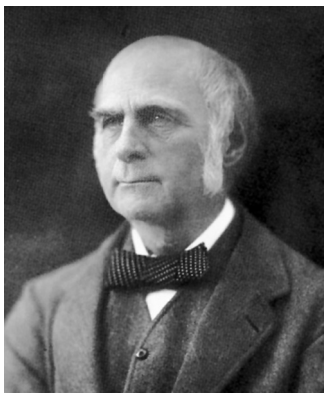
Chapter 7

Linear regression models

Chapter outline

Objective

In this chapter we will study linear relationships in sample data and use the method of least squares to estimate the necessary parameters.



Sir Francis Galton
(Source: http://en.wikipedia.org/wiki/Francis_Galton).

English scientist Sir Francis Galton (1822−1911), a cousin of Charles Darwin, made significant contributions to both genetics and psychology. He was the inventor of regression and a pioneer in applying statistics to biology. One of

Mathematical Statistics with Applications in R. https://doi.org/10.1016/B978-0-12-817815-7.00007-5

the data sets that he considered consisted of the heights of fathers and first sons. He was interested in predicting the height of a son based on the height of a father. Looking at the scatterplots of these heights, Galton saw that the trend was linear and increasing. After fitting a line to these data (using the techniques described in this chapter), he observed that for fathers whose heights were taller than the average, the regression line predicted that taller fathers tended to have shorter sons and shorter fathers tended to have taller sons. There is a regression toward the mean. That is how the method of this chapter got its name: regression.

7.1 Introduction

In earlier chapters, we were primarily concerned about inferences on population parameters. In this chapter, we examine the relationship between one or more variables and create a model that can be used for predictive purposes. For example, consider the question, "Is there statistical evidence to conclude that the countries with the highest average blood-cholesterol levels have the greatest incidence of heart disease?" It is important to answer this if we want to make appropriate lifestyle and medical choices. We will study the relationship between variables using regression analysis. Our aim is to create a model and study inferential procedures when one dependent and several independent variables are present. We denote by Y the random variable to be predicted, also called the *dependent* variable (or response variable) and by x_i the *independent* (or predictor) variables used to model (or predict) Y. For example, let (x, y) denote the height and weight of an adult male. Our interest may be to find the relationship between height and weight from sample measurements of n individuals. The process of finding a mathematical equation that best fits the noisy data is known as *regression analysis*. In his book *Natural Inheritance*, Sir Francis Galton introduced the word *regression* in 1889 to describe certain genetic relationships. The technique of regression is one of the most popular statistical tools to study the dependence of one variable with respect to another. There are different forms of regression: *simple linear, nonlinear, multiple,* and others. The primary use of a regression model is prediction. When using a model to predict Y for a particular set of values of $x_1, \ldots, x_k$, one may want to know how large the error of prediction might be. Regression analysis, in general after collecting the sample data, involves the following steps.

Procedure for regression modeling

1. Hypothesize the form of the model as $Y = f(x_1, \ldots, x_k; \beta_0, \beta_1, \ldots, \beta_k) + \varepsilon$. Here ε represents the random error term. We assume that $E(\varepsilon) = 0$ but $Var(\varepsilon) = \sigma^2$ is unknown. From this we can obtain $E(Y) = f(x_1, \ldots, x_k; \beta_0, \beta_1, \ldots, \beta_k)$.
2. Use the sample data to estimate unknown parameters in the model.
3. Check for goodness of fit of the proposed model.
4. Use the model for prediction.

The function $f(x_1, \ldots, x_k; \beta_0, \beta_1, \ldots, \beta_k)$ $(k \geq 1)$ contains the independent or predictor variables $x_1, \ldots, x_n$ (assumed to be nonrandom) and unknown parameters or weights $\beta_0, \beta_1, \ldots, \beta_k$ and ε representing the random or error variable. We now proceed to introduce the simplest form of a regression model, called simple linear regression.

7.2 The simple linear regression model

Consider a random sample of n observations of the form $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where X is the independent variable and Y is the dependent variable, both being scalars. A preliminary descriptive technique for determining the form of relationship between X and Y is the *scatter diagram* or the *scatterplot*. A scatter diagram is drawn by plotting the sample observations in Cartesian coordinates. The pattern of the points gives an indication of a linear relationship, nonlinear relationship, or no relationship between the variables. A no relationship may indicate that events are happening randomly and any effort to predict based on those data will be futile. Thus, we can consider the scatterplots as visualization and discovery tools. In practice with very large data sets, scatterplots may show trends, clusters, patterns, and relationships among the data points. In this chapter, we will use the scatterplots for identifying only possible linear or nonlinear relationships.

In Fig. 7.1A, the relationship between x and y is fairly linear, whereas the relationship is somewhat like a parabola in Fig. 7.1B, and in Fig. 7.1C there is no obvious relationship between the variables.

Once the scatter diagram reveals a linear relationship, the problem then is to find the linear model that best fits the given data. To this end, we will first give a general definition of a linear statistical model, called a multiple linear regression model.

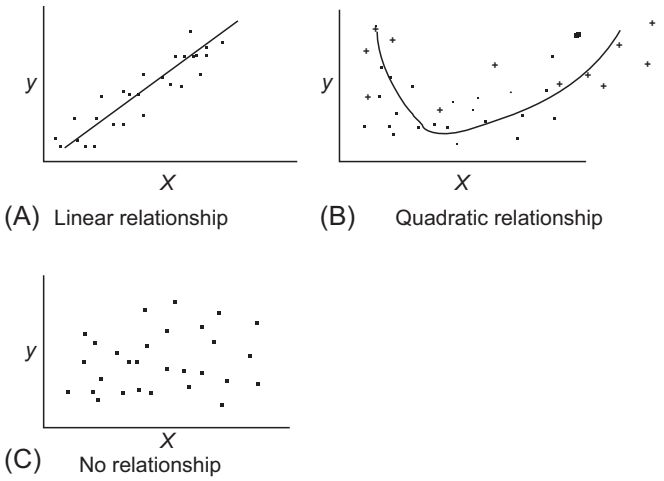


FIGURE 7.1 Scatter diagrams.

Definition 7.2.1 *A* **multiple linear regression model** *relating a random response* Y *to a set of predictor variables* $x_1, \ldots, x_k$ *is an equation of the form*

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon,$$

where $\beta_0, \ldots, \beta_k$ are unknown parameters, $x_1, \ldots, x_k$ are the independent nonrandom variables, and ε is a random variable representing an error term. We assume that $E(\varepsilon) = 0$, or equivalently,

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k.$$

To understand the basic concepts of regression analysis, we shall consider a single dependent variable Y and a single independent nonrandom variable x. We assume that there are no measurement errors in x_i. The possible measurement errors in y and the uncertainties in the assumed model are expressed through the random error ε. Our inability to provide an exact model for a natural phenomenon is expressed through the random term ε, which will have a specified probability distribution (such as a normal) with mean zero. Thus, one can think of Y as having a deterministic component, $E(Y)$, and a random component, ε. If we take $k = 1$ in the multiple linear regression model, we have a simple linear regression model.

Definition 7.2.2 *If* $Y = \beta_0 + \beta_1 x + \varepsilon$, this is called a **simple linear regression model.** *Here,* β_0 *is the y-intercept of the line and* β_1 *is the slope of the line. The term* ε *is the error component.*

This basic linear model assumes the existence of a linear relationship between the variables x and y that is disturbed by a random error ε. The known data points are the pairs $(x_1, y_2), (x_2, y_2), \ldots, (x_n, y_n)$; the problem of simple linear regression is to fit a straight line optimal in some sense to the set of data, as shown in Fig. 7.2.

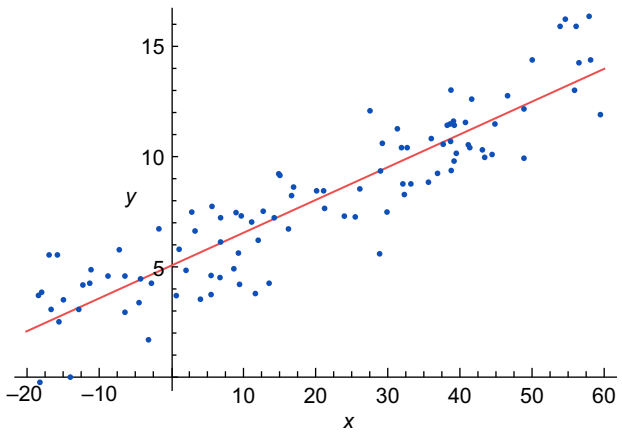


FIGURE 7.2 Scatterplot and least-squares regression line.

Now the problem becomes one of finding estimators for β_0 and β_1. Once we obtain the "good" estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$, we can fit a line to the data given by the prediction equation $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$. Unlike the single-variable estimation problems, now the response variable is dependent on the independent variables and thus estimators have to reflect this aspect. The question then becomes whether this predicted line gives the "best" (in some sense) description of the data. This necessitates a new method of estimation. We now describe the most widely used technique, called the method of least squares, to obtain the estimators or weights of the parameters.

7.2.1 The method of least squares

As stated, (x_1, y_1), (x_2, y_2), …, (x_n, y_n) are the n observed data points, with corresponding errors ε_i, $i = 1, \ldots, n$. That is,

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$

We assume that the errors ε_i, $i = 1, \ldots, n$ are independent and identically distributed with $E(\varepsilon_i) = 0$, $i = 1, \ldots, n$, and $Var(\varepsilon_i) = \sigma^2$, $i = 1, \ldots, n$. One of the ways to decide on how well a straight line fits the set of data is to determine the extent to which the data points deviate from the line. The straight line model for the response Y for a given x is

$$Y = \beta_0 + \beta_1 x + \varepsilon.$$

Because we assumed that $E(\varepsilon) = 0$, the expected value of Y is given by

$$E(Y) = \beta_0 + \beta_1 x.$$

The estimator of the $E(Y)$, denoted by $\widehat{Y}$, can be obtained by using the estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ of the parameters β_0 and β_1, respectively. Then, the fitted regression line we are looking for is given by

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 x.$$

For observed values (x_i, y_i), we obtain the estimated value of y_i as

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i.$$

The deviation of observed y_i from its predicted value $\widehat{y}_i$, called the ith residual, is defined by

$$e_i = (y_i - \widehat{y}_i) = \left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)\right].$$

The residuals, or errors e_i, are the vertical distances between observed and predicted values of $y_i's$ (Fig. 7.3).

Definition 7.2.3 The sum of squares for errors (SSE) or sum of squares of the residuals for all of the n data points is

$$SSE = \sum_{i=1}^{n} e_1^2 = \sum_{i=1}^{n} \left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)\right]^2.$$

The least-squares approach to estimation is to find $\widehat{\beta}_0$ and $\widehat{\beta}_1$ that minimize the sum of squared residuals, SSE. Thus, in the method of least squares, we choose β_0 and β_1 so that SSE is a minimum. The quantities $\widehat{\beta}_0$ and $\widehat{\beta}_1$ that make the SSE a minimum are called the *least-squares estimates* of the parameters β_0 and β_1, and the corresponding line $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$ is called the *least-squares line.*

Definition 7.2.4 The *least-squares* line $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$ is one that satisfies the following property:

$$SSE = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2,$$

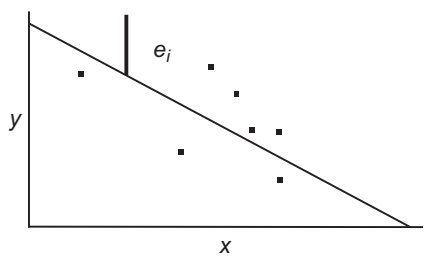


FIGURE 7.3 Illustration of e_i.

is a minimum for any other straight line model with the sum of errors (SE) being

$$SE = \sum_{i=1}^{n}(y_i - \widehat{y}_i) = 0.$$

Thus, the least-squares line is a line of the form $y = b_0 + b_1 x$ for which the error sum of squares $\sum_{i=1}^{n}(y_i - b_0 - b_1 x)^2$ is a minimum. The minimum is taken over all values of b_0 and b_1, and (x_1, y_1), (x_2, y_2), …, (x_n, y_n) are observed data pairs.

The problem of fitting a least-squares line now reduces to finding the quantities $\widehat{\beta}_0$ and $\widehat{\beta}_1$ that minimize the error sum of squares.

