COPYRIGHT

Abraham, B. and Ledolter, J.

Introduction to Regression Modeling

Belmont, CA: Duxbury Press, 2006

2 Simple Linear Regression

2.1 THE MODEL

In this chapter, we consider the linear regression model with a single predictor (regressor) variable. The model is given as

$$y = \mu + \epsilon, \text{ where } \mu = \beta_0 + \beta_1 x \tag{2.1}$$

It is commonly referred to as the **simple linear regression model** because only one predictor variable is involved. Suppose we have n pairs of observations (x_i, y_i) $i = 1, 2, \ldots, n$. Then we can characterize these observations as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, 2, \ldots, n$$

For the hardness data in Example 1.2.4 of Chapter 1, we have

$$\begin{aligned} 55.8 &= \beta_0 + \beta_1 .30 + \epsilon_1 \\ 59.1 &= \beta_0 + \beta_1 .30 + \epsilon_2 \\ \vdots \quad & \quad \vdots \\ 16.9 &= \beta_0 + \beta_1 .60 + \epsilon_{14} \end{aligned} \tag{2.2}$$

2.1.1 IMPORTANT ASSUMPTIONS

The standard analysis is based on the following assumptions about the regressor variable x and the random errors ϵ_i, $i = 1, \ldots, n$:

1. The regressor variable is under the experimenter's control, who can set the values $x_1, \ldots, x_n$. This means that x_i, $i = 1, 2, \ldots, n$, can be taken as constants; they are not random variables.
2. $E(\epsilon_i) = 0$, $i = 1, 2, \ldots, n$.
   This implies that $\mu_i = E(y_i) = \beta_0 + \beta_1 x_i$, $i = 1, 2, \ldots, n$.
3. $V(\epsilon_i) = \sigma^2$ is constant for all $i = 1, 2, \ldots, n$.
   This implies that the variances $V(y_i) = \sigma^2$ are all the same. All observations have the same precision.

4. Different errors ϵ_i and ϵ_j, and hence different responses y_i and y_j, are independent. This implies that $\text{Cov}(\epsilon_i, \epsilon_j) = 0$, for $i \neq j$.

The model implies that the response variable observations y_i are drawn from probability distributions with means $\mu_i = E(y_i) = \beta_0 + \beta_1 x_i$ and constant variance σ^2. In addition, any two observations y_i and y_j, for $i \neq j$, are independent.

2.1.2 OBJECTIVES OF THE ANALYSIS

Given a set of observations, the following questions usually arise:

1. Can we establish a relationship between y and x?
2. Can we predict y from x? To what extent can we predict y from x?
3. Can we control y by using x?

In order to answer these questions within the context of the simple linear regression model with mean $\mu = \beta_0 + \beta_1 x$, we need to estimate β_0, β_1, and σ^2 from available data (x_i, y_i), $i = 1, 2, \dots, n$. The slope β_1 is of particular interest, because a zero slope indicates the absence of a linear association.

2.2 ESTIMATION OF PARAMETERS

2.2.1 MAXIMUM LIKELIHOOD ESTIMATION

This is a common method of estimating the parameters. Maximum likelihood estimation selects the estimates of the parameters such that the likelihood function is maximized. The likelihood function of the parameters β_0, β_1, σ^2 is the joint probability density function of $y_1, y_2, \dots, y_n$, viewed as a function of the parameters. One looks for values of the parameters that give us the greatest probability of observing the data at hand.

A probability distribution for y must be specified if one wants to use this approach. In addition to the assumptions made earlier, we assume that ϵ_i has a normal distribution with mean zero and variance σ^2. This in turn implies that y_i has a normal distribution with mean $\mu_i = \beta_0 + \beta_1 x_i$ and variance σ^2. We write $\epsilon_i \sim N(0, \sigma^2)$ and $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$.

The probability density function for the ith response y_i is

$$p(y_i|\beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right] \tag{2.3}$$

and the joint probability density function of $y_1, y_2, \dots, y_n$ is

$$p(y_1, y_2, \dots, y_n|\beta_0, \beta_1, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right]$$

Treating this as a function of the parameters leads us to the likelihood function $L(\beta_0, \beta_1, \sigma^2|y_1, y_2, \ldots, y_n)$, and its logarithm

$$l(\beta_0, \beta_1, \sigma^2) = lnL(\beta_0, \beta_1, \sigma^2) = K - nln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \quad (2.4)$$

Here, $K = (-n/2)ln(2\pi)$ is a constant that does not depend on the parameters. The maximum likelihood estimators (MLEs) of $\beta_0, \beta_1, \sigma^2$ maximize $l(\beta_0, \beta_1, \sigma^2)$. Maximizing the log-likelihood $l(\beta_0, \beta_1, \sigma^2)$ with respect to β_0 and β_1 is equivalent to minimizing $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$. The method of estimating β_0 and β_1 by minimizing $S(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$ is referred to as the **method of least squares**.

2.2.2 LEAST SQUARES ESTIMATION

This discussion shows that maximum likelihood estimation, with the assumption of a normal distribution, leads to least squares estimation. However, least squares can be motivated in its own right, without having to refer to a normal distribution. One wants to obtain a line $\mu_i = \beta_0 + \beta_1 x_i$ that is "closest" to the points (x_i, y_i). The errors $\epsilon_i = y_i - \mu_i = y_i - \beta_0 - \beta_1 x_i$ $(i = 1, 2, \ldots, n)$ should be as small as possible. One approach to achieve this is to minimize the function

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n}\epsilon_i^2 = \sum(y_i - \mu_i)^2 = \sum(y_i - \beta_0 - \beta_1 x_i)^2 \qquad (2.5)$$

with respect to β_0 and β_1. This approach uses the squared distance as a measure of closeness. Note that other measures could be used, such as the absolute value of the difference, or some other power of the absolute difference. We use a symmetric loss function where positive and negative differences are treated the same. One could also think of nonsymmetric loss functions where over- and underpredictions are weighted differently. The squared error loss is the function that arises from the maximum likelihood procedure.

Taking derivatives with respect to β_0 and β_1, and setting the derivatives to zero,

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = -2\sum(y_i - \beta_0 - \beta_1 x_i) = 0$$

and

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = -2\sum(y_i - \beta_0 - \beta_1 x_i)x_i = 0$$

leads to the two equations:

$$\begin{aligned} n\beta_0 + \left(\sum x_i\right)\beta_1 &= \sum y_i \\ \left(\sum x_i\right)\beta_0 + \left(\sum x_i^2\right)\beta_1 &= \sum x_i y_i \end{aligned} \qquad (2.6)$$

These are referred to as the **normal equations**. Suppose that $\hat{\beta}_0$ and $\hat{\beta}_1$ denote the solutions for β_0 and β_1 in the two-equation system (2.6). Simple algebra shows that these solutions are given by

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{s_{xy}}{s_{xx}} \tag{2.7}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \text{where } \bar{y} = \frac{\sum y_i}{n} \quad \text{and} \quad \bar{x} = \frac{\sum x_i}{n}$$

They are called the **least squares estimates** (LSEs) of β_0 and β_1, respectively.

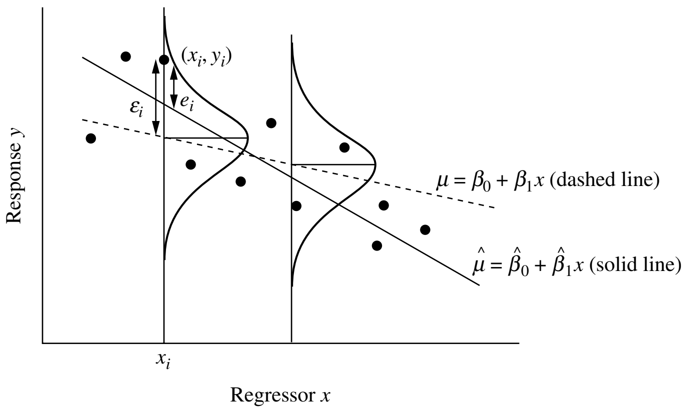
2.3 FITTED VALUES, RESIDUALS, AND THE ESTIMATE OF σ^2

The expression $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is called the **fitted value** that corresponds to the ith observation with x_i as the value for the explanatory variable. It is the value that is implied by the fitted model. Some textbooks refer to it as $\hat{y}_i$.

The difference between y_i and $\hat{\mu}_i$, $y_i - \hat{\mu}_i = e_i$, is referred to as the **residual**. It is the vertical distance between the observation y_i and the estimated line $\hat{\mu}_i$ evaluated at x_i.

The simple linear regression model, sample data, and the fitted line are illustrated in Figure 2.1. The broken line represents the mean $E(y_i) = \mu_i = \beta_0 + \beta_1 x_i$. The data are generated from distributions with densities sketched on the graph. The resulting data are used to determine the LSEs $\hat{\beta}_0$ and $\hat{\beta}_1$. The solid line on the graph represents the estimated line $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Imagine repeating this experiment with another set of n observations y_i at these specified x's. Due to the random component ϵ_i in the model, the observations will be different, and the estimates and the fitted line would change.

FIGURE 2.1 Mean Response and Estimated Regression Line: Simple Linear Regression


(x_i, y_i)
ε_i
e_i
$\mu = \beta_0 + \beta_1 x$ (dashed line)
$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x$ (solid line)
Response y
x_i
Regressor x

2.3.1 CONSEQUENCES OF THE LEAST SQUARES FIT

Least squares estimates set the derivatives of $S(\beta_0, \beta_1)$ equal to zero. The equations, evaluated at the least squares estimates,

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = -2\sum[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)] = 0$$

and

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = -2\sum[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]x_i = 0$$

imply certain restrictions:

i. $\sum_{i=1}^{n} e_i = 0$. This can be seen from the derivative with respect to β_0.

ii. $\sum_{i=1}^{n} e_i x_i = 0$. This follows from the derivative with respect to β_1.

iii. $\sum_{i=1}^{n} \hat{\mu}_i e_i = 0$. This is because

$$\sum \hat{\mu}_i e_i = \sum(\hat{\beta}_0 + \hat{\beta}_1 x_i)e_i = \hat{\beta}_0 \sum e_i + \hat{\beta}_1 \sum x_i e_i = 0$$

due to the results in (i) and (ii).

iv. $(\bar{x}, \bar{y})$ is a point on the line $\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x$. Evaluating the fitted model at $\bar{x}$ leads to $\hat{\beta}_0 + \hat{\beta}_1 \bar{x} = (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 \bar{x} = \bar{y}$.

v. $S(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} e_i^2$ is the minimum of $S(\beta_0, \beta_1)$.

2.3.2 ESTIMATION OF σ^2

Minimization of the log-likelihood function $l(\beta_0, \beta_1, \sigma^2)$ in Eq. (2.4) with respect to σ^2 leads to the MLE

$$\hat{\sigma}^2 = \frac{S(\hat{\beta}_0, \hat{\beta}_1)}{n} \tag{2.8}$$

The numerator $S(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x))^2 = \sum_{i=1}^{n} e_i^2$ is called the **residual sum of squares**; it is the minimum of $S(\beta_0, \beta_1)$.

The LSE of σ^2 is slightly different. It is obtained as

$$s^2 = \frac{S(\hat{\beta}_0, \hat{\beta}_1)}{n-2} \tag{2.9}$$

It is also called the **mean square error** (MSE). The only difference between the estimates in Eqs. (2.8) and (2.9) is in the denominator. The MLE divides by n, whereas the LSE divides by $n - 2$.

The residual sum of squares, $S(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} e_i^2$ consists of n squared residuals. However, the minimization of $S(\beta_0, \beta_1)$ has introduced two constraints among these n residuals; see (i) and (ii) given previously. Hence, only $n - 2$ residuals are needed for its computation. The remaining two residuals can always be calculated from $\sum e_i = \sum e_i x_i = 0$. One says that the residual sum of squares has $n - 2$ "degrees of freedom." The number of degrees of freedom symbolizes the number of independent components that are needed to determine the sum of squares.

The difference between the ML and the LS estimate of σ^2 is small, especially if n is reasonably large. In practice, one prefers the LSE s^2 because it is an **unbiased** estimate of σ^2; this is discussed further in Chapter 4.

2.3.3 LEAST SQUARES CALCULATIONS FOR THE HARDNESS EXAMPLE

Here we have $n = 14$, $\sum_{i=1}^{14} x_i = 630$, $\sum_{i=1}^{14} y_i = 520.2$, $\sum_{i=1}^{14} x_i y_i = 20{,}940$, $\sum_{i=1}^{14} x_i^2 = 30{,}300$:

$$\hat{\beta}_1 = \frac{20{,}940 - (630 \times 520.2/14)}{30{,}300 - (630^2/14)} = -1.266$$

$$\hat{\beta}_0 = \frac{520.2}{14} - (-1.266)\frac{630}{14} = 94.123$$

$$s^2 = \frac{1}{12}\sum_{i=1}^{14} e_i^2 = 2.235$$

The slope estimate $\hat{\beta}_1$ is negative. It implies lower than average hardness for items produced under higher than average temperatures. Is the estimate $\hat{\beta}_1 = -1.266$ extreme enough to claim that the unknown (population) slope β_1 is different from zero? For the answer to this question, one needs to understand the sampling properties of the estimators. In other words, if the true slope were zero and if we repeated the experiment many times at the same given temperature values, what would be the natural variability in the estimates $\hat{\beta}_1$? Would the one observed estimate $\hat{\beta}_1 = -1.266$ appear like an extreme realization from this sampling distribution? If our estimate is large compared to the sampling distribution that can be expected, then the estimate suggests that β_1 is different from zero.

