(n-2,\alpha/2)} \times \text{s.e.}(\hat{\beta}_0), \tag{2.33}$$

where $t_{(n-2,\alpha/2)}$ is the $(1-\alpha/2)$ percentile of a t distribution with $(n-2)$ degrees of freedom. Similarly, limits of the $(1-\alpha)\times 100\%$ confidence interval for β_1 are given by

$$\hat{\beta}_1 \pm t_{(n-2,\alpha/2)} \times \text{s.e.}(\hat{\beta}_1). \tag{2.34}$$

The confidence interval in (2.34) has the usual interpretation, namely, if we were to take repeated samples of the same size at the same values of X and construct for example 95% confidence intervals for the slope parameter for each sample, then 95% of these intervals would be expected to contain the true value of the slope.

From Table 2.9 we see that a 95% confidence interval for β_1 is

$$15.509 \pm 2.18 \times 0.505 = (14.408, 16.610). \tag{2.35}$$

That is, the incremental time required for each broken unit is between 14 and 17 minutes. The calculation of confidence interval for β_0 in this example is left as an exercise for the reader.

Note that the confidence limits in (2.33) and (2.34) are constructed for each of the parameters β_0 and β_1, separately. This does not mean that a simultaneous (joint) confidence region for the two parameters is rectangular. Actually, the simultaneous confidence region is elliptical. This region is given for the general case of multiple regression in the Appendix to Chapter 3 in (A.15), of which the simultaneous confidence region for β_0 and β_1 is a special case.

2.8 PREDICTIONS

The fitted regression equation can be used for prediction. We distinguish between two types of predictions:

1. The prediction of the value of the response variable Y which corresponds to any chosen value, x_0, of the predictor variable, or

2. The estimation of the mean response μ_0, when $X = x_0$.

For the first case, the predicted value $\hat{y}_0$ is

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0. \tag{2.36}$$

The standard error of this prediction is

$$\text{s.e.}(\hat{y}_0) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}\,. \tag{2.37}$$

Hence, the confidence limits for the predicted value with confidence coefficient $(1 - \alpha)$ are given by

$$\hat{y}_0 \pm t_{(n-2,\alpha/2)}\ \text{s.e.}(\hat{y}_0). \tag{2.38}$$

For the second case, the mean response μ_0 is estimated by

$$\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0. \tag{2.39}$$

The standard error of this estimate is

$$\text{s.e.}(\hat{\mu}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}\,, \tag{2.40}$$

from which it follows that the confidence limits for μ_0 with confidence coefficient $(1 - \alpha)$ are given by

$$\hat{\mu}_0 \pm t_{(n-2,\alpha/2)}\ \text{s.e.}(\hat{\mu}_0). \tag{2.41}$$

Note that the point estimate of μ_0 is identical to the predicted response $\hat{y}_0$. This can be seen by comparing (2.36) with (2.39). The standard error of $\hat{\mu}_0$ is, however, smaller than the standard error of $\hat{y}_0$ and can be seen by comparing (2.37) with (2.40). Intuitively, this makes sense. There is greater uncertainty (variability) in predicting one observation (the next observation) than in estimating the mean response when $X = x_0$. The averaging that is implied in the mean response reduces the variability and uncertainty associated with the estimate.

To distinguish between the limits in (2.38) and (2.41), the limits in (2.38) are sometimes referred to as the *prediction* or *forecast* limits, whereas the limits given in (2.41) are called the *confidence limits.*

Suppose that we wish to predict the length of a service call in which four components had to be repaired. If $\hat{y}_4$ denotes the predicted value, then from (2.36) we get

$$\hat{y}_4 = 4.162 + 15.509 \times 4 = 66.20,$$

with a standard error that is obtained from (2.37) as

$$\text{s.e.}(\hat{y}_4) = 5.392\sqrt{1 + \frac{1}{14} + \frac{(4-6)^2}{114}} = 5.67.$$

On the other hand, if the service department wishes to estimate the expected (mean) service time for a call that needed four components repaired, we would use (2.39) and (2.40), respectively. Denoting by μ_4, the expected service time for a call that needed four components to be repaired, we have:

$$\hat{\mu}_4 = 4.162 + 15.509 \times 4 = 66.20,$$

with a standard error

$$\text{s.e.}(\hat{\mu}_4) = 5.392\sqrt{\frac{1}{14} + \frac{(4-6)^2}{114}} = 1.76.$$

With these standard errors we can construct confidence intervals using (2.38) and (2.41), as appropriate.

