3
LEAST SQUARES REGRESSION

3.1 INTRODUCTION

This chapter examines the computation of the least squares regression model. A useful understanding of what is being computed when one uses least squares to compute the coefficients of the model can be developed before we turn to the statistical aspects. Section 3.2 will detail the computations of least squares regression. We then examine two particular aspects of the fitted equation:

- The crucial feature of the multiple regression model is its ability to provide the analyst a device for "holding other things constant." In an earlier example, we considered the "partial effect" of an additional year of education, holding age constant in

$$Earnings = \gamma_1 + \gamma_2\, Education + \gamma_3\, Age + \varepsilon.$$

  The theoretical exercise is simple enough. How do we do this in practical terms? How does the actual computation of the linear model produce the interpretation of partial effects? An essential insight is provided by the notion of partial regression coefficients. Sections 3.3 and 3.4 use the **Frisch–Waugh theorem** to show how the regression model controls for (i.e., holds constant) the effects of intervening variables.
- The model is proposed to describe the movement of an explained variable. In broad terms, $y = \mu(\mathbf{x}) + \varepsilon$. How well does the model do this? How can we measure the success? Sections 3.5 and 3.6 examine fit measures for the linear regression.

3.2 LEAST SQUARES REGRESSION

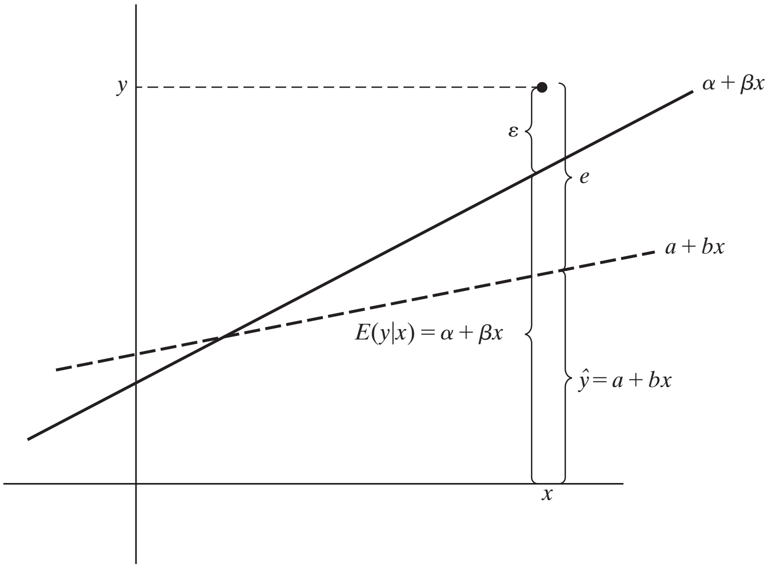
Consider a simple (the simplest) version of the model in the introduction,

$$Earnings = \alpha + \beta\, Education + \varepsilon.$$

The unknown parameters of the stochastic relationship, $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$, are the objects of estimation. It is necessary to distinguish between unobserved population quantities, such as $\boldsymbol{\beta}$ and ε_i, and sample estimates of them, denoted $\mathbf{b}$ and e_i. The **population regression** is $E[y_i|\mathbf{x}_i] = \mathbf{x}_i'\boldsymbol{\beta}$, whereas our estimate of $E[y_i|\mathbf{x}_i]$ is denoted $\hat{y}_i = \mathbf{x}_i'\mathbf{b}$. The **disturbance** associated with the ith data point is $\varepsilon_i = y_i - \mathbf{x}_i'\boldsymbol{\beta}$. For any value of $\mathbf{b}$, we shall estimate ε_i with the **residual**

$$e_i = y_i - \mathbf{x}_i'\mathbf{b}.$$

FIGURE 3.1 Population and Sample Regression.



From the two definitions,

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i = \mathbf{x}_i'\mathbf{b} + e_i.$$

These results are summarized for a two-variable regression in Figure 3.1.

The **population quantity**, $\boldsymbol{\beta}$, is a vector of unknown parameters of the joint probability distribution of $(y, \mathbf{x})$ whose values we hope to estimate with our sample data, $(y_i, \mathbf{x}_i), i = 1, \dots, n$. This is a problem of statistical inference that is discussed in Chapter 4 and much of the rest of the book. It is useful, however, to begin by considering the algebraic problem of choosing a vector $\mathbf{b}$ so that the fitted line $\mathbf{x}_i'\mathbf{b}$ is close to the data points. The measure of closeness constitutes a **fitting criterion**. The one used most frequently is **least squares**.[1]

3.2.1 THE LEAST SQUARES COEFFICIENT VECTOR

The least squares coefficient vector minimizes the sum of squared residuals:

$$\sum_{i=1}^{n} e_{i0}^2 = \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\mathbf{b}_0)^2, \tag{3-1}$$

where $\mathbf{b}_0$ denotes a choice for the coefficient vector. In matrix terms, minimizing the sum of squares in (3-1) requires us to choose $\mathbf{b}_0$ to

$$\text{Minimize}_{\mathbf{b}_0}\, S(\mathbf{b}_0) = \mathbf{e}_0'\mathbf{e}_0 = (\mathbf{y} - \mathbf{X}\mathbf{b}_0)'\,(\mathbf{y} - \mathbf{X}\mathbf{b}_0). \tag{3-2}$$

[1] We have yet to establish that the practical approach of fitting the line as closely as possible to the data by least squares leads to estimators with good statistical properties. This makes intuitive sense and is, indeed, the case. We shall return to the statistical issues in Chapter 4.

Expanding this gives

$$\mathbf{e}_0'\mathbf{e}_0 = \mathbf{y}'\mathbf{y} - \mathbf{b}_0'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\mathbf{b}_0 + \mathbf{b}_0'\mathbf{X}'\mathbf{X}\mathbf{b}_0 \tag{3-3}$$

or

$$S(\mathbf{b}_0) = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\mathbf{b}_0 + \mathbf{b}_0'\mathbf{X}'\mathbf{X}\mathbf{b}_0.$$

The necessary condition for a minimum is

$$\frac{\partial S(\mathbf{b}_0)}{\partial \mathbf{b}_0} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b}_0 = \mathbf{0}.^{2} \tag{3-4}$$

Let **b** be the solution (assuming it exists). Then, after manipulating (3-4), we find that **b** satisfies the **least squares normal equations**,

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}. \tag{3-5}$$

If the inverse of $\mathbf{X}'\mathbf{X}$ exists, which follows from the full column rank assumption (Assumption A2 in Section 2.3), then the solution is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{3-6}$$

For this solution to minimize the sum of squares, the second derivatives matrix,

$$\frac{\partial^2 S(\mathbf{b}_0)}{\partial \mathbf{b}_0\, \partial \mathbf{b}_0'} = 2\mathbf{X}'\mathbf{X},$$

must be a positive definite matrix. Let $q = \mathbf{c}'\mathbf{X}'\mathbf{X}\mathbf{c}$ for some arbitrary nonzero vector **c**. (The multiplication by 2 is irrelevant.) Then

$$q = \mathbf{v}'\mathbf{v} = \sum_{i=1}^{n} v_i^2, \quad \text{where } \mathbf{v} = \mathbf{X}\mathbf{c}.$$

Unless every element of **v** is zero, q is positive. But if **v** could be zero, then **v** would be a linear combination of the columns of **X** that equals **0**, which contradicts Assumption A2, that **X** has full column rank. Because **c** is arbitrary, q is positive for every nonzero **c**, which establishes that $2\mathbf{X}'\mathbf{X}$ is positive definite. Therefore, if **X** has full column rank, then the least squares solution **b** is unique and minimizes the sum of squared residuals.

3.2.2 APPLICATION: AN INVESTMENT EQUATION

To illustrate the computations in a multiple regression, we consider an example based on the macroeconomic data in Appendix Table F3.1. To estimate an investment equation, we first convert the investment series in Table F3.1 to real terms by dividing them by the GDP deflator and then scale the series so that they are measured in trillions of dollars. The real GDP series is the quantity index reported in the Economic Report of the President (2016). The other variables in the regression are a time trend (1, 2, . . .), an interest rate (the prime rate), and the yearly rate of inflation in the Consumer Price Index. These produce the data matrices listed in Table 3.1. Consider first a regression of real investment on a constant, the time trend, and real GDP, which correspond to x_1, x_2,

[2] See Appendix A.8 for discussion of calculus results involving matrices and vectors.

TABLE 3.1 Data Matrices

Real Investment (Y)	*Constant (1)*	*Trend (T)*	*Real GDP (G)*	*Interest Rate (R)*	*Inflation Rate (P)*
2.484	1	1	87.1	9.23	3.4
2.311	1	2	88.0	6.91	1.6
2.265	1	3	89.5	4.67	2.4
2.339	1	4	92.0	4.12	1.9
2.556	1	5	95.5	4.34	3.3
2.759	1	6	98.7	6.19	3.4
2.828	1	7	101.4	7.96	2.5
y = 2.717	**X** = 1	8	103.2	8.05	4.1
2.445	1	9	102.9	5.09	0.1
1.878	1	10	100.0	3.25	2.7
2.076	1	11	102.5	3.25	1.5
2.168	1	12	104.2	3.25	3.0
2.356	1	13	105.6	3.25	1.7
2.482	1	14	109.0	3.25	1.5
2.637	1	15	111.6	3.25	0.8

Notes:

1. Data from 2000–2014 obtained from Tables B-3, B-10, and B17 from Economic Report of the President: https://www.whitehouse.gov/sites/default/files/docs/2015_erp_appendix_b.pdf.
2. Results are based on the values shown. Slightly different results are obtained if the raw data on investment and the GNP deflator in Table F3.1 are input to the computer program and used to compute real investment = gross investment/(0.01*GNP deflator) internally.

and x_3. (For reasons to be discussed in Chapter 21, this is probably not a well-specified equation for these macroeconomic variables. It will suffice for a simple numerical example, however.) Inserting the specific variables of the example into (3-5), we have

$$\begin{array}{lclclcl} b_1 n & + & b_2 \Sigma_i T_i & + & b_3 \Sigma_i G_i & = & \Sigma_i Y_i, \\ b_1 \Sigma_i T_i & + & b_2 \Sigma_i T_i^2 & + & b_3 \Sigma_i T_i G_i & = & \Sigma_i T_i Y_i, \\ b_1 \Sigma_i G_i & + & b_2 \Sigma_i T_i G_i & + & b_3 \Sigma_i G_i^2 & = & \Sigma_i G_i Y_i. \end{array}$$

A solution for b_1 can be obtained by dividing the first equation by n and rearranging it to obtain

$$\begin{aligned} b_1 &= \overline{Y} - b_2 \overline{T} - b_3 \overline{G} \\ &= 2.41882 - b_2 \times 8 - b_3 \times 99.4133. \end{aligned} \tag{3-7}$$