2.4 PROPERTIES OF LEAST SQUARES ESTIMATES

Let us write the LSE of β_1 in Eq. (2.7) in slightly more convenient form:

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i - \bar{y}\sum_{i=1}^{n}(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sum_{i=1}^{n} c_i y_i\end{aligned}$$

where $c_i = (x_i - \bar{x})/s_{xx}$ are constants; $s_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

The constants c_i have several interesting properties:

$$\begin{aligned} &\text{i.} \quad \textstyle\sum_{i=1}^{n} c_i = \sum_{i=1}^{n}(x_i - \bar{x})/s_{xx} = 0 \\ &\text{ii.} \quad \textstyle\sum_{i=1}^{n} c_i x_i = \sum_{i=1}^{n} x_i(x_i - \bar{x})/s_{xx} = 1 \qquad (2.10) \\ &\text{iii.} \quad \textstyle\sum_{i=1}^{n} c_i^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/s_{xx}^2 = 1/s_{xx} \end{aligned}$$

These results can be used to derive the expected values and the variances of the LSEs $\hat{\beta}_0$ and $\hat{\beta}_1$.

2.4.1 EXPECTED VALUES OF LEAST SQUARES ESTIMATES

1. $$\begin{aligned} E(\hat{\beta}_1) = E\left(\sum_{i=1}^{n} c_i y_i\right) &= \textstyle\sum_{i=1}^{n} c_i E(y_i) = \sum c_i(\beta_0 + \beta_1 x_i) \\ &= \beta_0 \textstyle\sum c_i + \beta_1 \sum c_i x_i = 0 + \beta_1 \times 1 = \beta_1 \qquad (2.11) \end{aligned}$$

   Since $E(\hat{\beta}_1) = \beta_1$, we say that $\hat{\beta}_1$ is an unbiased estimator of β_1. This implies that when the experiment is repeated a large number of times, the average of the estimates $\hat{\beta}_1$ [i.e., $E(\hat{\beta}_1)$] coincides with the true value β_1.
2. Similarly,

   $$E(\hat{\beta}_0) = E(\bar{y} - \hat{\beta}_1\bar{x}) = E(\bar{y}) - \bar{x}E(\hat{\beta}_1) = E(\bar{y}) - \beta_1\bar{x}$$

   However, $E(\bar{y}) = E\left(\sum y_i/n\right) = \left[\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i)\right]/n = \beta_0 + \beta_1\bar{x}$. Hence,

   $$E(\hat{\beta}_0) = \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} = \beta_0 \qquad (2.12)$$

   Thus, $\hat{\beta}_0$ is also unbiased for β_0.
3. The LSE of $\mu_0 = \beta_0 + \beta_1 x_0$ is given by $\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ and $E(\hat{\mu}_0) = \beta_0 + \beta_1 x_0 = \mu_0$. Hence, $\hat{\mu}_0$ is unbiased for μ_0.
4. It can also be shown that s^2 is an unbiased estimator of σ^2. That is,

   $$E(s^2) = \sigma^2 \qquad (2.13)$$

   This result will be proved in Chapter 4 for the general regression model.

2.4.2 VARIANCES OF LEAST SQUARES ESTIMATES

1. $V(\hat{\beta}_1) = V(\sum_{i=1}^{n} c_i y_i) = \sum c_i^2 V(y_i) = \sum c_i^2 \sigma^2$ since the y_i's are independent and $V(y_i) = \sigma^2$ is constant.
   Hence, from Eq. (2.10)

   $$V(\hat{\beta}_1) = \sigma^2/s_{xx} \qquad (2.14)$$
2. In order to obtain the variance of $\hat{\beta}_0$, we write the estimator $\hat{\beta}_0$ as follows:

   $$\begin{aligned} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} &= \sum_{i=1}^{n}(y_i/n) - \bar{x}\sum_{i=1}^{n}(x_i - \bar{x})y_i/s_{xx} \\ &= \sum_{i=1}^{n} k_i y_i, \quad \text{where } k_i = \frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{s_{xx}} \end{aligned}$$

Then,

$$V(\hat{\beta}_0) = \sum_{i=1}^{n} k_i^2 \sigma^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}} \right] \tag{2.15}$$

Simple algebra shows that $\sum k_i^2$ equals the second factor in the previous expression.

3. For the variance of $V(\hat{\mu}_0)$, we write

$$\begin{aligned}
\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 &= \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_0 = \bar{y} + \hat{\beta}_1 (x_0 - \bar{x}) \\
&= \sum_{i=1}^{n} \left\{ \frac{y_i}{n} + (x_0 - \bar{x}) \frac{(x_i - \bar{x}) y_i}{s_{xx}} \right\} \\
&= \sum_{i=1}^{n} \left\{ \frac{1}{n} + \frac{(x_0 - \bar{x})(x_i - \bar{x})}{s_{xx}} \right\} y_i \\
&= \sum_{i=1}^{n} d_i y_i, \text{ where } d_i = \left\{ \frac{1}{n} + \frac{(x_0 - \bar{x})(x_i - \bar{x})}{s_{xx}} \right\}
\end{aligned}$$

Then,

$$V(\hat{\mu}_0) = \sum_{i=1}^{n} d_i^2 \sigma^2 = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right] \tag{2.16}$$

Simple algebra shows that $\sum d_i^2$ equals the second factor in the previous expression.

2.5 INFERENCES ABOUT THE REGRESSION PARAMETERS

The objective of most statistical modeling is to say something about the parameters of the population from which the data were taken (sampled). Of course, the more data, the smaller the uncertainty about the estimates. This fact is reflected in the variances of the LSEs; the denominators in the variances in Eqs. (2.14)–(2.16) increase with the sample size.

The uncertainty in the estimates can be expressed through confidence intervals, and for that one needs to make assumptions about the distribution of the errors. In the following discussion, we assume that the errors, and hence the observations, are normally distributed. That is,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \text{where } \epsilon_i \sim N(0, \sigma^2)$$

2.5.1 INFERENCE ABOUT β_1

The question whether or not the slope β_1 is zero is of particular interest. The slope β_1 expresses the effect on $E(y)$ of a unit change in the x variable.

Linear combinations of normal random variables are again normally distributed. The estimator $\hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i$, where $c_i = (x_i - \bar{x})/s_{xx}$, is a linear combination of normal random variables and hence itself a normal random variable. This result is shown in Chapter 3 for a more general situation. The mean and the variance were obtained before. We find that

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{s_{xx}}\right)$$

or, after standardization,

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{s_{xx}}} \sim N(0, 1)$$

The factor σ^2 in the variance is unknown and must be estimated. For inferences about β_1 we replace the unknown σ^2 by its LSE s^2 in Eq. (2.9). We consider the ratio

$$T = \frac{(\hat{\beta}_1 - \beta_1)}{s/\sqrt{s_{xx}}} = \frac{(\hat{\beta}_1 - \beta_1)}{\sigma/\sqrt{s_{xx}}} \Bigg/ \sqrt{\frac{(n-2)s^2}{\sigma^2(n-2)}} \tag{2.17}$$

The last identity (which you can check easily) appears unnecessary. However, the motivation for writing it in this form is that it facilitates the derivation of the sampling distribution. It can be shown that

i. The first term $Z = \dfrac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{s_{xx}}}$ follows a standard normal distribution.

ii. $\dfrac{(n-2)s^2}{\sigma^2}$ follows a chi-square distributon with $n-2$ degrees of freedom, χ^2_{n-2} (see the appendix in Chapter 4 for the proof in the general case).

iii. s^2 and $\hat{\beta}_1$ are independent (this is proved for the general case in Chapter 4).

iv. If $Z \sim N(0, 1)$, $U \sim \chi^2_v$, and Z and U are independent, then it follows that $Z/\sqrt{U/v}$ has a Student t distribution with v degrees of freedom. We denote this distribution as $t(v)$.

From the results in (i)–(iv), it follows that

$$T = \frac{\hat{\beta}_1 - \beta_1}{s/\sqrt{s_{xx}}} \sim t(n-2) \tag{2.18}$$

Standardization of $\hat{\beta}_1 - \beta_1$ by the standard deviation of $\hat{\beta}_1$, $\sigma/\sqrt{s_{xx}}$, leads to a standard normal distribution. Standardization by the **estimated** standard deviation, $s/\sqrt{s_{xx}}$, leads to a t distribution. The estimated standard deviation of $\hat{\beta}_1$ is also referred to as the **standard error** of the estimate $\hat{\beta}_1$, and we sometimes write it as s.e.$(\hat{\beta}_1) = s/\sqrt{s_{xx}}$. The standard error tells us about the variability of the sampling distribution of $\hat{\beta}_1$; that is, the extent to which an estimate can differ from the true (population) value.

Confidence Interval for β_1

Let us use $t(1-\alpha/2; n-2)$ to denote the $100(1-\alpha/2)$ percentile of a t distribution with $n-2$ degrees of freedom. Since the t distribution is symmetric, we

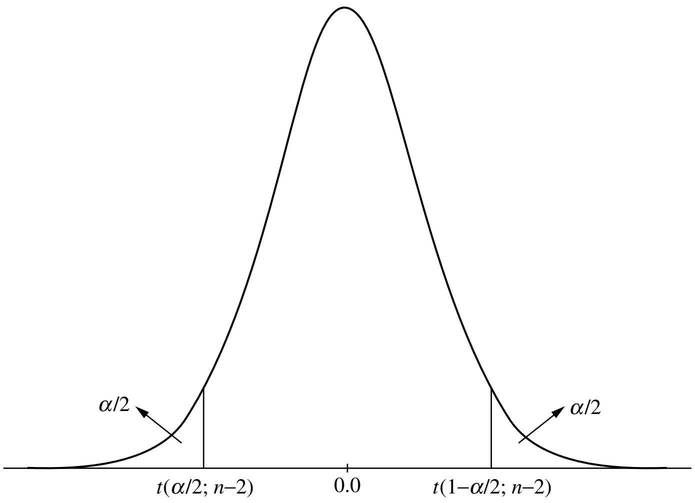


FIGURE 2.2
***t* Distribution and Confidence Intervals**

have that the $100(\alpha/2)$ percentile is given by $t(\alpha/2; n-2) = -t(1-\alpha/2; n-2)$ (Figure 2.2).

For example, the 97.5th and the 2.5th percentiles of the $t(12)$ distribution are given by $t(0.975; 12) = 2.18$ and $t(0.025; 12) = -2.18$, respectively.

The sampling distribution result in Eq. (2.18) implies

$$P\left[-t\left(1-\frac{\alpha}{2}; n-2\right) < \frac{\hat{\beta}_1 - \beta_1}{s/\sqrt{s_{xx}}} < t\left(1-\frac{\alpha}{2}; n-2\right)\right] = 1-\alpha$$

or

$$P\left[\hat{\beta}_1 - t\left(1-\frac{\alpha}{2}; n-2\right)\frac{s}{\sqrt{s_{xx}}} < \beta_1 < \hat{\beta}_1 + t\left(1-\frac{\alpha}{2}; n-2\right)\frac{s}{\sqrt{s_{xx}}}\right] = 1-\alpha$$

Hence, a $100(1-\alpha)$ percent confidence interval for β_1 is defined by the previous equation, and it is given by

$$\left[\hat{\beta}_1 - t\left(1-\frac{\alpha}{2}; n-2\right)\frac{s}{\sqrt{s_{xx}}},\quad \hat{\beta}_1 + t\left(1-\frac{\alpha}{2}; n-2\right)\frac{s}{\sqrt{s_{xx}}}\right] \tag{2.19}$$

Note the form of this interval. You get it by starting with the point estimate $\hat{\beta}_1$ and by adding and subtracting a certain multiple of its standard error. That is,

$$\text{Estimate} \pm (t \text{ value})(\text{standard error of estimate})$$

where the standard error of the estimate is the estimated standard deviation of the sampling distribution of $\hat{\beta}_1$, given by $s/\sqrt{s_{xx}}$. For a 95% confidence interval and $\alpha = 0.05$, one needs to look up the 97.5th percentile $t(0.975; n-2)$ and the 2.5th percentile, $t(0.025; n-2) = -t(0.975; n-2)$.

Testing a Hypothesis about β_1

When testing $H_0 : \beta_1 = 0$ against the alternative $H_1 : \beta_1 \neq 0$, one assesses the magnitude of the t ratio

$$t_0 = \frac{\hat{\beta}_1}{\text{s.e.}(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{s/\sqrt{s_{xx}}} \tag{2.20}$$

The t ratio is the standardized difference of the estimate $\hat{\beta}_1$ from the null hypothesis (which in this case is zero). The issue is whether the observed t ratio is large enough in absolute value so that one can also claim that the population parameter β_1 is different from zero. A comment on notation: The subscript zero in the observed t ratio $t_0 = \hat{\beta}_1/\text{s.e.}(\hat{\beta}_1)$ makes reference to the zero constraint in the null hypothesis $\beta_1 = 0$. We also write this t ratio as $t_0(\hat{\beta}_1)$ or simply as $t(\hat{\beta}_1)$.