As can be seen from (2.37), the standard error of prediction increases the farther the value of the predictor variable is from the center of the actual observations. Care should be taken when predicting the value of Minutes corresponding to a value for Units that does not lie close to the observed data. There are two dangers in such predictions. First, there is substantial uncertainty due to the large standard error. More important, the linear relationship that has been estimated may not hold outside the range of observations. Therefore, care should be taken in employing fitted regression lines for prediction far outside the range of observations. In our example we would not use the fitted equation to predict the service time for a service call which requires that 25 components be replaced or repaired. This value lies too far outside the existing range of observations.

2.9 MEASURING THE QUALITY OF FIT

After fitting a linear model relating Y to X, we are interested not only in knowing whether a linear relationship exits, but also in measuring the quality of the fit of the model to the data. The quality of the fit can be assessed by one of the following highly related (hence, somewhat redundant) ways:

1. When using the tests in (2.26) or (2.32), if H_0 is rejected, the magnitude of the values of the test (or the corresponding p-values) gives us information about the *strength* (not just the existence) of the linear relationship between Y and X. Basically, the larger the t (in absolute value) or the smaller the corresponding p-value, the stronger the linear relationship between Y and X. These tests are objective but they require all the assumptions stated earlier, specially the assumption of normality of the ε's.

2. The strength of the linear relationship between Y and X can also be assessed directly from the examination of the scatter plot of Y versus X together with the corresponding value of the correlation coefficient $\text{Cor}(Y, X)$ in (2.6). The closer the set of points to a straight line (the closer $\text{Cor}(Y, X)$ to 1 or -1), the stronger the linear relationship between Y and X. This approach is informal and subjective but it requires only the linearity assumption.

3. Examine the scatter plot of Y versus $\hat{Y}$. The closer the set of points to a straight line, the stronger the linear relationship between Y and X. One can measure the strength of the linear relationship in this graph by computing the

correlation coefficient between Y and $\hat{Y}$, which is given by

$$\text{Cor}(Y, \hat{Y}) = \frac{\sum(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum(y_i - \bar{y})^2 \sum(\hat{y}_i - \bar{\hat{y}})^2}}, \tag{2.42}$$

where $\bar{y}$ is the mean of the response variable Y and $\bar{\hat{y}}$ is the mean of the fitted values. In fact, the scatter plot of Y versus X and the scatter plot of Y versus $\hat{Y}$ are redundant because the patterns of points in the two graphs are identical. The two corresponding values of the correlation coefficient are related by the following equation:

$$\text{Cor}(Y, \hat{Y}) = |\text{Cor}(Y, X)|. \tag{2.43}$$

Note that $\text{Cor}(Y, \hat{Y})$ cannot be negative (why?), but $\text{Cor}(Y, X)$ can be positive or negative ($-1 \leq \text{Cor}(Y, X) \leq 1$). Therefore, in simple linear regression, the scatter plot of Y versus $\hat{Y}$ is redundant. However, in multiple regression, the scatter plot of Y versus $\hat{Y}$ is not redundant. The graph is very useful because, as we shall see in Chapter 3, it is used to assess the strength of the relationship between Y and the set of predictor variables $X_1, X_2, \ldots, X_p$.