Insert this solution in the second and third equations, and rearrange terms again to yield a set of two equations:

$$\begin{array}{lclcl} b_2 \Sigma_i (T_i - \overline{T})^2 & + & b_3 \Sigma_i (T_i - \overline{T})(G_i - \overline{G}) & = & \Sigma_i (T_i - \overline{T})(Y_i - \overline{Y}), \\ b_2 \Sigma_i (G_i - \overline{G})(T_i - \overline{T}) & + & b_3 \Sigma_i (G_i - \overline{G})^2 & = & \Sigma_i (G_i - \overline{G})(Y_i - \overline{Y}). \end{array}$$

This result shows the nature of the solution for the slopes, which can be computed from the sums of squares and cross products of the deviations of the variables from their means. Letting lowercase letters indicate variables measured as deviations from the sample means, we find that the normal equations are

$$\begin{array}{ccccc} b_2\Sigma_i t_i^2 & + & b_3\Sigma_i t_i g_i & = & \Sigma_i t_i y_i, \\ b_2\Sigma_i g_i t_i & + & b_3\Sigma_i g_i^2 & = & \Sigma_i g_i y_i, \end{array}$$

and the least squares solutions for b_2 and b_3 are

$$b_2 = \frac{\Sigma_i t_i y_i \Sigma_i g_i^2 - \Sigma_i g_i y_i \Sigma_i t_i g_i}{\Sigma_i t_i^2 \Sigma_i g_i^2 - (\Sigma_i g_i t_i)^2} = \frac{-1.6351(792.857) - 4.22255(451.9)}{280(792.857) - (451.9)^2} = -0.180169,$$

$$b_3 = \frac{\Sigma_i g_i y_i \Sigma_i t_i^2 - \Sigma_i t_i y_i \Sigma_i t_i g_i}{\Sigma_i t_i^2 \Sigma_i g_i^2 - (\Sigma_i g_i t_i)^2} = \frac{4.22255(280) - (-1.6351)(451.9)}{280(792.857) - (451.9)^2} = 0.1080157. \qquad \textbf{(3-8)}$$

With these solutions in hand, b_1 can now be computed using (3-7); $b_1 = -6.8780284$.

Suppose that we just regressed investment on the constant and GDP, omitting the time trend. At least some of the correlation between real investment and real GDP that we observe in the data will be explainable because both variables have an obvious time trend. (The trend in investment clearly has two parts, before and after the crash of 2007–2008.) Consider how this shows up in the regression computation. Denoting by "b_{yx}" the slope in the simple, **bivariate regression** of variable y on a constant and the variable x, we find that the slope in this reduced regression would be

$$b_{YG} = \frac{\Sigma_i g_i y_i}{\Sigma_i g_i^2} = 0.00533. \qquad \textbf{(3-9)}$$

By manipulating the earlier expression for b_3 and using the definition of the sample correlation between G and T, $r_{GT}^2 = (\Sigma_i g_i t_i)^2/(\Sigma_i g_i^2 \Sigma_i t_i^2)$, we obtain

$$b_{YG|T} = \frac{b_{YG}}{1 - r_{GT}^2} - \frac{b_{YT}b_{TG}}{1 - r_{GT}^2} = b_{YG} - \left(\frac{b_{YT}b_{TG} - r_{GT}^2 b_{YG}}{1 - r_{GT}^2}\right) = 0.1080157. \qquad \textbf{(3-10)}$$

(The notation "$b_{YG|T}$" used on the left-hand side is interpreted to mean the slope in the regression of Y on G and a constant "in the presence of T.") The slope in the **multiple regression** differs from that in the simple regression by a factor of 20, by including a correction that accounts for the influence of the additional variable T on both Y and G. For a striking example of this effect, in the simple regression of real investment on a time trend, $b_{YT} = -1.6351/280 = -0.00584$. But, in the multiple regression, after we account for the influence of GNP on real investment, the slope on the time trend is -0.180169. The general result for a three-variable regression in which x_1 is a constant term is

$$b_{Y2|3} = \frac{b_{Y2} - b_{Y3}b_{32}}{1 - r_{23}^2}. \qquad \textbf{(3-11)}$$

It is clear from this expression that the magnitudes of $b_{y2|3}$ and b_{y2} can be quite different. They need not even have the same sign. The result just seen is worth emphasizing; the coefficient on a variable in the simple regression [e.g., Y on $(1,G)$] will generally not be the same as the one on that variable in the multiple regression [e.g., $>Y$ on $(1,T,$G)] if the new variable and the old one are correlated. But, note that b_{YG} in (3-9) *will* be the same as $b_3 = b_{YG|T}$ in (3-8) if $\Sigma_i t_i g_i = 0$, that is, if T and G are not correlated.

In practice, you will never actually compute a multiple regression by hand or with a calculator. For a regression with more than three variables, the tools of matrix algebra are indispensable (as is a computer). Consider, for example, an enlarged model of investment that includes—in addition to the constant, time trend, and GDP—an interest rate and the rate of inflation. Least squares requires the simultaneous solution of five normal equations. Letting **X** and **y** denote the full data matrices shown previously, the normal equations in (3-5) are

$$\begin{bmatrix} 15.000 & 120.00 & 1491.2 & 76.05 & 33.90 \\ 120.000 & 1240.0 & 12381.5 & 522.06 & 244.10 \\ 1491.2 & 12381.5 & 149038 & 7453.03 & 3332.83 \\ 76.06 & 522.06 & 7453.03 & 446.323 & 186.656 \\ 33.90 & 244.10 & 3332.83 & 186.656 & 93.33 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix} = \begin{bmatrix} 36.28230 \\ 288.624 \\ 3611.17 \\ 188.176 \\ 82.7731 \end{bmatrix}.$$

The solution is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (-6.25441, -0.161342, 0.0994684, 0.0196656, -0.0107206)'.$$

3.2.3 ALGEBRAIC ASPECTS OF THE LEAST SQUARES SOLUTION

The normal equations are

$$\mathbf{X}'\mathbf{X}\mathbf{b} - \mathbf{X}'\mathbf{y} = -\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) = -\mathbf{X}'\mathbf{e} = \mathbf{0}. \qquad \textbf{(3-12)}$$

Hence, for every column $\mathbf{x}_k$ of **X**, $\mathbf{x}_k'\mathbf{e} = 0$. If the first column of **X** is a column of 1s, which we denote **i**, then there are three implications.

1. *The least squares residuals sum to zero*. This implication follows from $\mathbf{x}_1'\mathbf{e} = \mathbf{i}'\mathbf{e} = \Sigma_i e_i = 0$.
2. *The regression hyperplane passes through the point of means of the data*. The first normal equation implies that $\bar{y} = \bar{\mathbf{x}}'\mathbf{b}$. This follows from $\Sigma_i e_i = \Sigma_i (y_i - \mathbf{x}_i'\mathbf{b}) = 0$ by dividing by n.
3. *The mean of the fitted values from the regression equals the mean of the actual values*. This implication follows from point 2 because the fitted values are $\mathbf{x}_i'\mathbf{b}$.

It is important to note that none of these results need hold if the regression does not contain a constant term.

3.2.4 PROJECTION

The vector of least squares residuals is

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b}. \qquad \textbf{(3-13)}$$

Inserting the result in (3-6) for **b** gives

$$\mathbf{e} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y} = \mathbf{M}\mathbf{y}. \qquad \textbf{(3-14)}$$

The $n \times n$ matrix **M** defined in (3-14) is fundamental in regression analysis. You can easily show that **M** is both symmetric ($\mathbf{M} = \mathbf{M}'$) and idempotent ($\mathbf{M} = \mathbf{M}^2$). In view of (3-13), we can interpret **M** as a matrix that produces the vector of least squares residuals in the regression of $\mathbf{y}$ on $\mathbf{X}$ when it premultiplies any vector $\mathbf{y}$. It will be convenient later to refer to this matrix as a "**residual maker**." Matrices of this form will appear repeatedly in our development to follow.

DEFINITION 3.1: Residual Maker
Let the $n \times K$ full column rank matrix, $\mathbf{X}$ be composed of columns $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K)$, and let $\mathbf{y}$ be an $n \times 1$ column vector. The matrix, $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is a "residual maker" in that when $\mathbf{M}$ premultiplies a vector, $\mathbf{y}$, the result, $\mathbf{My}$, is the column vector of residuals in the least squares regression of $\mathbf{y}$ on $\mathbf{X}$.

It follows from the definition that

$$\mathbf{MX} = \mathbf{0}, \tag{3-15}$$

because if a column of $\mathbf{X}$ is regressed on $\mathbf{X}$, a perfect fit will result and the residuals will be zero.

Result (3-13) implies that $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$, which is the sample analog to Assumption A1, (2-3). (See Figure 3.1 as well.) The least squares results partition $\mathbf{y}$ into two parts, the fitted values $\hat{\mathbf{y}} = \mathbf{Xb}$ and the residuals, $\mathbf{e} = \mathbf{My}$. [See Section A.3.7, especially (A-54).] Because $\mathbf{MX} = \mathbf{0}$, these two parts are orthogonal. Now, given (3-13),

$$\hat{\mathbf{y}} = \mathbf{y} - \mathbf{e} = \mathbf{Iy} - \mathbf{My} = (\mathbf{I} - \mathbf{M})\mathbf{y} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{Py}. \tag{3-16}$$

The matrix $\mathbf{P}$ is a **projection matrix**. It is the matrix formed from $\mathbf{X}$ such that when a vector $\mathbf{y}$ is premultiplied by $\mathbf{P}$, the result is the fitted values in the least squares regression of $\mathbf{y}$ on $\mathbf{X}$. This is also the **projection** of the vector $\mathbf{y}$ into the column space of $\mathbf{X}$. (See Sections A3.5 and A3.7.) By multiplying it out, you will find that, like $\mathbf{M}$, $\mathbf{P}$ is symmetric and idempotent. Given the earlier results, it also follows that $\mathbf{M}$ and $\mathbf{P}$ are orthogonal;

$$\mathbf{PM} = \mathbf{MP} = \mathbf{0}.$$

As might be expected from (3-15),

$$\mathbf{PX} = \mathbf{X}.$$

As a consequence of (3-14) and (3-16), we can see that least squares partitions the vector $\mathbf{y}$ into two orthogonal parts,

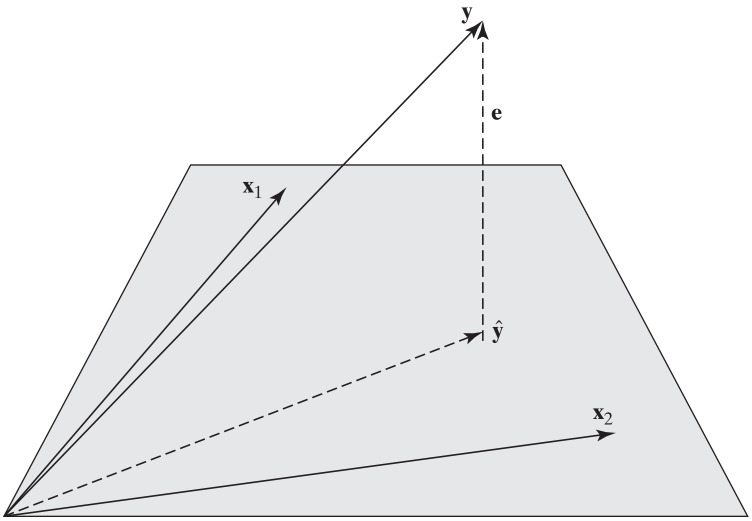
$$\mathbf{y} = \mathbf{Py} + \mathbf{My} = \textbf{projection} + \textbf{residual}.$$