7.2.2 Derivation of $\widehat{\beta}_0$ and $\widehat{\beta}_1$

Now we derive expressions for $\widehat{\beta}_0$ and $\widehat{\beta}_1$ using the methods of calculus. If *SSE* attains a minimum, then the partial derivatives of *SSE* with respect to β_0 and β_1 are zeros. That is,

$$\frac{\partial SSE}{\partial \beta_0} = \frac{\partial\left\{\sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)]^2\right\}}{\partial \beta_0}$$

$$= -\sum_{i=1}^{n} 2[y_i - (\beta_0 + \beta_1 x_i)] \tag{7.1}$$

$$= 2\left(\sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i\right) = 0,$$

and

$$\frac{\partial SSE}{\partial \beta_1} = \frac{\partial\left\{\sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)]^2\right\}}{\partial \beta_1}$$

$$= -\sum_{i=1}^{n} 2[y_i - (\beta_0 + \beta_1 x_i)]x_i \tag{7.2}$$

$$= -2\left(\sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2\right) = 0.$$

Eqs. (7.1) and (7.2) are called the *least-squares equations* for estimating the parameters of a line. From (7.1) and (7.2) we obtain a set of linear equations called the *normal equations*,

$$\sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i, \tag{7.3}$$

and

$$\sum_{i=1}^{n} x_i y_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2. \tag{7.4}$$

Solving for β_0 and β_1 from Eqs. (7.3) and (7.4), we obtain

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_1^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n} x_i y_i - \dfrac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_1^2 - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}, \tag{7.5}$$

and

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}. \tag{7.6}$$

To simplify the formula for $\widehat{\beta}_1$, set

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}, S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n},$$

we can rewrite Eq. (7.5) as

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}}.$$

It can be shown (by using the second derivatives) that Eqs. (7.5) and (7.6) do indeed minimize *SSE*. Now we will summarize the procedure for fitting a least-squares line.

Procedure for fitting a least-squares line

1. Form the n data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, and compute the following quantities: $\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i^2, \sum_{i=1}^{n} y_i, \sum_{i=1}^{n} y_i^2$, and $\sum_{i=1}^{n} x_i y_i$.

   Also compute the sample means, $\overline{x} = (1/n)\sum_{i=1}^{n} x_i$ and $\overline{y} = (1/n)\sum_{i=1}^{n} y_i$.

2. Compute

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

and

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}).$$

3. Compute $\widehat{\beta}_0$ and $\widehat{\beta}_1$ by substituting the computed quantities from step 1 into the equations

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

and

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}.$$

4. The fitted least-squares line is

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

For a graphical representation, in the xy-plane, plot all the data points and draw the least-squares line obtained in step 4.

Once we have accomplished the best-fit combination of the two parameters β_0 and β_1, any deviation of either parameter away from its optimum value will cause the sum of squares error to increase. Thus, the optimum combination of the pairs $\left(\widehat{\beta}_0, \widehat{\beta}_1\right)$ forms a global minimum point of the error sum of squares among all possible values of β_0 and β_1 for the given data set.

EXAMPLE 7.2.1

Use the method of least squares to fit a straight line to the accompanying data points. Give the estimates of β_0 and β_1. Plot the points and sketch the fitted least-squares line. The observed data values are given in the following table.

x	−1	0	2	−2	5	6	8	11	12	−3
Y	−5	−4	2	−7	6	9	13	21	20	−9

Solution

Form a table to compute various terms.

x_i	y_i	x_iy_i	x_i^2
−1	−5	5	1
0	−4	0	0
2	2	4	4
−2	−7	14	4
5	6	30	25
6	9	54	36
8	13	104	64
11	21	231	121
12	20	240	144
−3	−9	27	9
$\sum x_i = 38$	$\sum y_i = 46$	$\sum x_iy_i = 709$	$\sum x_i^2 = 408$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = 408 - \frac{(38)^2}{10} = 263.6,$$

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} = 709 - \frac{(38)(46)}{10} = 534.2,$$

$$\overline{x} = 3.8 \quad \text{and} \quad \overline{y} = 4.6.$$

Therefore,

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{534.2}{263.6} = 2.0266.$$

and

$$\begin{aligned}\widehat{\beta}_0 &= \overline{y} - \widehat{\beta}_1 \overline{x} \\ &= 4.6 - (2.0266)\,(3.8) = -3.1011.\end{aligned}$$

Hence, the least-squares line for these data is

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x = -3.1011 + 2.0266x$$

and its plot is shown in Fig. 7.4.

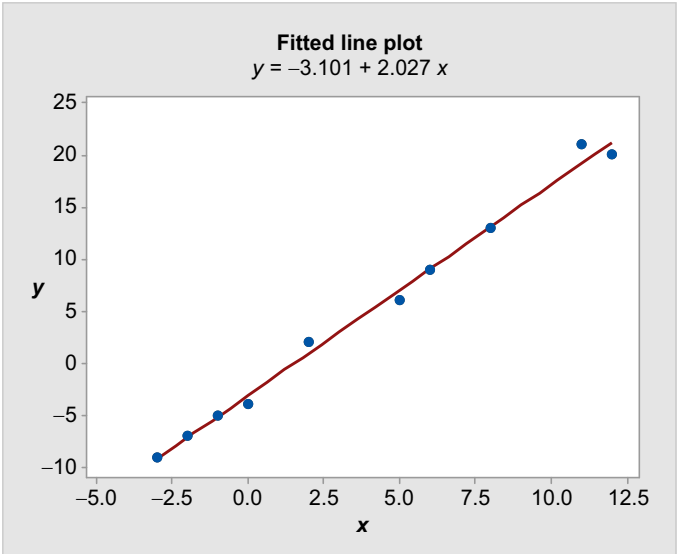


FIGURE 7.4 Simple regression line.

Recall that for the regression line $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$. we have defined *SSE* to be

$$SSE = \sum_{i=1}^{n} \left(y_i - \widehat{y}_i\right)^2 = \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2.$$

We now show that

$$SSE = S_{yy} - \widehat{\beta}_1 S_{xy,} \text{ where } S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2.$$

We know that

$$\begin{aligned} SEE &= \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2 \\ &= \sum_{i=1}^{n} \left(y_i - \overline{y} + \widehat{\beta}_1 \overline{x} - \widehat{\beta}_1 x_i\right)^2 \\ &= \sum_{i=1}^{n} \left[\left(y_i - \overline{y}\right) - \widehat{\beta}_1 (x_i - \overline{x})\right]^2 \\ &= \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2 + \widehat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \overline{x})^2 - 2\widehat{\beta}_1 \sum_{i=1}^{n} (x_i - \overline{x})\left(y_i - \overline{y}\right) \\ &= S_{yy} + \widehat{\beta}_1^2 S_{xx} - 2\widehat{\beta}_1 S_{xy}. \end{aligned}$$

Recall that $\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$.

Substituting for $\widehat{\beta}_1$, we obtain

$$\begin{aligned} SSE &= S_{yy} - \left(\frac{S_{xy}}{S_{xx}}\right)^2 S_{xx} - 2\frac{S_{xy}}{S_{xx}} S_{xy} \\ &= S_{yy} - \frac{S_{xy}}{S_{xx}} S_{xy} \\ &= S_{yy} - \widehat{\beta}_1 S_{xy}. \end{aligned}$$

7.2.3 Quality of the regression

Once we obtain the linear model, the question is, how well does this line fit the data? We could make use of the residuals, that is,

$$\widehat{e}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i,$$

to answer the question and to assess the quality of the fit. If our model is good, then the residual $\widehat{e}_i$ should be close to the random error ε with mean zero. Furthermore, the residuals should contain little or no information about the model, and there should be no recognizable pattern. If we plot the residuals versus the independent variables on the x-axis, ideally, the plot should look like a horizontal blur, the residuals showing no relationship to the x-values, as shown by Fig. 7.5. Otherwise, these plots reveal a not very good fit of the given data, as shown by Fig. 7.6, and we need to improve our model specifications. Thus, a symmetric trend in the plot of residuals e_i versus x_i or $\widehat{y}_i (i = 1, \ldots, n)$ indicates that the assumed regression model is not correct.

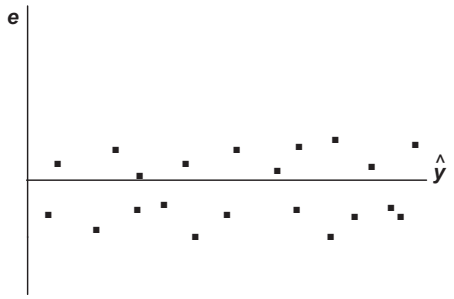


FIGURE 7.5 Good fit.

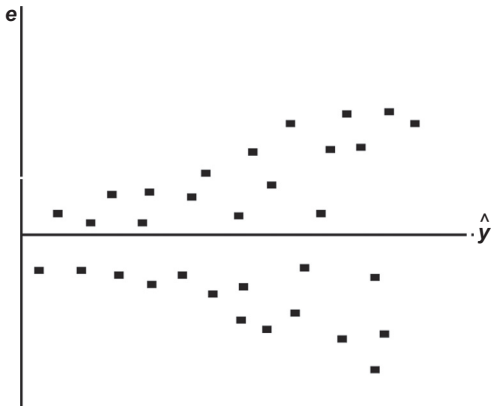


FIGURE 7.6 Not a good fit.

Whereas the residual plots give us a visual representation of the quality of fit, a numerical measure of how well the regression explains the data is obtained by calculating the *coefficient of determination*, also called the R^2 of the regression. Particular (observed) value of realized R^2 is

$$r^2 = \frac{S_{yy} - SSE}{S_{yy}} = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}.$$

Further discussion is given in Project 7B. Regression analysis with any of the standard statistical software packages will contain an output value of the R^2. This value will be between 0 and 1; closer to 11 means a better fit. For example, if the value of R^2 is 0.85, the regression captures 85% of the variation in the dependent variable. This is generally considered good regression.

7.2.4 Properties of the least-squares estimators for the model $Y = \beta_0 + \beta_1 x + \varepsilon$

We discussed in Chapter 4 the concept of sampling distribution of sample statistics such as that of $\overline{X}$. Similarly, knowledge of the distributional properties of the least-squares estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ is necessary to allow any statistical inferences to be made about them. The following result gives the sampling distribution of the least-squares estimators.

Theorem 7.2.1 Let $Y = \beta_0 + \beta_1 x + \varepsilon$ be a simple linear regression model with $\varepsilon \sim N(0, \sigma^2)$, and let the errors ε_i associated with different observations y_i $(i = 1, \ldots, N)$ be independent. Then

(a) $\widehat{\beta}_0$ and $\widehat{\beta}_1$ have normal distributions.
(b) The mean and variance are given by

$$E\left(\widehat{\beta}_0\right) = \beta_0, \quad Var\left(\widehat{\beta}_0\right) = \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\sigma^2,$$

and

$$E\left(\widehat{\beta}_1\right) = \beta_1, \quad Var\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{S_{xx}},$$

where $S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$. In particular, the least-squares estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are unbiased estimators of β_0 and β_1, respectively.

Proof.
We know that

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

$$= \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \overline{x})\left(Y_i - \overline{Y}\right)$$

$$= \frac{1}{S_{xx}} \left[\sum_{i=1}^{n} (x_i - \overline{x})\ Y_i - \overline{Y} \ \sum_{i=1}^{n} (x_i - \overline{x}) \right]$$

$$= \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \overline{x})\ Y_i,$$

where the last equality follows from the fact that $\sum_{i=1}^{n}(x_i - \overline{x}) = \sum_{i=1}^{n} x_i - n\overline{x} = 0$. Because Y_i is normally distributed, the sum $\frac{1}{S_{xx}} \sum_{i=1}^{n}(x_i - \overline{x})Y_i$ is also normal. Furthermore,

$$E\left[\widehat{\beta}_1\right] = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \overline{x}) E[Y_i]$$

$$= \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \overline{x})(\beta_0 + \beta_1 x_i)$$

$$= \frac{\beta_0}{S_{xx}} \sum_{i=1}^{n} (x_i - \overline{x}) + \frac{\beta_1}{S_{xx}} \sum_{i=1}^{n} (x_i - \overline{x}) x_i$$

$$= \beta_1 \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \overline{x}) x_i$$

$$= \beta_1 \frac{1}{S_{xx}} \left[\sum_{i=1}^{n} x_i^2 - \overline{x} \sum_{i=1}^{n} x_i \right]$$

$$= \beta_1 \frac{1}{S_{xx}} \left[\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right) \left(\frac{\sum_{i=1}^{n} x_i}{n} \right) \right]$$

$$= \beta_1 \frac{1}{S_{xx}} \left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i \right)^2}{n} \right]$$

$$= \beta_1 \frac{1}{S_{xx}} S_{xx} = \beta_1.$$

For the variance we have,

$$Var\left[\widehat{\beta}_1\right] = Var\left[\frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\overline{x})Y_i\right]$$

$$= \frac{1}{S_{xx}^2}\sum_{i=1}^{n}(x_i-\overline{x})^2\, Var\,[Y_i] \quad \left(\text{since the } Y_i' \text{s are independent}\right)$$

$$= \sigma^2\frac{1}{S_{xx}^2}\sum_{i=1}^{n}(x_i-\overline{x})^2\left(Var(Y_i) = Var(\beta_0+\beta_1+\varepsilon_i) = Var(\varepsilon_i) = \sigma^2\right)$$

$$= \frac{\sigma^2}{S_{xx}}.$$

Note that both $\overline{Y}$ and $\widehat{\beta}_1$ are normal random variables. It can be shown that they are also independent (see Exercise 7.3.3). Because $\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x}$ is a linear combination of $\overline{Y}$ and $\widehat{\beta}_1$, it is also normal. Now,

$$E\left[\widehat{\beta}_0\right] = E\left[\overline{Y} - \widehat{\beta}_1\overline{x}\right] = E\left[\overline{Y}\right] - \overline{x}E\left[\widehat{\beta}_1\right]$$

$$= E\left[\frac{1}{n}\sum_{i=1}^{n}Y_i\right] - \overline{x}\beta_1 = \frac{1}{n}\sum_{i=1}^{n}(\beta_0+\beta_1 x) - \overline{x}\beta_1$$

$$= \beta_0 + \overline{x}\beta_1 - \overline{x}\beta_1 = \beta_0.$$

The variance of $\widehat{\beta}_0$ is given by

$$Var\left[\widehat{\beta}_0\right] = Var\left[\overline{Y} - \widehat{\beta}_1\overline{x}\right]$$

$$= Var\left[\overline{Y}\right] + \overline{x}^2\, Var\left[\widehat{\beta}_1\right] \quad \left(\text{since } \overline{Y} \text{ and } \widehat{\beta}_1 \text{ are independent}\right)$$

$$= \frac{\sigma^2}{n} + \frac{\overline{x}^2\sigma^2}{S_{xx}} = \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\sigma^2.$$