4. Although scatter plots of Y versus $\hat{Y}$ and $\text{Cor}(Y, \hat{Y})$ are redundant in simple linear regression, they give us an indication of the quality of the fit in both simple and multiple regression. Furthermore, in both simple and multiple regressions, $\text{Cor}(Y, \hat{Y})$ is related to another useful measure of the quality of fit of the linear model to the observed data. This measure is developed as follows. After we compute the least squares estimates of the parameters of a linear model, let us compute the following quantities:

$$\begin{aligned} \text{SST} &= \sum(y_i - \bar{y})^2, \\ \text{SSR} &= \sum(\hat{y}_i - \bar{y})^2, \\ \text{SSE} &= \sum(y_i - \hat{y}_i)^2, \end{aligned} \tag{2.44}$$

where SST stands for the total sum of squared deviations in Y from its mean $\bar{y}$, SSR denotes the sum of squares due to regression, and SSE represents the sum of squared residuals (errors). The quantities $(\hat{y}_i - \bar{y})$, $(\hat{y}_i - \bar{y})$, and $(y_i - \hat{y}_i)$ are depicted in Figure 2.7 for a typical point (x_i, y_i). The line $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is the fitted regression line based on all data points (not shown on the graph) and the horizontal line is drawn at $Y = \bar{y}$. Note that for every point (x_i, y_i), there are two points, $(x_i, \hat{y}_i)$, which lies on the fitted line, and $(x_i, \bar{y})$ which lies on the line $Y = \bar{y}$.

A fundamental equality, in both simple and multiple regressions, is given by

$$\text{SST} = \text{SSR} + \text{SSE}. \tag{2.45}$$

This equation arises from the description of an observation as

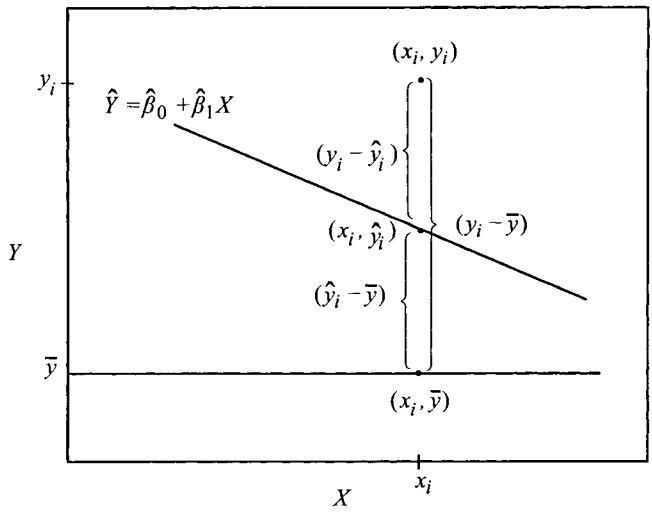

Figure 2.7 A graphical illustration of various quantities computed after fitting a regression line to data.

$$\begin{array}{ccccc} y_i & = & \hat{y}_i & + & (y_i - \hat{y}_i) \\ \text{Observed} & = & \text{Fit} & + & \text{Deviation from fit.} \end{array}$$

Subtracting $\bar{y}$ from both sides, we obtain

$$\begin{array}{ccccc} y_i - \bar{y} & = & (\hat{y}_i - \bar{y}) & + & (y_i - \hat{y}_i) \\ \text{Deviation from mean} & = & \text{Deviation due to fit} & + & \text{Residual.} \end{array}$$

Accordingly, the total sum of squared deviations in Y can be decomposed into the sum of two quantities, the first, SSR, measures the quality of X as a predictor of Y, and the second, SSE, measures the error in this prediction. Therefore, the ratio $R^2 = \text{SSR}/\text{SST}$ can be interpreted as the proportion of the total variation in Y that is accounted for by the predictor variable X. Using (2.45), we can rewrite R^2 as

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} . \tag{2.46}$$

Additionally, it can be shown that

$$[\text{Cor}(Y, X)]^2 = [\text{Cor}(Y, \hat{Y})]^2 = R^2. \tag{2.47}$$

In simple linear regression, R^2 is equal to the square of the correlation coefficient between the response variable Y and the predictor X or to the square of the correlation coefficient between the response variable Y and the fitted values $\hat{Y}$. The definition given in (2.46) provides us with an alternative

interpretation of the squared correlation coefficients. The *goodness-of-fit index*, R^2, may be interpreted as the proportion of the total variability in the response variable Y that is accounted for by the predictor variable X. Note that $0 \leq R^2 \leq 1$ because SSE $\leq$ SST. If R^2 is near 1, then X accounts for a large part of the variation in Y. For this reason, R^2 is known as the *coefficient of determination* because it gives us an idea of how the predictor variable X accounts for (determines) the response variable Y. The same interpretation of R^2 will carry over to the case of multiple regression.

