3.1 Least squares in matrix form

☜ Uses Appendix A.2–A.4, A.6, A.7.

3.1.1 Introduction

More than one explanatory variable

In the foregoing chapter we considered the simple regression model where the dependent variable is related to one explanatory variable. In practice the situation is often more involved in the sense that there exists more than one variable that influences the dependent variable.

As an illustration we consider again the salaries of 474 employees at a US bank (see Example 2.2 (p. 77) on bank wages). In Chapter 2 the variations in salaries (measured in logarithms) were explained by variations in education of the employees. As can be observed from the scatter diagram in Exhibit 2.5(*a*) (p. 85) and the regression results in Exhibit 2.6 (p. 86), around half of the variance can be explained in this way. Of course, the salary of an employee is not only determined by the number of years of education because many other variables also play a role. Apart from salary and education, the following data are available for each employee: begin or starting salary (the salary that the individual earned at his or her first position at this bank), gender (with value zero for females and one for males), ethnic minority (with value zero for non-minorities and value one for minorities), and job category (category 1 consists of administrative jobs, category 2 of custodial jobs, and category 3 of management jobs). The begin salary can be seen as an indication of the qualities of the employee that, apart from education, are determined by previous experience, personal characteristics, and so on. The other variables may also affect the earned salary.

Simple regression may be misleading

Of course, the effect of each variable could be estimated by a simple regression of salaries on each explanatory variable separately. For the explanatory variables education, begin salary, and gender, the scatter diagrams with regression lines are shown in Exhibit 3.1 (*a*–*c*). However, these results may be misleading, as the explanatory variables are mutually related. For

