3
LEAST SQUARES

3.1 INTRODUCTION

Chapter 2 defined the linear regression model as a set of characteristics of the population that underlies an observed sample of data. There are a number of different approaches to estimation of the parameters of the model. For a variety of practical and theoretical reasons that we will explore as we progress through the next several chapters, the method of least squares has long been the most popular. Moreover, in most cases in which some other estimation method is found to be preferable, least squares remains the benchmark approach, and often, the preferred method ultimately amounts to a modification of least squares. In this chapter, we begin the analysis of this important set of results by presenting a useful set of algebraic tools.

3.2 LEAST SQUARES REGRESSION

The unknown parameters of the stochastic relation $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$ are the objects of estimation. It is necessary to distinguish between population quantities, such as $\boldsymbol{\beta}$ and ε_i, and sample estimates of them, denoted $\mathbf{b}$ and e_i. The **population regression** is $E[y_i \mid \mathbf{x}_i] = \mathbf{x}_i'\boldsymbol{\beta}$, whereas our estimate of $E[y_i \mid \mathbf{x}_i]$ 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**

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

From the definitions,

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

These equations are summarized for the two variable regression in Figure 3.1.

The **population quantity** $\boldsymbol{\beta}$ is a vector of unknown parameters of the probability distribution of y_i whose values we hope to estimate with our sample data, $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$. This is a problem of statistical inference. It is instructive, however, to begin by considering the purely 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.**

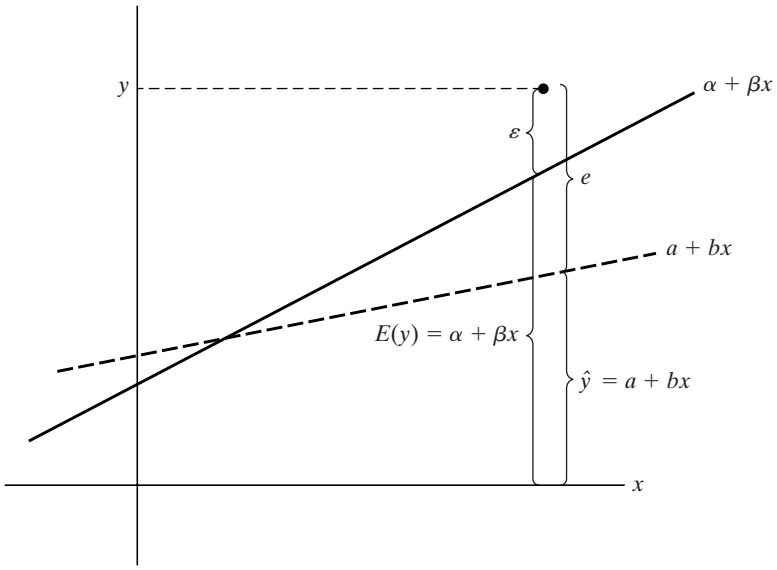


FIGURE 3.1 Population and Sample Regression.

Although numerous candidates have been suggested, 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 the choice for the coefficient vector. In matrix terms, minimizing the sum of squares in (3-1) requires us to choose $\mathbf{b}_0$ to

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

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

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

[1] We shall have to establish that the practical approach of fitting the line as closely as possible to the data by least squares leads to estimates with good statistical properties. This makes intuitive sense and is, indeed, the case. We shall return to the statistical issues in Chapters 4 and 5.

Let **b** be the solution. Then **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 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,

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

must be a positive definite matrix. Let $q = \mathbf{c}'\mathbf{X}'\mathbf{X}\mathbf{c}$ for some arbitrary nonzero vector **c**. 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 the assumption that **X** has full rank. Since **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 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 Data Table F3.1. To estimate an investment equation, we first convert the investment and GNP series in Table F3.1 to real terms by dividing them by the CPI, and then scale the two series so that they are measured in trillions of dollars. The other variables in the regression are a time trend $(1, 2, \ldots)$, an interest rate, and the rate of inflation computed as the percentage change in the CPI. 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 GNP, which correspond to x_1, x_2, and x_3. (For reasons to be discussed in Chapter 20, 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, we have

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

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

$$\begin{aligned}
b_1 &= \bar{Y} - b_2 \bar{T} - b_3 \bar{G} \\
&= 0.20333 - b_2 \times 8 - b_3 \times 1.2873.
\end{aligned} \tag{3-7}$$