If an estimator $\widehat{\theta}$ is a linear combination of the sample observations and has a variance that is less than or equal to that of any other estimator that is also a linear combination of the sample observations, then $\widehat{\theta}$ is said to be a *best linear unbiased estimator* (BLUE) for θ. The following result states that among all unbiased estimators for β_0 and β_1 that are linear in Y_i, the least-square estimators have the smallest variance.

Gauss–Markov theorem

Theorem 7.2.2 Let $Y = \beta_0 + \beta_1 x + \varepsilon$ be the simple regression model such that for each x_i fixed, each Y_i is an observable random variable and each $\varepsilon = \varepsilon_i$, $i = 1, 2, \ldots, n$ is an unobservable random variable. Also, let the random variable ε_i be such that $E[\varepsilon_i] = 0$, $Var(\varepsilon_i) = \sigma^2$ and $Cov(\varepsilon_i, \varepsilon_j) = 0$, if $i \neq j$. Then the least-squares estimators for β_0 and β_1 are best linear unbiased estimators.

It is important to note that even when the error variances are not constant, there still can exist unbiased least-square estimators, but the least-squares estimators do not have minimum variance.

7.2.5 Estimation of error variance σ^2

The greater the variance, σ^2, of the random error ε, the larger will be the errors in the estimation of model parameters β_0 and β_1. We can use already-calculated quantities to estimate this variability of errors. It can be shown that (see Exercise 7.2.1(b)) that

$$E(SSE) = (n-2)\sigma^2.$$

Thus, an unbiased estimator of the error variance, σ^2, is $\widehat{\sigma}^2 = (SSE)/(n-2)$. We will denote $(SSE)/(n-2)$ by mean square error (MSE).

Exercises 7.2

7.2.1. For a random sample of size n,

(a) Show that the error sum of squares can be expressed by

$$SSE = S_{yy} - \widehat{\beta}_1 S_{xy}.$$

(b) Show that $E[SSE] = (n-2)\sigma^2$.

7.2.2. The following are midterm and final examination test scores for 10 students from a calculus class, where x denotes the midterm score and y denotes the final score for each student.

x	68	87	75	91	82	77	86	82	75	79
y	74	79	80	93	88	79	97	95	89	92

(a) Calculate the least-squares regression line for these data.

(b) Plot the points and the least-squares regression line on the same graph.

7.2.3. The following data give the annual incomes (in thousands of dollars) and amounts (in thousands of dollars) of life insurance policies for eight persons.

Annual income	42	58	27	36	70	24	53	37
Life insurance	150	175	25	75	250	50	250	100

(a) Calculate the least-squares regression line for these data.

(b) Plot the points and the least-squares regression line on the same graph.

7.2.4. Consider a simple linear model $Y = \beta_0 + \beta_1 x + \varepsilon$, with $\varepsilon \sim N(0, \sigma^2)$. Show that

$$cov\left(\widehat{\beta}_0, \widehat{\beta}_1\right) = \frac{-\sigma^2 \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}.$$

7.2.5. **(a)** Show that the least-squares estimates of β_0 and β_1 of a line can be expressed as

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

and

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}.$$

(b) Using part (a), show that the line fitted by the method of least squares passes through the point $(\overline{x}, \overline{y})$.

7.2.6. Crickets make their chirping sounds by rapidly sliding one wing over the other. The faster they move their wings, the higher the number of chirping sounds that are produced. Scientists have noticed that crickets move their wings faster in warm temperatures than in cold temperatures (they also do this when they are threatened). Therefore, by listening to the pitch of the chirp of crickets, it is possible to tell the temperature of the air. The following table gives the number of cricket chirps per 13 s recorded at 10 different temperatures. Assume that the crickets are not threatened.

Temperature	60	66	70	73	78	80	82	87	90	92
Number of chirps	20	25	31	33	36	39	42	48	49	52

Calculate the least-squares regression line for these data and discuss its usefulness.

7.2.7. Consider the regression model

$$Y = \beta_1 x + \varepsilon$$

where $\varepsilon \sim N(0, \sigma^2)$. Show that

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

7.2.8. A farmer collected the following data, which show crop yields for various amounts of fertilizer used.

Fertilizer (pounds/100 sq. ft)	0	4	8	10	15	18	20	25
Yield (bushels)	6	7	10	13	17	18	22	23

(a) Calculate the least-squares regression line for these data.
(b) Plot the points and the least-squares regression line on the same graph.

7.2.9. An economist desires to estimate a line that relates personal disposable income (DI) to consumption expenditures (CE). Both DI and CE are in thousands of dollars. The following gives the data for a random sample of nine households of size four.

DI	25	22	19	36	40	47	28	52	60
CE	21	20	17	28	34	41	25	45	51

(a) Calculate the least-squares regression line for these data.
(b) Plot the points and the least-squares regression line on the same graph.

7.2.10. The following data represent systolic blood pressure readings on 10 randomly selected females between ages 41 and 82.

Age (x)	63	70	74	82	60	44	80	71	71	41
Systolic (y)	151	149	164	157	144	130	157	160	121	125

(a) Calculate the least-squares regression line for these data.
(b) Plot the points and the least-squares regression line on the same graph.

7.2.11. It is believed that exposure to solar radiation increases the pathogenesis of melanoma. Suppose that the following data give sunspot relative number and age-adjusted total incidence (incidence is the number of cases per 100,000 population) for 8 different years in a certain region.

Sunspot relative number	104	12	40	75	110	180	175	30
Incidence total	4.7	1.9	3.8	2.9	0.9	2.7	3.9	1.6

(a) Calculate the least-squares regression line for these data.
(b) Plot the points and the least-squares regression line on the same graph.

TABLE 7.1 Adult Mass and Gestation Period for Mammals.

Species	Adult mass (kg)	Gestation period (weeks)
African elephant	6000	88
Horse	400	48
Grizzly bear	400	30
Lion	200	17
Wolf	34	9
Badger	12	8
Rabbit	2	4.5
Squirrel	0.5	3.5

TABLE 7.2 Gestation Period of Mammals.

Species	Gestation period (weeks)
Indian elephant	89.0
Camel	57.0
Sea lion	51.4
Dog	8.7
Rat	3.0
Hamster	2.3

7.2.12. It is believed that the average size of a mammal species is a major factor in the period of gestation (the period of development in the uterus from conception until birth). In general, it is observed that the bigger the mammal is, the longer the gestation period. Table 7.1 gives adult mass in kilograms and gestation period in weeks of some species (source: http://www.saburchill.com/chapters/chap0037.html).

(a) Calculate the least-squares regression line for these data with adult mass as the independent variable.

(b) Plot the points and the least-squares regression line on the same graph.

(c) Calculate the least-squares regression line for these data with gestation period as the independent variable.

(d) Assuming that the regression model of part (c) holds for all mammals, estimate the adult mass in kilograms for the mammals given in Table 7.2.

7.2.13. Using the Internet, obtain home sales data relating square footage to sale price for 10 randomly selected homes for your area of interest and obtain a least-squares regression line for these data. Test for all the assumptions for this analysis and see if your data satisfy these assumptions.

7.2.14. The following data represent sales volume as a fraction of number of visits to company website.

Average number of visits per month x	100	150	175	200	240	464	530	480	598	650
Sales volume ($1000) y	16	25	27	31	34	88	108	95	132	165

(a) Calculate the least-squares regression line for these data with adult mass as the independent variable.

(b) Plot the points and the least-squares regression line on the same graph.

(c) Calculate the least-squares regression line for these data with average number of visits as the independent variable.

(d) Predict sales volume if the number visits $x = 490$.

7.3 Inferences on the least-squares estimators

Once we obtain the estimators of the slope β_1 and intercept β_0 of the model regression line, we are in a position to use Theorem 7.2.1 to make inferences regarding these model parameters. Using the properties of $\widehat{\beta}_0$ and $\widehat{\beta}_1$, in this section we study the confidence intervals and hypothesis tests concerning these parameters.

From Theorem 7.2.1, we can write

$$Z_1 = \frac{\widehat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{S_{xx}}}} \sim N(0,1).$$

Also, it can be shown that SSE/σ^2 is independent of $\widehat{\beta}_1$ and has a chi-square distribution with $n-2$ degrees of freedom. Let the *mean square error* be defined by

$$MSE = \frac{SSE}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n}\left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)\right]^2.$$

Then using Definition 4.2.2, we have

$$t_{\beta_1} = \frac{Z}{\sqrt{\frac{\left(\frac{SSE}{\sigma^2}\right)}{n-2}}} = \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\frac{MSE}{S_{xx}}}},$$

which follows the t-distribution with $n-2$ degrees of freedom.

Similarly, let

$$Z_0 = \frac{\widehat{\beta}_0 - \widehat{\beta}_0}{\sigma\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{yy}}\right)} \sim N(0,1).$$

Also, it can be shown that $\widehat{\beta}_0$ and SSE are independent. Hence,

$$t_{\beta 0} = \frac{z_0}{\sqrt{\frac{SSE}{\frac{\sigma^2}{n-2}}}} = \frac{\widehat{\beta}_0 - \beta_0}{\left[MSE\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\right]^{1/2}},$$

follows the t-distribution with $n-2$ degrees of freedom.

From these derivations, we can obtain the following procedure about the confidence intervals for the slopes β_1 and for the intercept β_0.

Procedure for obtaining confidence intervals for β_0 and β_1

1. Compute S_{xx}, S_{xy}, S_{yy}, $\overline{y}$, and $\overline{x}$ as in the procedure for fitting a least-squares line.
2. Compute $\widehat{\beta}_1, \widehat{\beta}_0$ using equations $\widehat{\beta}_1 = (S_{xy})/(S_{xx})$ and $\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x}$, respectively.
3. Compute SSE by $SSE = S_{yy} - \widehat{\beta}_1 S_{xy}$.
4. Define MSE to be

$$MSE = \frac{SSE}{n-2},$$

where n = number of pairs of observations $(x_1, y_1), \ldots, (x_n, y_n)$.

5. A$(1-\alpha)$100% confidence interval for β_1 is given by

$$\left(\widehat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{MSE}{S_{xx}}}, \widehat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{MSE}{S_{xx}}}\right)$$

where $t_{a/2}$ is the upper tail $\alpha/2$-point based on a t-distribution with $(n-2)$ degrees of freedom.

6. A$(1-\alpha)$100% confidence interval for β_0 is given by

$$\left(\widehat{\beta}_0 - t_{\alpha/2,n-2}\left[MSE\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\right]^{1/2},\right.$$
$$\left.\widehat{\beta}_0 + t_{\alpha/2,n-2}\left[MSE\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\right]^{1/2}\right).$$

We illustrate this procedure for obtaining confidence limits with the following example.

EXAMPLE 7.3.1

For the data of Example 7.2.1:
(a) Construct a 95% confidence interval for β_0 and interpret.
(b) Construct a 95% confidence interval for β_1 and interpret.