Under the null hypothesis ($\beta_1 = 0$), the t ratio, $T = \hat{\beta}_1/\text{s.e.}(\hat{\beta}_1)$, follows a $t(n-2)$ distribution. Hence, one can calculate the probability

$$p = P[|T| \geq |t_0|] = 2P[T \geq |t_0|] \tag{2.21}$$

This is referred to as the p value, or the **probability value**. If this p value is small (smaller than a selected significance level, usually 0.05), then it is unlikely that the observed t ratio has come from the null hypothesis. In such case, one would not believe in the null hypothesis and reject the hypothesis that $\beta_1 = 0$. On the other hand, if this p value is large (larger than the significance level), one would conclude that the observed t value may have originated from the null distribution. In this case, one has no reason to reject H_0.

2.5.2 INFERENCE ABOUT $\mu_0 = \beta_0 + \beta_1 x_0$

We saw that $\hat{\mu}_0 = \sum_{i=1}^{n} d_i y_i$, where $d_i = \dfrac{1}{n} + \dfrac{(x_0 - \bar{x})(x_i - \bar{x})}{s_{xx}}$. Since $\hat{\mu}_0$ is a linear combination of normal random variables, $\hat{\mu}_0$ is normal. Earlier we derived the mean $E(\hat{\mu}_0) = \mu_0 = \beta_0 + \beta_1 x_0$ and the variance

$$V(\hat{\mu}_0) = \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right]\sigma^2$$

Hence, the standardized random variable

$$\frac{\hat{\mu}_0 - \mu_0}{\sigma\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right]^{1/2}} \sim N(0, 1)$$

Substitution of the estimate s for σ changes the sampling distribution from a normal to a $t(n-2)$ distribution. As before, it can be shown that

$$T = \frac{\hat{\mu}_0 - \mu_0}{s\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right]^{1/2}} = \frac{(\hat{\beta}_0 + \hat{\beta}_1 x_0) - (\beta_0 + \beta_1 x)}{s\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right]^{1/2}} \sim t(n-2) \tag{2.22}$$

Using a t distribution with $n-2$ degrees of freedom (d.f.), a $100(1-\alpha)$ percent confidence interval for μ_0 is given by

$$\underbrace{(\hat{\beta}_0+\hat{\beta}_1 x_0)}_{\text{estimate}} \pm \underbrace{t\left(1-\frac{\alpha}{2};\, n-2\right)}_{t\text{ value}} \underbrace{s\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{s_{xx}}\right]^{1/2}}_{\text{s.e.}(\hat{\beta}_0+\hat{\beta}_1 x_0)} \tag{2.23}$$

Recall our rule about the construction of such intervals. Start with the point estimate and add/subtract a multiple of the estimated standard deviation of the point estimate, which is also referred to as the standard error of the estimate.

For the special case when $x_0=0$, μ_0 simplifies to $\mu_0=\beta_0$ and we can obtain a $100(1-\alpha)$ percent confidence interval for β_0 by setting $x_0=0$ in the previous interval Eq. (2.23). This turns out to be

$$\hat{\beta}_0 \pm t\left(1-\frac{\alpha}{2};\, n-2\right)s\left[\frac{1}{n}+\frac{\bar{x}^2}{s_{xx}}\right]^{1/2} \tag{2.24}$$

2.5.3 HARDNESS EXAMPLE CONTINUED

$\hat{\beta}_0=94.134$, $\hat{\beta}_1=-1.266$, $s_{xx}=\sum(x_i-\bar{x})^2=1{,}950$, $n=14$, $s^2=2.235$. The relevant degrees of freedom are $n-2=12$. For a 95% confidence interval for β_1, we need $t(0.975,\,12)=2.18$; and $\dfrac{s}{\sqrt{s_{xx}}}=\sqrt{\dfrac{2.235}{1{,}950}}=0.034$. The 95% confidence interval for β_1 is

$$\begin{gathered}\hat{\beta}_1 \pm t(0.975;\,12)\frac{s}{\sqrt{s_{xx}}}\\ -1.266 \pm (2.18)(0.034)\\ -1.266 \pm 0.072\end{gathered}$$

The confidence interval for β_1 extends from -1.338 to -1.194. Since “zero” is not in this interval, the data provide substantial evidence to reject $\beta_1=0$. Temperature appears to have a significant effect on hardness. Since $\hat{\beta}_1$ is negative, the hardness decreases as temperature increases.

Formally, one can test the null hypothesis $\beta_1=0$ by calculating the t ratio

$$t_0=\frac{\hat{\beta}_1}{\text{s.e.}(\hat{\beta}_1)}=\frac{-1.266}{0.034}=-37.4$$

and its probability value, $P(|T|>37.4)\approx 0.0001$. Since this is extremely small, there is overwhelming evidence against the hypothesis $\beta_1=0$. Temperature has a major impact on hardness.

A 95% confidence interval for β_0 uses the standard error

$$\text{s.e.}(\hat{\beta}_0)=\sqrt{s^2\left(\frac{1}{n}+\frac{\bar{x}^2}{s_{xx}}\right)}=1.575$$

The interval is given by

$$\hat{\beta}_0 \pm t(0.975;\, 12)\text{s.e.}(\hat{\beta}_0)$$
$$94.134 \pm (2.18)(1.575) \text{ or } 94.134 \pm 3.434$$
$$90.700 < \beta_0 < 97.578$$

The 95% confidence interval for the mean response $\mu_0 = \beta_0 + \beta_1 x_0$ when $x_0 = 55$ is centered at

$$\hat{\mu}_0 = 94.134 + (-1.266)(55) = 24.504$$

The standard error is

$$\begin{aligned}\text{s.e.}(\hat{\mu}_0) &= \sqrt{s^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right)} = \sqrt{2.235\left(\frac{1}{14} + \frac{100}{1950}\right)} \\ &= \sqrt{0.2742} = 0.524\end{aligned}$$

The 95% confidence interval for the mean response at $x_0 = 55$ is

$$24.504 \pm (2.18)(0.524),\ 24.504 \pm 1.142, \text{ or}$$
$$23.362 < \mu_0 < 25.646$$

We are 95% confident that our interval from 23.362 to 25.646 covers the average hardness for parts produced with temperature set at 55 degrees.

2.6 PREDICTION

We consider now the prediction of a **single** observation y, resulting from a **new** case with level x_p on the regressor variable. For illustration, in the hardness example one may be interested in the next run with temperature 55, and one may wish to predict the resulting hardness. Here, the emphasis is on a single observation and not on the mean (average) response for a given x_p.

The new observation y_p is the result of a new trial that is independent of the trials on which the regression analysis is based. However, we continue to assume that the model used for the sample data is also appropriate for the new observation.

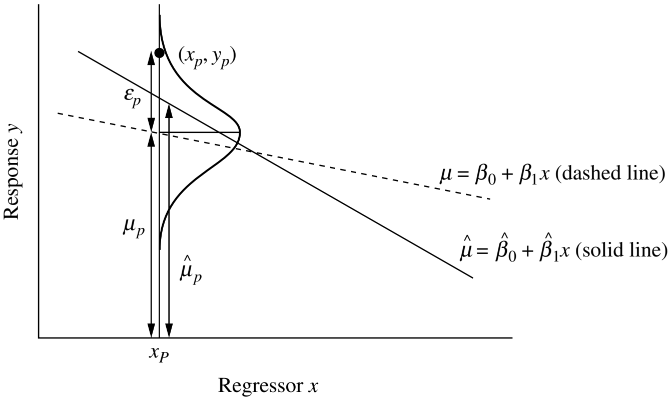
The distinction between drawing inferences about the mean response $\mu_p = \beta_0 + \beta_1 x_p$ and predicting a new observation y_p must be emphasized. In the former case, we discuss the mean of the probability distribution of all responses at x_p. In the latter case, we are concerned about an individual outcome from this probability distribution (Figure 2.3).

The new observation can be written as

$$y_p = \beta_0 + \beta_1 x_p + \epsilon_p$$

where ϵ_p is a future unknown random error, and x_p is assumed known. Initially, let us assume that β_0 and β_1 are known. Then the "best" prediction of y_p is obtained

FIGURE 2.3
Prediction: Simple Linear Regression


(x_p, y_p)
ε_p
$\mu = \beta_0 + \beta_1 x$ (dashed line)
$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x$ (solid line)
μ_p
$\hat{\mu}_p$
Response y
x_P
Regressor x

by replacing ϵ_p with its expected value, namely zero. If β_0 and β_1 are known, the prediction is given by

$$\hat{y}_p = \beta_0 + \beta_1 x_p$$

and the prediction error by

$$y_p - \hat{y}_p = \epsilon_p$$

The variance of the prediction error is

$$V(y_p - \hat{y}_p) = V(\epsilon_p) = \sigma^2 \tag{2.25}$$

In this case, the uncertainty about the prediction comes only through the random error ϵ_p.

Next, suppose that β_0 and β_1 are unknown and that they are estimated by $\hat{\beta}_0$ and $\hat{\beta}_1$. Then the best prediction is obtained by

$$\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p$$

and the prediction error is

$$y_p - \hat{y}_p = (\beta_0 + \beta_1 x_p) - (\hat{\beta}_0 + \hat{\beta}_1 x_p) + \epsilon_p = \mu_p - \hat{\mu}_p + \epsilon_p \tag{2.26}$$

The prediction error is the sum of two components: the new random error ϵ_p, and the error in estimating the mean response at x_p, $\mu_p - \hat{\mu}_p$. The LSEs $\hat{\beta}_0$ and $\hat{\beta}_1$, and hence the estimation error $\hat{\mu}_p - \mu_p$, are independent of ϵ_p since ϵ_p is a future random error that is unrelated to the data at hand. Hence, using the variance $V(\hat{\mu}_p)$ in Eq. (2.16), we find that the variance of the forecast error is

$$\begin{aligned} V(y_p - \hat{y}_p) &= V(\hat{\mu}_p) + V(\epsilon_p) \\ &= \sigma^2\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{s_{xx}}\right] + \sigma^2 \\ &= \left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{s_{xx}}\right]\sigma^2 \end{aligned} \tag{2.27}$$

Our uncertainty about the prediction $\hat{y}_p$ comes from two sources: (i) the random future error ϵ_p and (ii) the estimation of β_0 and β_1.

So far, we have discussed the properties (expectation and variance) of the prediction error. For prediction intervals we need to study the distribution of the error. Since the prediction error $y_p - \hat{y}_p$ is a linear combination of normal random variables, it is also normal. The mean

$$E(y_p - \hat{y}_p) = \mu_p - E(\hat{\mu}_p) + E(\epsilon_p) = \mu_p - \mu_p + 0 = 0$$

implying an unbiased forecast. The variance is given in Eq. (2.27). Therefore,

$$\frac{y_p - \hat{y}_p}{\sigma\left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{s_{xx}}\right]^{1/2}} \sim N(0, 1)$$

Replacing σ with its LSE s leads us to the ratio

$$T = \frac{y_p - \hat{y}_p}{\text{s.e.}(y_p - \hat{y}_p)} \tag{2.28}$$

where $\text{s.e.}(y_p - \hat{y}_p) = s\left(1 + \dfrac{1}{n} + \dfrac{(x_p - \bar{x})^2}{s_{xx}}\right)^{1/2}$. Following similar arguments as before, it can be shown that T in Eq. (2.28) has a Student t distribution with $n - 2$ degrees of freedom. Hence,

$$P\left[\hat{y}_p - t\left(1 - \frac{\alpha}{2}; n - 2\right)\text{s.e.}(y_p - \hat{y}_p) < y_p < \hat{y}_p + t\left(1 - \frac{\alpha}{2}; n - 2\right)\text{s.e.}(y_p - \hat{y}_p)\right] = 1 - \alpha$$

and a $100(1 - \alpha)$ percent **prediction interval** for y_p is given by

$$\hat{y}_p \pm t\left(1 - \frac{\alpha}{2}; n - 2\right)s\left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{s_{xx}}\right]^{1/2} \tag{2.29}$$

This interval is usually much wider than the confidence interval for the mean response μ_p at $x = x_p$. This is because of the random error ϵ_p reflecting the fact that individual observations vary around the mean level μ_p.

2.6.1 HARDNESS EXAMPLE CONTINUED

Suppose we are interested in predicting the hardness in the forthcoming run with temperature 55. Our prediction is $\hat{y}_p = 94.134 + (-1.266) \times (55) = 24.504$. The prediction error variance is estimated as

$$V(y_p - \hat{y}_p) = \left[1 + \frac{1}{14} + \frac{(55 - 45)^2}{1950}\right]2.235 = 2.5093$$

and a 95% prediction interval is given by

$$\hat{y}_p \pm t(0.975; 12)\sqrt{2.5093}$$
$$24.504 \pm (2.18)(1.584), \text{ or } 24.504 \pm 3.453$$

We are 95% confident that the interval from 21.051 to 27.957 will cover the hardness of the next run at temperature 55 degrees.

Note that this interval is considerably wider than the interval for the mean response μ_0 in Section 2.5. This is because of the additional uncertainty that comes through ϵ_p.

2.7 ANALYSIS OF VARIANCE APPROACH TO REGRESSION

In this section, we develop an approach for assessing the strength of the linear regression relationship. This approach can be extended quite easily to the more general regression models discussed in subsequent chapters.