The result is illustrated in Figure 3.2 for the two-variable case. The gray-shaded plane is the column space of $\mathbf{X}$. The projection and residual are the orthogonal dashed rays. We can also see the Pythagorean theorem at work in the sums of squares,

$$\begin{aligned}\mathbf{y}'\mathbf{y} &= \mathbf{y}'\mathbf{P}'\mathbf{Py} + \mathbf{y}'\mathbf{M}'\mathbf{My}\\ &= \hat{\mathbf{y}}'\hat{\mathbf{y}} + \mathbf{e}'\mathbf{e}.\end{aligned}$$

The sample linear projection of y on $\mathbf{x}$, $Proj(y|\mathbf{x})$, is an extremely useful device in empirical research. Linear least squares regression is often the starting point for model development. We will find in developing the regression model that if the population conditional mean function in Assumption A1, $E[y|\mathbf{x}]$, is linear in $\mathbf{x}$, then $E[y|\mathbf{x}]$ is also

FIGURE 3.2 Projection of **y** into the Column Space of **X.**



the population counterpart to the projection of y on $\mathbf{x}$. We will be able to show that $Proj(y|\mathbf{x})$ estimates $\mathbf{x}'\ \{E[\mathbf{xx}']\}^{-1}E[\mathbf{x}y]$, which appears implicitly in (3-16), is also $E[y|\mathbf{x}]$. If the conditional mean function is not linear in $\mathbf{x}$, then the projection of y on $\mathbf{x}$ will still estimate a useful descriptor of the joint distribution of y and $\mathbf{x}$.

3.3 PARTITIONED REGRESSION AND PARTIAL REGRESSION

It is common to specify a multiple regression model when, in fact, interest centers on only one or a subset of the full set of variables—the remaining variables are often viewed as "controls." Consider the earnings equation discussed in the Introduction. Although we are primarily interested in the effect of education on earnings, age is, of necessity, included in the model. The question we consider here is what computations are involved in obtaining, in isolation, the coefficients of a subset of the variables in a multiple regression (e.g., the coefficient of education in the aforementioned regression).

Suppose that the regression involves two sets of variables, $\mathbf{X}_1$ and $\mathbf{X}_2$. Thus,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}.$$

What is the algebraic solution for $\mathbf{b}_2$? The **normal equations** are

$$\begin{matrix}(1)\\(2)\end{matrix}\quad \begin{bmatrix}\mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2\\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2\end{bmatrix}\begin{bmatrix}\mathbf{b}_1\\ \mathbf{b}_2\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1'\mathbf{y}\\ \mathbf{X}_2'\mathbf{y}\end{bmatrix}. \tag{3-17}$$

A solution can be obtained by using the partitioned inverse matrix of (A-74). Alternatively, (1) and (2) in (3-17) can be manipulated directly to solve for $\mathbf{b}_2$. We first solve (1) for $\mathbf{b}_1$:

$$\begin{aligned}&\mathbf{X}_1'\mathbf{X}_1\mathbf{b}_1 + \mathbf{X}_1'\mathbf{X}_2\mathbf{b}_2 = \mathbf{X}_1'\mathbf{y},\\ &\mathbf{b}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} - (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\mathbf{b}_2 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{y} - \mathbf{X}_2\mathbf{b}_2).\end{aligned} \tag{3-18}$$

This solution states that $\mathbf{b}_1$ is the set of coefficients in the regression of $\mathbf{y}$ on $\mathbf{X}_1$, minus a correction vector. We digress briefly to examine an important result embedded in (3-18). Suppose that $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$. Then, $\mathbf{b}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\,\mathbf{y}$, which is simply the coefficient vector in the regression of $\mathbf{y}$ on $\mathbf{X}_1$. The general result is given in the following theorem.

THEOREM 3.1 Orthogonal Partitioned Regression
In the linear least squares multiple regression of $\mathbf{y}$ *on two sets of variables* $\mathbf{X}_1$ *and* $\mathbf{X}_2$, *if the two sets of variables are orthogonal, then the separate coefficient vectors can be obtained by separate regressions of* $\mathbf{y}$ *on* $\mathbf{X}_1$ *alone and* $\mathbf{y}$ *on* $\mathbf{X}_2$ *alone.*
Proof: *The assumption of the theorem is that* $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$ *in the normal equations in (3-17). Inserting this assumption into (3-18) produces the immediate solution for* $\mathbf{b}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$ *and likewise for* $\mathbf{b}_2$.

If the two sets of variables $\mathbf{X}_1$ and $\mathbf{X}_2$ are not orthogonal, then the solutions for $\mathbf{b}_1$ and $\mathbf{b}_2$ found by (3-17) and (3-18) are more involved than just the simple regressions in Theorem 3.1. The more general solution is suggested by the following theorem:

THEOREM 3.2 Frisch–Waugh (1933)–Lovell (1963) Theorem[3]
In the linear least squares regression of vector $\mathbf{y}$ *on two sets of variables,* $\mathbf{X}_1$ *and* $\mathbf{X}_2$, *the subvector* $\mathbf{b}_2$ *is the set of coefficients obtained when the residuals from a regression of* $\mathbf{y}$ *on* $\mathbf{X}_1$ *alone are regressed on the set of residuals obtained when each column of* $\mathbf{X}_2$ *is regressed on* $\mathbf{X}_1$.

To prove Theorem 3.2, begin from equation (2) in (3-17), which is

$$\mathbf{X}_2'\mathbf{X}_1\mathbf{b}_1 + \mathbf{X}_2'\mathbf{X}_2\mathbf{b}_2 = \mathbf{X}_2'\mathbf{y}.$$

Now, insert the result for $\mathbf{b}_1$ that appears in (3-18) into this result. This produces

$$\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} - \mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\mathbf{b}_2 + \mathbf{X}_2'\mathbf{X}_2\mathbf{b}_2 = \mathbf{X}_2'\mathbf{y}.$$

After collecting terms, the solution is

$$\begin{aligned}\mathbf{b}_2 &= [\mathbf{X}_2'(\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1')\mathbf{X}_2]^{-1}[\mathbf{X}_2'(\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1')\mathbf{y}]\\ &= (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}(\mathbf{X}_2'\mathbf{M}_1\mathbf{y}). \qquad \textbf{(3-19)}\end{aligned}$$

[3] The theorem, such as it was, appeared in the first volume of *Econometrica*, in the introduction to the paper: "The partial trend regression method can never, indeed, achieve anything which the individual trend method cannot, because the two methods lead by definition to identically the same results." Thus, Frisch and Waugh were concerned with the (lack of) difference between a regression of a variable $\mathbf{y}$ on a time trend variable, $\mathbf{t}$, and another variable, $\mathbf{x}$, compared to the regression of a detrended $\mathbf{y}$ on a detrended $\mathbf{x}$, where detrending meant computing the residuals of the respective variable on a constant and the time trend, $\mathbf{t}$. A concise statement of the theorem and its matrix formulation were added later by Lovell (1963).

The $\mathbf{M}_1$ matrix appearing in the parentheses inside each set of parentheses is the "residual maker" defined in (3-14) and Definition 3.1, in this case defined for a regression on the columns of $\mathbf{X}_1$. Thus, $\mathbf{M}_1\mathbf{X}_2$ is a matrix of residuals; each column of $\mathbf{M}_1\mathbf{X}_2$ is a vector of residuals in the regression of the corresponding column of $\mathbf{X}_2$ on the variables in $\mathbf{X}_1$. By exploiting the fact that $\mathbf{M}_1$, like $\mathbf{M}$, is symmetric and idempotent, we can rewrite (3-19) as

$$\mathbf{b}_2 = (\mathbf{X}_2^{*\prime}\mathbf{X}_2^{*})^{-1}\mathbf{X}_2^{*\prime}\mathbf{y}_*, \tag{3-20}$$

where $\mathbf{X}_2^* = \mathbf{M}_1\mathbf{X}_2$ and $\mathbf{y}_* = \mathbf{M}_1\mathbf{y}$. This result is fundamental in regression analysis.

This process is commonly called **partialing out** or **netting out** the effect of $\mathbf{X}_1$. For this reason, the coefficients in a multiple regression are often called the **partial regression coefficients**. The application of Theorem 3.2 to the computation of a single coefficient as suggested at the beginning of this section is detailed in the following: Consider the regression of $\mathbf{y}$ on a set of variables $\mathbf{X}$ and an additional variable $\mathbf{z}$. Denote the coefficients $\mathbf{b}$ and c, respectively.

COROLLARY 3.2.1 Individual Regression Coefficients
The coefficient on $\mathbf{z}$ *in a multiple regression of* $\mathbf{y}$ *on* $\mathbf{W} = [\mathbf{X}, \mathbf{z}]$ *is computed as* $c = (\mathbf{z}'\mathbf{M}_\mathbf{X}\mathbf{z})^{-1}(\mathbf{z}'\mathbf{M}_\mathbf{X}\mathbf{y}) = (\mathbf{z}_*'\mathbf{z}_*)^{-1}\,\mathbf{z}_*'\mathbf{y}_*$ *where* $\mathbf{z}_*$ *and* $\mathbf{y}_*$ *are the residual vectors from least squares regressions of* $\mathbf{z}$ *and* $\mathbf{y}$ *on* $\mathbf{X}$; $\mathbf{z}_* = \mathbf{M}_\mathbf{X}\mathbf{z}$ *and* $\mathbf{y}_* = \mathbf{M}_\mathbf{X}\mathbf{y}$ *where* $\mathbf{M}_\mathbf{X}$ *is defined in (3-14).*
Proof: *This is an application of Theorem 3.2 in which* $\mathbf{X}_1$ *is* $\mathbf{X}$ *and* $\mathbf{X}_2$ *is* $\mathbf{z}$.

In terms of Example 2.2, we could obtain the coefficient on education in the multiple regression by first regressing earnings and education on age (or age and age squared) and then using the residuals from these regressions in a simple regression. In the classic application of this latter observation, Frisch and Waugh (1933) noted that in a time-series setting, the same results were obtained whether a regression was fitted with a time-trend variable or the data were first "detrended" by netting out the effect of time, as noted earlier, and using just the detrended data in a simple regression.

Consider the case in which $\mathbf{X}_1$ is $\mathbf{i}$, a constant term that is a column of 1s in the first column of $\mathbf{X}$, and $\mathbf{X}_2$ is a set of variables. The solution for $\mathbf{b}_2$ in this case will then be the slopes in a regression that contains a constant term. Using Theorem 3.2 the vector of residuals for any variable, $\mathbf{x}$, in $\mathbf{X}_2$ will be

$$\begin{aligned}\mathbf{x}_* &= \mathbf{x} - \mathbf{i}(\mathbf{i}'\mathbf{i})^{-1}\mathbf{i}'\mathbf{x}\\ &= \mathbf{x} - \mathbf{i}(1/n)\mathbf{i}'\mathbf{x}\\ &= \mathbf{x} - \mathbf{i}\overline{x}\\ &= \mathbf{M}^0\mathbf{x}.\end{aligned} \tag{3-21}$$

(See Section A.5.4 where we have developed this result purely algebraically.) For this case, then, the residuals are deviations from the sample mean. Therefore, each column of $\mathbf{M}_1\mathbf{X}_2$ is the original variable, now in the form of deviations from the mean. This general result is summarized in the following corollary.