Solution
The following calculations were obtained in Example 7.2.1:

$$S_{xx} = 263.6, S_{xy} = 534.2, \overline{y} = 4.6 \text{ and } \overline{x} = 3.8.$$

Also,

$$\widehat{\beta}_1 = 2.0266, \widehat{\beta}_0 = -3.1011.$$

In addition to those calculations, we can compute

$$\sum_{i=1}^{n} y_i^2 = 1302 \text{ and } S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = 1302 - \frac{(46)^2}{10} = 1090.4.$$

Now,

$$\begin{aligned} SSE &= S_{yy} - \widehat{\beta}_1 S_{xy} \\ &= 1090.4 - (2.0266)(534.2) \\ &= 7.79028. \end{aligned}$$

Hence,

$$MSE = \frac{SSE}{n-2} = \frac{7.79028}{8} = 0.973785.$$

Now from the t-table, we have $t_{0.025,8} = 2.306$.
(a) A 95% confidence interval for β_0 is given by

$$\left(\widehat{\beta}_0 - t_{\alpha|2,n-2}\left[MSE\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\right]^{1/2}, \widehat{\beta}_0 + t_{\alpha|2,n-2}\left[MSE\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\right]^{1/2}\right)$$

$$= \left(-3.1011 - (2.306)\left[\left(0.973785\right)\left(\frac{1}{10} + \frac{(3.8)^2}{263.6}\right)\right]^{1/2}\right.$$

$$\left.-3.1011 + (2.306)\left[\left(0.973785\right)\left(\frac{1}{10} + \frac{(3.8)^2}{263.6}\right)\right]^{1/2}\right).$$

From which we obtain a 95% confidence interval for β_0 as $(-3.9846, -2.2176)$. Thus, we can conclude with at least 95% confidence that the true value of the intercept, β_0, is between -3.9846 and -2.2176.
(b) A 95% confidence interval for β_1 is given by

$$\left(\widehat{\beta}_1 - t_{\alpha|2,n-2}\sqrt{\frac{MSE}{S_{xx}}}, \widehat{\beta}_1 + t_{\alpha|2,n-2}\sqrt{\frac{MSE}{S_{xx}}}\right)$$

$$= \left(2.0266 - (2.306)\sqrt{\frac{0.973785}{236.6}}, 2.0266 + (2.306)\sqrt{\frac{0.973785}{236.6}}\right)$$

from which we obtain a 95% confidence interval for β_1 as (1.8864, 2.1668). Thus, we can conclude with 95% confidence that the true value of the slope of the linear regression model is between 1.8864 and 2.1663.

One of the assumptions for linear regression models that we have made is that the variance of the errors is a constant and independent of x. Errors with this property are called *homoscedastic*. If the variance of the errors is not constant, the errors are

called *heteroscedastic*. In the heteroscedastic case, standard errors and confidence intervals based on the assumption that s^2 is an estimate of σ^2 may be somewhat deceptive.

Now we introduce hypothesis testing concerning the slope and intercept of the fitted least-squares line. We use t_{β_0} and t_{β_1} defined earlier as the test statistic for testing hypotheses concerning β_0 and β_1, respectively. The usual one- and two-sided alternatives apply. We proceed to summarize these test procedures.

Hypothesis test for β_0

One-sided test	**Two-sided test**
H_0: $\beta_0 = \beta_{00}$ (β_{00} is a specific value of β_0)	H_0: $\beta_0 = \beta_{00}$
H_a: $\beta_0 > \beta_{00}$ or $\beta_0 < \beta_{00}$	H_a: $\beta_0 \neq \beta_{00}$
Test statistic:	Test statistic:
$t_{\beta_0} = \frac{\hat{\beta}_0 - \beta_{00}}{\left[MSE\left(\frac{1}{n} + \frac{\overline{x}}{S_{xx}}\right)\right]^{1/2}}$	$t_{\beta_0} = \frac{\hat{\beta}_0 - \beta_{00}}{\left[MSE\left(\frac{1}{n} + \frac{\overline{x}}{S_{xx}}\right)\right]^{1/2}}$
Rejection region:	Rejection region:
$t > t_{\alpha,(n-2)}$, (upper tail region)	$\lvert t\rvert > t_{\alpha/2,(n-2)}$
$t < -t_{\alpha,(n-2)}$, (lower tail region)	
Decision: If t_{β_0} falls in the rejection region, reject the null hypothesis at level of significance α.	
Assumptions: Assume that the errors ε_i, $i = 1, \ldots, n$ are independent and normally distributed with $E(\varepsilon_i) = 0$, $i = 1, \ldots, n$, and $Var(\varepsilon_i) = \sigma^2$, $i = 1, \ldots, n$.	

We now illustrate this procedure with the following example.

EXAMPLE 7.3.2

Using the data given in Example 7.2.1, test the hypothesis H_0: $\beta_0 = -3$ versus H_a: $\beta_0 \neq -3$ using the 0.05 level of significance.

Solution

We test H_0: $\beta_0 = -3$ versus H_a: $\beta_0 \neq -3$.

Here $\beta_{00} = -3$. The rejection region is $t < -2.306$ or $t > 2.306$.

From the calculations of the previous example, we have

$$t_{\beta_0} = \frac{\widehat{\beta}_0 - \beta_{00}}{\left[MSE\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\right]^{1/2}}$$

$$= \frac{-3.1011 - (-3)}{\left[(0.973785)\left(\frac{1}{10} + \frac{(3.8)^2}{263.2}\right)\right]^{1/2}}$$

$$= -0.26041.$$

Because the test statistic does not fall in the rejection region, at $\alpha = 0.05$, we do not reject H_0.

Hypothesis test for β_1

One-sided test	**Two-sided test**
H_0: $\beta_1 = \beta_{10}$ (β_{10} is a specific value of β_1)	H_0: $\beta_1 = \beta_{10}$
H_a: $\beta_1 > \beta_{10}$ or $\beta_1 < \beta_{10}$	H_a: $\beta_1 \neq \beta_{10}$
Test statistic:	Test statistic:
$t_{\beta_1} = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{MSE}{S_{xx}}}}$	$t_{\beta_1} = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{MSE}{S_{xx}}}}$
Rejection region:	Rejection region:
$t > t_{\alpha,(n-2)}$ (upper tail region) $t < -t_{\alpha,(n-2)}$ (lower tail region)	$\lvert t\rvert > t_{\alpha/2,\ (n-2)}$
Decision: If t_{β_1} falls in the rejection region, reject the null hypothesis at level of significance α.	
Assumptions: Assume that the errors ε_i, $i = 1, \ldots, n$ are independent and normally distributed with $E(\varepsilon_i) = 0$, $i = 1, \ldots, n$, and $Var(\varepsilon_i) = \sigma^2$, $i = 1, \ldots, n$.	

The test of hypothesis H_0: $\beta_1 = 0$ answers the question, is the regression significant? If $\beta_1 = 0$, we conclude that there is no significant linear relationship between X and Y, and hence, the independent variable X is not important in predicting the values of Y if the relationship of Y and X is not linear. Note that if $\beta_1 = 0$, then the model becomes $y = \beta_0 + \varepsilon$. Thus, the question of the importance of the independent variable in the regression model translates into a narrower question of the test of hypothesis H_0: $\beta_1 = 0$. That is, the regression line is actually a horizontal line through the intercept, β_0.

EXAMPLE 7.3.3

Using the data given in Example 7.2.1, test the hypothesis H_0: $\beta_1 = 2$ versus H_a: $\beta_1 \neq 2$ using the 0.05 level of significance.

Solution

We test

$$H_0: \beta_1 = 2 \text{ vs. } H_a: \beta_1 \neq 2.$$

We know that $\widehat{\beta}_1 = 2.0266$.

For $\alpha = 0.05$ and $n = 10$, the rejection region is $t_{\beta_1} < -2.306$ *or* $t_{\beta_1} > -2.306$. The test statistic is

$$t_{\beta_1} = \frac{\widehat{\beta}_1 - \beta_{10}}{\sqrt{\frac{MSE}{S_{xx}}}}$$

$$= \frac{2.0266 - 2}{\sqrt{\frac{2.0266 - 2}{263.6}}} = 0.4376.$$

Because the test statistic does not fall in the rejection region, at $\alpha = 0.05$, we do not reject H_0. Thus, for $\alpha = 0.05$, the given data support the null hypothesis that the true value of the slope, β_1, of the regression line is equal to 2.

As we already know, estimates of the regression coefficients β_0 and β_1 are subject to sampling uncertainty. Therefore, we will *never* estimate the true value accurately of these parameters from sample data. However, we may construct confidence intervals for the intercept and the slope parameters. Thus, a problem closely related to the problem of estimating the regression coefficients β_0 and β_1 is that of estimating the mean of the distribution of Y for a given value of x, that is, estimating $\beta_0 + \beta_1 x$. For a fixed value of x, say x_0, we have the following confidence limits.

A $(1 - \alpha)100\%$ confidence interval for $\beta_0 + \beta_1 x$ is given by

$$\left(\widehat{\beta}_0 + \widehat{\beta}_1 x\right) \pm t_{\alpha/2} s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}$$

where

$$s_e = \sqrt{\frac{S_{yy} - (S_{xy})^2}{(n-2)S_{xx}}}.$$

We could use the data from the previous example to easily calculate a confidence interval for $\beta_0 + \beta_1 x$.

7.3.1 Analysis of variance approach to regression

Another approach to hypothesis testing is based on analysis of variance (ANOVA). A detailed explanation of this approach is given in Chapter 9. Here we present necessary steps for regression. The main reason for this presentation is the fact that

most of the major statistical software outputs for regression analysis (see Section 7.9) are given in the form of ANOVA tables.

It can be verified that (see Exercise 7.3.7),

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2.$$

Denoting

$$SST = \sum_{i=1}^{n}(y_i-\bar{y})^2, SSE = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2, \text{ and } SSR = \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2,$$

the foregoing equation can be written as

$$SST = SSR + SSE.$$

Note that the total sum of squares (SST) is a measure of the variation of y_i's around the mean $\bar{y}$, and SSE is the residual or error sum of squares that measures the lack of fit of the regression model. Hence, sum of squares of regression or model (SSR) measures the variation that can be explained by the regression model.

We saw that to test the hypothesis H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$, the statistic

$$t_{\beta_1} = \frac{\hat{\beta}_1}{\sqrt{\frac{MSE}{S_{xx}}}},$$

was used, where t_{β_1} follows a t-distribution with $(n-2)$ degrees of freedom. From Exercise 4.2.18, we know that

$$t_{\beta_1}^2 = \frac{\hat{\beta}_1^2}{\left(\frac{MSS}{S_{xx}}\right)},$$

follows an F-distribution with numerator degrees of freedom 1 and denominator degrees of freedom $(n-2)$. We can also verify that

$$t_{\beta_1}^2 = \frac{MSR}{MSE}.$$

Thus, to test H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$, we could use the statistic

$$\frac{MSR}{MSE} \sim F(1, n-2),$$

and reject H_0 if

$$\frac{MSR}{MSE} \geq F_\alpha(1, n-2).$$

TABLE 7.3 ANOVA Table for Simple Regression.

Source of variation	Degrees of freedom	Sum of squares	Mean sum of squares	F-ratio
Regression (model)	1	SSR	$MSR = \frac{SSR}{d.f.}$	$\frac{MSR}{MSE}$
Error (residuals)	$n-2$	SSE	$\frac{SSE}{d.f.}$	
Total	$n-1$	SST		

The procedure is summarized in Table 7.3, known as the ANOVA table.

The last column in the ANOVA table gives the statistic $(MSR)/(MSE)$. It is also customary to give another column with the p value of the test.

EXAMPLE 7.3.4

In a study of baseline characteristics of 20 patients with foot ulcers, we want to see the relationship between the stage of ulcer that is determined using the Yarkony−Kirk scale, a higher number indicating a more severe stage, with range 1−6, and duration of the ulcer in days. Suppose we have the data shown in Table 7.4.

(a) Develop an ANOVA table to test H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$. What is the conclusion of the test based on $\alpha = 0.05$?

(b) Write down the expression for the least-squares line.

Solution

(a) We test H_0: $\beta_1 = 0$ vs. H_a: $\beta_1 \neq 0$. We will use Minitab to generate the ANOVA table (Table 7.5). Because the p value is less than 0.001, for $\alpha = 0.05$, we reject the null hypothesis that $\beta_1 = 0$ and conclude that there is a relationship between the stage of ulcer and its duration.

(b) Again, using the Minitab output, we obtain the least-squares line as

$$d = 4.61x - 2.40.$$

TABLE 7.4 Stage and Duration of Foot Ulcers.

Stage of ulcer (x)	4	3	5	4	4	3	3	4	6	3
Duration (d)	18	6	20	15	16	15	10	18	26	15
Stage of ulcer (x)	3	4	3	2	3	2	2	3	5	6
Duration (d)	8	16	17	6	7	7	8	11	21	24

TABLE 7.5 Anova Table for Foot Ulcer Data.