Variability among the y_i's is usually measured by their deviations from the mean, $y_i - \bar{y}$. Thus, a measure of the total variation about the mean is provided by the **total sum of squares** (SST):

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

If SST = 0, all observations are the same. The greater is SST, the greater is the variation among the y observations. The standard deviation of the y's is obtained through

$$s_y = \sqrt{\text{SST}/(n-1)}$$

The objective of the analysis of variance is to partition the total variation SST into two parts: (i) the variation that is accounted for by the model and (ii) the variation that is left unexplained by the model. We can write the deviation of the response observation from its mean as

$$y_i - \bar{y} = (\hat{\mu}_i - \bar{y}) + (y_i - \hat{\mu}_i), \qquad i = 1, 2, \ldots, n$$

where $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is the estimated (fitted) mean (Figure 2.4). The total sum

FIGURE 2.4
Decomposition of the Variation: Simple Linear Regression

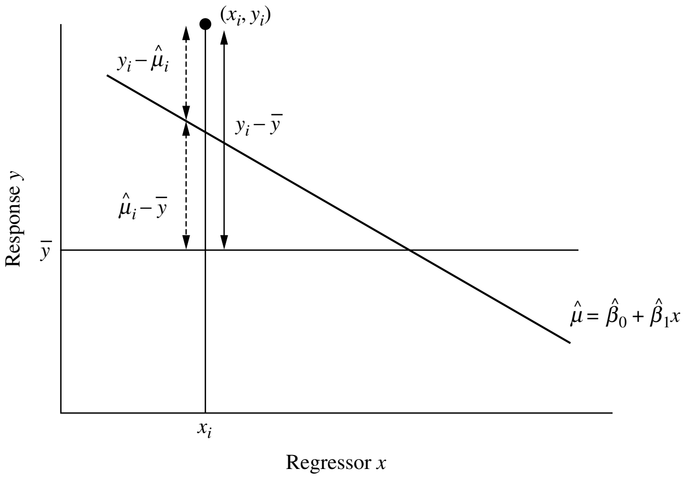


TABLE 2.1 THE ANALYSIS OF VARIANCE TABLE

Source	df	SS	MS	F
Regression (Model)	1	$\text{SSR} = \sum(\hat{\mu}_i - \bar{y})^2$	$\text{MSR} = \frac{\text{SSR}}{1}$	$\frac{\text{SSR}/1}{s^2}$
Residual	$n-2$	$\text{SSE} = \sum(y_i - \hat{\mu}_i)^2$	$\text{MSE} = \frac{\text{SSE}}{n-2} = s^2$	
Total (corrected)	$n-1$	$\text{SST} = \sum(y_i - \bar{y})^2$		

of squares can be written as

$$\begin{aligned}\text{SST} = \sum(y_i - \bar{y})^2 &= \sum(\hat{\mu}_i - \bar{y})^2 + \sum(y_i - \hat{\mu}_i)^2 + 2\sum(\hat{\mu}_i - \bar{y})(y_i - \hat{\mu}_i)\\ &= \sum(\hat{\mu}_i - \bar{y})^2 + \sum(y_i - \hat{\mu}_i)^2\\ &= \text{SSR} + \text{SSE} \qquad (2.30)\end{aligned}$$

since the cross-product term $\sum(\hat{\mu}_i - \bar{y})(y_i - \hat{\mu}_i) = \sum e_i(\hat{\mu}_i - \bar{y}) = \sum e_i\hat{\mu}_i - \bar{y}\sum e_i = 0$; this follows from properties (i) and (iii) in Section 2.3.

The difference $(y_i - \hat{\mu}_i) = e_i$ is the residual, and it reflects the component in the response that could not be explained by the regression model. The second term in Eq. (2.30), $\text{SSE} = \sum(y_i - \hat{\mu}_i)^2 = \sum e_i^2$, is known as the **residual (error) sum of squares**. It measures the variability in the response that is unexplained by the regression model. The first component in Eq. (2.30), $\text{SSR} = \sum(\hat{\mu}_i - \bar{y})^2$, is referred to as the **regression sum of squares**. The $\hat{\mu}_i$ are the fitted values of the response variable that are implied by the model. SSR measures the variability in the response variable that is accounted for by the model. SSR can also be written in the following equivalent way:

$$\begin{aligned}\text{SSR} = \sum(\hat{\mu}_i - \bar{y})^2 = \sum(\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2 &= \sum(\bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1 x_i - \bar{y})^2\\ &= \hat{\beta}_1^2\sum(x_i - \bar{x})^2 \qquad (2.31)\end{aligned}$$

This expression will be useful later on.

Equation (2.30) shows that the SST can be partitioned into these two components: SST = SSR + SSE. The first component, SSR, expresses the variability that is explained by the model; the second component, SSE, is the variability that could **not** be explained. The decomposition of the SST is usually displayed in a table, the so-called analysis of variance (ANOVA) table (Table 2.1).

Column 2 in the ANOVA table contains the **degrees of freedom** of the sum of squares contributions. The degrees of freedom are the number of independent components that are needed to calculate the respective sum of squares.

The total sum of squares, $\text{SST} = \sum(y_i - \bar{y})^2$, is the sum of n squared components. However, since $\sum(y_i - \bar{y}) = 0$, only $n - 1$ components are needed for its calculation. The remaining one can always be calculated from $(y_n - \bar{y}) = -\sum_{i=1}^{n-1}(y_i - \bar{y})$. Hence, SST has $n - 1$ degrees of freedom.

$\text{SSE} = \sum e_i^2$ is the sum of the n squared residuals. However, there are two restrictions among the residuals, coming from the two normal equations

$[\sum e_i = \sum e_i x_i = 0]$. Hence, only $n-2$ residuals are needed to calculate SSE because the remaining two can be computed from the restrictions. One says that SSE has $n-2$ degrees of freedom: the number of observations minus the number of estimated regression coefficients β_0 and β_1.

This leaves $\text{SSR} = \hat{\beta}_1^2 \sum (x_i - \bar{x})^2$. Only one (linear) function of the responses, $\hat{\beta}_1 = \sum c_i y_i$, is needed for its calculation. Hence, the degrees of freedom for SSR is one. Observe that also the degrees of freedom add up: d.f. (SST) $=$ d.f. (SSR) $+$ d.f. (SSE).

The sums of squares in column 3 are divided by their degrees of freedom in column 2. The resulting ratios are called the **mean squares**: MSR $=$ SSR/1 is the mean square due to regression; $s^2 = \text{MSE} = \text{SSE}/(n-2)$ is the mean square due to residual; it is also called the mean square error (see our discussion in Section 2.3).

The last column in the ANOVA table contains the F ratio:

$$F = \text{MSR/MSE} = \text{SSR}/s^2 \tag{2.32}$$

It will soon become clear why this is called the F ratio.

In Eq. (2.13) we mentioned that $E(s^2) = \sigma^2$; $\text{MSE} = s^2$ is an unbiased estimate of $V(\epsilon_i) = \sigma^2$. One can show that the expectation of $E(\text{MSR})$ is given by

$$\begin{aligned} E(\text{MSR}) = E(\text{SSR}) &= E\left[\hat{\beta}_1^2 \sum (x_i - \bar{x})^2\right] = \left[\sum (x_i - \bar{x})^2\right] E\left(\hat{\beta}_1^2\right) \\ &= \left[\sum (x_i - \bar{x})^2\right]\left[V(\hat{\beta}_1) + \left[E(\hat{\beta}_1)\right]^2\right] \\ &= \left[\sum (x_i - \bar{x})^2\right]\left[\frac{\sigma^2}{\sum (x_i - \bar{x})^2} + \beta_1^2\right] \\ &= \sigma^2 + \beta_1^2 \sum (x_i - \bar{x})^2 \end{aligned} \tag{2.33}$$

When $\beta_1 = 0$, then also $E(\text{MSR}) = \sigma^2$. On the other hand, when $\beta_1 \neq 0$, $E(\text{MSR})$ is greater than σ^2 since the term $\beta_1^2 \sum (x_i - \bar{x})^2$ is always positive. Thus, a test whether $\beta_1 = 0$ can be constructed by comparing the MSR and the mean square due to residuals MSE. A MSR substantially larger than s^2 (the mean square of residuals) suggests that $\beta_1 \neq 0$. This is the basic idea behind the **analysis of variance test** which is discussed next.

Let us consider the ratio in the last column of the ANOVA table. We note the following:

i. The variance of $\hat{\beta}_1$, $V(\hat{\beta}_1) = \sigma^2 / \sum (x_i - x)^2$ was derived in Eq. (2.14). Since $\dfrac{(\hat{\beta}_1 - \beta_1)}{\sigma / \{\sum (x_i - \bar{x})^2\}^{1/2}} \sim N(0, 1)$, it follows that its square $\dfrac{(\hat{\beta}_1 - \beta_1)^2}{\sigma^2} \sum (x_i - \bar{x})^2 \sim \chi_1^2$. Hence, for $\beta_1 = 0$, we have that $\dfrac{\hat{\beta}_1^2 \sum (x_i - \bar{x})^2}{\sigma^2} = \dfrac{\text{SSR}}{\sigma^2} \sim \chi_1^2$.

ii. $\frac{\text{SSE}}{\sigma^2} = \frac{(n-2)s^2}{\sigma^2} \sim \chi^2_{n-2}$ (This is shown in the appendix in Chapter 4).

iii. SSR and SSE are independent. (This is proved in the appendix in Chapter 4).

These facts imply the following result for the ratio:

$$F = \frac{\frac{\text{SSR}}{\sigma^2}/1}{\frac{\text{SSE}}{\sigma^2}/(n-2)} = \frac{\text{SSR}}{s^2}$$

If $\beta_1 = 0$ (i.e., there is no linear relationship between y and x), the F ratio (Eq. 2.32) in the last column of the ANOVA table follows an F distribution with 1 and $n-2$ d.f. We write $F \sim F(1, n-2)$. The degrees of freedom are easy to remember because they stand in the d.f. column next to the respective sum of squares. For $\beta_1 \neq 0$, we expect larger values for F. For testing the hypothesis that $\beta_1 = 0$, we calculate the probability value

$$p = P(F \geq f_0)$$

where f_0 is the observed value of the F statistic. If the p value is small (smaller than a preselected significance level, usually 0.05), then there is evidence against the hypothesis $\beta_1 = 0$. If the p value is large, then there is no evidence to reject the null hypothesis that $\beta_1 = 0$.

2.7.1 COEFFICIENT OF DETERMINATION: R^2

We now discuss a descriptive measure that is commonly used in practice to describe the degree of linear association between y and x. Consider the identity

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

The ratio

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} \tag{2.34}$$

is used to assess the “fit” of a regression model. It expresses the proportion of the total variation of the response around the mean that is explained by the regression model.

R^2 must always be between 0 and 1: $0 \leq R^2 \leq 1$. $R^2 = 0$ indicates that none of the variability in the y is explained by the regression model. SSE $= 0$ and $R^2 = 1$ indicate that all observations fall on the fitted line exactly.

Given the observations $y_1, y_2, \ldots, y_n$, SST is a certain fixed quantity. However, SSR (and SSE) change with the choice of the model. Models with larger SSR (smaller SSE) and larger R^2 are usually preferable to models with smaller SSR (larger SSE) and smaller R^2. However, a large R^2 does not necessarily imply that a particular model fits the data well. Also, a small R^2 does not imply that the model is a poor fit. Thus one should use R^2 with caution. The coefficient of determination R^2 does not capture the essential information as to whether a given relation is useful in a particular application. We will discuss this more fully in Chapter 4.

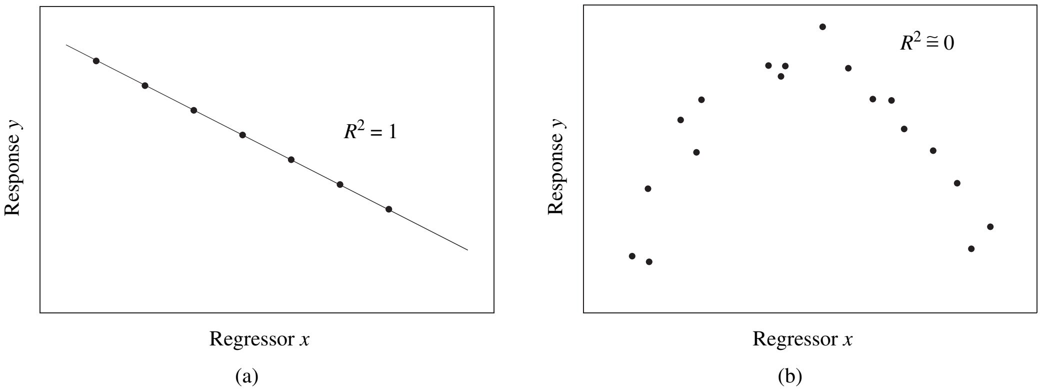


FIGURE 2.5 R^2 **for Different Situations**

It should also be emphasized that R^2 is a measure of the **linear** association between y and x. A small R^2 does not always imply a poor relationship between y and x. As indicated in Figure 2.5, the relation between y and x may be quadratic and R^2 could be a small value.