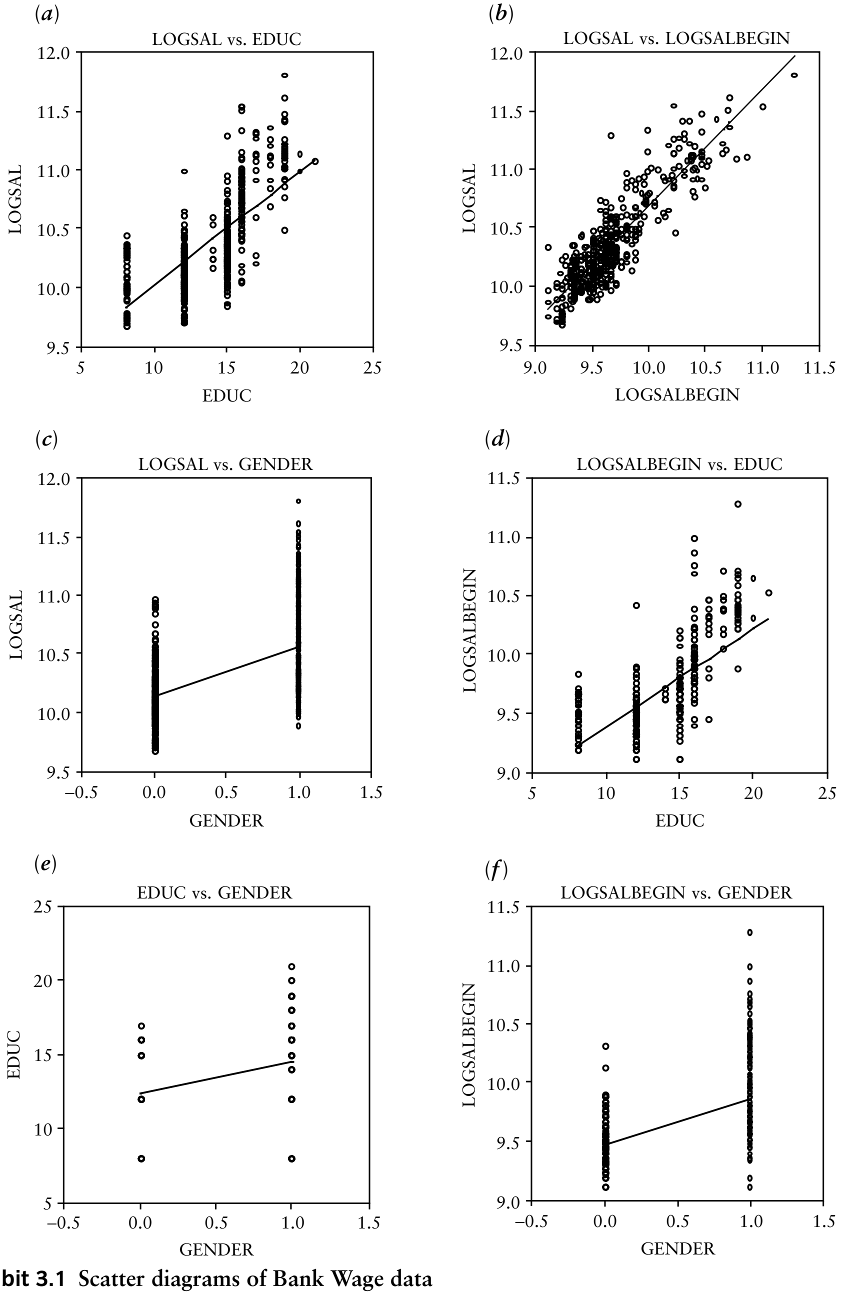

Exhibit 3.1 Scatter diagrams of Bank Wage data

Scatter diagrams with regression lines for several bivariate relations between the variables LOGSAL (logarithm of yearly salary in dollars), EDUC (finished years of education), LOGSALBEGIN (logarithm of yearly salary when employee entered the firm) and GENDER (0 for females, 1 for males), for 474 employees of a US bank.

example, the gender effect on salaries (c) is partly caused by the gender effect on education (e). Similar relations between the explanatory variables are shown in (d) and (f). This mutual dependence is taken into account by formulating a multiple regression model that contains more than one explanatory variable.

3.1.2 Least squares

☜ Uses Appendix A.7.

Regression model in matrix form

The linear model with several explanatory variables is given by the equation

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki} + \varepsilon_i \quad (i = 1, \cdots, n). \tag{3.1}$$

From now on we follow the convention that the constant term is denoted by β_1 rather than α. The first explanatory variable x_1 is defined by $x_{1i} = 1$ for every $i = 1, \cdots, n$, and for simplicity of notation we write β_1 instead of $\beta_1 x_{1i}$. For purposes of analysis it is convenient to express the model (3.1) in *matrix form*. Let

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{21} & \cdots & x_{k1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{2n} & \cdots & x_{kn} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}. \tag{3.2}$$

Note that in the $n \times k$ matrix $X = (x_{ji})$ the first index j $(j = 1, \cdots, k)$ refers to the variable number (in columns) and the second index i $(i = 1, \cdots, n)$ refers to the observation number (in rows). The notation in (3.2) is common in econometrics (whereas in books on linear algebra the indices i and j are often reversed). In our notation, we can rewrite (3.1) as

$$y = X\beta + \varepsilon. \tag{3.3}$$

Here β is a $k \times 1$ vector of unknown parameters and ε is an $n \times 1$ vector of unobserved disturbances.

Residuals and the least squares criterion

If b is a $k \times 1$ vector of estimates of β, then the estimated model may be written as

$$y = Xb + e. \tag{3.4}$$

Here e denotes the $n \times 1$ vector of residuals, which can be computed from the data and the vector of estimates b by means of

$$e = y - Xb. \tag{3.5}$$

We denote transposition of matrices by primes ($'$)—for instance, the transpose of the residual vector e is the $1 \times n$ matrix $e' = (e_1, \cdots, e_n)$. To determine the least squares estimator, we write the sum of squares of the residuals (a function of b) as

$$\begin{aligned} S(b) &= \sum e_i^2 = e'e = (y - Xb)'(y - Xb) \\ &= y'y - y'Xb - b'X'y + b'X'Xb. \end{aligned} \tag{3.6}$$

Derivation of least squares estimator

The minimum of $S(b)$ is obtained by setting the derivatives of $S(b)$ equal to zero. Note that the function $S(b)$ has scalar values, whereas b is a column vector with k components. So we have k first order derivatives and we will follow the convention to arrange them in a column vector. The second and third terms of the last expression in (3.6) are equal (a 1×1 matrix is always symmetric) and may be replaced by $-2b'X'y$. This is a linear expression in the elements of b and so the vector of derivatives equals $-2X'y$. The last term of (3.6) is a quadratic form in the elements of b. The vector of first order derivatives of this term $b'X'Xb$ can be written as $2X'Xb$. The proof of this result is left as an exercise (see Exercise 3.1). To get the idea we consider the case $k = 2$ and we denote the elements of $X'X$ by c_{ij}, $i, j = 1, 2$, with $c_{12} = c_{21}$. Then $b'X'Xb = c_{11}b_1^2 + c_{22}b_2^2 + 2c_{12}b_1b_2$. The derivative with respect to b_1 is $2c_{11}b_1 + 2c_{12}b_2$ and the derivative with respect to b_2 is $2c_{12}b_1 + 2c_{22}b_2$. When we arrange these two partial derivatives in a 2×1 vector, this can be written as $2X'Xb$. See Appendix A (especially Examples A.10 and A.11 in Section A.7) for further computational details and illustrations.