TABLE 3.1 Data Matrices

Real Investment (Y)	*Constant (1)*	*Trend (T)*	*Real GNP (G)*	*Interest Rate (R)*	*Inflation Rate (P)*
0.161	1	1	1.058	5.16	4.40
0.172	1	2	1.088	5.87	5.15
0.158	1	3	1.086	5.95	5.37
0.173	1	4	1.122	4.88	4.99
0.195	1	5	1.186	4.50	4.16
0.217	1	6	1.254	6.44	5.75
0.199	1	7	1.246	7.83	8.82
$\mathbf{y} =$ 0.163	$\mathbf{X} =$ 1	8	1.232	6.25	9.31
0.195	1	9	1.298	5.50	5.21
0.231	1	10	1.370	5.46	5.83
0.257	1	11	1.439	7.46	7.40
0.259	1	12	1.479	10.28	8.64
0.225	1	13	1.474	11.77	9.31
0.241	1	14	1.503	13.42	9.44
0.204	1	15	1.475	11.02	5.99

Note: Subsequent results are based on these values. Slightly different results are obtained if the raw data in Table F3.1 are input to the computer program and transformed internally.

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

$$\begin{aligned} b_2\Sigma_i(T_i-\bar{T})^2 &+ b_3\Sigma_i(T_i-\bar{T})(G_i-\bar{G}) = \Sigma_i(T_i-\bar{T})(Y_i-\bar{Y}),\\ b_2\Sigma_i(T_i-\bar{T})(G_i-\bar{G}) &+ b_3\Sigma_i(G_i-\bar{G})^2 = \Sigma_i(G_i-\bar{G})(Y_i-\bar{Y}). \end{aligned} \quad \textbf{(3-8)}$$

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. Letting lowercase letters indicate variables measured as deviations from the sample means, we find that the least squares solutions for b_2 and b_3 are

$$b_2 = \frac{\Sigma_i t_i y_i \Sigma_i g_i^2 - \Sigma_i g_i y_i \Sigma_i t_i g_i}{\Sigma_i t_i^2 \Sigma_i g_i^2 - (\Sigma_i g_i t_i)^2} = \frac{1.6040(0.359609) - 0.066196(9.82)}{280(0.359609) - (9.82)^2} = -0.0171984,$$

$$b_3 = \frac{\Sigma_i g_i y_i \Sigma_i t_i^2 - \Sigma_i t_i y_i \Sigma_i t_i g_i}{\Sigma_i t_i^2 \Sigma_i g_i^2 - (\Sigma_i g_i t_i)^2} = \frac{0.066196(280) - 1.6040(9.82)}{280(0.359609) - (9.82)^2} = 0.653723.$$

With these solutions in hand, the intercept can now be computed using (3-7); $b_1 = -0.500639$.

Suppose that we just regressed investment on the constant and GNP, omitting the time trend. At least some of the correlation we observe in the data will be explainable because both investment and real GNP have an obvious time trend. 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

$$b_{yg} = \frac{\Sigma_i g_i y_i}{\Sigma_i g_i^2} = 0.184078. \quad \textbf{(3-9)}$$

Now divide both the numerator and denominator in the expression for b_3 by $\Sigma_i t_i^2 \Sigma_i g_i^2$. By manipulating it a bit and using the definition of the sample correlation between G and T, $r_{gt}^2 = (\Sigma_i g_i t_i)^2/(\Sigma_i g_i^2 \Sigma_i t_i^2)$, and defining b_{yt} and b_{tg} likewise, we obtain

$$b_{yg\cdot t} = \frac{b_{yg}}{1 - r_{gt}^2} - \frac{b_{yt} b_{tg}}{1 - r_{gt}^2} = 0.653723. \tag{3-10}$$

(The notation "$b_{yg\cdot t}$" used on the left-hand side is interpreted to mean the slope in the regression of y on g "in the presence of t.") The slope in the **multiple regression** differs from that in the simple regression 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, $b_{yt} = 1.604/280 = 0.0057286$, a positive number that reflects the upward trend apparent in the data. But, in the multiple regression, after we account for the influence of GNP on real investment, the slope on the time trend is -0.0171984, indicating instead a downward trend. The general result for a three-variable regression in which x_1 is a constant term is

$$b_{y2\cdot 3} = \frac{b_{y2} - b_{y3} b_{32}}{1 - r_{23}^2}. \tag{3-11}$$

It is clear from this expression that the magnitudes of $b_{y2\cdot 3}$ and b_{y2} can be quite different. They need not even have the same sign.

As a final observation, note what becomes of $b_{yg\cdot t}$ in (3-10) if r_{gt}^2 equals zero. The first term becomes b_{yg}, whereas the second becomes zero. (If G and T are not correlated, then the slope in the regression of G on T, b_{tg}, is zero.) Therefore, we conclude the following.