There is a simple relationship between the R^2 in simple linear regression and the correlation coefficient between y and x. R^2 is the square of the sample correlation coefficient r. You can see this by writing

$$\begin{aligned} R^2 &= \frac{\text{SSR}}{\text{SST}} = \frac{\hat{\beta}_1^2 \sum(x_i-\bar{x})^2}{\sum(y_i-\bar{y})^2} \\ &= \frac{\left[\sum(x_i-\bar{x})(y_i-\bar{y})\right]^2}{\left[\sum(x_i-\bar{x})^2\right]^2} \frac{\sum(x_i-\bar{x})^2}{\sum(y_i-\bar{y})^2} = \left[\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2)}}\right]^2 \\ &= r^2 \end{aligned} \tag{2.35}$$

2.7.2 HARDNESS EXAMPLE CONTINUED

ANOVA TABLE FOR HARDNESS EXAMPLE

Source	df	SS	MS	F
Regression (Model)	1	3,126.134	3,126.134	1,398.7
Residual	12	26.820	2.235	
Total (corrected)	13	3,152.954		

Here, the F statistic has 1 and 12 df. From the F tables we find that

$$p = P(F \geq 1,398.7) < 0.0001$$

is tiny. Thus, there is considerable evidence in the data against the hypothesis $\beta_1 = 0$. We can safely reject $\beta_1 = 0$. There is a strong relationship between hardness and temperature.

Note

$$f_0 = \frac{\text{SSR}}{s^2} = \frac{\hat{\beta}_1^2 \sum (x_i - \bar{x})^2}{s^2} = \left[\frac{\hat{\beta}_1}{s/s_{xx}} \right]^2 = t_0^2 \tag{2.36}$$

where t_0 is the t ratio in Eq. (2.20) in Section 2.5. The F statistic obtained here is the square of the t statistic that was used earlier for testing $\beta_1 = 0$. We know that in general the square of a $t(n-2)$ random variable follows an $F(1, n-2)$ distribution; see Exercise 2.2. Hence, the F test discussed here and the t ratio discussed earlier represent two equivalent tests of the null hypothesis $\beta_1 = 0$.

The coefficient of determination in this example is $R^2 = 3,126.134/3,152.954 = 0.991$. Thus, 99.1% of the total sum of squares is explained by the regression model. Temperature is an important predictor of hardness.

2.8 ANOTHER EXAMPLE

This example addresses the variation in achievement test scores among Iowa elementary schools. The test scores are the average "core" scores from the Iowa Test of Basic Skills, a commonly used standardized test for elementary schools in Iowa. The core score includes vocabulary, reading comprehension, spelling, capitalization, punctuation, usage, expression, and math concepts and problems. The data set in Table 2.2 contains the average fourth-grade tests scores of all elementary schools in the six largest Iowa communities.

Average test scores vary from school to school. The average test score of a school depends on, among other factors, the natural ability of the children attending the school, the quality of the educational programs offered by the school, and the support the children get from their parents. These explanatory factors tend to be related to the economic situation that surrounds each school. Causal links are complicated, but one can expect that richer communities tend to have more resources, children of economically secure parents have more opportunities, and well-to-do parents select their residences such that their children can go to "better" schools. We use the percentage of students in the federal free and reduced-price breakfast and lunch program as an economic indicator; it serves as a proxy for "poverty." Students qualify for this program if they come from families with incomes at or below 130% of the poverty level. The information on $n = 133$ schools is taken from an article in the *Des Moines Register,* November 2000. Poverty and test scores are from the 1999–2000 school year.

A scatter plot of test scores against poverty is shown in Figure 2.6. One notices that average test scores of schools with a high percentage of children in the subsidized lunch program are considerably lower than those of schools with small percentages. The relationship between test scores and the proportion of children on subsidized lunch is roughly linear, which leads us to the simple linear regression model, $y = \beta_0 + \beta_1 x + \epsilon$, that we study in this chapter.

TABLE 2.2 IOWA TEST OF BASIC SKILLS DATA [DATA FILE: iowatest]

School	Poverty	Test Scores	City
Coralville Cen.	20	65	Iowa City
Hills	42	35	Iowa City
Hoover	10	84	Iowa City
Horn	5	83	Iowa City
Kirkwood	34	49	Iowa City
Lemme	17	69	Iowa City
Lincoln	3	88	Iowa City
Longfellow	24	63	Iowa City
Lucas	21	65	Iowa City
Mann	34	58	Iowa City
Penn	24	52	Iowa City
Roosevelt	35	61	Iowa City
Shimek	4	81	Iowa City
Twain	57	43	Iowa City
Weber	24	66	Iowa City
Wickham	10	62	Iowa City
Wood	31	65	Iowa City
Black Hawk	35	46	Waterloo
Edison	62	41	Waterloo
Elk Run	56	48	Waterloo
Grant	81	36	Waterloo
Irving	45	52	Waterloo
Jewett	50	44	Waterloo
Kingsley	15	76	Waterloo
Kittrell	40	48	Waterloo
Lincoln	74	30	Waterloo
Longfellow	99	27	Waterloo
Lowell	82	28	Waterloo
McKinstry	81	20	Waterloo
Orange	38	56	Waterloo
Roosevelt	80	23	Waterloo
Arthur	13	75	Cedar Rapids
Cleveland	27	55	Cedar Rapids
Coolidge	10	72	Cedar Rapids
Erskine	25	67	Cedar Rapids
Garfield	39	46	Cedar Rapids
Grant Wood	44	55	Cedar Rapids
Harrison	55	35	Cedar Rapids
Hiawatha	27	56	Cedar Rapids
Hoover	30	66	Cedar Rapids
Jackson	7	69	Cedar Rapids
Johnson	59	51	Cedar Rapids
Kenwood	41	75	Cedar Rapids
Madison	16	70	Cedar Rapids
Nixon	21	62	Cedar Rapids
Pierce	3	75	Cedar Rapids
Polk	80	54	Cedar Rapids
Taylor	78	36	Cedar Rapids

(*Continued*)

TABLE 2.2 (Continued)

School	Poverty	Test Scores	City
Truman	10	57	Cedar Rapids
Van Buren	52	43	Cedar Rapids
Wilson	39	41	Cedar Rapids
Wright	27	53	Cedar Rapids
Adams	17	52	Davenport
Blue Grass	9	53	Davenport
Buchanan	57	37	Davenport
Buffalo	31	43	Davenport
Eisenhower	40	58	Davenport
Fillmore	57	39	Davenport
Garfield	49	43	Davenport
Grant	38	47	Davenport
Harrison	22	56	Davenport
Hayes	61	30	Davenport
Jackson	58	34	Davenport
Jefferson	89	21	Davenport
Johnson	53	40	Davenport
Lincoln	59	56	Davenport
Madison	87	29	Davenport
McKinley	50	49	Davenport
Monroe	73	36	Davenport
Perry	51	20	Davenport
Truman	40	48	Davenport
Walcott	23	59	Davenport
Washington	71	38	Davenport
Wilson	39	53	Davenport
Adams	50	58	Des Moines
Brooks/Lucas	79	32	Des Moines
Cattell	49	50	Des Moines
Douglas	37	54	Des Moines
Edmunds	77	28	Des Moines
Findley	61	51	Des Moines
Garton	55	27	Des Moines
Granger	47	49	Des Moines
Greenwood	32	67	Des Moines
Hanawalt	12	79	Des Moines
Hills	31	57	Des Moines
Howe	50	50	Des Moines
Hubbell	22	81	Des Moines
Jackson	40	45	Des Moines
Jefferson	3	74	Des Moines
Longfellow	80	50	Des Moines
Lovejoy	62	40	Des Moines
Madison	52	45	Des Moines
Mann	65	32	Des Moines
McKee	57	31	Des Moines
McKinley	78	36	Des Moines
Mitchell	54	46	Des Moines

TABLE 2.2 (Continued)

School	Poverty	Test Scores	City
Monroe	45	53	Des Moines
Moore	40	53	Des Moines
Moulton	83	30	Des Moines
Oak Park	52	49	Des Moines
Park Avenue	42	36	Des Moines
Perkins	65	51	Des Moines
Phillips	29	61	Des Moines
Pleasant Hill	17	68	Des Moines
Stowe	53	47	Des Moines
Strudebaker	25	53	Des Moines
Wallace	77	24	Des Moines
Watrous	39	47	Des Moines
Willard	84	42	Des Moines
Windsor	32	62	Des Moines
Woodlawn	35	59	Des Moines
Wright	28	60	Des Moines
Bryant	32	56	Sioux City
Clark	4	78	Sioux City
Crescent Park	49	65	Sioux City
Emerson	53	40	Sioux City
Everett	79	48	Sioux City
Grant	50	45	Sioux City
Hunt	72	43	Sioux City
Irving	86	27	Sioux City
Joy	33	65	Sioux City
Leeds	46	42	Sioux City
Lincoln	14	76	Sioux City
Longfellew	34	40	Sioux City
Lowell	54	57	Sioux City
McKinley	84	37	Sioux City
Nodland	10	74	Sioux City
Riverview	60	59	Sioux City
Roosevelt	48	50	Sioux City
Smith	72	39	Sioux City
Sunnyside	14	73	Sioux City
Washington	20	57	Sioux City
Whittier	39	48	Sioux City

In a subsequent chapter, we will use this data set to illustrate model checking. The question whether the model can be improved by adding a quadratic component of poverty will be addressed in Exercise 5.17. In addition, we will investigate whether “city” information adds explanatory power. It may be that irrespective of the proportion of children on subsidized lunch, students in college communities (e.g., Iowa City) score higher. If that were true, one would need to look for plausible explanations.

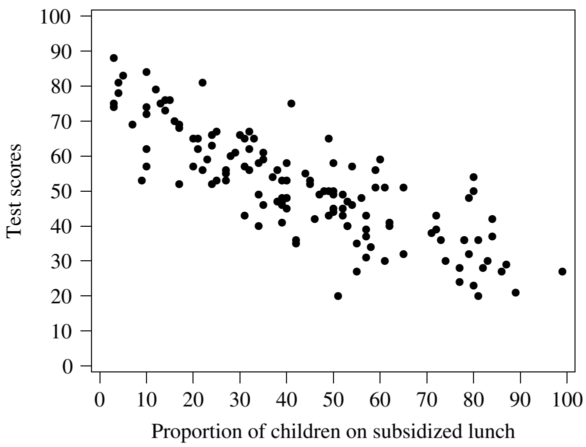


FIGURE 2.6 Scatterplot of Tests Scores Against Proportion of Children on Subsidized Lunch

TABLE 2.3 MINITAB OUTPUT OF TEST SCORES AGAINST PROPORTION OF CHILDREN ON SUBSIDIZED LUNCH

Test Scores = 74.6 −0.536 Poverty

Predictor	Coef	SE Coef	*T*	*P*
Constant	74.606	1.613	46.25	0.000
Poverty	−0.53578	0.03262	−16.43	0.000

$s = 8.766$ $R^2 = 67.3\%$

Analysis of Variance

Source	DF	SS	MS	*F*	*P*
Regression	1	20731	20731	269.79	0.000
Residual Error	131	10066	77		
Total	132	30798			

The output from a standard regression program is listed below. Most computer packages, such as Minitab, SPSS, and SAS, provide very similar output. In Table 2.3, we show the output from Minitab, a popular statistics software.

The fitted regression equation,

$$\text{Test scores} = 74.6 - 0.536 \text{ poverty} \tag{2.37}$$

implies that with each additional unit (1%) increase in the proportion on subsidized lunch, average test scores decrease by 0.54 points.

The LSEs $\hat{\beta}_0 = 74.606$ and $\hat{\beta}_1 = -0.536$, their standard errors s.e.$(\hat{\beta}_0) = 1.613$ and s.e.$(\hat{\beta}_1) = 0.033$, and the t ratios $74.606/1.613 = 46.25$ and $-0.536/0.033 = –16.43$ are shown in the columns labeled Coef, SE Coef, and T. The last column labeled "P" contains the probability value of the regression coefficients. The t ratio for the slope β_1 is -16.43. It leads to a very small probability value. Under the null hypothesis of no relationship ($\beta_1 = 0$), there is almost no chance to get such an extreme value. Hence, we reject—very soundly—the null hypothesis

$\beta_1 = 0$. Yes, there is a very strong, and negative, association among test scores and the proportion of children on subsidized lunch.

The sum of squares, the degrees of freedom ($n - 1 = 133 - 1 = 132$ for total, $n - 2 = 133 - 2 = 131$ for error, and 1 for regression), and the mean squares are shown in the ANOVA table. The R^2 from the regression is 67.3%. We obtain this value by dividing the regression sum of squares SSR = 20,731 by the total sum of squares SST = 30,798. It says that 67.3% of the variation in average test scores can be explained through the linear association with poverty.

Another interpretation of "model fit" focuses on standard deviations. The standard deviation of the test scores, not keeping track of poverty, is given by $s_y = [\mathrm{SST}/(n-1)]^{1/2} = [30{,}798/132]^{1/2} = 15.275$. After factoring in (or adjusting the analysis for) poverty, the standard deviation of the yet unexplained deviations is given by $s = [\mathrm{SSE}/(n-2)]^{1/2} = [10{,}066/131]^{1/2} = 8.766$. This is the square root of the MSE. The reduction from $s_y = 15.275$ to $s = 8.766$ is 42.6%.