The least squares estimator

Combining the above results, we obtain

$$\frac{\partial S}{\partial b} = -2X'y + 2X'Xb. \tag{3.7}$$

The least squares estimator is obtained by minimizing $S(b)$. Therefore we set these derivatives equal to zero, which gives the *normal equations*

$$X'Xb = X'y. \tag{3.8}$$

Solving this for b, we obtain

$$b = (X'X)^{-1}X'y \tag{3.9}$$

provided that the inverse of $X'X$ exists, which means that the matrix X should have rank k. As X is an $n \times k$ matrix, this requires in particular that $n \geq k$—that is, the number of parameters is smaller than or equal to the number of observations. In practice we will almost always require that k is considerably smaller than n.

Proof of minimum

From now on, if we write b, we always mean the expression in (3.9). This is the classical formula for the *least squares estimator* in matrix notation. If the matrix X has rank k, it follows that the Hessian matrix

$$\frac{\partial^2 S}{\partial b \partial b'} = 2X'X \tag{3.10}$$

is a positive definite matrix (see Exercise 3.2). This implies that (3.9) is indeed the minimum of (3.6). In (3.10) we take the derivatives of a vector $\left(\frac{\partial S}{\partial b}\right)$ with respect to another vector (b') and we follow the convention to arrange these derivatives in a matrix (see Exercise 3.2). An alternative proof that b minimizes the sum of squares (3.6) that makes no use of first and second order derivatives is given in Exercise 3.3.

Summary of computations

The least squares estimates can be computed as follows.

Least squares estimation

- *Step 1: Choice of variables*. Choose the variable to be explained (y) and the explanatory variables ($x_1, \cdots, x_k$, where x_1 is often the constant that always takes the value 1).
- *Step 2: Collect data*. Collect n observations of y and of the related values of $x_1, \cdots, x_k$ and store the data of y in an $n \times 1$ vector and the data on the explanatory variables in the $n \times k$ matrix X.
- *Step 3: Compute the estimates*. Compute the least squares estimates by the OLS formula (3.9) by using a regression package.

☜ **Exercises:** T: 3.1, 3.2.

3.1.3 Geometric interpretation

☜ Uses Sections 1.2.2, 1.2.3; Appendix A.6.

Least squares seen as projection

The least squares method can be given a geometric interpretation, which we discuss now. Using the expression (3.9) for b, the residuals may be written as

$$e = y - Xb = y - X(X'X)^{-1}X'y = My \qquad (3.11)$$

where

$$M = I - X(X'X)^{-1}X'. \qquad (3.12)$$

The matrix M is symmetric ($M' = M$) and idempotent ($M^2 = M$). Since it also has the property $MX = 0$, it follows from (3.11) that

$$X'e = 0. \qquad (3.13)$$

We may write the explained component $\hat{y}$ of y as

$$\hat{y} = Xb = Hy \qquad (3.14)$$

where

$$H = X(X'X)^{-1}X' \qquad (3.15)$$

is called the 'hat matrix', since it transforms y into $\hat{y}$ (pronounced: 'y-hat'). Clearly, there holds $H' = H$, $H^2 = H$, $H + M = I$ and $HM = 0$. So

$$y = Hy + My = \hat{y} + e$$

where, because of (3.11) and (3.13), $\hat{y}'e = 0$, so that the vectors $\hat{y}$ and e are orthogonal to each other. Therefore, the least squares method can be given the following interpretation. The sum of squares $e'e$ is the square of the length of the residual vector $e = y - Xb$. The length of this vector is minimized by choosing Xb as the orthogonal *projection* of y onto the space spanned by the columns of X. This is illustrated in Exhibit 3.2. The projection is characterized by the property that $e = y - Xb$ is orthogonal to all columns of X, so that $0 = X'e = X'(y - Xb)$. This gives the normal equations (3.8).

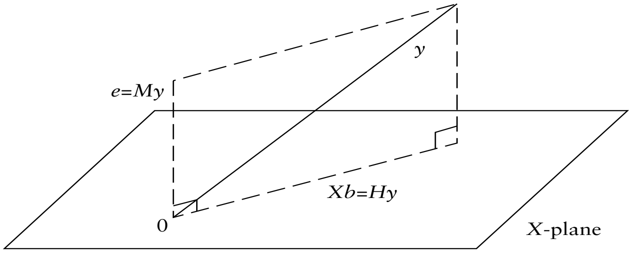

Exhibit 3.2 Least squares

Three-dimensional geometric impression of least squares, the vector of observations on the dependent variable y is projected onto the plane of the independent variables X to obtain the linear combination Xb of the independent variables that is as close as possible to y.