THEOREM 3.1 Orthogonal Regression
If the variables in a multiple regression are not correlated (i.e., are orthogonal), then the multiple regression slopes are the same as the slopes in the individual simple regressions.

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 GNP—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

$$\begin{bmatrix} 15.000 & 120.00 & 19.310 & 111.79 & 99.770 \\ 120.000 & 1240.0 & 164.30 & 1035.9 & 875.60 \\ 19.310 & 164.30 & 25.218 & 148.98 & 131.22 \\ 111.79 & 1035.9 & 148.98 & 953.86 & 799.02 \\ 99.770 & 875.60 & 131.22 & 799.02 & 716.67 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix} = \begin{bmatrix} 3.0500 \\ 26.004 \\ 3.9926 \\ 23.521 \\ 20.732 \end{bmatrix}.$$

The solution is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (-0.50907, -0.01658, 0.67038, -0.002326, -0.00009401)'.$$

3.2.3 ALGEBRAIC ASPECTS OF THE LEAST SQUARES SOLUTION

The normal equations are

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

Hence, for every column $\mathbf{x}_k$ of $\mathbf{X}$, $\mathbf{x}_k'\mathbf{e} = 0$. If the first column of $\mathbf{X}$ is a column of 1s, 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 $\bar{y} = \bar{\mathbf{x}}'\mathbf{b}$.
3. *The mean of the fitted values from the regression equals the mean of the actual values*. This implication follows from point 1 because the fitted values are just $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$.

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

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b}. \tag{3-13}$$

Inserting the result in (3-6) for $\mathbf{b}$ gives

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

The $n \times n$ matrix $\mathbf{M}$ defined in (3-14) is fundamental in regression analysis. You can easily show that $\mathbf{M}$ is both symmetric ($\mathbf{M} = \mathbf{M}'$) and idempotent ($\mathbf{M} = \mathbf{M}^2$). In view of (3-13), we can interpret $\mathbf{M}$ 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 on to refer to this matrix as a "**residual maker.**") It follows that

$$\mathbf{M}\mathbf{X} = \mathbf{0}. \tag{3-15}$$

One way to interpret this result is that if $\mathbf{X}$ is regressed on $\mathbf{X}$, a perfect fit will result and the residuals will be zero.

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

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

The matrix $\mathbf{P}$, which is also symmetric and idempotent, 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 **y** into the column space of **X**. (See Sections A3.5 and A3.7.) By multiplying it out, you will find that, like **M**, **P** is symmetric and idempotent. Given the earlier results, it also follows that **M** and **P** are orthogonal;

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

Finally, as might be expected from (3-15)

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

As a consequence of (3-15) and (3-16), we can see that least squares partitions the vector **y** into two orthogonal parts,

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

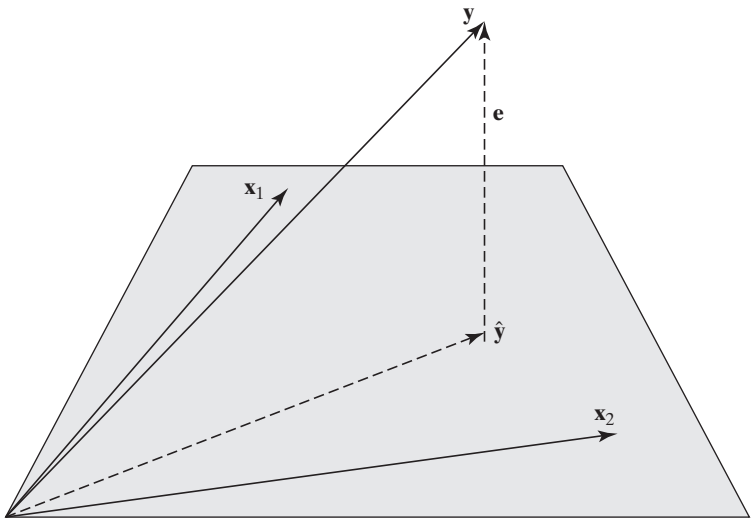
The result is illustrated in Figure 3.2 for the two variable case. The gray shaded plane is the column space of **X**. The projection and residual are the orthogonal dotted rays. We can also see the Pythagorean theorem at work in the sums of squares,

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

In manipulating equations involving least squares results, the following equivalent expressions for the sum of squared residuals are often useful:

$$\mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{M}'\mathbf{M}\mathbf{y} = \mathbf{y}'\mathbf{M}\mathbf{y} = \mathbf{y}'\mathbf{e} = \mathbf{e}'\mathbf{y},$$
$$\mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y} = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\mathbf{b}.$$

FIGURE 3.2 Projection of **y** into the column space of **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. Consider the earnings equation discussed in Example 2.2. Although we are primarily interested in the association of earnings and education, 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 (for example, 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 $\mathbf{b}_1$:

$$\mathbf{b}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} - (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\mathbf{b}_2 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{y} - \mathbf{X}_2\mathbf{b}_2). \tag{3-18}$$

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, which we have just proved is the following theorem.