Source of variation	Degrees of freedom	Sum of squares	Mean sum of squares	F-ratio	p Value
Regression (model)	1	570.04	570.04	77.05	0.000
Error (residuals)	18	133.16	7.40		
Total	19	703.20			

Exercises 7.3

7.3.1. An experiment was conducted to observe the effect of an increase in temperature on the potency of an antibiotic. Three one-ounce portions of the antibiotic were stored for equal lengths of time at each of the following Fahrenheit temperatures: 40 degrees, 55 degrees, 70 degrees, and 90 degrees. The potency readings observed at the end of the experimental period were

Potency reading, y	49	38	27	24	38	33	19	28	16	18	23	
Temperature, x	40 degrees			55 degrees			70 degrees			90 degrees		

(a) Find the least-squares line appropriate for these data.
(b) Plot the points and graph the line as a check for your calculations.
(c) Calculate the 95% confidence intervals for β_0 and β_1, respectively.

7.3.2. Consider the data

x	38	26	48	22	40	15	30	33
y	10	11	16	8	12	5	10	11

(a) Find the least-squares line appropriate for these data.
(b) Plot the points and graph the line as a check for your calculations.
(c) Calculate the 95% confidence intervals for β_0 and β_1, respectively.

7.3.3. Show that $\overline{Y}$ and $\widehat{\beta}_1$ are independent, under the usual assumptions of a simple linear regression model.

7.3.4. Using the data of Exercise 7.2.10, calculate the 95% confidence intervals for β_0 and β_1, respectively.

7.3.5. The following data represent survival time in days after a heart transplant and patient age in years at the time of transplant for 10 randomly selected patients.

Age at transplant	28	41	46	53	39	36	47	29	48	44
Survival time, in days	7	278	44	48	406	382	1995	176	323	1846

(a) Find the least-squares line appropriate for these data.
(b) Plot the points and graph the line.
(c) Calculate the 95% confidence intervals for β_0 and β_1, respectively.

7.3.6. The following data represent weights of cigarettes (g) from different manufacturers and their nicotine contents (mg).

Weight	15.8	14.9	9.0	4.5	15.0	17.0	8.6	12.0	4.1	16.0
Nicotine	0.957	0.886	0.852	0.911	0.889	0.919	0.969	1.118	0.946	1.094

(a) Find the least-squares line appropriate for these data.
(b) Plot the points and graph the line. Do you think the linear regression is appropriate?
(c) Calculate the 95% confidence intervals for β_0 and β_1, respectively.

7.3.7. The following data represent total CO_2 emissions per vehicle (in metric tons per vehicle) (http://corporate.ford.com/microsites/sustainability-report-2012-13/environment-data-energy).

Year	2007	2008	2009	2010	2011	2012
Total	1.01	1.09	1.07	1.01	0.91	0.90

(a) Find the least-squares line appropriate for this data.
(b) Plot the points and graph the line.
(c) Calculate the 95% confidence intervals for β_0 and β_1, respectively.

7.3.8. Show that

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 + \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2.$$

7.4 Predicting a particular value of *Y*

In the earlier sections, we have seen how to fit a least-squares line for a given set of data. Also using this line, we could find $E(Y)$, for any given value of x. Instead of obtaining this mean value, we may be interested in predicting the particular value of Y for a given x. In fact, one of the primary uses of the estimated regression line is to predict the response value of Y for a

given value of x. Prediction problems are very important in several real-world problems; for example, in economics one may be interested in predicting a particular gain associated with an investment.

Let $\widehat{Y}_0$ denote a predictor of a particular value of $Y = Y_0$ and let the corresponding values of x be x_0. We shall choose $\widehat{Y}_0$ to be $E(\widehat{Y}|x_0)$. Let $\widehat{Y}$ denote a predictor of a particular value of Y. Then the error η of the predictor in comparison to a particular value of Y is

$$\eta = Y - \widehat{Y}_0.$$

Both Y and $\widehat{Y}$ are normal random variables, and the error is a linear function of Y and $\widehat{Y}$. This means that η itself is normally distributed. Also, because $E(\widehat{Y}) = E(Y)$, we have

$$E(\eta) = E(Y|x_0) - E(\widehat{Y}) = 0.$$

Furthermore,

$$Var(\eta) = Var(Y - \widehat{Y}) = Var(Y) + Var(\widehat{Y}) - 2Cov(Y, \widehat{Y}).$$

We can consider Y and $\widehat{Y}$ as independent, because we are predicting a different value of Y, not used in the calculation of $\widehat{Y}$. Therefore, $Cov(Y, \widehat{Y}) = 0$. In that case,

$$\begin{aligned} Var(\eta) &= Var(Y_0) + Var(\widehat{Y}_0) \\ &= \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}\right] \\ &= \left[1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}\right]\sigma^2. \end{aligned}$$

Hence, the error of predicting a particular value of Y, given x, is normally distributed with mean zero and variance $\left[1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}\right]\sigma^2$.

That is,

$$\eta \sim N\left(0, \left[1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}\right]\sigma^2\right),$$

and

$$Z = \frac{Y - \widehat{Y}}{\sigma\sqrt{\left[1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}\right]}} \sim N(0,1).$$

If we substitute the sample standard deviation S for σ, then we can show that

$$T = \frac{Y - \widehat{Y}}{S\sqrt{\left[1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}\right]}},$$

follows the t-distribution with $[n-(k+1)]$ degrees of freedom. Using this fact, we now give a prediction interval for the random variable Y, the response of a given situation.

We know that

$$P\left(-t_{\alpha/2} < T < t_{\alpha/2}\right) = 1 - \alpha.$$

Substituting for T, we have

$$P\left(-t_{\alpha/2} < \frac{Y - \widehat{Y}}{S\sqrt{\left[1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{S_{xx}}\right]}} < t_{\alpha/2} \right) = 1 - \alpha,$$

which implies that

$$P\left[\widehat{Y} - t_{\alpha/2}S\sqrt{\left[1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{S_{xx}}\right]} < Y < \widehat{Y} + t_{\alpha/2}S\sqrt{\left[1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{S_{xx}}\right]}\right] = 1 - \alpha.$$

Hence, we have the following.

A $(1 - \alpha)100\%$ *prediction interval* for Y is

$$\widehat{Y} \pm t_{\alpha/2}S\sqrt{\left[1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{S_{xx}}\right]}$$

where $t_{\alpha/2}$ is based on $(n - 2)$ degrees of freedom and $S^2 = \frac{SSE}{n-2} = \sqrt{MSE}$.

We illustrate this statistical procedure with the following example.

EXAMPLE 7.4.1

Using the data given in Example 7.2.1, obtain a 95% prediction interval at $x = 5$.

Solution

We have shown that $\widehat{y} = -3.1011 + 2.0266x$. Hence, at $x = 5$, $\widehat{y} = 7.0319$.

Also, $\overline{x} = 3.8$, $S_{xx} = 263.6$, $SSE = 7.79{,}028$, and $S = \sqrt{\frac{7.79028}{8}} = 2.306$.

From the t-table, $t_{0.025,8} = 2.306$.

Thus, we have

$$7.0319 \pm (2.306)(0.98681)\sqrt{\left[1 + \frac{1}{10} + \frac{(5 - 3.8)^2}{263.6}\right]},$$

which gives the 95% prediction interval as (4.6393, 9.4245).

We can conclude with at least 95% confidence that the true value of Y at the point $x = 5$ will be somewhere between 4.6393 and 9.4245.

Exercises 7.4

7.4.1. The following are midterm and final examination test scores for 10 calculus students, where x denotes the midterm score and y denotes the final score for each student.

x	68	87	75	91	82	77	86	82	75	79
y	74	89	80	93	88	79	97	95	89	92

Obtain a 95% prediction interval for $x = 92$ and interpret its meaning.

7.4.2. The following data give the annual incomes (in thousands of dollars) and amounts (in thousands of dollars) of life insurance policies for eight persons.

Annual income	42	58	27	36	70	24	53	37
Life insurance	150	175	25	75	250	50	250	100

Obtain a 90% prediction interval for $x = 59$ and interpret its meaning.

7.4.3. For the following data, construct a 95% prediction interval for $x = 12$.

x	1	3	5	7	9	11
Y	16	36	43	65	80	88

7.4.4. The data given below are from a random sample of height (in inches) and weight (in pounds) of seven basketball players.

Height	73	83	77	80	85	71	80
Weight	186	234	208	237	265	190	220

Construct a 99% prediction interval for height equal to 90. Interpret the result and state any assumptions.

7.4.5. For the data in Exercise 7.2.10, obtain a 95% prediction interval for the age, $x = 85$, interpret and state any assumptions.

7.4.6. For the CO_2 emission data of Exercise 7.3.7, construct a 95% prediction interval for the year 2013 emission.

7.5 Correlation analysis

Using the regression model, we can evaluate the magnitude of change in the dependent variable due to certain changes in the independent variables. One of the main assumptions we have used is that the independent variables are known. However, there are problems where the x-values as well as the y-values are assumed to be random variables. This would be the case, for example, if we study the relationship between secondhand smoking and the incidence of a certain disease. Here, basically, one treats X as random, and hence the simple linear regression model is

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

This implies that

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

and one looks for dependence of X and Y. Once we have determined that there is a relationship between the variables, the next question that arises is how closely the variables are associated. A measure of the amount of linear dependency of the two random variables is the *correlation*. The correlation coefficient tells us how strongly two variables are linearly related. The statistical method used to measure the degree of correlation is referred to as the *correlation analysis*. We will assume that the vector random variable (X, Y) has a bivariate normal distribution. In this case, it can be shown that

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

At times, our interest may not be in the linear relationship; rather, we may merely want to know whether X and Y are independent random variables. If (X, Y) has a bivariate normal distribution, then testing for independence is equivalent to testing that the correlation coefficient, $\rho = \sigma_{xy}/(\sigma_x \sigma_y)$, is equal to zero. Note that ρ is positive if X and Y increase together and ρ is negative if Y decreases as X increases. If $\rho = 0$, there is no relation between X and Y; if $\rho > 0$, there is a positive relation between X and Y (increasing slope); and when $\rho < 0$, we have a negative relationship (decreasing slope). Thus, the correlation coefficient can be used to measure how well the linear regression model fits the data.

Let (X_1, Y_1), (X_2, Y_2), …, (X_n, Y_n) be a random sample from a bivariate normal distribution. The maximum likelihood estimator of ρ is the sample correlation coefficient defined by $\widehat{\rho}$ or r,

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \tag{7.7}$$

$$= \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}.$$

Equivalently, we can rewrite (7.7) by

$$r = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{\left[n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right]\left[n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2\right]}}.$$

We can see that $-1 \leq r \leq 1$. The value of r could readily be obtained by the calculations one already has performed for the regression analysis. Observe that the numerator of r is exactly the same as the numerator of $\widehat{\beta}_1$ derived in Section 7.2. Because the denominators of both $\widehat{\beta}_1$ and r are nonnegative, they have the same sign. It can be shown that this estimator is not unbiased. If the value of r is near or equal to zero, this implies little or no linear relationship between x and y. On the other hand, the closer r is to 1 or -1, the stronger the linear relationship between x and y. When $r > 0$, values of y increase as the values of x increase, and the data set is said to be *positively correlated*. When $r < 0$, values of y decrease as the values of x increase, and the data set is said to be *negatively correlated*. $r = 0$ indicates no linear relationship between x and y, however, there can be a nonlinear relationship in this case. In this book, we use the term *correlation* only when referring to linear relationships. In actual practice we can use the value of r to decide whether it is appropriate to develop linear regression models in a given situation. As a rule of thumb, if $r > 0.30$ or $r < -0.30$, we proceed with developing a linear regression model. However, a much higher or lower value is desirable. For example, if in a given problem where $r = 0.77$, it conveys to us that approximately 77% of the data we have are linearly related.

The probability distribution for r is difficult to obtain. For large samples, this difficulty could be overcome by using the fact that the Fisher z-transform, given by

$$z = (1/2)\ln[(1+r)/(1-r)],$$

is approximately normally distributed with mean $\mu_z = (1/2)\ln[(1+\rho)(1-\rho)]$ and variance $\sigma_z = 1/(n-3)$. Thus, for large random samples, we can test hypotheses about ρ using the approximate test statistic:

$$Z = \frac{z - \mu_z}{\sigma_z}$$

$$= \frac{(1/2)\ln\left(\frac{1+r}{1-r}\right) - (1/2)\ln\left(\frac{1+\rho}{1-\rho}\right)}{\frac{1}{\sqrt{n-3}}}.$$

For example, suppose we are interested in testing the hypothesis that the true value of ρ is a specific number, say, ρ_0, with a certain value of α. We can proceed to make a decision by following the procedure given below.