COROLLARY 3.2.2 Regression with a Constant Term
The slopes in a multiple regression that contains a constant term can be obtained by transforming the data to deviations from their means and then regressing the variable y in deviation form on the explanatory variables, also in deviation form.

[We used this result in (3-8).] Having obtained the coefficients on $\mathbf{X}_2$, how can we recover the coefficients on $\mathbf{X}_1$ (the constant term)? One way is to repeat the exercise while reversing the roles of $\mathbf{X}_1$ and $\mathbf{X}_2$. But there is an easier way. We have already solved for $\mathbf{b}_2$. Therefore, we can use (3-18) in a solution for $\mathbf{b}_1$. If $\mathbf{X}_1$ is just a column of 1s, then the first of these produces the familiar result

$$b_1 = \overline{y} - \overline{x}_2 b_2 - \cdots - \overline{x}_K b_K$$

[which is used in (3-7)].

Theorem 3.2 and Corollaries 3.2.1 and 3.2.2 produce a useful interpretation of the **partitioned regression** when the model contains a constant term. According to Theorem 3.1, if the columns of $\mathbf{X}$ are orthogonal, that is, $\mathbf{X}_k'\mathbf{x}_m = 0$ for columns k and m, then the separate regression coefficients in the regression of $\mathbf{y}$ on $\mathbf{X}$ when $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K]$ are simply $\mathbf{x}_k'\mathbf{y}/\mathbf{x}_k'\mathbf{x}_k$. When the regression contains a constant term, we can compute the multiple regression coefficients by regression of $\mathbf{y}$ in mean deviation form on the columns of $\mathbf{X}$, also in deviations from their means. In this instance, the *orthogonality* of the columns means that the sample covariances (and correlations) of the variables are zero. The result is another theorem:

THEOREM 3.3 Orthogonal Regression
If the multiple regression of **y** *on* **X** *contains a constant term and the variables in the regression are uncorrelated, then the multiple regression slopes are the same as the slopes in the individual simple regressions of* **y** *on a constant and each variable in turn.*
Proof: *The result follows from Theorems 3.1 and 3.2.*

3.4 PARTIAL REGRESSION AND PARTIAL CORRELATION COEFFICIENTS

The use of multiple regression involves a conceptual experiment that we might not be able to carry out in practice, the *ceteris paribus* analysis familiar in economics. To pursue the earlier example, a regression equation relating earnings to age and education enables us to do the experiment of comparing the earnings of two individuals of the same age with different education levels, *even if the sample contains no such pair of individuals*. It is this characteristic of the regression that is implied by the term partial regression coefficients. The way we obtain this result, as we have seen, is first to regress income and education on age and then to compute the residuals from this regression. By construction, age will not have any power in explaining variation in these residuals. Therefore, any correlation between income and education after this "purging" is independent of (or after removing the effect of) age.

The same principle can be applied to the correlation between two variables. To continue our example, to what extent can we assert that this correlation reflects a direct relationship rather than that both income and education tend, on average, to rise as individuals become older? To find out, we would use a **partial correlation coefficient**, which is computed along the same lines as the partial regression coefficient. In the context of our example, the partial correlation coefficient between income and education, controlling for the effect of age, is obtained as follows:

1. $y_* =$ the residuals in a regression of income on a constant and age.
2. $z_* =$ the residuals in a regression of education on a constant and age.
3. The partial correlation r^*_{yz} is the simple correlation between y_* and z_*.

This calculation might seem to require a large amount of computation. Using Corollary 3.2.1, the two residual vectors in points 1 and 2 are $\mathbf{y}_* = \mathbf{My}$ and $\mathbf{z}_* = \mathbf{Mz}$ where $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the residual maker defined in (3-14). We will assume that there is a constant term in $\mathbf{X}$ so that the vectors of residuals $\mathbf{y}_*$ and $\mathbf{z}_*$ have zero sample means. Then, the square of the partial correlation coefficient is

$$r^{*2}_{yz} = \frac{(\mathbf{z}'_*\mathbf{y}_*)^2}{(\mathbf{z}'_*\mathbf{z}_*)(\mathbf{y}'_*\mathbf{y}_*)}.$$

There is a convenient shortcut. Once the multiple regression is computed, the t ratio in (5-13) for testing the hypothesis that the coefficient equals zero (e.g., the last column of Table 4.6) can be used to compute

$$r^{*2}_{yz} = \frac{t_z^2}{t_z^2 + \text{degrees of freedom}}, \tag{3-22}$$

where the **degrees of freedom** is equal to $n-(K+1)$; $K+1$ is the number of variables in the regression plus the constant term. The proof of this less than perfectly intuitive result will be useful to illustrate some results on partitioned regression. We will rely on two useful theorems from least squares algebra. The first isolates a particular diagonal element of the inverse of a **moment matrix** such as $(\mathbf{X}'\mathbf{X})^{-1}$.

THEOREM 3.4 Diagonal Elements of the Inverse of a Moment Matrix
Let $\mathbf{W}$ *denote the partitioned matrix* $[\mathbf{X}, \mathbf{z}]$*—that is, the K columns of* $\mathbf{X}$ *plus an additional column labeled* $\mathbf{z}$*. The last diagonal element of* $(\mathbf{W}'\mathbf{W})^{-1}$ *is* $(\mathbf{z}'\mathbf{M_X}\mathbf{z})^{-1} = (\mathbf{z}'_*\mathbf{z}_*)^{-1}$ *where* $\mathbf{z}_* = \mathbf{M_X}\mathbf{z}$ *and* $\mathbf{M_X} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.
Proof: *This is an application of the partitioned inverse formula in (A-74) where* $A_{11} = \mathbf{X}'\mathbf{X}$, $A_{12} = \mathbf{X}'\mathbf{z}$, $A_{21} = \mathbf{z}'\mathbf{X}$ *and* $A_{22} = \mathbf{z}'\mathbf{z}$*. Note that this theorem generalizes the development in Section A.2.8, where* $\mathbf{X}$ *contains only a constant term,* $\mathbf{i}$.

We can use Theorem 3.4 to establish the result in (3-22). Let c and $\mathbf{u}$ denote the coefficient on $\mathbf{z}$ and the vector of residuals in the multiple regression of $\mathbf{y}$ on $\mathbf{W} = [\mathbf{X}, \mathbf{z}]$, respectively. Then, by definition, the squared t ratio that appears in (3-22) is

$$t_z^2 = \frac{c^2}{\left[\dfrac{\mathbf{u}'\mathbf{u}}{n-(K+1)}\right](\mathbf{W}'\mathbf{W})^{-1}_{K+1,\,K+1}},$$

where $(\mathbf{W}'\mathbf{W})^{-1}_{K+1,\,K+1}$ is the $(K+1)$ (last) diagonal element of $(\mathbf{W}'\mathbf{W})^{-1}$. [The bracketed term appears in (4-17).] The theorem states that this element of the matrix equals $(\mathbf{z}_*'\mathbf{z}_*)^{-1}$. From Corollary 3.2.1, we also have that $c^2 = [(\mathbf{z}_*'\mathbf{y}_*)/(\mathbf{z}_*'\mathbf{z}_*)]^2$. For convenience, let $DF = n - (K+1)$. Then, $t_z^2 = \dfrac{(\mathbf{z}_*'\mathbf{y}_*/\mathbf{z}_*'\mathbf{z}_*)^2}{(\mathbf{u}'\mathbf{u}/DF)(\mathbf{z}_*'\mathbf{z}_*)^{-1}} = \dfrac{(\mathbf{z}_*'\mathbf{y}_*)^2 DF}{(\mathbf{u}'\mathbf{u})(\mathbf{z}_*'\mathbf{z}_*)^{\ -1}}$. It follows that the result in (3-22) is equivalent to

$$\frac{t_z^2}{t_z^2 + DF} = \frac{\dfrac{(\mathbf{z}_*'\mathbf{y}_*)^2 DF}{(\mathbf{u}'\mathbf{u})(\mathbf{z}_*'\mathbf{z}_*)}}{\dfrac{(\mathbf{z}_*'\mathbf{y}_*)^2 DF}{(\mathbf{u}'\mathbf{u})(\mathbf{z}_*'\mathbf{z}_*)} + DF} = \frac{\dfrac{(\mathbf{z}_*'\mathbf{y}_*)^2}{(\mathbf{u}'\mathbf{u})(\mathbf{z}_*'\mathbf{z}_*)}}{\dfrac{(\mathbf{z}_*'\mathbf{y}_*)^2}{(\mathbf{u}'\mathbf{u})(\mathbf{z}_*'\mathbf{z}_*)} + 1} = \frac{(\mathbf{z}_*'\mathbf{y}_*)^2}{(\mathbf{z}_*'\mathbf{y}_*)^2 + (\mathbf{u}'\mathbf{u})(\mathbf{z}_*'\mathbf{z}_*)}.$$

Divide numerator and denominator by $(\mathbf{z}_*'\mathbf{z}_*)(\mathbf{y}_*'\mathbf{y}_*)$ to obtain

$$\frac{t_z^2}{t_z^2 + DF} = \frac{(\mathbf{z}_*'\mathbf{y}_*)^2/((\mathbf{z}_*'\mathbf{z}_*)(\mathbf{y}_*'\mathbf{y}_*))}{(\mathbf{z}_*'\mathbf{y}_*)^2/((\mathbf{z}_*'\mathbf{z}_*)(\mathbf{y}_*'\mathbf{y}_*)) + ((\mathbf{u}'\mathbf{u})(\mathbf{z}_*'\mathbf{z}_*))/((\mathbf{z}_*'\mathbf{z}_*)(\mathbf{y}_*'\mathbf{y}_*))} = \frac{r_{yz}^{*2}}{r_{yz}^{*2} + (\mathbf{u}'\mathbf{u})/(\mathbf{y}_*'\mathbf{y}_*)}. \qquad \textbf{(3-23)}$$

We will now use a second theorem to manipulate $\mathbf{u}'\mathbf{u}$ and complete the derivation. The result we need is given in Theorem 3.5.

Returning to the derivation, then, $\mathbf{e}'\mathbf{e} = \mathbf{y}_*'\mathbf{y}_*$ and $c^2(\mathbf{z}_*'\mathbf{z}_*) = (\mathbf{z}_*'\mathbf{y}_*)^2/(\mathbf{z}_*'\mathbf{z}_*)$. Therefore,

$$\frac{\mathbf{u}'\mathbf{u}}{\mathbf{y}_*'\mathbf{y}_*} = \frac{\mathbf{y}_*'\mathbf{y}_* - (\mathbf{z}_*'\mathbf{y}_*)^2/\mathbf{z}_*'\mathbf{z}_*}{\mathbf{y}_*'\mathbf{y}_*} = 1 - r_{yz}^{*2}.$$

Inserting this in the denominator of (3-23) produces the result we sought.