The last column of the ANOVA table contains the F ratio, $F = (\mathrm{SSR}/1)/(\mathrm{MSE}) = 269.79$. It serves as a test of the null hypothesis $\beta_1 = 0$. The probability value to the right of this number is the probability that an $F(1, 131)$ random variable exceeds this value. The probability is virtually zero, implying a solid rejection of the null hypothesis. Note that the F test in the simple linear regression model is equivalent to the test that looks at the t ratio. The square of the t ratio, $(-16.43)^2 = 269.79$, is identical to the F ratio.

2.9 RELATED MODELS

2.9.1 REGRESSION THROUGH THE ORIGIN

Theory or the patterns in the scatter plot may imply that the straight line should pass through the origin. Theory may suggest that the relationship between y and x is strictly proportional, implying that the line in the model $y_i = \beta x_i + \epsilon_i$ passes through the origin. The slope coefficient β can be estimated by minimizing the sum of squares

$$S(\beta) = \sum (y_i - \beta x_i)^2 \tag{2.38}$$

The minimization leads to

$$\hat{\beta} = \sum x_i y_i \Big/ \sum x_i^2 \tag{2.39}$$

and the residuals $y_i - \hat{\mu}_i = y_i - \hat{\beta} x_i$. The LSE of σ^2 is

$$s^2 = \sum (y_i - \hat{\beta} x_i)^2/(n-1) \tag{2.40}$$

Note that we divide by $(n - 1)$ degrees of freedom, because there is only one restriction among the residuals, $\sum e_i x_i = 0$.

One can show that $\hat{\beta}$ is unbiased [i.e., $E(\hat{\beta}) = \beta$] and that its variance is given by

$$V(\hat{\beta}) \equiv V\left[\frac{\sum x_i y_i}{\sum x_i^2}\right] = \frac{V\left(\sum x_i y_i\right)}{\left[\sum x_i^2\right]^2} = \frac{\sigma^2 \sum x_i^2}{\left[\sum x_i^2\right]^2} = \frac{\sigma^2}{\sum x_i^2} \tag{2.41}$$

Inference about β is similar to the one in the model with intercept, $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, except that now s^2 has $(n-1)$ rather than $(n-2)$ degrees of freedom.

A $100(1-\alpha)$ percent confidence interval for β is given by

$$\hat{\beta} \pm t\left(1 - \frac{\alpha}{2}; n-1\right)\text{s.e.}(\hat{\beta}) \tag{2.42}$$

where $\text{s.e.}(\hat{\beta}) = s/\sqrt{\sum x_i^2}$.

A $100(1-\alpha)$ percent confidence interval for the mean response at x_0, $\mu_0 = \beta x_0$, is

$$\hat{\mu}_0 \pm t\left(1 - \frac{\alpha}{2}; n-1\right) \text{s.e. } (\hat{\mu}_0) \tag{2.43}$$

where

$$\hat{\mu}_0 = \hat{\beta} x_0 \text{ and s.e. } (\hat{\mu}_0) = s|x_0|/\sqrt{\sum x_i^2}$$

A $100(1-\alpha)$ percent prediction interval for a future observation y_p at x_p is

$$\hat{y}_p \pm t\left(1 - \frac{\alpha}{2}; n-1\right) s\sqrt{1 + \left(x_p^2/\sum x_i^2\right)} \tag{2.44}$$

where $\hat{y}_p = \hat{\mu}_p = \hat{\beta} x_p$.

2.9.2 THE CASE OF RANDOM *x*'S

In our discussions so far we have assumed that the x's are fixed constants. Thus, our inferences are based on repeated sampling, when the x's are kept the same from sample to sample.

Frequently, this assumption is not appropriate since fixed x's may not be possible. It may be preferable to consider both y and x as random variables having some joint distribution. Do the results of the previous sections still hold true in this situation?

We assume that y and x are jointly distributed as a bivariate normal distribution

$$f(y, x) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{(1-\rho^2)} Q\right\} \tag{2.45}$$

where

$$Q = \left(\frac{y-\mu_y}{\sigma_y}\right)^2 + \left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{y-\mu_y}{\sigma_y}\right)\left(\frac{x-\mu_x}{\sigma_x}\right)$$

Here, $\mu_y = E(y)$, $\mu_x = E(x)$ are the means; $\sigma_y^2 = V(y)$, $\sigma_x^2 = V(x)$ are the variances, and $\rho = E(y-\mu_y)(x-\mu_x)/\sigma_x\sigma_y$ is the correlation between y and x.

It can be shown that the conditional distribution of y given x is also normal with conditional mean $E(y|x) = \beta_0 + \beta_1 x$ and conditional variance $V(y|x) = \sigma_y^2(1-\rho^2)$. The regression coefficients β_0 and β_1 are related to the parameters of the bivariate normal distribution: $\beta_1 = (\sigma_y/\sigma_x)\rho$ and $\beta_0 = \mu_y - \rho(\sigma_y/\sigma_x)\mu_x$. Zero correlation ($\rho = 0$) implies $\beta_1 = 0$; then there is no linear association between y and x.

In this more general setup, one can also show that the maximum likelihood estimates of β_0 and β_1 are given by

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}, \qquad \hat{\beta}_1 = s_{xy}/s_{xx}$$

which are exactly the previous estimates. Furthermore, $\hat{\rho} = s_{xy}/\sqrt{s_{xx}s_{yy}} = \hat{\beta}_1\sqrt{\frac{s_{xx}}{s_{yy}}}$. Hence, the regression model in which y and x are jointly normally distributed can be analyzed using the methods that treat x as fixed.

APPENDIX: UNIVARIATE DISTRIBUTIONS

1. THE NORMAL DISTRIBUTION

We say Y is a normal random variable if the density function of Y is given by

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

We use the notation $Y \sim N(\mu, \sigma^2)$. Note that $E(Y) = \mu$ and $V(Y) = \sigma^2$. The density function of the standard normal distribution with mean 0 and variance 1 is shown in Figure 2.7a.

To calculate probabilities for Y, we use the representation $Y = \mu + \sigma Z$, where Z is the standard normal distribution. Hence,

$$P(Y \le y) = P(\mu + \sigma Z \le y) = P\left(Z \le \frac{y-\mu}{\sigma}\right)$$

Table A at the end of the book gives the cumulative probabilities $P(Z \le z)$ for the standard normal distribution. For example, $P(Z \le 0) = 0.5$, $P(Z \le 1) = 0.8413$, $P(Z \le -0.85) = 0.1977$. The 97.5th percentile of the standard normal is 1.96; the 95th percentile is 1.645.

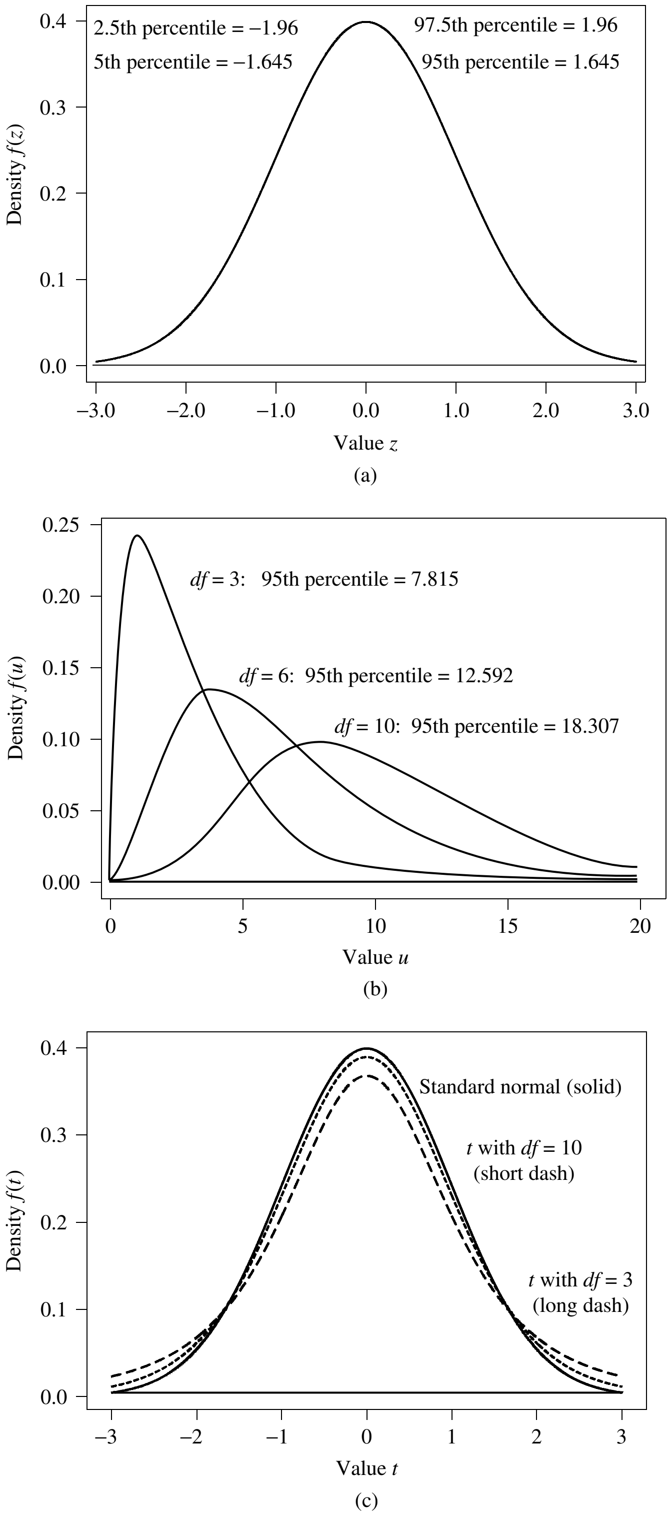
2. THE χ^2 DISTRIBUTION

We say that U follows a chi-square distribution with v degrees of freedom if the density function of U is given by

$$f(u) = cu^{(v/2)-1}e^{-u/2} \quad \text{for} \quad u \ge 0$$

c is a constant that makes the density integrate to 1. The parameter v is called the degrees of freedom. We write $U \sim \chi^2_v$. The density functions of three chi-square distributions ($v = 3$, 6, 10) are shown in Figure 2.7b. The mean of the chi-square distribution is given by $E(U) = v$, and the variance is given by $V(U) = 2v$. Table B (at the end of the book) gives selected percentiles. For example, in a χ^2_5 distribution, the 50, 95, and 99th percentiles are 4.3515, 11.0705, and 15.0863, respectively.

FIGURE 2.7 Densities of (a) Standard Normal Distribution, (b) Three Chi-Square Distributions, (c) Standard Normal and Two *t* Distributions, and (d) Four *F* Distributions

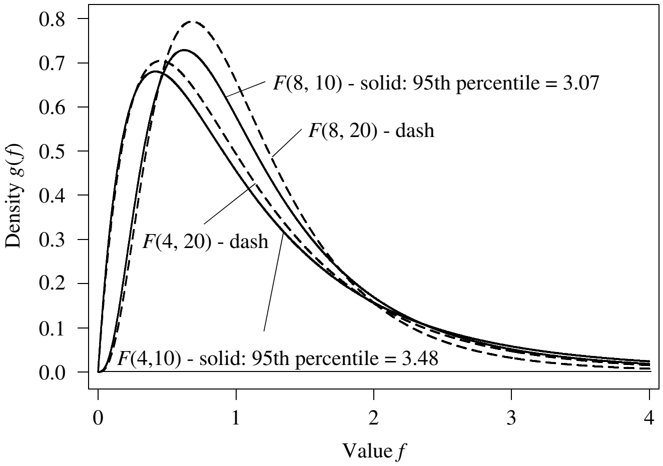


(a)

(b)

(c)

FIGURE 2.7 (Continued)



(d)

Suppose $Z_1, \ldots, Z_v$ are independent $N(0, 1)$ variables. Then Z_1^2 has a χ^2 distribution with 1 degree of freedom, and $U = Z_1^2 + Z_2^2 + \cdots + Z_v^2$ has a χ^2 distribution with v degrees of freedom.

3. THE STUDENT *t* DISTRIBUTION

We say that T follows a Student t distribution if the density function of T is given by

$$f(t) = \frac{c}{\left(1 + \frac{t^2}{v}\right)^{(v+1)/2}}, \quad -\infty < t < \infty$$

c is a constant that makes the density integrate to 1. We write $T \sim t(v)$. The parameter v is called the degrees of freedom. The density functions of two t distributions with $v = 3$ and $v = 10$ degrees of freedom are shown in Figure 2.7c. The t distribution is symmetric with mean $E(T) = 0$ and variance $V(T) = v/(v-2)$ and is similar to the standard normal distribution, but with slightly heavier tails. As can be seen from Figure 2.7c, the t distribution is close to the standard normal distribution when the degrees of freedom (v) is large. Table C (at the end of the book) gives selected percentiles of Student t distributions. For example, for $v = 10$ d.f., the 90th percentile is 1.372, and the 5th percentile is -1.812.

Suppose $Z \sim N(0, 1)$ and $U \sim \chi_v^2$, with Z and U independent. Then $T = Z/\sqrt{U/v}$ has a Student t distribution with v degrees of freedom.