Using the Computer Repair data, the fitted values, and the residuals in Table 2.7, the reader can verify that $\text{Cor}(Y, X) = \text{Cor}(Y, \hat{Y}) = 0.994$, from which it follows that $R^2 = (0.994)^2 = .987$. The same value of R^2 can be computed using (2.46). Verify that SST = 27768.348 and SSE = 348.848. So that

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{348.848}{27768.348} = 0.987.$$

The value $R^2 = 0.987$ indicates that nearly 99% of the total variability in the response variable (Minutes) is accounted for by the predictor variable (Units). The high value of R^2 indicates a strong linear relationship between servicing time and the number of units repaired during a service call.

We reemphasize that the regression assumptions should be checked before drawing statistical conclusions from the analysis (e.g., conducting tests of hypothesis or constructing confidence or prediction intervals) because the validity of these statistical procedures hinges on the validity of the assumptions. Chapter 4 presents a collection of graphical displays that can be used for checking the validity of the assumptions. We have used these graphs for the computer repair data and found no evidence that the underlying assumptions of regression analysis are not in order. In summary, the 14 data points in the Computer Repair data have given us an informative view of the repair time problem. Within the range of observed data, we are confident of the validity of our inferences and predictions.

2.10 REGRESSION LINE THROUGH THE ORIGIN

We have considered fitting the model

$$Y = \beta_0 + \beta_1 X + \varepsilon, \tag{2.48}$$

which is a regression line with an intercept. Sometimes, it may be necessary to fit the model

$$Y = \beta_1 X + \varepsilon, \tag{2.49}$$

a line passing through the origin. This model is also called the *no-intercept* model. The line may be forced to go through the origin because of subject matter theory or other physical and material considerations. For example, distance traveled as a function of time should have no constant. Thus, in this case, the regression model

in (2.49) is appropriate. Many other practical applications can be found where model (2.49) is more appropriate than (2.48). We shall see some of these examples in Chapter 7.

The least squares estimate of β_1 in (2.49) is

$$\hat{\beta}_1 = \frac{\sum y_i x_i}{\sum x_i^2}. \tag{2.50}$$

The ith fitted value is

$$\hat{y}_i = \hat{\beta}_1 x_i, \quad i = 1, 2, \ldots, n, \tag{2.51}$$

and the corresponding residual is

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \ldots, n. \tag{2.52}$$

The standard error of the $\hat{\beta}_1$ is

$$s.e(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum x_i^2}}, \tag{2.53}$$

where

$$\hat{\sigma} = \sqrt{\frac{\sum e_i^2}{n-1}} = \sqrt{\frac{\text{SSE}}{n-1}}. \tag{2.54}$$

Note that the degrees of freedom for SSE is $n - 1$, not $n - 2$, as is the case for a model with an intercept.

Note that the residuals in (2.52) do not necessarily add up to zero as is the case for a model with an intercept (see Exercise 2.11(c)). Also, the fundamental identity in (2.45) is no longer true in general. For this reason, some quality measures for models with an intercept such as R^2 in (2.46), are no longer appropriate for models with no-intercept. The appropriate identity for the case of models with no intercept is obtained by replacing $\bar{y}$ in (2.44) by zero. Hence, the fundamental identity becomes

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} e_i^2, \tag{2.55}$$

from which R^2 is redefined as

$$R^2 = \frac{\sum \hat{y}_i^2}{\sum y_i^2} = 1 - \frac{\sum e_i^2}{\sum y_i^2}. \tag{2.56}$$

This is the appropriate form of R^2 for models with no intercept. Note, however, that the interpretations for the two formulas of R^2 are different. In the case of models with an intercept, R^2 can be interpreted as the proportion of the variation in Y that is accounted for by the predictor variable X after adjusting Y by its mean. For models without an intercept, no adjustment of Y is made. For example, if we fit (2.49) but use the formula for R^2 in (2.46), it is possible for R^2 to be negative in

some cases (see Exercise 2.11(d)). Therefore, the correct formula and the correct interpretation should be used.