Geometry of least squares

Let $S(X)$ be the space spanned by the columns of X (that is, the set of all $n \times 1$ vectors that can be written as Xa for some $k \times 1$ vector a) and let $S^{\perp}(X)$ be the space orthogonal to $S(X)$ (that is, the set of all $n \times 1$ vectors z with the property that $X'z = 0$). The matrix H projects onto $S(X)$ and the matrix M projects onto $S^{\perp}(X)$. In $y = \hat{y} + e$, the vector y is decomposed into two orthogonal components, with $\hat{y} \in S(X)$ according to (3.14) and $e \in S^{\perp}(X)$ according to (3.13). The essence of this decomposition is given in Exhibit 3.3, which can be seen as a two-dimensional version of the three-dimensional picture in Exhibit 3.2.

Geometric interpretation as a tool in analysis

This geometric interpretation can be helpful to understand some of the algebraic properties of least squares. As an example we consider the effect of applying linear transformations on the set of explanatory variables. Suppose that the $n \times k$ matrix X is replaced by $X_* = XA$ where A is a $k \times k$ invertible matrix. Then the least squares fit ($\hat{y}$), the residuals (e), and the projection matrices (H and M) remain unaffected by this transformation. This is immediately evident from the geometric pictures in Exhibits 3.2 and 3.3, as $S(X_*) = S(X)$.

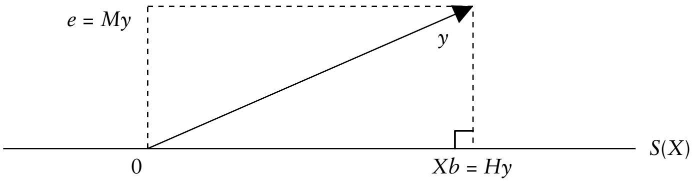

Exhibit 3.3 Least squares

Two-dimensional geometric impression of least squares where the k-dimensional plane $S(X)$ is represented by the horizontal line, the vector of observations on the dependent variable y is projected onto the space of the independent variables $S(X)$ to obtain the linear combination Xb of the independent variables that is as close as possible to y.

The properties can also be checked algebraically, by working out the expressions for $\hat{y}$, e, H, and M in terms of X_*. The least squares estimates change after the transformation, as $b_* = (X_*'X_*)^{-1}X_*'y = A^{-1}b$. For example, suppose that the variable x_k is measured in dollars and x_k^* is the same variable measured in thousands of dollars. Then $x_{ki}^* = x_{ki}/1000$ for $i = 1, \cdots, n$, and $X_* = XA$ where A is the diagonal matrix diag$(1, \cdots, 1, 0.001)$. The least squares estimates of β_j for $j \neq k$ remain unaffected—that is, $b_j^* = b_j$ for $j \neq k$, and $b_k^* = 1000 b_k$. This also makes perfect sense, as one unit increase in x_k^* corresponds to an increase of a thousand units in x_k.

☜ **Exercises:** T: 3.3.

3.1.4 Statistical properties

☜ Uses Sections 1.2.2, 1.3.2.

Seven assumptions on the multiple regression model

To analyse the statistical properties of least squares estimation, it is convenient to use as conceptual background again the simulation experiment described in Section 2.2.1 (p. 87–8). We first restate the seven assumptions of Section 2.2.3 (p. 92) for the multiple regression model (3.3) and use the matrix notation introduced in Section 3.1.2.

- *Assumption 1: fixed regressors.* All elements of the $n \times k$ matrix X containing the observations on the explanatory variables are non-stochastic. It is assumed that $n \geq k$ and that the matrix X has rank k.
- *Assumption 2: random disturbances, zero mean.* The $n \times 1$ vector ε consists of random disturbances with zero mean so that $E[\varepsilon] = 0$, that is, $E[\varepsilon_i] = 0$ $(i = 1, \cdots, n)$.
- *Assumption 3: homoskedasticity.* The covariance matrix of the disturbances $E[\varepsilon\varepsilon']$ exists and all its diagonal elements are equal to σ^2, that is, $E[\varepsilon_i^2] = \sigma^2$ $(i = 1, \cdots, n)$.
- *Assumption 4: no correlation.* The off-diagonal elements of the covariance matrix of the disturbances $E[\varepsilon\varepsilon']$ are all equal to zero, that is, $E[\varepsilon_i \varepsilon_j] = 0$ for all $i \neq j$.
- *Assumption 5: constant parameters.* The elements of the $k \times 1$ vector β and the scalar σ are fixed unknown numbers with $\sigma > 0$.
- *Assumption 6: linear model.* The data on the explained variable y have been generated by the data generating process (DGP)

$$y = X\beta + \varepsilon. \tag{3.16}$$

- *Assumption 7: normality.* The disturbances are jointly normally distributed.