THEOREM 3.5 Change in the Sum of Squares When a Variable Is Added to a Regression

If $\mathbf{e}'\mathbf{e}$ *is the sum of squared residuals when* $\mathbf{y}$ *is regressed on* $\mathbf{X}$ *and* $\mathbf{u}'\mathbf{u}$ *is the sum of squared residuals when* $\mathbf{y}$ *is regressed on* $\mathbf{X}$ *and* $\mathbf{z}$, *then*

$$\mathbf{u}'\mathbf{u} = \mathbf{e}'\mathbf{e} - c^2\,(\mathbf{z}_*'\mathbf{z}_*) \le \mathbf{e}'\mathbf{e}, \qquad \textbf{(3-24)}$$

where c is the coefficient on $\mathbf{z}$ *in the long regression of* $\mathbf{y}$ *on* $[\mathbf{X}, \mathbf{z}]$ *and* $\mathbf{z}_* = \mathbf{Mz}$ *is the vector of residuals when* $\mathbf{z}$ *is regressed on* $\mathbf{X}$.

Proof: *In the long regression of* $\mathbf{y}$ *on* $\mathbf{X}$ *and* $\mathbf{z}$, *the vector of residuals is* $\mathbf{u} = \mathbf{y} - \mathbf{Xd} - \mathbf{z}c$. *Note that unless* $\mathbf{X}'\mathbf{z} = \mathbf{0}$, $\mathbf{d}$ *will not equal* $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. *(See Section 4.3.2.) Moreover, unless* $c = 0$, $\mathbf{u}$ *will not equal* $\mathbf{e} = \mathbf{y} - \mathbf{Xb}$. *From Corollary 3.2.1,* $c = (\mathbf{z}_*'\mathbf{z}_*)^{-1}(\mathbf{z}_*'\mathbf{y}_*)$. *From (3-18), we also have that the coefficients on* $\boldsymbol{X}$ *in this long regression are*

$$\mathbf{d} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \mathbf{z}c) = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}c.$$

Inserting this expression for **d** *in that for* **u** *gives*

$$\mathbf{u} = \mathbf{y} - \mathbf{Xb} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}c - \mathbf{z}c = \mathbf{e} - \mathbf{M_X z}c = \mathbf{e} - \mathbf{z}_* c.$$

Then,

$$\mathbf{u}'\mathbf{u} = \mathbf{e}'\mathbf{e} + c^2\,(\mathbf{z}_*'\mathbf{z}_*) - 2c(\mathbf{z}_*'\mathbf{e}).$$

But, $\mathbf{e} = \mathbf{M_x y} = \mathbf{y}_*$ *and* $\mathbf{z}_*'\mathbf{e} = \mathbf{z}_*'\mathbf{y}_* = c(\mathbf{z}_*'\mathbf{z}_*)$. *Inserting this result in* $\mathbf{u}'\mathbf{u}$ *immediately above gives the result in the theorem.*

Example 3.1 Partial Correlations

For the data in the application in Section 3.2.2, the simple correlations between investment and the regressors, r_{Yk}, and the partial correlations, r_{Yk}^*, between investment and the four regressors (given the other variables) are listed in Table 3.2. As is clear from the table, there is no necessary relation between the simple and partial correlation coefficients. One thing worth noting is that the signs of the partial correlations are the same as those of the coefficients, but not necessarily the same as the signs of the raw correlations. Note the difference in the coefficient on *Inflation*.

3.5 GOODNESS OF FIT AND THE ANALYSIS OF VARIANCE

The original fitting criterion, the sum of squared residuals, suggests a measure of the fit of the regression line to the data. However, as can easily be verified, the sum of squared residuals can be scaled arbitrarily just by multiplying all the values of y by the desired scale factor. Because the fitted values of the regression are based on the values of $\mathbf{x}$, we might ask instead whether *variation* in $\mathbf{x}$ is a good predictor of *variation* in y. Figure 3.3 shows three possible cases for a simple linear regression model, $y = \beta_1 + \beta_2 x + \varepsilon$. The measure of fit described here embodies both the fitting criterion and the covariation of y and $\mathbf{x}$.

Variation of the dependent variable is defined in terms of deviations from its mean, $(y_i - \bar{y})$. The **total variation** in y is the sum of squared deviations:

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

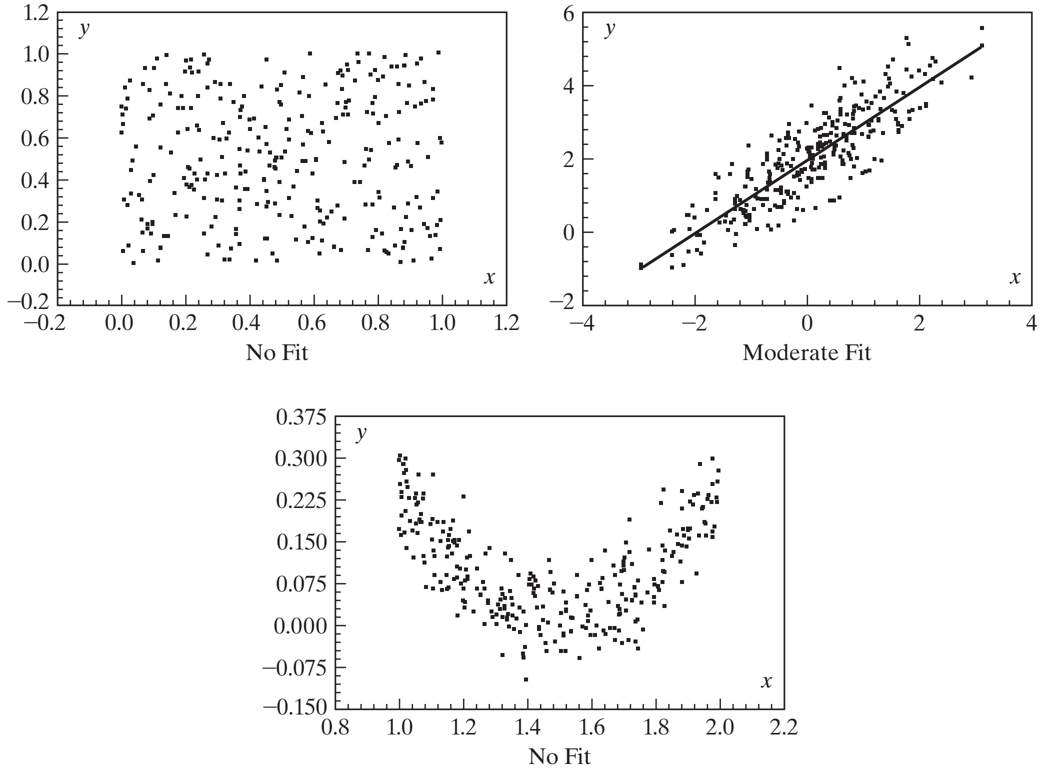
In terms of the regression equation, we may write the full set of observations as

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e} = \hat{\mathbf{y}} + \mathbf{e}.$$

TABLE 3.2 Correlations of Investment with Other Variables (DF = 10)

Variable	*Coefficient*	*t Ratio*	*Simple Correlation*	*Partial Correlation*
Trend	–0.16134	–3.42	–0.09965	–0.73423
RealGDP	0.09947	4.12	0.15293	0.79325
Interest	0.01967	0.58	0.55006	0.18040
Inflation	–0.01072	–0.27	0.19332	–0.08507

FIGURE 3.3 Sample Data.



For an individual observation, we have

$$y_i = \hat{y}_i + e_i = \mathbf{x}_i'\mathbf{b} + e_i.$$

If the regression contains a constant term, then the residuals will sum to zero and the mean of the predicted values of y_i will equal the mean of the actual values. Subtracting $\overline{y}$ from both sides and using this result and result 2 in Section 3.2.3 gives

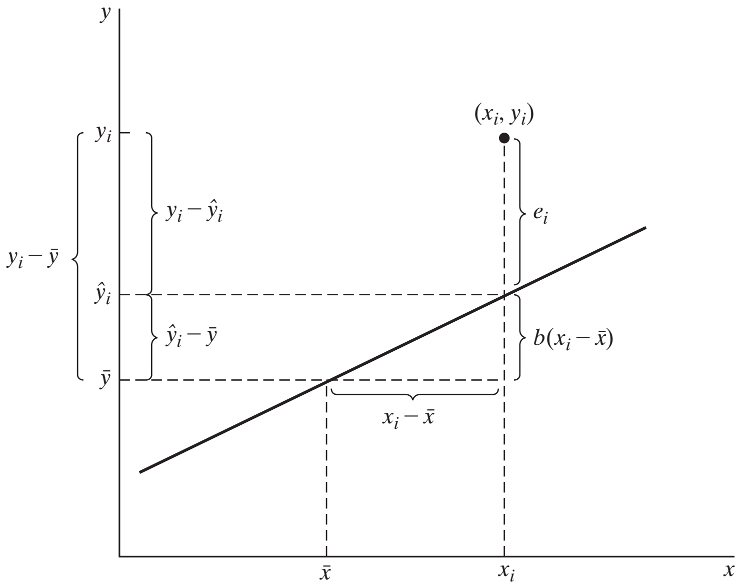
$$y_i - \overline{y} = \hat{y}_i - \overline{y} + e_i = (\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{b} + e_i.$$

Figure 3.4 illustrates the computation for the two-variable regression. Intuitively, the regression would appear to fit well if the deviations of y from its mean are more largely accounted for by deviations of x from its mean than by the residuals. Since both terms in this decomposition sum to zero, to quantify this fit, we use the sums of squares instead. For the full set of observations, we have

$$\mathbf{M}^0\mathbf{y} = \mathbf{M}^0\mathbf{X}\mathbf{b} + \mathbf{M}^0\mathbf{e},$$

where $\mathbf{M}^0$ is the $n \times n$ idempotent matrix that transforms observations into deviations from sample means. [See (3-21)and Section A.2.8; $\mathbf{M}^0$ is a residual maker for $\mathbf{X} = \mathbf{i}$.] The column of $\mathbf{M}^0\mathbf{X}$ corresponding to the constant term is zero, and, since the residuals

FIGURE 3.4 Decomposition of y_i.



already have mean zero, $\mathbf{M}^0\mathbf{e} = \mathbf{e}$. Then, since $\mathbf{e}'\mathbf{M}^0\mathbf{X} = \mathbf{e}'\mathbf{X} = \mathbf{0}$, the total sum of squares is

$$\mathbf{y}'\mathbf{M}^0\mathbf{y} = \mathbf{b}'\mathbf{X}'\mathbf{M}^0\mathbf{X}\mathbf{b} + \mathbf{e}'\mathbf{e}.$$

Write this as total sum of squares = regression sum of squares + error sum of squares, or

$$\text{SST} = \text{SSR} + \text{SSE}. \tag{3-25}$$

(Note that this is the same partitioning that appears at the end of Section 3.2.4.)