THEOREM 3.2 Orthogonal Partitioned Regression
In the multiple linear least squares 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.*

Note that Theorem 3.2 encompasses Theorem 3.1.

Now, inserting (3-18) in equation (2) of (3-17) 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

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

The matrix appearing in the parentheses inside each set of square brackets is the "residual maker" defined in (3-14), 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 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

$$\mathbf{X}_2^* = \mathbf{M}_1\mathbf{X}_2 \quad \text{and} \quad \mathbf{y}^* = \mathbf{M}_1\mathbf{y}.$$

This result is fundamental in regression analysis.

THEOREM 3.3 Frisch–Waugh Theorem
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* $\mathbf{X}_1$*.*

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 this theorem 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.

COROLLARY 3.3.1 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* $c = (\mathbf{z}'\mathbf{M}\mathbf{z})^{-1}(\mathbf{z}'\mathbf{M}\mathbf{y}) = (\mathbf{z}^{*\prime}\mathbf{z}^*)^{-1}\mathbf{z}^{*\prime}\mathbf{y}^*$ *where* $\mathbf{z}^*$ *and* $\mathbf{y}^*$ *are the residual vectors from least squares regressions of* $\mathbf{z}$ *and* $\mathbf{y}$ *on* $\mathbf{X}$*;* $\mathbf{z}^* = \mathbf{M}\mathbf{z}$ *and* $\mathbf{y}^* = \mathbf{M}\mathbf{y}$ *where* $\mathbf{M}$ *is defined in (3-14).*

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 a classic application of this latter observation, Frisch and Waugh (1933) (who are credited with the result) 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.[2]

[2]Recall our earlier investment example.

As an application of these results, consider the case in which $\mathbf{X}_1$ is **i**, a column of 1s in the first column of $\mathbf{X}$. The solution for $\mathbf{b}_2$ in this case will then be the slopes in a regression with a constant term. The coefficient in a regression of any variable **z** on **i** is $[\mathbf{i}'\mathbf{i}]^{-1}\mathbf{i}'\mathbf{z} = \bar{z}$, the fitted values are $\mathbf{i}\bar{z}$, and the residuals are $z_i - \bar{z}$. When we apply this to our previous results, we find the following.

COROLLARY 3.3.2 Regression with a Constant Term
The slopes in a multiple regression that contains a constant term are 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

$$b_1 = \bar{y} - \bar{x}_2 b_2 - \cdots - \bar{x}_K b_K \qquad \textbf{(3-21)}$$

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

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 Example 2.2, a regression equation relating earnings to age and education enables us to do the conceptual 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 formidable amount of computation. There is, however, a convenient shortcut. Once the multiple regression is computed, the t ratio in (4-13) and (4-14) for testing the hypothesis that the coefficient equals zero (e.g., the last column of Table 4.2) can be used to compute

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

The proof of this less than perfectly intuitive result will be useful to illustrate some results on partitioned regression and to put into context two very useful results from least squares algebra. As in Corollary 3.3.1, let $\mathbf{W}$ denote the $n \times (K + 1)$ regressor matrix $[\mathbf{X}, \mathbf{z}]$ and let $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. We assume that there is a constant term in $\mathbf{X}$, so that the vectors of residuals $\mathbf{y}_* = \mathbf{M}\mathbf{y}$ and $\mathbf{z}_* = \mathbf{M}\mathbf{z}$ will have zero sample means. The squared partial correlation is

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

Let c and $\mathbf{u}$ denote the coefficient on $\mathbf{z}$ and the vector of residuals in the multiple regression of $\mathbf{y}$ on $\mathbf{W}$. The squared t ratio 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 partitioned inverse formula in (A-74) can be applied to the matrix $[\mathbf{X}, \mathbf{z}]'[\mathbf{X}, \mathbf{z}]$. This matrix appears in (3-17), with $\mathbf{X}_1 = \mathbf{X}$ and $\mathbf{X}_2 = \mathbf{z}$. The result is the inverse matrix that appears in (3-19) and (3-20), which implies the first important result.