Assumptions 3 and 4 can be summarized in matrix notation as

$$E[\varepsilon\varepsilon'] = \sigma^2 I, \tag{3.17}$$

where I denotes the $n \times n$ identity matrix. If in addition Assumption 7 is satisfied, then ε follows the multivariate normal distribution

$$\varepsilon \sim \mathrm{N}(0, \sigma^2 I).$$

Assumptions 4 and 7 imply that the disturbances $\varepsilon_i, i = 1, \cdots, n$ are mutually independent.

Least squares is unbiased

The expected value of b is obtained by using Assumptions 1, 2, 5, and 6. Assumption 6 implies that the least squares estimator $b = (X'X)^{-1}X'y$ can be written as

$$b = (X'X)^{-1}X'(X\beta + \varepsilon) = \beta + (X'X)^{-1}X'\varepsilon.$$

Taking expectations is a linear operation — that is, if z_1 and z_2 are two random variables and A_1 and A_2 are two non-random matrices of appropriate dimensions so that $z = A_1z_1 + A_2z_2$ is well defined, then $E[z] = A_1E[z_1] + A_2E[z_2]$. From Assumptions 1, 2, and 5 we obtain

$$E[b] = E[\beta + (X'X)^{-1}X'\varepsilon] = \beta + (X'X)^{-1}X'E[\varepsilon] = \beta. \tag{3.18}$$

So b is *unbiased*.

The covariance matrix of *b*

Using the result (3.18), we obtain that under Assumptions 1–6 the *covariance matrix* of b is given by

$$\begin{aligned} \mathrm{var}(b) &= E[(b-\beta)(b-\beta)'] = E[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}] \\ &= (X'X)^{-1}X'E[\varepsilon\varepsilon']X(X'X)^{-1} = (X'X)^{-1}X'(\sigma^2 I)X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}. \end{aligned} \tag{3.19}$$

The diagonal elements of this matrix are the variances of the estimators of the individual parameters, and the off-diagonal elements are the covariances between these estimators.

Least squares is best linear unbiased

The Gauss–Markov theorem, proved in Section 2.2.5 (p. 97–8) for the simple regression model, also holds for the more general model (3.16). It states that, among all linear unbiased estimators, b has *minimal variance*—that is, b is the best linear unbiased estimator (BLUE) in the sense that, if $\hat{\beta} = Ay$ with A a $k \times n$ non-stochastic matrix and $E[\hat{\beta}] = \beta$, then $\text{var}(\hat{\beta}) - \text{var}(b)$ is a positive semidefinite matrix. This means that for every $k \times 1$ vector c of constants there holds $c'(\text{var}(\hat{\beta}) - \text{var}(b))c \geq 0$, or, equivalently, $\text{var}(c'b) \leq \text{var}(c'\hat{\beta})$. Choosing for c the jth unit vector, this means in particular that for the jth component $\text{var}(b_j) \leq \text{var}(\hat{\beta}_j)$ so that the least squares estimators are efficient. This result holds true under Assumptions 1–6, the assumption of normality is not needed.

Proof of Gauss–Markov theorem

T

To prove the result, first note that the condition that $E[\hat{\beta}] = E[Ay] = AE[y] = AX\beta = \beta$ for all β implies that $AX = I$, the $k \times k$ identity matrix. Now define $D = A - (X'X)^{-1}X'$, then $DX = AX - (X'X)^{-1}X'X = I - I = 0$ so that

$$\text{var}(\hat{\beta}) = \text{var}(Ay) = \text{var}(A\varepsilon) = \sigma^2 AA' = \sigma^2 DD' + \sigma^2 (X'X)^{-1},$$

where the last equality follows by writing $A = D + (X'X)^{-1}X'$ and working out AA'. This shows that $\text{var}(\hat{\beta}) - \text{var}(b) = \sigma^2 DD'$, which is positive semidefinite, and zero if and only if $D = 0$—that is, $A = (X'X)^{-1}X'$. So $\hat{\beta} = b$ gives the minimal variance.

☞ **Exercises:** T: 3.4.