We can now obtain a measure of how well the regression line fits the data by using the

$$\textbf{coefficient of determination: } \frac{\text{SSR}}{\text{SST}} = \frac{\mathbf{b}'\mathbf{X}'\mathbf{M}^0\mathbf{X}\mathbf{b}}{\mathbf{y}'\mathbf{M}^0\mathbf{y}} = 1 - \frac{\mathbf{e}'\mathbf{e}}{\mathbf{y}'\mathbf{M}^0\mathbf{y}} = 1 - \frac{\Sigma_{i=1}^n e_i^2}{\Sigma_{i=1}^n (y_i - \bar{y})^2}. \tag{3-26}$$

The coefficient of determination is denoted R^2. As we have shown, it must be between 0 and 1, and it measures the proportion of the total variation in y that is accounted for by variation in the regressors. It equals zero if the regression is a horizontal line, that is, if all the elements of **b** except the constant term are zero. In this case, the predicted values of y are always $\bar{y}$, so deviations of **x** from its mean do not translate into different predictions for y. As such, **x** has no explanatory power. The other extreme, $R^2 = 1$, occurs if the values of **x** and y all lie in the same hyperplane (on a straight line for a two-variable regression) so that the residuals are all zero. If all the values of y_i lie on a vertical line, then R^2 has no meaning and cannot be computed.

Regression analysis is often used for forecasting. In this case, we are interested in how well the regression model predicts movements in the dependent variable. With this in mind, an equivalent way to compute R^2 is also useful. First, the sum of squares for the predicted values is

$$\Sigma_{i=1}^{n}(\hat{y}_i - \hat{\bar{y}})^2 = \hat{\mathbf{y}}'\mathbf{M}^0\hat{\mathbf{y}} = \mathbf{b}'\mathbf{X}'\mathbf{M}^0\mathbf{X}\mathbf{b},$$

but $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$, $\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e}$, $\mathbf{M}^0\mathbf{e} = \mathbf{e}$, and $\mathbf{X}'\mathbf{e} = \mathbf{0}$, so $\hat{\mathbf{y}}'\mathbf{M}^0\hat{\mathbf{y}} = \hat{\mathbf{y}}'\mathbf{M}^0\mathbf{y} = \Sigma_{i=1}^{n}(\hat{y}_i - \hat{\bar{y}})(y_i - \bar{y})$. Multiply $R^2 = \hat{\mathbf{y}}'\mathbf{M}^0\hat{\mathbf{y}}/\mathbf{y}'\mathbf{M}^0\mathbf{y} = \hat{\mathbf{y}}'\mathbf{M}^0\mathbf{y}/\mathbf{y}'\mathbf{M}^0\mathbf{y}$ by $1 = \hat{\mathbf{y}}'\mathbf{M}^0\mathbf{y}/\hat{\mathbf{y}}'\mathbf{M}^0\hat{\mathbf{y}}$ to obtain

$$R^2 = \frac{[\Sigma_i(y_i - \bar{y})(\hat{y}_i - \hat{\bar{y}})]^2}{[\Sigma_i(y_i - \bar{y})^2][\Sigma_i(\hat{y}_i - \hat{\bar{y}})^2]}, \tag{3-27}$$

which is the squared correlation between the observed values of y and the predictions produced by the estimated regression equation.

Example 3.2 Fit of a Consumption Function

The data plotted in Figure 2.1 are listed in Appendix Table F2.1. For these data, where y is C and x is X, we have $\bar{y} = 273.2727$, $\bar{x} = 323.2727$, $S_{yy} = 12{,}618.182$, $S_{xx} = 12{,}300.182$, and $S_{xy} = 8{,}423.182$, so $\text{SST} = 12{,}618.182$, $b = 8{,}423.182/12{,}300.182 = 0.6848014$, $\text{SSR} = b^2S_{xx} = 5{,}768.2068$, and $\text{SSE} = \text{SST} - \text{SSR} = 6{,}849.975$. Then $R^2 = b^2S_{xx} = 0.457135$. As can be seen in Figure 2.1, this is a moderate fit, although it is not particularly good for aggregate time-series data. On the other hand, it is clear that not accounting for the anomalous wartime data has degraded the fit of the model. This value is the R^2 for the model indicated by the solid line in the figure. By simply omitting the years 1942–1945 from the sample and doing these computations with the remaining seven observations—the dashed line—we obtain an R^2 of 0.93379. Alternatively, by creating a variable WAR which equals 1 in the years 1942–1945 and zero otherwise and including this in the model, which produces the model shown by the two dashed lines, the R^2 rises to 0.94450.

We can summarize the calculation of R^2 in an **analysis of variance** table, which might appear as shown in Table 3.3.

Example 3.3 Analysis of Variance for the Investment Equation

The analysis of variance table for the investment equation of Section 3.2.2 is given in Table 3.4.

3.5.1 THE ADJUSTED *R*-SQUARED AND A MEASURE OF FIT

There are some problems with the use of R^2 in analyzing **goodness of fit**. The first concerns the number of degrees of freedom used up in estimating the parameters.

TABLE 3.3 Analysis of Variance Table

Source	*Sum of Squares*	*Degrees of Freedom*	*Mean Square*
Regression	$\mathbf{b}'\mathbf{X}'\mathbf{y} - n\bar{y}^2$	$K - 1$ (assuming a constant term)	
Residual	$\mathbf{e}'\mathbf{e}$	$n - K$ (including the constant term)	s^2
Total	$\mathbf{y}'\mathbf{y} - n\bar{y}^2$	$n - 1$	s_y^2
R^2	$1 - \mathbf{e}'\mathbf{e}/(\mathbf{y}'\mathbf{y} - n\bar{y}^2)$		

TABLE 3.4 Analysis of Variance for the Investment Equation

Source	*Sum of Squares*	*Degrees of Freedom*	*Mean Square*
Regression	0.75621	4	
Residual	0.20368	10	0.02037
Total	0.95989	14	0.06856
R^2	0.78781		

[See (3-22) and Table 3.3.] *R^2 will never decrease when another variable is added to a regression equation*. Equation (3-24) provides a convenient means for us to establish this result. Once again, we are comparing a regression of $\mathbf{y}$ on $\mathbf{X}$ with sum of squared residuals $\mathbf{e}'\mathbf{e}$ to a regression of $\mathbf{y}$ on $\mathbf{X}$ and an additional variable $\mathbf{z}$, which produces sum of squared residuals $\mathbf{u}'\mathbf{u}$. Recall the vectors of residuals $\mathbf{z}_* = \mathbf{Mz}$ and $\mathbf{y}_* = \mathbf{My} = \mathbf{e}$, which implies that $\mathbf{e}'\mathbf{e} = (\mathbf{y}_*'\mathbf{y}_*)$. Let c be the coefficient on $\mathbf{z}$ in the longer regression. Then $c = (\mathbf{z}_*'\mathbf{z}_*)^{-1}(\mathbf{z}_*'\mathbf{y}_*)$, and inserting this in (3-24) produces

$$\mathbf{u}'\mathbf{u} = \mathbf{e}'\mathbf{e} - \frac{(\mathbf{z}_*'\mathbf{y}_*)^2}{(\mathbf{z}_*'\mathbf{z}_*)} = \mathbf{e}'\mathbf{e}(1 - r_{yz}^{*2}), \qquad \textbf{(3-28)}$$

where r_{yz}^* is the partial correlation between $\mathbf{y}$ and $\mathbf{z}$, controlling for $\mathbf{X}$. Now divide through both sides of the equality by $\mathbf{y}'\mathbf{M}^0\mathbf{y}$. From (3-26), $\mathbf{u}'\mathbf{u}/\mathbf{y}'\mathbf{M}^0\mathbf{y}$ is $(1 - R_{\mathbf{X}z}^2)$ for the regression on $\mathbf{X}$ and $\mathbf{z}$ and $\mathbf{e}'\mathbf{e}/\mathbf{y}'\mathbf{M}^0\mathbf{y}$ is $(1 - R_{\mathbf{X}}^2)$. Rearranging the result produces the following:

THEOREM 3.6 Change in R^2 When a Variable Is Added to a Regression
Let $R_{\mathbf{X}z}^2$ be the coefficient of determination in the regression of $\mathbf{y}$ on $\mathbf{X}$ and an additional variable $\mathbf{z}$, let $R_{\mathbf{X}}^2$ be the same for the regression of $\mathbf{y}$ on $\mathbf{X}$ alone, and let r_{yz}^* be the partial correlation between $\mathbf{y}$ and $\mathbf{z}$, controlling for $\mathbf{X}$. Then

$$R_{\mathbf{X}z}^2 = R_{\mathbf{X}}^2 + (1 - R_{\mathbf{X}}^2)\, r_{yz}^{*2}. \qquad \textbf{(3-29)}$$

Thus, the R^2 in the longer regression cannot be smaller. It is tempting to exploit this result by just adding variables to the model; R^2 will continue to rise to its limit of 1.[4] The **adjusted R^2** (for degrees of freedom), which incorporates a penalty for these results, is computed as follows:

$$\overline{R}^2 = 1 - \frac{\mathbf{e}'\mathbf{e}/(n-K)}{\mathbf{y}'\mathbf{M}^0\mathbf{y}/(n-1)}. \qquad \textbf{(3-30)}$$

For computational purposes, the connection between R^2 and $\overline{R}^2$ is

$$\overline{R}^2 = 1 - \frac{n-1}{n-K}(1 - R^2).$$

[4] This result comes at a cost, however. The parameter estimates become progressively less precise as we do so. We will pursue this result in Chapter 4.

The adjusted R^2 may decline when a variable is added to the set of independent variables. Indeed, $\overline{R}^2$ could even be negative. To consider an admittedly extreme case, suppose that **x** and **y** have a sample correlation of zero. Then the adjusted R^2 will equal $-1/(n-2)$. Whether $\overline{R}^2$ rises or falls when a variable is added to the model depends on whether the contribution of the new variable to the fit of the regression more than offsets the correction for the loss of an additional degree of freedom. The general result (the proof of which is left as an exercise) is as follows.

THEOREM 3.7 Change in $\overline{R}^2$ When a Variable Is Added to a Regression
In a multiple regression, $\overline{R}^2$ will fall (rise) when the variable x is deleted from the regression if the square of the t ratio associated with this variable is greater (less) than 1.