THEOREM 3.4 Diagonal Elements of the Inverse of a Moment Matrix

If $\mathbf{W} = [\mathbf{X}, \mathbf{z}]$, *then the last diagonal element of* $(\mathbf{W}'\mathbf{W})^{-1}$ *is* $(\mathbf{z}'\mathbf{M}\mathbf{z})^{-1} = (\mathbf{z}'_*\mathbf{z}_*)^{-1}$, *where* $\mathbf{z}_* = \mathbf{M}\mathbf{z}$ *and* $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

(Note that this result generalizes the development in Section A.2.8 where $\mathbf{X}$ is only the constant term.) If we now use Corollary 3.3.1 and Theorem 3.4 for c, after some manipulation, we obtain

$$\frac{t_z^2}{t_z^2 + [n - (K+1)]} = \frac{(\mathbf{z}'_*\mathbf{y}_*)^2}{(\mathbf{z}'_*\mathbf{y}_*)^2 + (\mathbf{u}'\mathbf{u})(\mathbf{z}'_*\mathbf{z}_*)} = \frac{r^{*2}_{yz}}{r^{*2}_{yz} + (\mathbf{u}'\mathbf{u})/(\mathbf{y}'_*\mathbf{y}_*)},$$

where

$$\mathbf{u} = \mathbf{y} - \mathbf{X}\mathbf{d} - \mathbf{z}c$$

is the vector of residuals when $\mathbf{y}$ is regressed on $\mathbf{X}$ *and* $\mathbf{z}$. 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 8.2.1.) Moreover, unless $c = 0$, $\mathbf{u}$ will not equal $\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b}$. Now we have shown in Corollary 3.3.1 that $c = (\mathbf{z}_*'\mathbf{z}_*)^{-1}(\mathbf{z}_*'\mathbf{y}_*)$. We also have, from (3-18), that the coefficients on $\mathbf{X}$ in the regression of $\mathbf{y}$ on $\mathbf{W} = [\mathbf{X}, \mathbf{z}]$ are

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

So, inserting this expression for $\mathbf{d}$ in that for $\mathbf{u}$ gives

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

Now

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

But $\mathbf{e} = \mathbf{M}\mathbf{y} = \mathbf{y}_*$ and $\mathbf{z}_*'\mathbf{e} = \mathbf{z}_*'\mathbf{y}_* = c(\mathbf{z}_*'\mathbf{z}_*)$. Inserting this in $\mathbf{u}'\mathbf{u}$ gives our second useful result.

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*

$$\mathbf{u}'\mathbf{u} = \mathbf{e}'\mathbf{e} - c^2(\mathbf{z}_*'\mathbf{z}_*) \leq \mathbf{e}'\mathbf{e}, \qquad \textbf{(3-23)}$$

where c *is the coefficient on* $\mathbf{z}$ *in the long regression and* $\mathbf{z}_* = [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{z}$ *is the vector of residuals when* $\mathbf{z}$ *is regressed on* $\mathbf{X}$.

Returning to our derivation, we note that $\mathbf{e}'\mathbf{e} = \mathbf{y}_*'\mathbf{y}_*$ and $c^2(\mathbf{z}_*'\mathbf{z}_*) = (\mathbf{z}_*'\mathbf{y}_*)^2/(\mathbf{z}_*'\mathbf{z}_*)$. Therefore, $(\mathbf{u}'\mathbf{u})/(\mathbf{y}_*'\mathbf{y}_*) = 1 - r_{yz}^{*2}$, and we have our result.

Example 3.1 Partial Correlations

For the data 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

TABLE 3.2 Correlations of Investment with Other Variables

	Simple Correlation	*Partial Correlation*
Time	0.7496	−0.9360
GNP	0.8632	0.9680
Interest	0.5871	−0.5167
Inflation	0.4777	−0.0221

noting is the signs of the coefficients. The signs of the partial correlation coefficients are the same as the signs of the respective regression coefficients, three of which are negative. All the simple correlation coefficients are positive because of the latent "effect" of time.