3.1.5 Estimating the disturbance variance

Derivation of unbiased estimator

T

Next we consider the estimation of the unknown variance σ^2. As in the previous chapter we make use of the sum of squared residuals $e'e$. Intuition could suggest to estimate $\sigma^2 = E[\varepsilon_i^2]$ by the sample mean $\frac{1}{n}\sum e_i^2 = \frac{1}{n}e'e$, but this estimator is not unbiased. It follows from (3.11) and (3.16) and the fact that $MX = 0$ that $e = My = M(X\beta + \varepsilon) = M\varepsilon$. So

$$E[e] = 0, \tag{3.20}$$

$$\text{var}(e) = E[ee'] = E[M\varepsilon\varepsilon' M] = ME[\varepsilon\varepsilon']M = \sigma^2 M^2 = \sigma^2 M. \tag{3.21}$$

To evaluate $E[e'e]$ it is convenient to use the trace of a square matrix, which is defined as the sum of the diagonal elements of this matrix. Because the trace and the expectation operator can be interchanged, we find, using the property that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, that

$$E[e'e] = E[\mathrm{tr}(ee')\,] = \mathrm{tr}(E[ee']\,) = \sigma^2 \mathrm{tr}(M).$$

Using the property that $\mathrm{tr}(A + B) = \mathrm{tr}(A) + \mathrm{tr}(B)$ we can simplify this as

$$\begin{aligned}\mathrm{tr}(M) &= \mathrm{tr}(I_n - X(X'X)^{-1}X') = n - \mathrm{tr}(X(X'X)^{-1}X') \\ &= n - \mathrm{tr}(X'X(X'X)^{-1}) \;= n - \mathrm{tr}(I_k) = n - k,\end{aligned}$$

where the subscripts denote the order of the identity matrices.

The least squares estimator s^2 and standard errors

This shows that $E[e'e] = (n - k)\sigma^2$ so that

$$s^2 = \frac{e'e}{n-k} \tag{3.22}$$

is an unbiased estimator of σ^2. The square root s of (3.22) is called the *standard error of the regression*. If in the expression (3.19) we replace σ^2 by s^2 and if we denote the jth diagonal element of $(X'X)^{-1}$ by a_{jj}, then $s\sqrt{a_{jj}}$ is called the *standard error* of the estimated coefficient b_j. This is an estimate of the standard deviation $\sigma\sqrt{a_{jj}}$ of b_j.

Intuition for the factor $1/(n - k)$

The result in (3.22) can also be given a more intuitive interpretation. Suppose we would try to explain y by a matrix X with $k = n$ columns and rank k. Then we would obtain $e = 0$, a perfect fit, but we would not have obtained any information on σ^2. Of course this is an extreme case. In practice we confine ourselves to the case $k < n$. The very fact that we choose b in such a way that the sum of squared residuals is minimized is the cause of the fact that the squared residuals are smaller (on average) than the squared disturbances. Let us consider a diagonal element of (3.21),

$$\mathrm{var}(e_i) = \sigma^2(1 - h_i), \tag{3.23}$$

where h_i is the ith diagonal element of the matrix $H = I - M$ in (3.15). As H is positive semidefinite, it follows that $h_i \geq 0$. If the model contains a constant term (so that the matrix X contains a column of ones), then $h_i > 0$ (see Exercise 3.7). So each single element e_i of the residual vector has a variance that is smaller than σ^2, and therefore the sum of squares $\sum e_i^2$ has an expected value less than $n\sigma^2$. This effect becomes stronger when we have more parameters to obtain a good fit for the data. If one would like to use a small residual variance as a criterion for a good model, then the denominator $(n-k)$ of the estimator (3.22) gives an automatic penalty for choosing models with large k.

Intuition for the number of degrees of freedom ($n - k$)

As $e = M\varepsilon$, it follows under Assumptions 1–7 that $e'e/\sigma^2 = \varepsilon' M\varepsilon/\sigma^2$ follows the χ^2-distribution with $(n-k)$ degrees of freedom. This follows from the results in Section 1.2.3 (p. 32), using the fact that M is an idempotent matrix with rank $(n-k)$. The term *degrees of freedom* refers to the restrictions $X'e = 0$. We may partition this as $X_1'e_1 + X_2'e_2 = 0$, where X_1' is a $k \times (n-k)$ matrix and X_2' a $k \times k$ matrix. If the matrix X_2' has a rank less than k, we may rearrange the columns of X' in such a way that X_2' has rank k. The restrictions imply that, once we have freely chosen the $n-k$ elements of e_1, the remaining elements are dictated by $e_2 = -(X_2')^{-1}X_1'e_1$. This is also clear from Exhibit 3.3. For given matrix X of explanatory variables, the residual vector lies in $S^{\perp}(X)$ and this space has dimension $(n-k)$. That is, k degrees of freedom are lost because β has been estimated.

☞ **Exercises:** T: 3.5, 3.7a.