We have shown that R^2 will never fall when a variable is added to the regression. We now consider this result more generally. The change in the residual sum of squares when a set of variables $\mathbf{X}_2$ is added to the regression is

$$\mathbf{e}_1'\mathbf{e}_1 - \mathbf{e}_{1,2}'\mathbf{e}_{1,2} = \mathbf{b}_2'\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\mathbf{b}_2,$$

where $\mathbf{e}_1$ is the residuals when **y** is regressed on $\mathbf{X}_1$ alone and $\mathbf{e}_{1,2}$ indicates regression on *both* $\mathbf{X}_1$ and $\mathbf{X}_2$. The coefficient vector $\mathbf{b}_2$ is the coefficients on $\mathbf{X}_2$ in the multiple regression of **y** on $\mathbf{X}_1$ and $\mathbf{X}_2$. [See (3-19) and (3-20) for definitions of $\mathbf{b}_2$ and $\mathbf{M}_1$.] Therefore,

$$R_{1,2}^2 = 1 - \frac{\mathbf{e}_1'\mathbf{e}_1 - \mathbf{b}_2'\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\mathbf{b}_2}{\mathbf{y}'\mathbf{M}^0\mathbf{y}} = R_1^2 + \frac{\mathbf{b}_2'\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\mathbf{b}_2}{\mathbf{y}'\mathbf{M}^0\mathbf{y}},$$

which is greater than R_1^2 unless $\mathbf{b}_2$ equals zero. ($\mathbf{M}_1\mathbf{X}_2$ could not be zero unless $\mathbf{X}_2$ is a linear function of $\mathbf{X}_1$, in which case the regression on $\mathbf{X}_1$ and $\mathbf{X}_2$ could not be computed.) This equation can be manipulated a bit further to obtain

$$R_{1,2}^2 = R_1^2 + \frac{\mathbf{y}'\mathbf{M}_1\mathbf{y}}{\mathbf{y}'\mathbf{M}^0\mathbf{y}}\,\frac{\mathbf{b}_2'\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\mathbf{b}_2}{\mathbf{y}'\mathbf{M}_1\mathbf{y}}.$$

But $\mathbf{y}'\mathbf{M}_1\mathbf{y} = \mathbf{e}_1'\mathbf{e}_1$, so the first term in the product is $1 - R_1^2$. The second is the **multiple correlation** in the regression of $\mathbf{M}_1\mathbf{y}$ on $\mathbf{M}_1\mathbf{X}_2$, or the partial correlation (after the effect of $\mathbf{X}_1$ is removed) in the regression of **y** on $\mathbf{X}_2$. Collecting terms, we have

$$R_{1,2}^2 = R_1^2 + (1 - R_1^2)r_{y2.1}^{*2}. \qquad \textbf{(3-31)}$$

[This is the multivariate counterpart to (3-29).]

It is possible to push R^2 as high as desired (up to one) just by adding regressors to the model. This possibility motivates the use of the adjusted R^2 in (3-30), instead of R^2 as a method of choosing among alternative models. Since $\overline{R}^2$ incorporates a penalty for reducing the degrees of freedom while still revealing an improvement in fit, one possibility is to choose the specification that maximizes $\overline{R}^2$. It has been suggested that the adjusted R^2 does not penalize the loss of degrees of freedom heavily enough.[5] Some alternatives that have been proposed for comparing models (which we index by j) are a modification of the adjusted R squared, that minimizes Amemiya's (1985) **prediction criterion**,

$$PC_j = \frac{\mathbf{e}_j'\mathbf{e}_j}{n - K_j}\left(1 + \frac{K_j}{n}\right) = s_j^2\left(1 + \frac{K_j}{n}\right),$$

$$\overline{R}_j^2 = 1 - \frac{n + K_j}{n - K_j}(1 - R_j^2).$$

Two other fitting criteria are the Akaike and Bayesian information criteria discussed in Section 5.10.1,

$$AIC_j = \ln\left(\frac{\mathbf{e}_j'\mathbf{e}_j}{n}\right) + \frac{2K}{n},$$

$$BIC_j = \ln\left(\frac{\mathbf{e}_j'\mathbf{e}_j}{n}\right) + \frac{K \ln n}{n}.$$

3.5.2 *R*-SQUARED AND THE CONSTANT TERM IN THE MODEL

A second difficulty with R^2 concerns the constant term in the model. The proof that $0 \leq R^2 \leq 1$ requires $\mathbf{X}$ to contain a column of 1s. If not, then (1) $\mathbf{M}^0\mathbf{e} \neq \mathbf{e}$ and (2) $\mathbf{e}'\mathbf{M}^0\mathbf{X} \neq \mathbf{0}$, and the term $2\mathbf{e}'\mathbf{M}^0\mathbf{X}\mathbf{b}$ in $\mathbf{y}'\mathbf{M}^0\mathbf{y} = (\mathbf{M}^0\mathbf{X}\mathbf{b} + \mathbf{M}^0\mathbf{e})'(\mathbf{M}^0\mathbf{X}\mathbf{b} + \mathbf{M}^0\mathbf{e})$ in the expansion preceding (3-25) will not drop out. Consequently, when we compute

$$R^2 = 1 - \frac{\Sigma_{i=1}^n e_i^2}{\Sigma_{i=1}^n (y_i - \overline{y})^2},$$

the result is unpredictable. It will never be higher and can be far lower than the same figure computed for the regression with a constant term included. It can even be negative. Computer packages differ in their computation of R^2. An alternative computation,

$$R^2 = \frac{\Sigma_{i=1}^n (\hat{y}_i - \hat{\overline{y}})^2}{\Sigma_{i=1}^n (y_i - \overline{y})^2},$$

is equally problematic. Again, this calculation will differ from the one obtained with the constant term included; this time, R^2 may be larger than 1. Some computer packages bypass these difficulties by reporting a third "R^2," the squared sample correlation between the actual values of y and the fitted values from the regression. If the regression contains a constant term, then all three computations give the same answer. Even if not, this last one will always produce a value between zero and one. But it is not a proportion of variation explained. On the other hand, for the purpose of comparing models, this squared correlation might well be a useful descriptive device. It is important for users of computer packages to be aware of how the reported R^2 is computed.

[5] See, for example, Amemiya (1985, pp. 50–51).

3.5.3 COMPARING MODELS

The value of R^2 of 0.94450 that we obtained for the consumption function in Example 3.2 seems high in an absolute sense. Is it? Unfortunately, there is no absolute basis for comparison. In fact, in using aggregate time-series data, coefficients of determination this high are routine. In terms of the values one normally encounters in cross sections, an R^2 of 0.5 is relatively high. Coefficients of determination in cross sections of individual data as high as 0.2 are sometimes noteworthy. The point of this discussion is that whether a regression line provides a good fit to a body of data depends on the setting.

Little can be said about the relative quality of fits of regression lines in different contexts or in different data sets even if they are supposedly generated by the same data-generating mechanism. One must be careful, however, even in a single context, to be sure to use the same basis for comparison for competing models. Usually, this concern is about how the dependent variable is computed. For example, a perennial question concerns whether a linear or loglinear model fits the data better. Unfortunately, the question cannot be answered with a direct comparison. An R^2 for the linear regression model is different from an R^2 for the loglinear model. Variation in y is different from variation in ln y. The latter R^2 will typically be larger, but this does not imply that the loglinear model is a better fit in some absolute sense.

It is worth emphasizing that R^2 is a measure of *linear* association between x and y. For example, the third panel of Figure 3.3 shows data that might arise from the model

$$y_i = \alpha + \beta x_i + \gamma x_i^2 + \varepsilon_i.$$

The relationship between y and x in this model is nonlinear, and a linear regression of y on x would find no fit.

3.6 LINEARLY TRANSFORMED REGRESSION

As a final application of the tools developed in this chapter, we examine a purely algebraic result that is very useful for understanding the computation of linear regression models. In the regression of $\mathbf{y}$ on $\mathbf{X}$, suppose the columns of $\mathbf{X}$ are linearly transformed. Common applications would include changes in the units of measurement, say by changing units of currency, hours to minutes, or distances in miles to kilometers. Example 3.4 suggests a slightly more involved case. This is a useful practical, algebraic result. For example, it simplifies the analysis in the first application suggested, changing the units of measurement. If an independent variable is scaled by a constant, p, the regression coefficient will be scaled by $1/p$. There is no need to recompute the regression.

Example 3.4 Art Appreciation

Theory 1 of the determination of the auction prices of Monet paintings holds that the price is determined by the dimensions (width, *W,* and height, *H*) of the painting,

$$\begin{aligned}\ln \mathit{Price} &= \beta_1(1) + \beta_2 \ln W + \beta_3 \ln H + \varepsilon\\ &= \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon.\end{aligned}$$

Theory 2 claims, instead, that art buyers are interested specifically in surface area and aspect ratio,

$$\begin{aligned}\ln \mathit{Price} &= \gamma_1(1) + \gamma_2 \ln (WH) + \gamma_3 \ln (W/H) + \varepsilon\\ &= \gamma_1 z_1 + \gamma_2 z_2 + \gamma_3 z_3 + u.\end{aligned}$$

It is evident that $z_1 = x_1$, $z_2 = x_2 + x_3$, and $z_3 = x_2 - x_3$. In matrix terms, $\mathbf{Z} = \mathbf{XP}$ where

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & -1 \end{bmatrix}, \mathbf{P}^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 1/2 \\ 0 & 1/2 & -1/2 \end{bmatrix}.$$

The effect of a transformation on the linear regression of **y** on **X** compared to that of **y** on **Z** is given by Theorem 3.8. Thus, $\beta_1 = \gamma_1$, $\beta_2 = 1/2(\gamma_2 + \gamma_3)$, $\beta_3 = 1/2(\gamma_2 - \gamma_3)$.

THEOREM 3.8 Transformed Variables
In the linear regression of **y** *on* $\mathbf{Z} = \mathbf{XP}$ *where* **P** *is a nonsingular matrix that transforms the columns of* **X**, *the coefficients will equal* $\mathbf{P}^{-1}\mathbf{b}$ *where* **b** *is the vector of coefficients in the linear regression of* **y** *on* **X**, *and the* R^2 *will be identical.*
Proof: *The coefficients are*

$$\begin{aligned}\mathbf{d} &= (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} = [(\mathbf{XP})'(\mathbf{XP})]^{-1}(\mathbf{XP})'\mathbf{y} = (\mathbf{P}'\mathbf{X}'\mathbf{XP})^{-1}\mathbf{P}'\mathbf{X}'\mathbf{y} \\ &= \mathbf{P}^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{P}'^{-1}\mathbf{P}'\mathbf{X}'\mathbf{y} = \mathbf{P}^{-1}\mathbf{b}.\end{aligned}$$

The vector of residuals is $\mathbf{u} = \mathbf{y} - \mathbf{Z}(\mathbf{P}^{-1}\mathbf{b}) = \mathbf{y} - \mathbf{XPP}^{-1}\mathbf{b} = \mathbf{y} - \mathbf{Xb} = \mathbf{e}$. *Since the residuals are identical, the numerator of* $1 - R^2$ *is the same, and the denominator is unchanged. This establishes the result.*

3.7 SUMMARY AND CONCLUSIONS

This chapter has described the exercise of fitting a line (hyperplane) to a set of points using the method of least squares. We considered the primary problem first, using a data set of n observations on K variables. We then examined several aspects of the solution, including the nature of the projection and residual maker matrices and several useful algebraic results relating to the computation of the residuals and their sum of squares. We also examined the difference between gross or simple regression and correlation and multiple regression by defining partial regression coefficients and partial correlation coefficients. The Frisch–Waugh–Lovell Theorem (3.2) is a fundamentally useful tool in regression analysis that enables us to obtain the expression for a subvector of a vector of regression coefficients. We examined several aspects of the partitioned regression, including how the fit of the regression model changes when variables are added to it or removed from it. Finally, we took a closer look at the conventional measure of how well the fitted regression line predicts or "fits" the data.