The formula for the t-test in (2.29) for testing $H_0 : \beta_1 = \beta_1^0$ against the two-sided alternative $H_1 : \beta_1 \neq \beta_1^0$, continues to hold but with the new definitions of $\hat{\beta}_1$ and s.e.$(\hat{\beta}_1)$ in (2.50) and (2.53), respectively.

As we mentioned earlier, models with no intercept should be used whenever they are consistent with the subject matter (domain) theory or other physical and material considerations. In some applications, however, one may not be certain as to which model should be used. In these cases, the choice between the models given in (2.48) and (2.49) has to be made with care. First, the goodness of fit should be judged by comparing the residual mean squares ($\hat{\sigma}^2$) produced by the two models because it measures the closeness of the observed and predicted values for the two models. Second, one can fit model (2.48) to the data and use the t-test in (2.31) to test the significance of the intercept. If the test is significant, then use (2.48), otherwise use (2.49).

An excellent exposition of regression models through the origin is provided by Eisenhauer (2003) who also alerts the users of regression models through the origin to be careful when fitting these models using computer software programs because some of them give incorrect and confusing results for the case of regression models through the origin.

2.11 TRIVIAL REGRESSION MODELS

In this section we give two examples of trivial regression models, that is, regression equations that have no regression coefficients. The first example arises when we wish to test for the mean μ of a single variable Y based on a random sample of n observations $y_1, y_2, \ldots, y_n$. Here we have $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$. Assuming that Y is normally distributed with mean μ and variance σ^2, the well-known *one-sample t-test*

$$t = \frac{\bar{y} - 0}{\text{s.e.}(\bar{y})} = \frac{\bar{y}}{s_y/\sqrt{n}}, \tag{2.57}$$

can be used to test H_0, where s_y is sample standard deviation of Y. Alternatively, the above hypotheses can be formulated as

$$H_0(Model\ 1) : Y = \varepsilon \ \text{ against } \ H_1(Model\ 2) : Y = \beta_0 + \varepsilon, \tag{2.58}$$

where $\beta_0 = \mu_0$. Thus, Model 1 indicates that $\mu = 0$ and Model 2 indicates that $\mu \neq 0$. The least squares estimate of β_0 in Model 2 is $\bar{y}$, the ith fitted value is $\hat{y}_i = \bar{y}$, and the ith residual is $e_i = y_i - \bar{y}$. It follows then that an estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n-1} = \frac{\sum(y_i - \bar{y})^2}{n-1} = s_y^2, \tag{2.59}$$

which is the *sample variance* of Y. The standard error of $\hat{\beta}_0$ is then $\hat{\sigma}/\sqrt{n} = s_y/\sqrt{n}$, which is the familiar standard error of the sample mean $\bar{y}$. The t-test for

testing Model 1 against Model 2 is

$$t_1 = \frac{\hat{\beta}_0 - 0}{\text{s.e.}(\hat{\beta}_0)} = \frac{\bar{y}}{s_y/\sqrt{n}}, \tag{2.60}$$

which is the same as the one-sample t-test in (2.57).

The second example occurs in connection with the *paired two-sample t-test*. For example, to test whether a given diet is effective in weight reduction, a random sample of n people is chosen and each person in the sample follows the diet for a specified period of time. Each person's weight is measured at the beginning of the diet and at the end of the period. Let Y_1 and Y_2 denote the weight at the beginning and at the end of diet period, respectively. Let $Y = Y_1 - Y_2$ be the difference between the two weights. Then Y is a random variable with mean μ and variance σ^2. Consequently, testing whether or not the diet is effective is the same as testing $H_0 : \mu = 0$ against $H_1 : \mu > 0$. With the definition of Y and assuming that Y is normally distributed, the well-known paired two-sample t-test is the same as the test in (2.57). This situation can be modeled as in (2.58) and the test in (2.60) can be used to test whether the diet is effective in weight reduction.