3.1.6 Coefficient of determination

Derivation of R^2

The performance of least squares can be evaluated by the *coefficient of determination* R^2—that is, the fraction of the total sample variation $\sum (y_i - \bar{y})^2$ that is explained by the model.

In matrix notation, the total sample variation can be written as $y'Ny$ with

$$N = I - \frac{1}{n}\iota\iota',$$

where $\iota = (1, \cdots, 1)'$ is the $n \times 1$ vector of ones. The matrix N has the property that it takes deviations from the mean, as the elements of Ny are $y_i - \bar{y}$. Note that N is a special case of an M-matrix (3.12) with $X = \iota$, as $\iota'\iota = n$. So Ny can be interpreted as the vector of residuals and $y'Ny = (Ny)'Ny$ as the residual sum of squares from a regression where y is explained by $X = \iota$. If X in the multiple regression model (3.3) contains a constant term, then the fact that $X'e = 0$ implies that $\iota'e = 0$ and hence $Ne = e$. From $y = Xb + e$ we then obtain $Ny = NXb + Ne = NXb + e =$ 'explained' + 'residual', and

$$\begin{aligned} y'Ny &= (Ny)'Ny = (NXb + e)'(NXb + e) \\ &= b'X'NXb + e'e. \end{aligned}$$

Here the cross term vanishes because $b'X'Ne = 0$, as $Ne = e$ and $X'e = 0$. It follows that the total variation in y (SST) can be decomposed in an explained part $SSE = b'X'NXb$ and a residual part $SSR = e'e$.

Coefficient of determination: R^2

Therefore R^2 is given by

$$R^2 = \frac{SSE}{SST} = \frac{b'X'NXb}{y'Ny} = 1 - \frac{e'e}{y'Ny} = 1 - \frac{SSR}{SST}. \tag{3.24}$$

The third equality in (3.24) holds true if the model contains a constant term. If this is not the case, then SSR may be larger than SST (see Exercise 3.7) and R^2 is defined as SSE/SST (and not as $1 - SSR/SST$). If the model contains a constant term, then (3.24) shows that $0 \leq R^2 \leq 1$. It is left as an exercise (see Exercise 3.7) to show that R^2 is the squared sample correlation coefficient between y and its explained part $\hat{y} = Xb$. In geometric terms, R (the square root of R^2) is equal to the length of NXb divided by the length of Ny—that is, R is equal to the cosine of the angle between Ny and NXb. This is illustrated in Exhibit 3.4. A good fit is obtained when Ny is close to NXb—that is, when the angle between these two vectors is small. This corresponds to a high value of R^2.

Adjusted R^2

When explanatory variables are added to the model, then R^2 never decreases (see Exercise 3.6). The wish to penalize models with large k has motivated an adjusted R^2 defined by adjusting for the degrees of freedom.

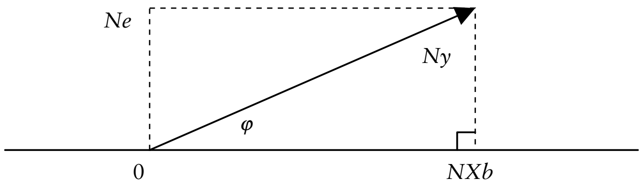

Exhibit 3.4 Geometric picture of R^2

Two-dimensional geometric impression of the coefficient of determination. The dependent variable and all the independent variables are taken in deviation from their sample means, with resulting vector of dependent variables Ny and matrix of independent variables NX. The explained part of Ny is NXb with residuals $Ne = e$, and the coefficient of determination is equal to the square of the cosine of the indicated angle φ.

$$\overline{R}^2 = 1 - \frac{e'e/(n-k)}{y'Ny/(n-1)} = 1 - \frac{n-1}{n-k}(1 - R^2). \qquad (3.25)$$

☞ **Exercises:** T: 3.6**a**, **b**, 3.7**b**, **c**.

3.1.7 Illustration: Bank Wages

E

XM301BWA

To illustrate the foregoing results we consider the data on salary and education discussed earlier in Chapter 2 and in Section 3.1.1. We will discuss (i) the data, (ii) the model, (iii) the normal equations and the least squares estimates, (iv) the interpretation of the estimates, (v) the sums of squares and R^2, and (vi) the orthogonality of residuals and explanatory variables.

(i) Data

The data consist of a cross section of 474 individuals working for a US bank. For each employee, the information consists of the following variables: salary (S), education (x_2), begin salary (B), gender ($x_4 = 0$ for females, $x_4 = 1$ for males), minority ($x_5 = 1$ if the individual belongs to a minority group, $x_5 = 0$ otherwise), job category ($x_6 = 1$ for clerical jobs, $x_6 = 2$ for custodial jobs, and $x_6 = 3$ for management positions), and some further job-related variables.