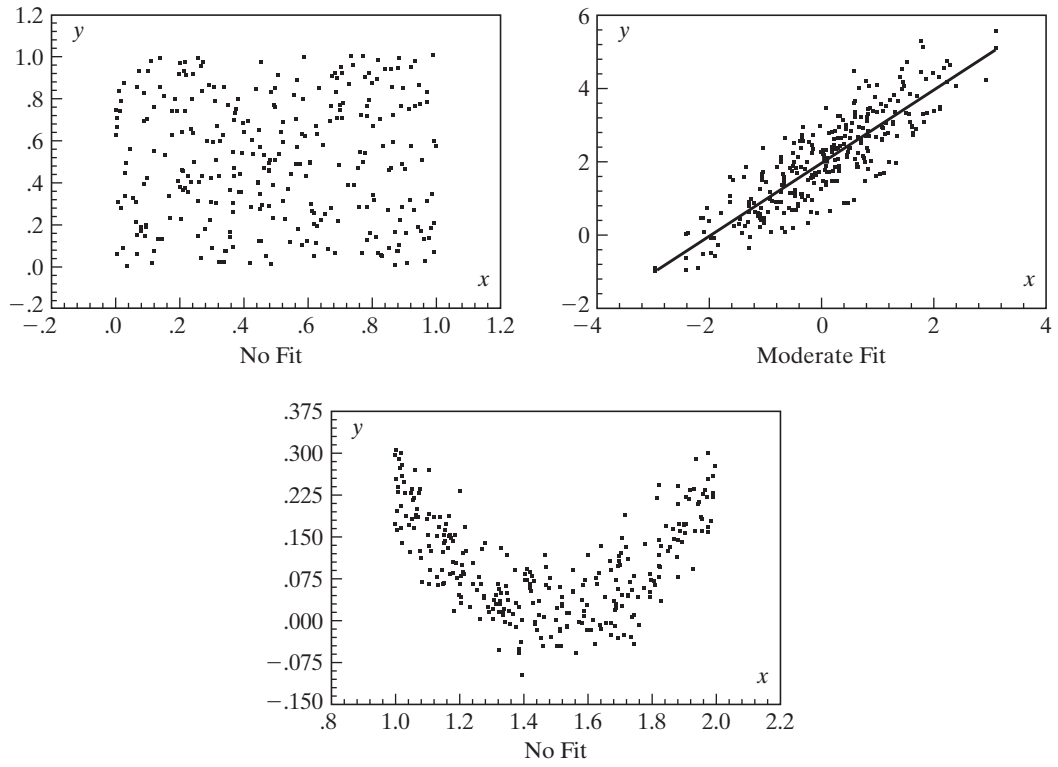
3.5 GOODNESS OF FIT AND THE ANALYSIS OF VARIANCE

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. Since 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. 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:

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

FIGURE 3.3 Sample Data.



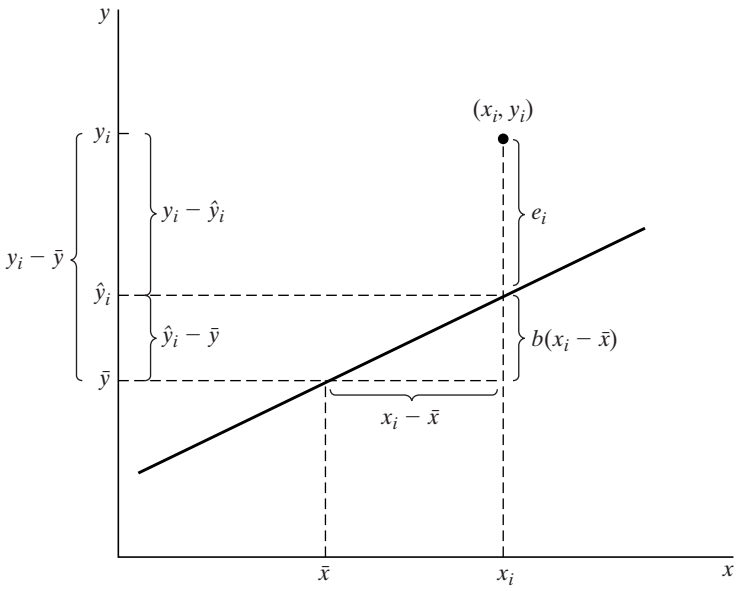


FIGURE 3.4 Decomposition of y_i.

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

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e} = \hat{\mathbf{y}} + \mathbf{e}. \tag{3-24}$$

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 $\bar{y}$ from both sides and using this result and result 2 in Section 3.2.3 gives

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

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

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

where $\mathbf{M}^0$ is the $n \times n$ idempotent matrix that transforms observations into deviations from sample means. (See Section A.2.8.) The column of $\mathbf{M}^0\mathbf{X}$ corresponding to the constant term is zero, and, since the residuals 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

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

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 precisely the 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

$$\textbf{coefficient of determination:}\ \frac{\text{SSR}}{\text{SST}} = \frac{\mathbf{b}'\mathbf{X}'\mathbf{M}^0\mathbf{X}\mathbf{b}}{\mathbf{y}'\mathbf{M}^0\mathbf{y}} = 1 - \frac{\mathbf{e}'\mathbf{e}}{\mathbf{y}'\mathbf{M}^0\mathbf{y}}. \tag{3-26}$$