Hypothesis test for ρ

One-sided test

H_0: $\rho = \rho_0$
H_a: $\rho > \rho_0$ or
H_a: $\rho < \rho_0$

Test statistic:

$$Z = \frac{(1/2)\ln\left(\frac{1+r}{1-r}\right) - (1/2)\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)}{\frac{1}{\sqrt{n-3}}}$$

Rejection region:
$z > z_\alpha$ (upper tail region)
$z < -z_\alpha$ (lower tail region)

Two-sided test

H_0: $\rho \neq \rho_0$
H_a: $\rho \neq \rho_0$
Test statistic:

$$Z = \frac{(1/2)\ln\left(\frac{1+r}{1-r}\right) - (1/2)\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)}{\frac{1}{\sqrt{n-3}}}$$

Rejection region:
$|z| > z_{a/2}$

Decision: If z falls in the rejection region, reject the null hypothesis at the level of significance α.
Assumption: (X, Y) follow the bivariate normal, and this test procedure is approximate.

EXAMPLE 7.5.1

For the data given in Example 7.2.1, would you say that the variables X and Y are independent? Use $\alpha = 0.05$.

Solution
We test

$$H_0: \rho = 0 \quad vs.\ H_a: \rho \neq 0.$$

From Example 7.2.1, for $n = 10$, we have the following summary:

$$\sum_{i=1}^{10} x_i = 38; \quad \sum_{i=1}^{10} y_i = 46; \quad \sum_{i=1}^{10} x_i y_i = 709,$$

and

$$\sum_{i=1}^{10} x_i^2 = 408; \quad \sum_{i=1}^{10} y_i^2 = 1302.$$

Hence,

$$r = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{\left[n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right]\left[n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2\right]}}$$

$$= \frac{(10)(709) - (38)(46)}{\sqrt{\left[(10)(408) - (38)^2\right]\left[(10)(1302) - (46)^2\right]}}$$

$$= 0.99641.$$

The test statistic is

$$z = \frac{(1/2)\ln\left(\frac{1+r}{1-r}\right) - (1/2)\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)}{\frac{1}{\sqrt{n-3}}}$$

$$= \frac{(1/2)\ln\left(\frac{1+0.99641}{1-0.99641}\right) - (1/2)\ln\left(\frac{1+0}{1-0}\right)}{\frac{1}{\sqrt{7}}}$$

$$= 8.3618.$$

For $z_{\alpha/2} = z_{0.025} = 1.96$, the rejection region is $|z| > 1.96$. Because the observed value of the test statistic falls in the rejection region, we reject the null hypothesis and conclude that at $\alpha = 0.05$, the variables X and Y are dependent.

Exercises 7.5

7.5.1. This table shows the midterm and final examination test scores for 10 students from a differential equations class, where x denotes the midterm scores and y denotes the final scores.

x	68	87	75	91	82	77	86	82	75	79
y	74	89	80	93	88	79	97	95	89	92

(a) At 95% confidence level, test whether X and Y are independent.
(b) Find the p value.
(c) State any assumptions you have made in solving the problem.

7.5.2. The following table gives the annual incomes (in thousands of dollars) and amounts (in thousands of dollars) of life insurance policies for eight persons.

Annual income	42	58	27	36	70	24	53	37
Life insurance	150	175	25	75	250	50	250	100

(a) At the 98% confidence level, test whether annual income and the amount of life insurance policies are independent.
(b) Find the attained significance level.
(c) State any assumptions you have made in solving the problem.

7.5.3. Show that

$$r = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{\left[n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right]\left[n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2\right]}},$$

is not an unbiased estimator of the population coefficient, ρ.

7.5.4. Using the data in Example 7.2.1:
(a) Compute r, the coefficient of correlation.
(b) Would you say that the variables X and Y are independent? Use $\alpha = 0.05$.
(c) State any assumptions you have made in solving the problem.

7.5.5. A new medication is tested for serum cholesterol-lowering properties on six randomly selected volunteers. The serum cholesterol values are given in the following table.

Before treatment:	232	254	220	200	213	222
After treatment:	212	240	225	205	204	218

(a) At 95% confidence level, test whether X and Y are independent.
(b) Find the p value.
(c) Calculate the least-squares regression line for these data.
(d) Interpret the usefulness of the model.
(e) State any assumptions you have made in solving the problem.

7.6 Matrix notation for linear regression

Most real-life applications of regression analysis use models that are more complex than the simple straight-line model. For example, a person's body weight may depend not just on the person's eating habits; it may depend on additional factors such as heredity, exercise, and type of work. Hence, we may want to incorporate other potential independent variables in the modeling. We now study the situation where k (>1) independent variables are used to predict the dependent variable. The model to be studied is of the form

$$Y = \beta_0 + \beta_1 x_1 + \beta_1 x_2 + \cdots + \beta_k x_k + \varepsilon.$$

Here, $\varepsilon \sim N(0, \sigma^2)$. This model is called a *multiple regression model.*

Let $y_1, y_2, \ldots, y_n$ be n independent observations on Y. Then each observation y_i can be written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon,$$

where x_{ij} is the jth independent variable for the ith observation, $i = 1, 2, \ldots, n$, and $\varepsilon_i's$ are independent as in the simple linear regression case. It is sometimes advantageous to introduce matrices to study the linear equations. Let $x_0 = 1$. Define the following matrices:

$$X = \begin{bmatrix} x_0 & x_{11} & x_{12} & . & . & x_{1k} \\ x_0 & x_{21} & x_{22} & . & . & x_{2k} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ x_0 & x_{n1} & x_{n2} & . & . & x_{nk} \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix},$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_n \end{bmatrix} \quad \text{and} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}.$$

Thus, the n equations representing the linear equations can be rewritten in the matrix form as

$$Y = X\beta + \varepsilon.$$

In particular, for the n observations from the simple linear model of the form

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

we can write

$$Y = X\beta + \varepsilon,$$

where

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & . \\ 1 & . \\ 1 & . \\ 1 & x_n \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}, \quad \text{and} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}.$$

We can see that

$$X'X = \begin{bmatrix} 1 & 1 & . & . & . & 1 \\ x_1 & x_2 & . & . & . & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_1^2 \end{bmatrix},$$

where $'$ denotes the transpose of a matrix.

Also,

$$X'Y = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}.$$

Let us now go back to the multiple regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_1 x_2 + \cdots + \beta_k x_k + \varepsilon.$$

The least-squares estimators $\widehat{\beta}_i$ of β_i for $i = 0, 1, 2, \ldots, k$ are the ones that minimize the sum of squares

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \cdots + \widehat{\beta}_k x_k \right) \right]^2$$

$$= \left(y - X\widehat{\beta} \right)' \left(y - X\widehat{\beta} \right)$$

$$= y'y - y'X\widehat{\beta} - \left(X\widehat{\beta} \right)' y + \left(\widehat{\beta} X \right)' X\widehat{\beta}.$$

To minimize *SSE* with respect to β, we differentiate *SSE* with respect to β and equate it to zero. Thus,

$$\frac{\partial}{\partial \beta} (y'y - y'X'\beta - \beta'X'y + X'\beta'X\beta) = 0,$$

yielding

$$(X'X)\widehat{\beta} = X'Y.$$

Assuming the matrix $(X'X)$ is invertible, we obtain

$$\widehat{\beta} = (X'X)^{-1} X'Y.$$

We now summarize the procedure to obtain a multiple linear regression equation.

Procedure to obtain a multiple linear regression equation

1. Rewrite the n observations as

   $$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_1 x_{2i} + \cdots + \beta_k x_{ki}, \quad i = 1, 2, \ldots, n$$

   in the matrix notation as

   $$Y = X\beta + \varepsilon$$

2. Compute $(X'X)^{-1}$ and obtain the estimators of β as

   $$\widehat{\beta} = (X'X)^{-1} X'Y.$$

3. Then the regression equation is

   $$\widehat{Y} = X\widehat{\beta}.$$

EXAMPLE 7.6.1

Using the data given in Example 7.2.1, use the matrix approach to solve the problem of operations.

Solution

From the data in Example 7.2.1, we have

$$Y = \begin{bmatrix} -9 \\ -7 \\ -5 \\ -4 \\ 2 \\ 6 \\ 9 \\ 13 \\ 21 \\ 20 \end{bmatrix} \quad \text{and} \quad X = \begin{bmatrix} 1 & -3 \\ 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 2 \\ 1 & 5 \\ 1 & 6 \\ 1 & 8 \\ 1 & 11 \\ 1 & 12 \end{bmatrix}.$$

Thus, we can write,

$$X'X = \begin{bmatrix} 10 & 38 \\ 38 & 408 \end{bmatrix}, \quad X'Y = \begin{bmatrix} 46 \\ 709 \end{bmatrix}, \quad (X'X)^{-1} = \begin{bmatrix} 0.1548 & -0.0144 \\ -0.0144 & 0.0038 \end{bmatrix}.$$

Hence,

$$\widehat{\beta} = (X'X)^{-1}(X'Y) = \begin{bmatrix} 0.1548 & -0.0144 \\ -0.0144 & 0.0038 \end{bmatrix} \begin{bmatrix} 46 \\ 709 \end{bmatrix}$$
$$= \begin{bmatrix} -3.1009 \\ 2.0266 \end{bmatrix} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{bmatrix}.$$

Thus, the least-squares line is given by

$$\widehat{y} = -3.1009 + 2.0266X,$$

which is identical to the regression line we obtained in Example 7.2.1.

EXAMPLE 7.6.2

The following data relate to the prices (Y) of five randomly chosen houses in a certain neighborhood, the corresponding ages of the houses (x_1), and square footage (x_2).

Price y in thousands of dollars	Age x_1 in years	Square footage x_2 in thousands of square feet
100	1	1
80	5	1
104	5	2
94	10	2
130	20	3

Fit a multiple linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

to the foregoing data.

Solution
We have,

$$Y = \begin{bmatrix} 100 \\ 80 \\ 104 \\ 94 \\ 130 \end{bmatrix}; \; X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 5 & 1 \\ 1 & 5 & 2 \\ 1 & 0 & 2 \\ 1 & 20 & 3 \end{bmatrix}; \; X'X = \begin{bmatrix} 5 & 41 & 9 \\ 41 & 551 & 96 \\ 9 & 96 & 19 \end{bmatrix};$$

$$X'Y = \begin{bmatrix} 508 \\ 4560 \\ 966 \end{bmatrix}$$

and

$$(X'X)^{-1} = \begin{bmatrix} 2.3076 & 0.1565 & -1.8840 \\ 0.1565 & 0.0258 & -0.2044 \\ -1.8840 & -0.2044 & 1.9779 \end{bmatrix}.$$

Hence,

$$(X'X)^{-1}(X'Y) = \begin{bmatrix} 66.1252 \\ -0.3794 \\ 21.4365 \end{bmatrix}.$$

Thus, the regression model is

$$y = 66.12 - 0.3794x_1 + 21.4365x_2.$$

Thus, for a given x_1 and x_2 we can estimate (predict) the value of the house.

7.6.1 ANOVA for multiple regression

As in Section 7.3, we can obtain an ANOVA table for multilinear regression (with k independent or explanatory variables) to test the hypothesis

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

versus,

$$H_a : \text{At least one of the parameters } \beta_j \neq 0, j = 1, \ldots, k.$$

The calculations for multiple regression are almost identical to those for simple linear regression, except that the test statistic $(MSR)/(MSE)$ has an $F(k, n-k-1)$ distribution. Note that the F-test does not indicate which of the parameters $\beta_j \neq 0$, except to say that at least one of them is not zero. The ANOVA table for multiple regression is given by Table 7.6.

EXAMPLE 7.6.3

For the data in Example 7.6.2, obtain an ANOVA table and test the hypothesis

TABLE 7.6 ANOVA Table for Multiple Regression.

Source of variation	Degrees of freedom	Sum of squares	Mean sum of squares	F-ratio
Regression (model)	K	SSR	$MSR = \frac{SSR}{d.f.}$	$\frac{MSR}{MSE}$
Error (residuals)	$n-k-1$	SSE	$\frac{SSE}{d.f.}$	
Total	$n-1$	SST		

TABLE 7.7 ANOVA Table for Home Price Data.

Source of variation	Degrees of freedom	Sum of squares	Mean sum of squares	*F*-ratio	*p* Value
Regression (model)	2	956.5	478.2	2.50	0.286
Error (residuals)	2	382.7	191.4		
Total	4	1339.2			

$$H_0 : \beta_1 = \beta_2 = 0 \text{ vs. } H_a\text{: at least one of the } \beta_i \neq 0,\ i = 1,\ 2.$$

Use $\alpha = 0.05$.

Solution

We test H_0: $\beta_1 = \beta_2 = 0$ vs. H_a: At least one of the $\beta_i \neq 0$, $i = 1, 2$. Here $n = 5$, $k = 2$. Using Minitab, we obtain the ANOVA table (Table 7.7). Based on the p value, we cannot reject the null hypothesis at $\alpha = 0.05$.

Exercises 7.6

7.6.1 Given the data

X_1	X_2	y
3	1	4
2	5	3
3	3	6
1	2	5

(a) Write the multiple regression model in matrix form.
(b) Find $X'X$, $(X'X)^{-1}$, and $X'y$.
(c) Estimate β.