4. THE *F* DISTRIBUTION

We say that F follows an F distribution if the density function is given by

$$g(f) = c\frac{f^{(v/2)-1}}{\left(1 + \frac{v}{w}f\right)^{(v+w)/2}}, \quad f \geq 0$$

c is a constant that makes the density integrate to 1. We write $F \sim F(v, w)$. The integer parameters v and w are called the degrees of freedom. The density functions of four F distributions are shown in Figure 2.7d. The mean of an F distribution is given by $E(F) = w/(w-2)$; it depends only on the second degrees of freedom and is slightly larger than 1. The variance depends on both v and w. Table D (at the end of the book) gives selected percentiles. For example, the 95 and 99th percentiles of an $F(5,8)$ distribution are given by 3.69 and 6.63, respectively.

Suppose $U \sim \chi_v^2$ and $V \sim \chi_w^2$, with U and V independent. Then $F = \frac{U/v}{V/w}$ has an F distribution with v and w degrees of freedom.

A comment on statistical tables: Most computer programs calculate cumulative probabilities and percentiles (the "inverse" of the cumulative probabilities) for a wide selection of distributions. For some programs (such as EXCEL) the calculation of percentiles requires the specification of the upper tail area.

EXERCISES

2.1. Determine the 95 and 99th percentiles of

a. The normal distribution with mean 10 and standard deviation 3;

b. The t distributions with 10 and 25 degrees of freedom;

c. The chi-square distributions with 1, 4, and 10 degrees of freedom;

d. The F distributions with 2 and 10, and 4 and 10 degrees of freedom.

2.2. It is a fact that two distributions are the same if (all) their percentiles are identical.

a. Convince yourself, by looking up several percentiles, that the square of a standard normal distribution is the same as a chi-square distribution with one degree of freedom. Determine the percentile of the χ_1^2 and the percentile of the square of a standard normal distribution, Z^2, and show that they are the same. Use the fact that $P(Z^2 \le z) = P(-\sqrt{z} \le Z \le \sqrt{z})$. Hence, for example, the 95th percentile of Z^2 is the same as the 97.5th percentile of Z.

b. Convince yourself, by looking up several percentiles, that the square of a t distribution with v degrees of freedom is the same as the $F(1, v)$ distribution.

2.3. For each of the four sets of data given below (see Anscombe, 1973), plot y versus x. The data are given in the file **anscombe**. Fit a straight line model to each of the data sets giving least squares estimates, ANOVA table, and R^2. Compute the correlation coefficient between y and x for each data set. Comment on your results. Would a linear regression of y on x be appropriate in all cases? Discuss.

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

2.4. A car dealer is interested in modeling the relationship between the weekly number of cars sold and the daily average number of salespeople who work on the showroom floor during that week. The dealer believes that the relationship between the two variables can be described by a straight line. The following data were supplied by the car dealer:

Week of	No. of Cars Sold y	Average. No. of Sales People on Duty x
January 30	20	6
June 29	18	6
March 2	10	4
October 26	6	2
February 7	11	3

a. Construct a scatter plot (y vs x) for the data.

b. Assuming that the relationship between the variables is described by a straight line, use the method of least squares to estimate the y intercept and the slope of the line.

c. Plot the least squares line on your scatter plot.

d. According to your least squares line, approximately how many cars should the dealer expect to sell in a week if an average of five salespeople are kept on the showroom floor each day?

e. Calculate the fitted value $\hat{\mu}$ for each observed x value. Use the fitted values to calculate the corresponding residuals. Plot the residuals against the fitted values. Are you satisfied with the fit?

f. Calculate an estimate of σ^2.

g. Construct a 95% confidence interval for β_1 and use it to assess the hypothesis that $\beta_1 = 0$.

h. Given the results of (a)–(g), what conclusions are you prepared to draw about the relationship between sales and number of salespeople on duty.

i. Would you be willing to use this model to help determine the number of salespeople to have on duty next year?

2.5. Use S-Plus or any other available statistics software for Exercise 2.4. Check your hand calculations with the results from these programs.

2.6. Dr. Joseph Hooker collected a set of 31 measurements on the boiling temperature of water (TEMP; in degrees Fahrenheit) and the atmospheric pressure (AP; in inches of mercury) at various locations in the Himalaya Mountains (see Weisberg, 1980). The data are given in the file **hooker**.

a. Plot TEMP vs AP. Does a linear model seem appropriate?

b. Repeat (a), plotting TEMP versus $x = 100ln(AP)$.

c. Fit a linear model

$$\text{TEMP} = \beta_0 + \beta_1 x + \epsilon$$

and calculate the estimates of β_0, β_1, and σ^2. Draw the fitted line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ on the plot in (b). Does the model seem appropriate?

d. Find a 95% confidence interval for

i. β_1;

ii. the average temperature when the pressure is 25.

e. Suppose the temperature had been measured in °C instead of °F. Explain (think about it, but don't compute) how the estimates in (c) and the confidence intervals in (d) would change.

2.7. a. Consider the model

$$y_i = \beta + \epsilon_i,\ E(\epsilon_i) = 0,\ V(\epsilon_i) = \sigma^2,$$
$$\text{Cov}(\epsilon_i, \epsilon_j) = 0 \text{ for } i \neq j,\ i = 1, 2, \ldots, n$$

Find the LSEs of β and σ^2.

b. Discuss the following statements:

i. For the linear model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1, 2, \ldots, n$$

a 95% confidence interval for $\mu_k = \beta_0 + \beta_1 x_k$ is narrower than a 95% prediction interval for y_k.

ii. For the linear model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, 2, \ldots, n$$

a 99% prediction interval for y_k is wider than a 95% prediction interval for y_k.

iii. For a certain regression situation it is reported that SST = 25, SSR = 30, and SSE = −5. These calculations are correct since SST = SSR + SSE.

2.8. Consider the annual number of cars sold and the revenues of the 10 largest car companies:

Company	Cars Sold (Millions)	Revenues (in Million Euros)
General Motors	8,149	1,996
Ford/Volvo	7,316	2,118
Renault/Nissan	4,778	1,174
Volkswagen	4,580	943
DaimlerChrysler	4,506	1,813
Toyota	4,454	1,175
Fiat	2,535	628
Honda	2,291	605
PSA	2,278	465
BMW	1,187	447

Consider the results of a simple linear regression model of y = revenues on x = sales:

a. Test whether the number of cars sold is an important predictor variable (use significance level 0.05).

b. Calculate a 95% confidence interval for the regression coefficient of number of cars sold.

c. Calculate a 90% confidence interval for the regression coefficient of number of cars sold.

d. Obtain the coefficient of determination.

e. Determine the standard deviation among revenues (y), after factoring in the explanatory variable sales (x). Compare this standard deviation to the standard deviation of y without considering the explanatory variable.

f. Estimate the revenues for BMW.

2.9. Grade point averages of 12 graduating MBA students, GPA, and their GMAT scores taken before entering the MBA program are given below. Use the GMAT scores as a predictor of GPA, and conduct a regression of GPA on GMAT scores.

x = GMAT	y = GPA
560	3.20
540	3.44
520	3.70
580	3.10
520	3.00
620	4.00
660	3.38
630	3.83
550	2.67
550	2.75
600	2.33
537	3.75

a. Obtain and interpret the coefficient of determination R^2.

b. Calculate the fitted value for the second person.

c. Test whether GMAT is an important predictor variable (use significance level 0.05).

2.10. The following are the results of a regression of fuel efficiency (gallons per 100 miles traveled) on the weight (in pounds) of the car. A total of 45 cars were considered.

The regression equation is

Gall/100 miles = 0.560 + 0.00102 Weight

Predictor	Coef	SE Coef	t	p
Constant	0.5598	0.1983	2.82	0.007
Weight	0.00102418	0.00007103	14.42	0.000

$R^2 = 82.9\%$ $R^2(\text{adj}) = 82.5\%$

Analysis of Variance

Source	DF	SS	MS	F	p
Regression	1	13.709	13.709	207.91	0.000
Error	43	2.835	0.066		
Total	44	16.544			

a. Determine an approximate 95% prediction interval for the fuel efficiency of an automobile weighing 2000 pounds.
The computer output does not give you the information to construct exact prediction intervals. Approximate the prediction intervals, assuming that the sample size n is large enough to allow you to ignore the parameter estimation uncertainty.

b. Determine an approximate 95% prediction interval for the fuel efficiency of an automobile weighing 1500 pounds.

2.11. Discuss the functional relationship between the coefficient of determination R^2 and the F ratio.

2.12. Occasionally, a model is considered in which the intercept is known to be zero a priori. Such a model is given by

$$y_i = \beta_1 x_i + \epsilon_i, i = 1, 2, \ldots, n$$

where the errors ϵ_i follow the usual assumptions.

a. Obtain the LSEs $\left(\hat{\beta}_1, s^2\right)$ of $\left(\beta_1, \sigma^2\right)$.

b. Define $e_i = y_i - \hat{\beta}_1 x_i$. Is it still true that $\sum_{i=1}^{n} e_i = 0$? Why or why not?

c. Show that $V\left(\hat{\beta}_1\right) = \sigma^2 / \sum_{i=1}^{n} x_i^2$.

2.13. The data listed in the file **sriver** include the water content of snow on April 1 (x) and the water yield from April to July (y) in the Snake River watershed in Wyoming. Information on $n = 17$ years (from 1919 to 1935) is listed (see Weisberg, 1980).

a. Fit a regression through the origin ($y = \beta_1 x + \epsilon$), and find $\hat{\beta}_1$ and s^2. Obtain a 95% confidence interval for β_1.

b. A more general model for the data includes an intercept,

$$y = \beta_0 + \beta_1 x + \epsilon.$$

Is there convincing evidence that suggests that the simpler model in (a) is an appropriate representation?

2.14. Often, researchers need to calibrate measurement processes. For that they use a set of known x's to obtain observed y's, then fit a model called the calibration model and use this model to convert future measured y's back into the corresponding x's.

The following is an example taken from analytical chemistry where the process is the assay of the element calcium. Determining calcium in the presence of other elements is quite tricky. The following table records the quantities of calcium in carefully prepared solutions (x) and the corresponding analytical results (y):

x	4	8	12.5	16	20	25	31	36	40	40
y	3.7	7.8	12.1	15.6	19.8	24.5	31.1	35.5	39.4	39.5

a. Fit a simple linear regression of y as a function of x. List the assumptions that you make.

b. Calculate a 95% confidence interval for the intercept of your model.

c. Calculate a 95% confidence interval for the slope of your model.

d. In this context two properties may be expected:

i. When $x = 0$, then $y = 0$; if there is no calcium present, your technique should not find any.

ii. If the empirical technique is any good at all, then the slope in the simple linear regression should be 1.

Is there evidence for (i)? For (ii)?

e. If you accept (i) as a condition to be imposed on the model a priori, then the model reduces to

$$y = \beta x + \epsilon$$

Redo part (c) and reexamine property (ii) for your new model.

f. Explain why the results in (d) and (e) are different.

2.15. The following data give the monthly machine maintenance cost (y) in hundreds of dollars and the number of machine hours per month (x), taken over the last 7 months.

Cost (y)	Hours (x)
26	110
25	98
20	121
18	116
30	90
40	88
30	84

a. Fit a linear regression. Construct the ANOVA table. Find R^2 and test the hypothesis that $\beta_1 = 0$ using the F ratio.

b. Obtain the standard errors of β_0 and β_1. Using the t distribution, test the hypothesis: (i) $\beta_0 = 0$; (ii) $\beta_1 = 0$. Construct a 99% confidence interval for β_1.

c. Find the fitted value $\hat{\mu}$ at $x = 100$ and estimate its standard error. Calculate the 95% confidence interval for $\beta_0 + \beta_1 100$.

d. Repeat (c), only this time take $x = 84$. Explain the change in the interval length.

2.16. A company builds custom electronic instruments and computer components. All jobs are manufactured to customer specifications. The firm wants to be able to estimate its overhead cost. As part of a preliminary investigation, the firm decides to focus on a particular department and investigates the relationship between total departmental overhead cost (y) and total direct labor hours (x). The data for the most recent 16 months are given below. They are also given in the file **overhead**.

Month Number	Total Departmental Overhead (y)	Total Direct Labor Hours (x)
1	25,835	878
2	27,451	1,088
3	28,611	1,281
4	32,361	1,340
5	28,967	1,090
6	24,817	1,067
7	29,795	1,188
8	26,135	928
9	31,361	1,319
10	26,006	790
11	27,812	934
12	28,612	871
13	22,992	781
14	31,836	1,236
15	26,252	902
16	26,977	1,140

The two objectives of this investigation are

a. Summarize for management the relationship between total departmental overhead and total direct labor hours.

b. Estimate the expected and predict the actual total departmental overhead from the total direct labor hours.

Analyze the overhead data and write a brief paragraph for your manager that summarizes the results that you have obtained about the two objectives. Include the computer output that you think is necessary for clarification of your discussion.

2.17. The following data, in the file **turtles**, are measurements on length and width (both in mm) of 10 painted female turtles (*Chrysemys picta marginta*):

Length (y)	Width (x)
100	81
103	86
109	88
123	94
133	102
134	100
137	98
141	105
150	107
155	115

a. Plot y against x on a scatter plot. Comment on this plot.

b. Assuming the model $y = \beta_0 + \beta_1 x + \epsilon$, obtain the LSEs of the coefficients and their corresponding 95% confidence limits.

c. Graph the fitted line on the plot in (a). Is this a good fit? Explain.

d. Predict the length of a female turtle if it is 100 mm wide, and obtain the 95% prediction limits.

e. Is the linear relationship a strong or a weak one? Explain.