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

$$\mathbf{b}'\mathbf{X}'\mathbf{M}^0\mathbf{X}\mathbf{b} = \hat{\mathbf{y}}'\mathbf{M}^0\hat{\mathbf{y}},$$

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 $\hat{\mathbf{y}}'\mathbf{M}^0\hat{\mathbf{y}} = \hat{\mathbf{y}}'\mathbf{M}^0\mathbf{y}$. 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 $\bar{y} = 273.2727$, $\bar{x} = 323.2727$, $S_{yy} = 12{,}618.182$, $S_{xx} = 12{,}300.182$, $S_{xy} = 8{,}423.182$, so SST $= 12{,}618.182$, $b = 8{,}423.182/12{,}300.182 = 0.6848014$, SSR $= b^2 S_{xx} = 5{,}768.2068$, and SSE $=$ SST$-$SSR $= 6{,}849.975$. Then $R^2 = b^2 S_{xx}/\text{SST} = 0.457135$. 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 dotted line in the figure. By simply omitting the years 1942–1945 from the sample and doing these computations with the remaining seven observations—the heavy solid line—we obtain an R^2 of 0.93697. 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 solid lines, the R^2 rises to 0.94639.

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 an Investment Equation

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

TABLE 3.3 Analysis of Variance

Source	Degrees of Freedom	Mean Square
Regression $\mathbf{b}'\mathbf{X}'\mathbf{y} - n\bar{y}^2$	$K-1$ (assuming a constant term)	
Residual $\mathbf{e}'\mathbf{e}$	$n-K$	s^2
Total $\mathbf{y}'\mathbf{y} - n\bar{y}^2$	$n-1$	$S_{yy}/(n-1) = s_y^2$
Coefficient of determination	$R^2 = 1 - \mathbf{e}'\mathbf{e}/(\mathbf{y}'\mathbf{y} - n\bar{y}^2)$	

TABLE 3.4 Analysis of Variance for the Investment Equation

Source	Degrees of Freedom	Mean Square
Regression 0.0159025	4	0.003976
Residual 0.0004508	10	0.00004508
Total 0.016353	14	0.0011681

$R^2 = 0.0159025/0.016353 = 0.97245.$

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. *R^2 will never decrease when another variable is added to a regression equation.* Equation (3-23) 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 $\mathbf{z}_* = \mathbf{M}\mathbf{z}$ and $\mathbf{y}_* = \mathbf{M}\mathbf{y} = \mathbf{e}$, 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 (3-23) produces

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

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

$$R_{\mathbf{X}z}^2 = R_{\mathbf{X}}^2 + \left(1 - R_{\mathbf{X}}^2\right) r_{yz}^{*2}. \tag{3-29}$$

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.[3] The **adjusted** R^2 (for degrees of freedom), which incorporates a penalty for these results is computed as follows:

$$\bar{R}^2 = 1 - \frac{\mathbf{e}'\mathbf{e}/(n-K)}{\mathbf{y}'\mathbf{M}^0\mathbf{y}/(n-1)}.^4 \tag{3-30}$$

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

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

The adjusted R^2 may decline when a variable is added to the set of independent variables. Indeed, $\bar{R}^2$ may even be negative. To consider an admittedly extreme case, suppose that $\mathbf{x}$ and $\mathbf{y}$ have a sample correlation of zero. Then the adjusted R^2 will equal $-1/(n-2)$. (Thus, the name "adjusted *R*-squared" is a bit misleading—as can be seen in (3-30), $\bar{R}^2$ is not actually computed as the square of any quantity.) Whether $\bar{R}^2$ rises or falls 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 $\bar{R}^2$ When a Variable Is Added to a Regression

In a multiple regression, $\bar{R}^2$ will fall (rise) when the variable x is deleted from the regression if 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

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

where we use subscript 1 to indicate the regression based on $\mathbf{X}_1$ alone and 1,2 to indicate the use of *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 $\mathbf{y}$ 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,

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

[3]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.