The above two examples show that the one-sample and the paired two-sample tests can be obtained as special cases using regression analysis.

2.12 BIBLIOGRAPHIC NOTES

The standard theory of regression analysis is developed in a number of good text books, some of which have been written to serve specific disciplines. Each provides a complete treatment of the standard results. The books by Snedecor and Cochran (1980), Fox (1984), and Kmenta (1986) develop the results using simple algebra and summation notation. The development in Searle (1971), Rao (1973), Seber (1977), Myers (1990), Sen and Srivastava (1990), Green (1993), Graybill and Iyer (1994), and Draper and Smith (1998) lean more heavily on matrix algebra.

EXERCISES

2.1 Using the data in Table 2.6:

(a) Compute $Var(Y)$ and $Var(X)$.

(b) Prove or verify that $\sum_{i=1}^{n}(y_i - \bar{y}) = 0$.

(c) Prove or verify that any standardized variable has a mean of 0 and a standard deviation of 1.

(d) Prove or verify that the three formulas for $\text{Cor}(Y, X)$ in (2.5), (2.6), and (2.7) are identical.

(e) Prove or verify that the three formulas for $\hat{\beta}_1$ in (2.14) and (2.20) are identical.

2.2 Explain why you would or wouldn't agree with each of the following statements:

(a) $\text{Cov}(Y, X)$ and $\text{Cor}(Y, X)$ can take values between $-\infty$ and $+\infty$.

(b) If $\text{Cov}(Y, X) = 0$ or $\text{Cor}(Y, X) = 0$, one can conclude that there is no relationship between Y and X.

(c) The least squares line fitted to the points in the scatter plot of Y versus $\hat{Y}$ has a zero intercept and a unit slope.

2.3 Using the regression output in Table 2.9, test the following hypotheses using $\alpha = 0.1$:

(a) $H_0 : \beta_1 = 15$ versus $H_1 : \beta_1 \neq 15$

(b) $H_0 : \beta_1 = 15$ versus $H_1 : \beta_1 > 15$

(c) $H_0 : \beta_0 = 0$ versus $H_1 : \beta_0 \neq 0$

(d) $H_0 : \beta_0 = 5$ versus $H_1 : \beta_0 \neq 5$

2.4 Using the regression output in Table 2.9, construct the 99% confidence interval for β_0.

2.5 When fitting the simple linear regression model $Y = \beta_0 + \beta_1 X + \varepsilon$ to a set of data using the least squares method, each of the following statements can be proven to be true. Prove each statement mathematically or demonstrate its correctness numerically (using the data in Table 2.5):

(a) The sum of the ordinary least squares residuals is zero.

(b) The two tests in (2.26) and (2.32) are equivalent.

(c) The scatter plot of Y versus X and the scatter plot of Y versus $\hat{Y}$ have identical patterns.

(d) The correlation coefficient between Y and $\hat{Y}$ must be nonnegative.

2.6 Using the data in Table 2.5, and the fitted values and the residuals in Table 2.7, verify that:

(a) $\text{Cor}(Y, X) = \text{Cor}(Y, \hat{Y}) = 0.994$

(b) SST = 27768.348

(c) SSE = 348.848

2.7 Verify that the four data sets in Table 2.4 give identical results for the following quantities:.

(a) $\hat{\beta}_0$ and $\hat{\beta}_1$

(b) $\text{Cor}(Y, X)$

(c) R^2

(d) The t-test

2.8 When fitting a simple linear regression model $Y = \beta_0 + \beta_1 X + \varepsilon$ to a set of data using the least squares method, suppose that $H_0 : \beta_1 = 0$ was not rejected. This implies that the model can be written simply as: $Y = \beta_0 + \varepsilon$. The least squares estimate of β_0 is $\hat{\beta}_0 = \bar{y}$. (Can you prove that?)

(a) What are the ordinary least squares residuals in this case?