2.18. The following data, in the file **bloodpressure**, are measurements of systolic blood pressure (SBP) and age for a sample of 15 individuals older than age 40 years:

SBP (y)	Age (x)
164	65
220	63
133	47
146	54
162	60
144	44
166	59
152	64
140	51
145	49
135	57
150	56
170	63
122	41
120	43

a. Plot systolic blood pressure against age. Comment on the plot.

b. Assuming the model $y = \beta_0 + \beta_1 x + \epsilon$, obtain the fitted equation.

c. Construct an ANOVA table for the simple linear model.

d. Use the results from the ANOVA table and the F ratio to test for a significant linear relationship at the 5% level.

e. Test the hypothesis $H_0 : \beta_1 = 0$ at the 5% level using a t test. Does your conclusion agree with the finding in part (d)?

f. Do you think that the individual with $x = 63$ and $y = 220$ is an unusual observation? Why? Check if this observation is influential. Remove it from the data set and redo steps (b)–(e). The observation is influential if there are substantial changes in the resulting fit. Do you think that there are substantial changes? Explain.

2.19. An experiment was conducted to determine the extent to which the growth rate of a certain fungus can be affected by filling test tubes containing the same medium at the same temperature with different inert gases. Three such experiments were performed for each of six different gases, and the average growth rate from these three tests was used as the response. The following table gives the molecular weight (x) of each gas used and the average growth rate (y) in milliliters per hour:

Gas	Average Growth Rate (y)	Molecular Weight (x)
A	3.85	4.0
B	3.48	20.2
C	3.27	28.2
D	3.08	39.9
E	2.56	83.8
F	2.21	131.3

a. Find the LSEs of the slope and the intercept for the linear model $y = \beta_0 + \beta_1 x + \epsilon$, and draw the fitted line on the scatter plot.

b. Is there a significant linear relationship between y and x at the 1% level? Comment on the fit of the line.

c. What information has not been used that may improve the sensitivity of the analysis?

d. Would it be appropriate to use this fitted line to estimate the growth rate of the fungus for a gas with a molecular weight of 200? Explain.

2.20. An investigation involving five factors has singled out temperature as having the greatest impact on the accelerated lifetime of a special type of heater. On the advice of the process engineer, temperatures 1,520, 1,620, 1,660, and 1,708° F were chosen.

Twenty-four heaters were selected at random from the current production and split randomly among the four temperatures. The life times of these heaters are given below.

Temperature						
T	Lifetime y (Hours)					
1,520	1,953	2,135	2,471	4,727	6,143	6,314
1,620	1,190	1,286	1,550	2,125	2,557	2,845
1,660	651	837	848	1,038	1,361	1,543
1,708	511	651	651	652	688	729

a. Plot the data and summarize the important features of the relationship.

b. Transform the y's to $\text{LY} = \ln y$ and replot the data. Comment on the functional relationship.

c. Fit the model

$$\text{LY} = \beta_0 + \beta_1 T + \epsilon$$

   i. Assess the fit by adding the fitted line to the scatter plot.

   ii. If you are not satisfied with the fit, state why. What other approach might you take to get a better fitting model?

2.21. The data are taken from Roberts, H. V., and Ling, R. F. *Conversational Statistics with IDA*. New York: Scientific Press/McGraw-Hill, 1982.

The iron content of crushed blast furnace slag needs to be determined. Two methods are available. One involves a chemical analysis in the laboratory, which is time-consuming and expensive. The other is a much cheaper and quicker magnetic test that can be carried out on-site. Measurements on 53 consecutive slags are listed below. The data are given in the file **ironcontent**.

Graph the results of the chemical test for iron (y) against the magnetic test (x). Fit a simple linear regression. Calculate and interpret the coefficient of determination R^2. Investigate the extent to which the results of the chemical tests of iron content can be predicted from a magnetic test of iron content.

y = Chemical	x = Magnetic
24	25
16	22
24	17
18	21
18	20
10	13
14	16
16	14
18	19
20	10
21	23
20	20
21	19
15	15
16	16
15	16
17	12
19	15
16	15
15	15
15	15
13	17
24	18
22	16
21	18
24	22
15	20
20	21
20	21
25	21
27	25
22	22
20	18
24	21
24	18
23	20
29	25
27	20
23	18
19	19
25	16
15	16
16	16
27	26

$y =$ Chemical	$x =$ Magnetic
27	28
30	28
29	30
26	32
25	28
25	36
32	40
28	33
25	33

2.22. The data are taken from Mosteller, F., Rourke, R. E. K. and Thomas, G. B.: *Probability with Statistical Applications*, (2nd ed.). Reading, MA: Addison-Wesley, 1970.

Average percentage memory retention was measured against passing time (in minutes). The measurements were taken five times during the first hour after the experimental subjects memorized a list of disconnected items and then at various times up to 1 week later. The data given in the file **memory**.

Graph memory retention (y) against time (x). Consider transformations such as the logarithm of y and/or the logarithm of x. Estimate and check the appropriate regression models.

A model such as $y = \alpha \exp(-\beta \text{Time})$ indicates geometric loss of memory. Discuss whether this is an appropriate model or whether there are other models that are equally (or better) suited to describe the data.

$x =$ Time (Minutes)	$y =$ Memory Retention (%)
1	0.84
5	0.71
15	0.61
30	0.56
60 (1 hour)	0.54
120	0.47
240	0.45
480	0.38
720	0.36
1,440	0.26
2,880	0.20
5,760	0.16
10,080	0.08

2.23. The data are taken from Gilchrist, W. *Statistical Modelling*. Chichester, UK: Wiley, 1984. These data give the distance by road and the straight line distance between 20 different pairs of points in Sheffield. The data are given in the file **distance**.

What is the relationship between the two variables? How well can you predict the road distance (y) from the linear distance (x)?

$x =$ Linear Distance	$y =$ Road Distance
9.5	10.7
5.0	6.5
23.0	29.4
15.2	17.2
11.4	18.4
11.8	19.7
12.1	16.6
22.0	29.0
28.2	40.5
12.1	14.2
9.8	11.7
19.0	25.6
14.6	16.3
8.3	9.5
21.6	28.8
26.5	31.2
4.8	6.5
21.7	25.7
18.0	26.5
28.0	33.1

2.24. The data are taken from Risebrough, R. W. Effects of environmental pollutants upon animals other than man. In *Proceedings of the 6th Berkeley Symposium on Mathematics and Statistics, VI*. Berkeley: University of California Press, 1972, pp. 443–463.

Polychlorinated biphenyl (PCB), an industrial pollutant, is thought to have harmful effects on the thickness of egg shells. The amount of PCB (in parts per million) and the thickness of the shell (in millimeters) of 65 Anacapa pelican eggs are given below. The data are also given in the file **pelicaneggs**.

Investigate the relationship between the thickness of the shell and the amount of PCB in pelican eggs. Construct a scatter plot and fit a linear regression model. Calculate a 95% confidence interval for the slope. Obtain the ANOVA table and the coefficient of determination R^2. Interpret the results and comment on the adequacy of the model.

x = Concentration of PCB	y = Thickness
452	0.14
139	0.21
166	0.23
175	0.24
260	0.26
204	0.28
138	0.29
316	0.29
396	0.30
46	0.31
218	0.34
173	0.36
220	0.37
147	0.39
216	0.42
216	0.46
206	0.49
184	0.19
177	0.22
246	0.23
296	0.25
188	0.26
89	0.28
198	0.29
122	0.30
250	0.30
256	0.31
261	0.34
132	0.36
212	0.37
171	0.40
164	0.42
199	0.46
115	0.20
214	0.22
177	0.23
205	0.25
208	0.26
320	0.28
191	0.29
305	0.30
230	0.30
204	0.32
143	0.35
175	0.36
119	0.39
216	0.41
185	0.42
236	0.47
315	0.20
356	0.22
289	0.23
324	0.26
109	0.27
265	0.29
193	0.29
203	0.30
214	0.30
150	0.34
229	0.35
236	0.37
144	0.39
232	0.41
87	0.44
237	0.49

2.25. The data are taken from Wallach, D., and Goffinet, B. Mean square error of prediction in models for studying ecological and agronomic systems. *Biometrics*, 43, 561–573, 1987.

The energy requirements (in Mcal/day) for a sample of 64 grazing merino sheep are given below, together with their body weights (kg). The data are given in the file **energyrequirement**. Construct a scatter plot and establish a model that explains the energy requirements as a linear function of body weight. Obtain a 95% confidence interval for the slope. Calculate and interpret the coefficient of determination R^2. Comment on the adequacy of the model. Discuss whether or not the variance of the measurements is constant across weight.

x = Weight	y = Energy Requirement
22.1	1.31
26.2	1.27
33.2	1.25
34.3	1.14
49.0	1.78
52.6	1.70
27.6	1.39
31.0	1.47
32.6	1.75
44.6	2.25
52.6	3.73
28.6	2.13
34.4	1.85
25.1	1.46
27.0	1.21
33.2	1.32
34.9	1.00
49.2	2.53
53.3	2.66
28.4	1.27
31.0	1.50
33.1	1.82
52.1	2.67
46.7	2.21
29.2	1.80
34.4	1.63
25.1	1.00
30.0	1.23
33.2	1.47
42.6	1.81
51.8	1.87
23.9	1.37
28.9	1.74
31.8	1.60
34.1	1.36
52.4	2.28
37.1	2.11
26.2	1.05
26.4	1.27
25.7	1.20
30.2	1.01
33.9	1.03
43.7	1.73
51.8	1.92
25.1	1.39
29.3	1.54
32.0	1.67
34.2	1.59
52.7	3.15
31.8	1.39
45.9	2.36
27.5	0.94
25.9	1.36
30.2	1.12
33.8	1.46
44.9	1.93
52.5	1.65
26.7	1.26
29.7	1.44
32.1	1.80
44.4	2.33
53.1	2.73
36.1	1.79
36.8	2.31

2.26. The data are taken from Atkinson, A. C. *Plots, Transformations, and Regression*. Oxford: Clarendon Press, 1985.

Here, we list 17 observations on the boiling point (°F) and the barometric pressure (in inches of mercury). The data are given in the file **boiling**. Relate boiling point to barometric pressure. Construct a scatter plot and establish a model that relates the boiling point to barometric pressure. Test the regression coefficients for their significance. Calculate and interpret the coefficient of determination R^2. Comment on the fit and the adequacy of the model.

Note that this exercise deals with the same problem as Exercise 2.6 but uses different data. Plot the data of the two exercises on the same graph, and add the two fitted regression lines that you found. Comment on the graph.

y = Boiling Point	x = Barometric Pressure
194.5	20.79
194.3	20.79
197.9	22.40
198.4	22.67

y = Boiling Point	x = Barometric Pressure
199.4	23.15
199.9	23.35
200.9	23.89
201.1	23.99
201.4	24.02
201.3	24.01
203.6	25.14
204.6	26.57
209.5	28.49
208.6	27.76
210.7	29.04
211.9	29.88
212.2	30.06

2.27. The data are taken from Bissell, A. F. Lines through the origin—IS NO INT the answer? *Journal of Applied Statistics*, 19, 193–210, 1992.

In a chemical process, batches of liquid are passed through a bed containing a certain ingredient. The ingredient gets absorbed by the liquid, and usually approximately 6–6.5% of the weight of the ingredient gets absorbed. In order to be sure that there is enough material, the bed is supplied with approximately 7.5% material. Excess material is costly and should be minimized because any excess cannot be recovered.

The interest is in the relationship between the material supplied (x) and the amount and/or the percentage of absorption. Develop appropriate regression models for both expressions of the response (in kg and as a percent) and comment on their fit and adequacy. The data are given in the file **absorption**.

x = Liquid	y = Take-up (kg)	y = Take-up (%)
310	14.0	4.52
330	17.1	5.18
370	21.3	5.76
400	20.4	5.10
450	27.4	6.09
490	27.2	5.55
520	28.4	5.46
560	32.5	5.80
580	31.9	5.50
650	34.1	5.25
650	39.8	6.12
650	38.5	5.92
760	50.4	6.63
800	43.8	5.48
810	50.4	6.22
910	53.5	5.88
1,020	71.3	6.99
1,020	64.3	6.30
1,160	79.6	6.86
1,200	80.8	6.73
1,230	78.5	6.38
1,380	98.9	7.17
1,460	105.6	7.23
1,490	98.6	6.62

2.28. Search the Web for useful regression applets. Many such applets are available.

After entering points on a scatter plot, these applets calculate the least squares estimates, draw in the fitted regression line, and calculate summary statistics, such as the correlation coefficient or the coefficient of determination R^2. Applets allow you to change points and they illustrate the effect of such changes on the regression results.

Applets also illustrate the standard errors of the estimates. They take repeated samples of a certain size from a given population of points and for each sample they calculate an estimate of the regression slope. The results of repeated draws from the population are displayed in the form of histograms. This illustrates the sampling variability of the estimate.

Experiment with these applets and write a short discussion of what you can learn from them. Note that these applets are designed for the bivariate regression situation mostly because it is difficult to draw observations in higher dimensional space.