7.6.2. A study is conducted to estimate the demand for housing (y) based on current interest rate X_1 and the rate of unemployment. The data in Table 7.8 are obtained.

(a) Fit the multiple regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_1 x_2 + \varepsilon.$$

(b) Test whether the model is significant.

7.6.3. The following data give the annual incomes (in thousands of dollars) and amounts (in thousands of dollars) of life insurance policies for eight persons.

TABLE 7.8 Housing Demand, Interest Rate, and Unemployment Rate.

Units sold	Interest rate (%)	Unemployment rate (%)
65	9.0	10.0
59	9.3	8.0
80	8.9	8.2
90	9.1	7.7
100	9.0	7.1
105	8.7	7.2

Annual income	42	58	27	36	70	24	53	37
Life insurance	150	175	25	75	250	50	250	100

Calculate the least-squares regression line for this data using matrix operations.

7.6.4. The following is a random sample of height (in inches) and weight (in pounds) of seven basketball players.

Height	73	83	77	80	85	71	80
Weight	186	234	208	237	265	190	220

Calculate the least-squares regression line for this data using matrix operations.

7.7 Regression diagnostics

In the previous sections, we derived least-squares estimators for the parameters in the linear regression model. These estimators are useful as long as we can determine (1) how well the model fits the data and (2) how good our estimates are in providing possible relationships between variables of interest. Some of these problems are discussed in Chapter 14 in a unified manner. We now briefly discuss some aspects of the adequacy of the simple linear regression model. In multiple regression, in addition to the problems discussed here, there are other problems, such as collinearity and model specification (inclusion of all relevant variables, as well as exclusion of irrelevant variables), that need to be examined. They are beyond the level of this text. Many graphical methods and numerical tests dealing with these problems are available in the literature and are often called regression diagnostics. Most of the major statistical software packages incorporate these tests, making it easier to perform regression diagnostics so as to detect potential problems.

We have seen that the (ordinary) least-squares regression model must meet the following assumptions.

1. **Linearity**. The existence of a linear relationship between x and y is the basis of the simple linear regression model. A simple method to test for linearity is to draw a scatterplot of data points. As we explained in Section 7.2, we could also plot residual e_i versus x_i or $\widehat{Y}_i$. A symmetric trend in the plot of the residuals versus the explanatory variable or the fitted values indicates there is a problem with the obtained regression model. For a correct model, the residuals should center around zero across the explanatory variables and the fitted values. The degree of linear relationship can be ascertained by the correlation coefficient, r, given in Section 7.5 or by using the value of the coefficient of determination r^2, explained in Project 8B. Most statistical software packages give the value of r^2 (refer to outputs given in Section 7.9). The closer the value of r^2 is to 1, the better the least-squares equation $\widehat{y} = \widehat{\beta}_1 x + \widehat{\beta}_0$ performs as a predictor of y.
2. **Homoscedasticity** (homogeneity of variance). This assumption says that the variance of the error term remains constant across all values of x. In this case we know by the Gauss–Markov theorem that the least-squares estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the best linear unbiased estimators of β_0 and β_1. A frequently used graphical method is to draw the residuals versus a fitted plot. This can be easily done using statistical software packages. The graph of residuals e_i versus fitted values $\widehat{Y}_i$ or explanatory variable x_i indicates a change in the spread of residuals as $\widehat{Y}$ or x changes. It may look like Fig. 7.7.

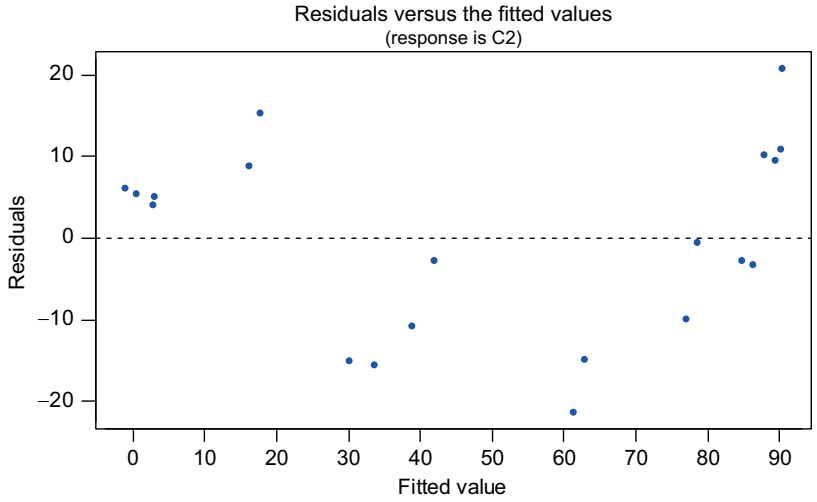


FIGURE 7.7 Scatterplot of fitted values versus residuals.

If the variances of y_i values are not constant, the inferences we made, such as confidence intervals on means, prediction, and so forth, are off. The severity of this discrepancy depends on the degree of the assumption violation. If we see that the pattern of data points only changes slightly, that will indicate a mild heteroscedasticity. Two numerical tests for heteroscedasticity are explained in Section 14.4.3.

3. **Independence of ε_i and ε_j, for $i \neq j$**. This assumption specifies that the errors associated with one observation should not be correlated with the errors of any other observation. In general, whether the two samples are independent of each other is decided by the structure of the experiment from which they arise. Violation of the independence assumption can occur in a variety of situations. For example, if we take a survey on a certain issue on children's education from one particular school, these observations may reflect some pattern, thus violating the independence assumption. If data are collected on the same variable over time, then the assumption of independence will be violated. Project 12B explains a run test for checking of this assumption. Also, see Section 14.4.4.
4. **Normality of the errors**. This assumption specifies that the distribution of the ε_i values should be normal. This assumption is crucial when sample size is small if the p value for the test is to be valid. For large samples, by the central limit theorem this assumption becomes less important unless the prediction of a single value of y is involved. Thus, a test of normality is necessary mainly when the t-test is used. Section 14.4.1 explains some of the tests for normality. A simple way is to draw a probability plot for the errors to conform to the assumption of normality. If we observe nonnormality, one of the ways to overcome the problem is to use data transformation such as logarithmic transformation, as explained in Section 14.4.2, and perform the regression analysis on the transformed data. Sometimes nonparametric methods may be more appropriate, but we will not deal with this topic in this book.

Another important issue is the existence of *influential observations*, individual observations that have a strong influence on estimated coefficients. If a single observation substantially changes our results, we need to do further investigation. The ordinary least-squares method is quite sensitive for outlying observations, both for independent variables and for dependent variables, and can have an adverse effect on the estimate. In higher dimensional data, these outlying observations can remain unnoticed. This aspect in one explanatory variable case is discussed in Project 8C. One of the simple ways to identify such observations is to draw a scatterplot. In the scatterplot, if we see a data point that is farther away from the rest of the data points, that is an indication of a possible influential point or an outlier.

The natural question is, if we find that the data violate one or more of the assumptions, what can we do about it? We have already explained that violation of the normality assumption in large samples is not an issue unless prediction is involved, because prediction depends on normality of an individual observation. Thus, if the inferences are based on the t- or F-tests or prediction is involved, we may be able to transform Y to Y' to achieve normality. If we have predicted Y', then back-transform to predict Y. If we observe nonlinearity of data, we may be able to transform x to $x' = h(x)$ such that Y is linear in x', or consider a polynomial model in x, in which case the ideas of multiple linear regression may be utilized. Robust estimates of variances of β_0 and β_1 or the method of weighted least squares may be used to deal with the case of nonconstant variance. Often careful experimental design could be done to remove possible correlation in errors. There are also robust methods available for correlation analysis. We refer to specialized books on regression methods for further details on these issues. If we detect influential observations, there are statistical techniques available, such as least-trimmed-squares estimators, to deal with outlying observations.

7.8 Chapter summary

In this chapter, we first derived the least-squares line and its properties. Then we learned about the confidence intervals for the coefficients in the regression model and did hypothesis tests on the values of the coefficients. We introduced the matrix notation for linear regression as well as for multiple regression. We discussed how to predict a particular value of Y for a given value of X. In order to study the dependence of X and Y, we presented correlation analysis.

The following are some of the key definitions we have used in this chapter:

- Predictors
- Response variable
- Regression analysis
- Multiple linear regression model
- Simple linear regression model
- Sum of squares for errors (SSE)
- Sum of squares of the residuals

- Least-squares line
- Least-squares equations
- Normal equations
- Best linear unbiased estimator (BLUE)
- Correlation analysis

The following important concepts and procedures were discussed in this chapter:

- Procedure for regression modeling
- Procedure for fitting a least-squares line
- Properties of the least-squares estimators for the model $Y = \beta_0 + \beta_1 x + \varepsilon$
- The Gauss−Markov theorem
- Procedure for obtaining confidence intervals of β_0 and β_1
- Procedure to obtain a multiple linear regression equation
- Prediction interval for the response variable Y
- Hypothesis testing for correlation, ρ
- Linearity
- Homoscedasticity
- Independence of ε_i and ε_j, for $i \neq j$
- Normality of the errors
- Influential observations

7.9 Computer examples

7.9.1 Examples using R

EXAMPLE 7.9.1 For the following data, use the method of least-squares regression to fit a straight line to the accompanying data points. Give the estimates of β_0 and β_1. Plot the points and sketch the fitted least-squares line.

Sample (x)	−1	0	2	−2	5	6	8	11	12	−3
Sample (y)	−5	−4	2	−7	6	9	13	21	20	−9

This example assumes you put the data into variables *x* and *y*. Please modify your code appropriately.

R code

```
model = lm(y ~ x);
summary(model);
```

Solution

From the output below the estimate of β_0 is −3.10091, and the estimate of β_1 is 2.02656. Hence, the regression line is $\hat{y} = -3.10091 + 2.02656x$.

Output

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.21775 -0.70220  0.03452  0.17394  1.80880

Coefficients:
            Estimate  Std. Error t value Pr(>|t|)
(Intercept) -3.10091     0.38882  -7.975 4.47e-05 ***
x            2.02656     0.06087  33.292 7.23e-10 ***
---
Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9883 on 8 degrees of freedom
Multiple R-squared: 0.9928, Adjusted R-squared: 0.9919
F-statistic: 1108 on 1 and 8 DF, p-value: 7.232e-10
```

EXAMPLE 7.9.2 Now obtain the fitted regression line, using results from the previous example.

This example assumes you have your linear model stored in the variable model from the previous example. This example also assumes you have the data from the previous example stored in *x* and *y*. Please modify your code appropriately.

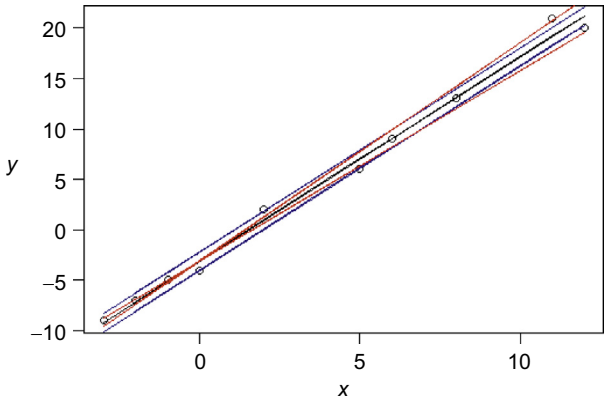
R Code:

```
yhat=predict(model,data=x);
plot(x,y);
lines(x,yhat);
c=confint(model);
m=model;
m$coefficients[1]=c[1];
lines(x,predict(m,data=x),col="blue");
m$coefficients[1]=c[3];
m=model;
m$coefficients[2]=c[2];
lines(x,predict(m,data=x),col="red");
m$coefficients[2]=c[4];
lines(x,predict(m,data=x),col="red");
```

New command for confidence interval of model estimates

Output:

We obtain a graph with confidence intervals for the intercept in blue and confidence intervals for the slope in red. The coefficient of determination r^2 is 0.9928, and the *p* value is small, suggesting the model fits pretty well.



EXAMPLE 7.9.3 In this example we'll be using matrix multiplication to perform linear regression. The following is a random sample of height (in inches) and weight (in pounds) of several basketball players.

Sample (*x*)	73	83	77	80	85	71	80
Sample (*y*)	186	234	208	237	265	190	220

Calculate the least-squares regression line for these data. This example assumes you've placed the data into variables *x* and *y*. Please modify your code appropriately.

R Code:

```
library('MASS');
x=cbind(c(1:length(x))*0+1,x);
b=ginv(t(x)%*%x)%*%t(x)%*%y;
yhat=x%*%b;
plot(x[,2],y);
lines(x[,2],yhat);
```

Required for ginv() function

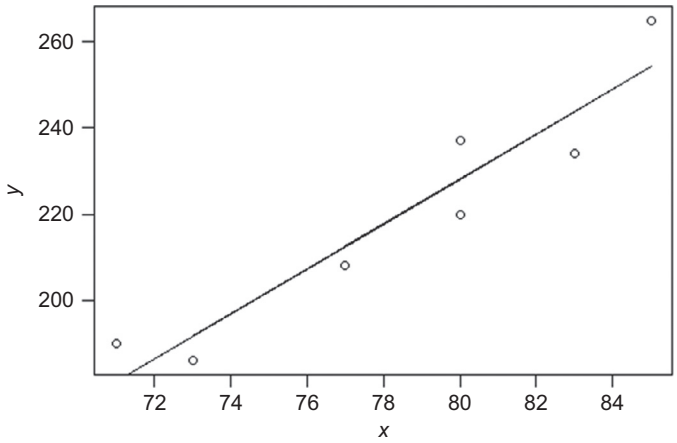
Creates a matrix with a column of 1's for the intercept

Store coefficients into *b*

Calculate *y*hat using the regression equation

Output:

Looking at the coefficients, we see that $\widehat{\beta}_0 = -188.476$ and $\widehat{\beta}_1 = 5.208$. Hence, the regression line is given by $\widehat{y} = -188.476 + 5.208x$. It is more difficult to perform confidence intervals and other tasks since this is done using matrices instead of model objects.



EXAMPLE 7.9.4 Consider the following advertisement expenses versus total sales data.

Year	Advertising Cost ($)	Yearly Sales Volume (Units)
1999	20,210	112,485
2000	22,469	118,332
2001	23,982	122,435
2002	24,645	125,569
2003	24,988	125,880
2004	25,250	127,362
2005	25,978	125,967
2006	26,556	127,252
2007	26,978	127,456
2008	27,125	127,789
2009	27,461	128,313
2010	28,120	128,662
2011	28,888	128,879
2012	29,200	129,290

Use the method of least-squares regression to fit a straight line to the accompanying data points. Plot the points and sketch the fitted least-squares line. Interpret the output.

R-code

```
> x < -c(22469, 23982, 24645, 24988, 25250, 25978, 26556, 26978, 27125, 27461, 28120, 28888, 29200).
> y < -c(118332, 122435, 125569, 125880, 127362, 125967, 127252, 127456, 127789, 128313, 128662,
128879,129290).
> model = lm(y ~ x).
> summary(model).
```

7.9.2 Minitab examples

EXAMPLE 7.9.5

For the data in Example 7.2.1, use the method of least squares to fit a straight line to the accompanying data points. Give the estimates of β_0 and β_1. Plot the points and sketch the fitted least-squares line.

Solution
Enter independent variable, *x*, in ***C1*** and the response variable, *y*, in ***C2***. Then:
Stat > ***Regression*** > ***Regression*** … > in ***Response*:** type ***C2,*** and in ***Predictors:*** type ***C1*** > click ***OK.***
Now to obtain the fitted regression line, use the following procedure:
Stat > ***Regression*** > ***Fitted Line Plot*** … > in ***Response(Y):*** type ***C2***, and in ***Predictors(X)***: type ***C1*** > click ***Linear OK.***
If in addition, we need, say, 95% confidence and predictor bands, then use:
Stat > ***Regression*** > ***Fitted Line Plot*** … > in ***Response(Y):*** type ***C2,*** and in ***Predictor(X)*:** type ***C1*** > click
Linear > click ***options*** … > click ***Display confidence bands*** and ***Display predictor bands*** > in ***Title:*** type a title for the graph and ***OK*** > ***OK.***

7.9.3 SPSS examples

A detailed explanation of regression methods including diagnostics using SPSS can be obtained at the site: http://www.ats.ucla.edu/stat/spss/webbooks/reg/. We will just demonstrate a simple case with an example.

EXAMPLE 7.9.6

The following is a random sample of height (in inches) and weight (in pounds) of seven basketball players.

Height	73	83	77	80	85	71	80
Weight	186	234	208	237	265	190	220

Calculate the least-squares regression line for these data using SPSS.

Solution
Enter height in column 1 and weight in column 2. Then, ***Analyze*** > ***Regression*** > ***Linear*** … > move ***var00002*** to ***dependent:***, and ***var00001*** to ***Independent(s):*** > click ***OK***.

7.9.4 SAS examples

For regression analysis, we can use the SAS command called GLM, which stands for general linear model, and REG, which stands for regression. In the following example we will give a simplified version of the foregoing procedure. A good explanation of regression methods including diagnostics using SAS can be obtained at http://www.ats.ucla.edu/stat/sas/webbooks/reg/.

EXAMPLE 7.9.7

Using the SAS commands, redo Example 7.9.1.

Solution
We can use the following commands.

```
options nodate nonumber;
data exreg;
INPUT x y @@;
datalines;
   −1 −5
   0 −4
   2 2
   −2 −7
   5 6
   6 9
   8 13
   11 21
   12 20
   −3 −9
;
```

```
proc reg data = exreg;
     title 'Regression of Y on X';
model y = x/p clm;
run;
```

We obtain the following output.

```
                                                    Regression of Y on X
                                                     The REG Procedure
                                                       Model: MODEL1
                                                  Dependent Variable: y
                                                   Analysis of Variance

  Source DF Sum of Squares Mean Square F Value Pr > F
  Model         1     1082.58589    1082.58589  1108.34  <.0001
Error         8      7.81411       0.97676
Corrected Total 9    1090.40000
Root MSE           0.98831           R-Square           0.9928
Dependent Mean     4.60000           Adj R-Sq           0.9919
       Coeff Var             21.48508

                        Parameter Estimates
         Parameter    Standard
Variable      DF      Estimate      Error      t Value      Pr > |t|
Intercept     1         -3.10091    0.38882     −7.98       <.0001
   x          1          2.02656    0.06087     33.29       <.0001

                    Regression of Y on X
                     The REG Procedure
                       Model: MODEL1
                    Dependent Variable: y
                Output Statistics
          Dep Var Predicted Std Error
Obs      y        Value    Mean    Predict    95% CL Mean     Residual
1    -5.0000    -5.1275  0.4278   -6.1141     -4.1409         0.1275
2    -4.0000    -3.1009  0.3888   -3.9975     -2.2043         0.8991
3     2.0000     0.9522  0.3312    0.1885      1.7159         1.0478
4    -7.0000    -7.1540  0.4715   -8.2413     -6.0667         0.1540
5     6.0000     7.0319  0.3210    6.2917      7.7720        -1.0319
6     9.0000     9.0584  0.3400    8.2743      9.8425        -0.0584
7    13.0000    13.1115  0.4038   12.1804     14.0427        -0.1115
8    21.0000    19.1912  0.5383   17.9499     20.4325         1.8088
9    20.0000    21.2178  0.5889   19.8597     22.5758        -1.2178
10   -9.0000    -9.1806  0.5187  -10.3766     -7.9845         0.1806

                  Sum of Residuals                        0
                  Sum of Squared Residuals          7.81411
                  Predicted Residual SS (PRESS)    14.18340
```

By looking at the parameter estimates in the foregoing output, we see that an intercept value of -3.10091 is the estimate of β_0, and the estimate of β_1 is 2.02656, corresponding to the variable x. For each value of x, the actual value and predicted value of y are given as the output statistics.

It is important to note that the presentation of results of analysis in a simple way is as important as the analysis itself. For example, if one is interested only in a simple linear regression, most of the output values in the foregoing output may not be necessary. All the values until the parameter estimates are giving us the analysis of variance results, and all the values in the REG procedure are dealing with prediction and confidence intervals. For clarity and simplicity of report, we may only need to report the regression line, and perhaps the graph of the line.

If we need the plot of the points (x, y), add the following commands to the previous program. We will not give the corresponding graph.

```
proc plot data = exreg;
title 'Plot of Y Vs. X';
plot y*x;
```

```
run;
```

If we need the graph of the regression line along with, say, 95% prediction and confidence intervals, we add the following.

```
proc gplot data = exreg;
plot y*x
y*x
y*x/overlay frame vaxis = axis1 haxis = axis2;
symbol1 v = −h = 1.5 i = none c = black;
symbol2 v = none i = rlclm95 c = red;
symbol3 v = none i = rlcli95 c = blue;
 axis1 order = (−5 to 14 by 1).
 offset = (1).
label = (h = 1.5 f = duplex);
 axis2 order = (−10 to 20 by 1).
 offset = (1).
label = (h = 1.5 f = duplex);
 title h = 1.5.
'Effect of X on Y';
 title2 h = 1.2 f = duplex.
'Common regression line with 95% confidence interval';
title3 h = 1.5 f = duplex
 'Regression line is predicted Y = −3.1011 +2.0266X';
run;
```

Projects for chapter 7

7A Checking the adequacy of the model by scatterplots

If the regression model is adequate, then the fitted equation can be used to make inferences. Otherwise, the inferences made will be practically useless. Note that the residuals give all the information on lack of fit. Figs. 7.5 and 7.6 give an indication of good fit and misfit.

(1) Collect a couple of real-life data and find a regression line for each.
(2) Draw the scatterplot for the residuals e_i versus x and determine whether the regression lines obtained in (1) are a good fit or not.

7B The coefficient of determination

One of the ways to measure the contribution of x in predicting y is to consider how much the prediction errors were reduced by using the information provided by the variable x. The quantity called the coefficient of determination measures how well the least-squares equation $\widehat{y} = \widehat{\beta}_1 x + \widehat{\beta}_0$ performs as a predictor of y. If x contributes no information for predicting y, then the best prediction for values of y is simply the sample mean $\overline{y}$. The resulting sum of squares of deviation for this model $\widehat{y} = \overline{y}$ is $S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2$. In the case where x contributes information for predicting y, then we have seen that the sum of squares of deviation for the model $\widehat{y} = \widehat{\beta}_1 x + \widehat{\beta}_0$ is $S_{yy} = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$. It can be shown that $\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \leq \sum_{i=1}^{n} (y_i - \overline{y})^2$.

The *coefficient of determination* is the proportion of the sum of squares of deviations of the y values that can be credited to a linear relationship between x and y. This is defined by

$$r^2 = \frac{S_{yy} - SSE}{S_{yy}}$$

$$= 1 - \frac{SSE}{S_{yy}}$$

$$= 1 - \frac{\sum_{i=1}^{n} \left(y_i - \widehat{y}_i\right)^2}{\sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2}.$$

We can see that $0 \leq r^2 \leq 1$. We can interpret r^2 to be the proportion of variability explained by the regression line. When x contributes no information for predicting y, S_{yy} and SSE will be nearly equal, and hence r^2 will be near to zero. If x contributes information for predicting y, S_{yy} will be larger than SSE, and hence r^2 will be greater than zero. Thus, $r^2 = 0.75$ means that use of $\widehat{y}$ instead of $\overline{y}$ to predict y reduced the sum of squares of deviations of the y values about their predicted values $\widehat{y}$ by 75%. This can also be interpreted as meaning that nearly 75% of the variation is explained by the independent variable x. In general, about $(r^2 \times 100)\%$ of the sample variation in y can be attributed to using x to predict y in the linear model. The *coefficient of nondetermination* is the percent of variation that is unexplained by the regression equation and is given by $1 - r^2$.

(1) For Exercises 7.2.2 and 7.2.3, find the coefficient of determination, and discuss the information contributed by x in predicting y.

(2) Collect a couple of real-life data and find the corresponding regression lines. Also draw the scatterplot for e_i versus $\widehat{y}$ and determine whether the regression line obtained is a good fit or not based on the coefficient of determination.

7C Outliers and high leverage points

One of the important aspects of residual analysis is to identify any existence of unusual observations in a data set. There are two possibilities for a data point to be unusual. It could be in the response variable (i.e., in the horizontal direction) representing model failure, or in the predictor variable (i.e., in the vertical direction). It should be noted that unusual observations in the horizontal direction occur when we assume that the independent variable X in the linear model is random. An observation that is unusual in the vertical direction is called an *outlier*. An observation that is unusual in the horizontal direction is called a *high leverage point* (or just *leverage point*).

Consider the following 10 points, which we will call base points, and three additional points representing an outlier (O), a high leverage point (H), and both (OH), respectively.

	10 base points										O	H	OH
x	−1	0	2	−2	5	6	8	11	12	−3	6	19	19
y	−5	−4	2	−7	6	9	13	21	20	−9	30	13	30

Investigate the effect of adding a single aberrant point by running four separate regressions: **(1)** regression for 10 base points; **(2)** regression for 10 base points plus O; **(3)** regression for 10 base points plus H; and (4) regression for 10 base points plus OH. For each of them, find $\widehat{\beta}_0$ and $\widehat{\beta}_1$ as well as the coefficient of determination. Discuss the effects of each type of outlier on the regression line.