Table 2.10 Regression Output When Y Is Regressed on X for the Labor Force Participation Rate of Women

Variable	Coefficient	$s.e.$	t-test	p-value
Constant	0.203311	0.0976	2.08	0.0526
X	0.656040	0.1961	3.35	< 0.0038
$n = 19$	$R^2 = 0.397$	$R_a^2 = 0.362$	$\hat{\sigma} = 0.0566$	$d.f. = 17$

(b) Show that the ordinary least squares residuals sum up to zero.

2.9 Let Y and X denote the labor force participation rate of women in 1972 and 1968, respectively, in each of 19 cities in the United States. The regression output for this data set is shown in Table 2.10. It was also found that SSR = 0.0358 and SSE = 0.0544. Suppose that the model $Y = \beta_0 + \beta_1 X + \varepsilon$ satisfies the usual regression assumptions.

(a) Compute $Var(Y)$ and $\text{Cor}(Y, X)$.

(b) Suppose that the participation rate of women in 1968 in a given city is 45%. What is the estimated participation rate of women in 1972 for the same city?

(c) Suppose further that the mean and variance of the participation rate of women in 1968 are 0.5 and 0.005, respectively. Construct the 95% confidence interval for the estimate in (b).

(d) Construct the 95% confidence interval for the slope of the true regression line, β_1.

(e) Test the hypothesis: $H_0 : \beta_1 = 1$ versus $H_1 : \beta_1 > 1$ at the 5% significance level.

(f) If Y and X were reversed in the above regression, what would you expect R^2 to be?

2.10 One may wonder if people of similar heights tend to marry each other. For this purpose, a sample of newly married couples was selected. Let X be the height of the husband and Y be the height of the wife. The heights (in centimeters) of husbands and wives are found in Table 2.11. The data can also be found in the book's Web site.

(a) Compute the covariance between the heights of the husbands and wives.

(b) What would the covariance be if heights were measured in inches rather than in centimeters?

(c) Compute the correlation coefficient between the heights of the husband and wife.

(d) What would the correlation be if heights were measured in inches rather than in centimeters?

(e) What would the correlation be if every man married a woman exactly 5 centimeters shorter than him?

(f) We wish to fit a regression model relating the heights of husbands and wives. Which one of the two variables would you choose as the response variable? Justify your answer.

(g) Using your choice of the response variable in (f), test the null hypothesis that the slope is zero.

(h) Using your choice of the response variable in (f), test the null hypothesis that the intercept is zero.

(i) Using your choice of the response variable in (f), test the null hypothesis that the both the intercept and the slope are zero.

(j) Which of the above hypotheses and tests would you choose to test whether people of similar heights tend to marry each other? What is your conclusion?

(k) If none of the above tests is appropriate for testing the hypothesis that people of similar heights tend to marry each other, which test would you use? What is your conclusion based on this test?

2.11 Consider fitting a simple linear regression model through the origin, $Y = \beta_1 X + \varepsilon$, to a set of data using the least squares method.

(a) Give an example of a situation where fitting the model (2.49) is justified by theoretical or other physical and material considerations.

(b) Show that least squares estimate of β_1 is as given in (2.50).

(c) Show that the residuals $e_1, e_2, \ldots, e_n$ will not necessarily add up to zero.

(d) Give an example of a data set Y and X in which R^2 in (2.46) but computed from fitting (2.49) to the data is negative.

(e) Which goodness of fit measures would you use to compare model (2.49) with model (2.48)?

2.12 In order to investigate the feasibility of starting a Sunday edition for a large metropolitan newspaper, information was obtained from a sample of 34 newspapers concerning their daily and Sunday circulations (in thousands) (*Source*: *Gale Directory of Publications*, 1994). The data are given in Table 2.12 and can be found in the book's Web site.

(a) Construct a scatter plot of Sunday circulation versus daily circulation. Does the plot suggest a linear relationship between Daily and Sunday circulation? Do you think this is a plausible relationship?

(b) Fit a regression line predicting Sunday circulation from Daily circulation.

(c) Obtain the 95% confidence intervals for β_0 and β_1.