Key Terms and Concepts

- Adjusted R^2
- Analysis of variance
- Bivariate regression
- Coefficient of determination
- Degrees of freedom
- Disturbance
- Fitting criterion
- Frisch–Waugh theorem
- Goodness of fit
- Least squares
- Least squares normal equations
- Moment matrix
- Multiple correlation
- Multiple regression
- Netting out
- Normal equations
- Orthogonal regression
- Partial correlation coefficient

- Partial regression coefficient
- Partialing out
- Partitioned regression
- Prediction criterion
- Population quantity
- Population regression
- Projection
- Projection matrix
- Residual
- Residual maker
- Total variation

Exercises

1. *The two-variable regression.* For the regression model $y = \alpha + \beta x + \varepsilon$,
   a. Show that the least squares normal equations imply $\Sigma_i e_i = 0$ and $\Sigma_i x_i e_i = 0$.
   b. Show that the solution for the constant term is $a = \bar{y} - b\bar{x}$.
   c. Show that the solution for b is $b = \left[\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right]/\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]$.
   d. Prove that these two values uniquely minimize the sum of squares by showing that the diagonal elements of the second derivatives matrix of the sum of squares with respect to the parameters are both positive and that the determinant is $4n\left[\left(\sum_{i=1}^{n}x_i^2\right) - n\bar{x}^2\right] = 4n\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]$, which is positive unless all values of x are the same.
2. *Change in the sum of squares.* Suppose that $\mathbf{b}$ is the least squares coefficient vector in the regression of $\mathbf{y}$ on $\mathbf{X}$ and that $\mathbf{c}$ is any other $K\times 1$ vector. Prove that the difference in the two sums of squared residuals is

$$(\mathbf{y} - \mathbf{Xc})'(\mathbf{y} - \mathbf{Xc}) - (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) = (\mathbf{c} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\mathbf{c} - \mathbf{b}).$$

   Prove that this difference is positive.
3. *Partial Frisch and Waugh.* In the least squares regression of $\mathbf{y}$ on a constant and $\mathbf{X}$, to compute the regression coefficients on $\mathbf{X}$, we can first transform $\mathbf{y}$ to deviations from the mean $\bar{y}$ and, likewise, transform each column of $\mathbf{X}$ to deviations from the respective column mean; second, regress the transformed $\mathbf{y}$ on the transformed $\mathbf{X}$ without a constant. Do we get the same result if we only transform $\mathbf{y}$? What if we only transform $\mathbf{X}$?
4. *Residual makers.* What is the result of the matrix product $\mathbf{M}_1\mathbf{M}$ where $\mathbf{M}_1$ is defined in (3-19) and $\mathbf{M}$ is defined in (3-14)?
5. *Adding an observation.* A data set consists of n observations contained in $\mathbf{X}_n$ and $\mathbf{y}_n$. The least squares estimator based on these n observations is $\mathbf{b}_n = (\mathbf{X}_n'\mathbf{X}_n)^{-1}\mathbf{X}_n'\mathbf{y}_n$. Another observation, $\mathbf{x}_s$ and y_s, becomes available. Prove that the least squares estimator computed using this additional observation is

$$\mathbf{b}_{n,s} = \mathbf{b}_n + \frac{1}{1 + \mathbf{x}_s'(\mathbf{X}_n'\mathbf{X}_n)^{-1}\mathbf{x}_s}(\mathbf{X}_n'\mathbf{X}_n)^{-1}\,\mathbf{x}_s\,(y_s - \mathbf{x}_s'\mathbf{b}_n).$$

   Note that the last term is e_s, the residual from the prediction of y_s using the coefficients based on $\mathbf{X}_n$ and $\mathbf{y}_n$. Conclude that the new data change the results of least squares only if the new observation on y cannot be perfectly predicted using the information already in hand.
6. *Deleting an observation.* A common strategy for handling a case in which an observation is missing data for one or more variables is to fill those missing variables with 0s and add a variable to the model that takes the value 1 for that one observation and 0 for all other observations. Show that this strategy is equivalent to discarding the observation as regards the computation of $\mathbf{b}$ but it does have an

effect on R^2. Consider the special case in which $\mathbf{X}$ contains only a constant and one variable. Show that replacing missing values of x with the mean of the complete observations has the same effect as adding the new variable.

7. *Demand system estimation.* Let Y denote total expenditure on consumer durables, nondurables, and services and E_d, E_n, and E_s are the expenditures on the three categories. As defined, $Y = E_d + E_n + E_s$. Now, consider the expenditure system

$$\begin{aligned} E_d &= \alpha_d + \beta_d Y + \gamma_{dd}P_d + \gamma_{dn}P_n + \gamma_{ds}P_s + \varepsilon_d, \\ E_n &= \alpha_n + \beta_n Y + \gamma_{nd}P_d + \gamma_{nn}P_n + \gamma_{ns}P_s + \varepsilon_n, \\ E_s &= \alpha_s + \beta_s Y + \gamma_{sd}P_d + \gamma_{sn}P_n + \gamma_{ss}P_s + \varepsilon_s. \end{aligned}$$

Prove that if all equations are estimated by ordinary least squares, then the sum of the expenditure coefficients will be 1 and the four other column sums in the preceding model will be zero.

8. *Change in adjusted* R^2. Prove that the adjusted R^2 in (3-30) rises (falls) when variable $\mathbf{x}_k$ is deleted from the regression if the square of the t ratio on $\mathbf{x}_k$ in the multiple regression is less (greater) than 1.

9. *Regression without a constant.* Suppose that you estimate a multiple regression first with, then without, a constant. Whether the R^2 is higher in the second case than the first will depend in part on how it is computed. Using the (relatively) standard method $R^2 = 1 - (\mathbf{e}'\mathbf{e}/\mathbf{y}'\mathbf{M}^0\mathbf{y})$, which regression will have a higher R^2?

10. Three variables, N, D, and Y, all have zero means and unit variances. A fourth variable is $C = N + D$. In the regression of C on Y, the slope is 0.8. In the regression of C on N, the slope is 0.5. In the regression of D on Y, the slope is 0.4. What is the sum of squared residuals in the regression of C on D? There are 21 observations and all moments are computed using $1/(n - 1)$ as the divisor.

11. Using the matrices of sums of squares and cross products immediately preceding Section 3.2.3, compute the coefficients in the multiple regression of real investment on a constant, GNP, and the interest rate. Compute R^2.

12. In the December 1969 *American Economic Review* (pp. 886–896), Nathaniel Leff reports the following least squares regression results for a cross section study of the effect of age composition on savings in 74 countries in 1964:

$$\begin{aligned} \ln S/Y &= 7.3439 + 0.1596 \ln Y/N + 0.0254 \ln G - 1.3520 \ln D_1 - 0.3990 \ln D_2, \\ \ln S/N &= 2.7851 + 1.1486 \ln Y/N + 0.0265 \ln G - 1.3438 \ln D_1 - 0.3966 \ln D_2, \end{aligned}$$

where S/Y = domestic savings ratio, S/N = per capita savings, Y/N = per capita income, D_1 = percentage of the population under 15, D_2 = percentage of the population over 64, and G = growth rate of per capita income. Are these results correct? Explain.[6]

13. *Is it possible to partition* R^2? The idea of "hierarchical partitioning" is to decompose R^2 into the contributions made by each variable in the multiple regression. That is, if $x_1, \ldots, x_K$ are entered into a regression one at a time, then c_k is the incremental contribution of x_k such that given the order entered, $\Sigma_k c_k = R^2$ and the incremental

[6] See Goldberger (1973) and Leff (1973) for discussion.

contribution of x_k is then c_k/R^2. Of course, based on (3-31), we know that this is not a useful calculation.

a. Argue based on (3-31) why it is not useful.

b. Show using (3-31) that the computation is sensible if (and only if) all variables are orthogonal.

c. For the investment example in Section 3.2.2, compute the incremental contribution of T if it is entered first in the regression. Now compute the incremental contribution of T if it is entered last.

Application

The data listed in Table 3.5 are extracted from Koop and Tobias's (2004) study of the relationship between wages and education, ability, and family characteristics. (See Appendix Table F3.2.) Their data set is a panel of 2,178 individuals with a total of 17,919 observations. Shown in the table are the first year and the time-invariant variables for the first 15 individuals in the sample. The variables are defined in the article.

Let $\mathbf{X}_1$ equal a constant, education, experience, and ability (the individual's own characteristics). Let $\mathbf{X}_2$ contain the mother's education, the father's education, and the number of siblings (the household characteristics). Let $\mathbf{y}$ be the log of the hourly wage.

a. Compute the least squares regression coefficients in the regression of $\mathbf{y}$ on $\mathbf{X}_1$. Report the coefficients.

b. Compute the least squares regression coefficients in the regression of $\mathbf{y}$ on $\mathbf{X}_1$ and $\mathbf{X}_2$. Report the coefficients.

TABLE 3.5 Subsample from Koop and Tobias Data

Person	*Education*	*ln Wage*	*Experience*	*Ability*	*Mother's Education*	*Father's Education*	*Siblings*
1	13	1.82	1	1.00	12	12	1
2	15	2.14	4	1.50	12	12	1
3	10	1.56	1	−0.36	12	12	1
4	12	1.85	1	0.26	12	10	4
5	15	2.41	2	0.30	12	12	1
6	15	1.83	2	0.44	12	16	2
7	15	1.78	3	0.91	12	12	1
8	13	2.12	4	0.51	12	15	2
9	13	1.95	2	0.86	12	12	2
10	11	2.19	5	0.26	12	12	2
11	12	2.44	1	1.82	16	17	2
12	13	2.41	4	−1.30	13	12	5
13	12	2.07	3	−0.63	12	12	4
14	12	2.20	6	−0.36	10	12	2
15	12	2.12	3	0.28	10	12	3

c. Regress each of the three variables in $\mathbf{X}_2$ on all the variables in $\mathbf{X}_1$ and compute the residuals from each regression. Arrange these new variables in the 15 × 3 matrix $\mathbf{X}_2^*$. What are the sample means of these three variables? Explain the finding.

d. Using (3-26), compute the R^2 for the regression of $\mathbf{y}$ on $\mathbf{X}_1$ and $\mathbf{X}_2$. Repeat the computation for the case in which the constant term is omitted from $\mathbf{X}_1$. What happens to R^2?

e. Compute the adjusted R^2 for the full regression including the constant term. Interpret your result.

f. Referring to the result in part c, regress $\mathbf{y}$ on $\mathbf{X}_1$ and $\mathbf{X}_2^*$. How do your results compare to the results of the regression of $\mathbf{y}$ on $\mathbf{X}_1$ and $\mathbf{X}_2$? The comparison you are making is between the least squares coefficients when $\mathbf{y}$ is regressed on $\mathbf{X}_1$ and $\mathbf{M}_1\mathbf{X}_2$ and when $\mathbf{y}$ is regressed on $\mathbf{X}_1$ and $\mathbf{X}_2$. Derive the result theoretically. (Your numerical results should match the theory, of course.)