[4]This measure is sometimes advocated on the basis of the unbiasedness of the two quantities in the fraction. Since the ratio is not an unbiased estimator of any population quantity, it is difficult to justify the adjustment on this basis.

which is greater than R_1^2 unless $\mathbf{b}_2$ equals zero. ($\mathbf{M}_1\mathbf{X}_2$ could not be zero unless $\mathbf{X}_2$ was 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 $\mathbf{y}$ on $\mathbf{X}_2$. Collecting terms, we have

$$R_{1,2}^2 = R_1^2 + (1 - R_1^2) r_{y2\cdot 1}^2.$$

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

Therefore, it is possible to push R^2 as high as desired just by adding regressors. This possibility motivates the use of the adjusted R-squared in (3-30), instead of R^2 as a method of choosing among alternative models. Since $\bar{R}^2$ 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 $\bar{R}^2$. It has been suggested that the adjusted R-squared 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

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

which minimizes Amemiya's (1985) **prediction criterion,**

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

and the Akaike and Bayesian information criteria which are given in (8-18) and (8-19).

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 preceding expansion will not drop out. Consequently, when we compute

$$R^2 = 1 - \frac{\mathbf{e}'\mathbf{e}}{\mathbf{y}'\mathbf{M}^0\mathbf{y}},$$

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,

$$R^2 = \frac{\mathbf{b}'\mathbf{X}'\mathbf{y}}{\mathbf{y}'\mathbf{M}^0\mathbf{y}},$$

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

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

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. This approach could be deceptive. If the regression contains a constant term, then, as we have seen, all three computations give the same answer. Even if not, this last one will still 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. Indeed, some packages will give a warning in the results when a regression is fit without a constant or by some technique other than linear least squares.

3.5.3 COMPARING MODELS

The value of R^2 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 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

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

(The constant γ allows x to be distributed about some value other than zero.) The relationship between y and x in this model is nonlinear, and a linear regression would find no fit.

A final word of caution is in order. The interpretation of R^2 as a proportion of variation explained is dependent on the use of least squares to compute the fitted values. It is always correct to write

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + e_i$$

regardless of how $\hat{y}_i$ is computed. Thus, one might use $\hat{y}_i = \exp(\widehat{\ln y}_i)$ from a loglinear model in computing the sum of squares on the two sides, however, the cross-product term vanishes only if least squares is used to compute the fitted values and if the model contains a constant term. Thus, ~~in the suggested example, it would still be unclear whether the linear or loglinear model fits better;~~ the cross-product term has been ignored

in computing R^2 for the loglinear model. Only in the case of least squares applied to a linear equation with a constant term can R^2 be interpreted as the proportion of variation in y explained by variation in $\mathbf{x}$. An analogous computation can be done without computing deviations from means if the regression does not contain a constant term. Other purely algebraic artifacts will crop up in regressions without a constant, however. For example, the value of R^2 will change when the same constant is added to each observation on y, but it is obvious that nothing fundamental has changed in the regression relationship. One should be wary (even skeptical) in the calculation and interpretation of fit measures for regressions without constant terms.

3.6 SUMMARY AND CONCLUSIONS

This chapter has described the purely algebraic 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 Theorem (3.3) is a fundamentally useful tool in regression analysis which enables us to obtain in closed form 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-squared
- Analysis of variance
- Bivariate regression
- Coefficient of determination
- 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 $b = [\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})]/[\sum_{i=1}^{n}(x_i - \bar{x})^2]$.

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 $4n[(\sum_{i=1}^{n} x_i^2) - n\bar{x}^2] = 4n[\sum_{i=1}^{n}(x_i - \bar{x})^2]$, 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. **Linear Transformations of the data.** Consider the least squares regression of $\mathbf{y}$ on K variables (with a constant) $\mathbf{X}$. Consider an alternative set of regressors $\mathbf{Z} = \mathbf{XP}$, where $\mathbf{P}$ is a nonsingular matrix. Thus, each column of $\mathbf{Z}$ is a mixture of some of the columns of $\mathbf{X}$. Prove that the residual vectors in the regressions of $\mathbf{y}$ on $\mathbf{X}$ and $\mathbf{y}$ on $\mathbf{Z}$ are identical. What relevance does this have to the question of changing the fit of a regression by changing the units of measurement of the independent variables?

4. **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 $\mathbf{X}$?

5. **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)?

6. **Adding an observation.** A data set consists of n observations on $\mathbf{X}_n$ 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 $\mathbf{b}_n$. 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.

7. **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.

8. **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

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

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.

9. **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.
10. **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?
11. 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.
12. 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, real GNP and the interest rate. Compute R^2.
13. 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:

$$\ln S/Y = 7.3439 + 0.1596 \ln Y/N + 0.0254 \ln G - 1.3520 \ln D_1 - 0.3990 \ln D_2$$
$$\ln S/N = 8.7851 + 1.1486 \ln Y/N + 0.0265 \ln G - 1.3438 \ln D_1 - 0.3966 \ln D_2$$

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.