(d) Is there a significant relationship between Sunday circulation and Daily circulation? Justify your answer by a statistical test. Indicate what hypothesis you are testing and your conclusion.

Table 2.11 Heights of Husband (H) and Wife (W) in (Centimeters)

Row	H	W	Row	H	W	Row	H	W
1	186	175	33	180	166	65	181	175
2	180	168	34	188	181	66	170	169
3	160	154	35	153	148	67	161	149
4	186	166	36	179	169	68	188	176
5	163	162	37	175	170	69	181	165
6	172	152	38	165	157	70	156	143
7	192	179	39	156	162	71	161	158
8	170	163	40	185	174	72	152	141
9	174	172	41	172	168	73	179	160
10	191	170	42	166	162	74	170	149
11	182	170	43	179	159	75	170	160
12	178	147	44	181	155	76	165	148
13	181	165	45	176	171	77	165	154
14	168	162	46	170	159	78	169	171
15	162	154	47	165	164	79	171	165
16	188	166	48	183	175	80	192	175
17	168	167	49	162	156	81	176	161
18	183	174	50	192	180	82	168	162
19	188	173	51	185	167	83	169	162
20	166	164	52	163	157	84	184	176
21	180	163	53	185	167	85	171	160
22	176	163	54	170	157	86	161	158
23	185	171	55	176	168	87	185	175
24	169	161	56	176	167	88	184	174
25	182	167	57	160	145	89	179	168
26	162	160	58	167	156	90	184	177
27	169	165	59	157	153	91	175	158
28	176	167	60	180	162	92	173	161
29	180	175	61	172	156	93	164	146
30	157	157	62	184	174	94	181	168
31	170	172	63	185	160	95	187	178
32	186	181	64	165	152	96	181	170

Table 2.12 Newspapers Data: Daily and Sunday Circulations (in Thousands)

Newspaper	Daily	Sunday
Baltimore Sun	391.952	488.506
Boston Globe	516.981	798.298
Boston Herald	355.628	235.084
Charlotte Observer	238.555	299.451
Chicago Sun Times	537.780	559.093
Chicago Tribune	733.775	1133.249
Cincinnati Enquirer	198.832	348.744
Denver Post	252.624	417.779
Des Moines Register	206.204	344.522
Hartford Courant	231.177	323.084
Houston Chronicle	449.755	620.752
Kansas City Star	288.571	423.305
Los Angeles Daily News	185.736	202.614
Los Angeles Times	1164.388	1531.527
Miami Herald	444.581	553.479
Minneapolis Star Tribune	412.871	685.975
New Orleans Times-Picayune	272.280	324.241
New York Daily News	781.796	983.240
New York Times	1209.225	1762.015
Newsday	825.512	960.308
Omaha World Herald	223.748	284.611
Orange County Register	354.843	407.760
Philadelphia Inquirer	515.523	982.663
Pittsburgh Press	220.465	557.000
Portland Oregonian	337.672	440.923
Providence Journal-Bulletin	197.120	268.060
Rochester Democrat & Chronicle	133.239	262.048
Rocky Mountain News	374.009	432.502
Sacramento Bee	273.844	338.355
San Francisco Chronicle	570.364	704.322
St. Louis Post-Dispatch	391.286	585.681
St. Paul Pioneer Press	201.860	267.781
Tampa Tribune	321.626	408.343
Washington Post	838.902	1165.567

(e) What proportion of the variability in Sunday circulation is accounted for by Daily circulation?

(f) Provide an interval estimate (based on 95% level) for the true average Sunday circulation of newspapers with Daily circulation of 500,000.

(g) The particular newspaper that is considering a Sunday edition has a Daily circulation of 500,000. Provide an interval estimate (based on 95% level) for the predicted Sunday circulation of this paper. How does this interval differ from that given in (f)?

(h) Another newspaper being considered as a candidate for a Sunday edition has a Daily circulation of 2,000,000. Provide an interval estimate for the predicted Sunday circulation for this paper? How does this interval compare with the one given in (g)? Do you think it is likely to be accurate?