(ii) Model

As a start, we will consider the model with $y = \log(S)$ as variable to be explained and with x_2 and $x_3 = \log(B)$ as explanatory variables. That is, we consider the regression model

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i \qquad (i = 1, \cdots, n).$$

(iii) Normal equations and least squares estimates

As before, to simplify the notation we define the first regressor by $x_{1i} = 1$. The normal equations (3.8) involve the cross product terms $X'X$ and $X'y$. For the data at hand they are given (after rounding) in Exhibit 3.5, Panel 1. Solving the normal equations (3.8) gives the least squares estimates shown in Panel 3 in Exhibit 3.5, so that (after rounding) $b_1 = 1.647$, $b_2 = 0.023$, and $b_3 = 0.869$. It may be checked from the cross products in Panel 1 in Exhibit 3.5 that $X'Xb = X'y$ (apart from rounding errors) — that is,

$$\begin{pmatrix} 474 & 6395 & 4583 \\ 6395 & 90215 & 62166 \\ 4583 & 62166 & 44377 \end{pmatrix} \begin{pmatrix} 1.647 \\ 0.023 \\ 0.869 \end{pmatrix} = \begin{pmatrix} 4909 \\ 66609 \\ 47527 \end{pmatrix}.$$

Panel 1	IOTA	LOGSAL	EDUC	LOGSALBEGIN
IOTA	474			
LOGSAL	4909	50917		
EDUC	6395	66609	90215	
LOGSALBEGIN	4583	47527	62166	44377

Panel 2	LOGSAL	EDUC	LOGSALBEGIN
LOGSAL	1.000000		
EDUC	0.696740	1.000000	
LOGSALBEGIN	0.886368	0.685719	1.000000

Panel 3: Dependent Variable: LOGSAL		
Method: Least Squares		
Sample: 1 474		
Included observations: 474		
Variable	Coefficient	Std. Error
C	1.646916	0.274598
EDUC	0.023122	0.003894
LOGSALBEGIN	0.868505	0.031835
R-squared	0.800579	
Adjusted R-squared	0.799733	
S.E. of regression	0.177812	
Sum squared resid	14.89166	
Total sum of squares	74.67462	
Explained sum of squares	59.78296	

Panel 4: Dependent Variable: RESID	
Method: Least Squares	
Sample: 1 474	
Included observations: 474	
Variable	Coefficient
C	3.10E-11
EDUC	2.47E-13
LOGSALBEGIN	−3.55E-12
R-squared	0.000000
Adjusted R-squared	−0.004246
S.E. of regression	0.177812
Sum squared resid	14.89166

Exhibit 3.5 Bank Wages (Section 3.1.7)

Panel 1 contains the cross product terms ($X'X$ and $X'y$) of the variables (iota denotes the constant term with all values equal to one), Panel 2 shows the correlations between the dependent and the two independent variables, and Panel 3 shows the outcomes obtained by regressing salary (in logarithms) on a constant and the explanatory variables education and the logarithm of begin salary. The residuals of this regression are denoted by RESID, and Panel 4 shows the result of regressing these residuals on a constant and the two explanatory variables (3.10E-11 means $3.10*10^{-11}$, and so on; these values are zero up to numerical rounding).

(iv) Interpretation of estimates

A first thing to note here is that the marginal relative effect of education on wage (that is, $\frac{dS/S}{dx_2} = \frac{d\log(S)}{dx_2} = \frac{dy}{dx_2} = \beta_2$) is estimated now as 0.023, whereas in Chapter 2 this effect was estimated as 0.096 with a standard error of 0.005 (see Exhibit 2.11, p. 103). This is a substantial difference. That is, an additional year of education corresponds on average with a 9.6 per cent increase in salary. But, if the begin salary is ‘kept fixed’, an additional year of education gives only a 2.3 per cent increase in salary. The cause of this difference is that the variable ‘begin salary’ is strongly related to the variable ‘education’. This is clear from Panel 2 in Exhibit 3.5, which shows that x_2 and x_3 have a correlation of around 69 per cent. We refer also to Exhibit 3.1 (*d*), which shows a strong positive relation between x_2 and x_3. This means that in Chapter 2, where we have excluded the begin salary from the model, part of the positive association between education and salary is due to a third variable, begin salary. This explains why the estimated effect in Chapter 2 is larger.

(v) Sums of squares and R^2

The sums of squares for this model are reported in Panel 3 in Exhibit 3.5, with values $SST = 74.675$, $SSE = 59.783$, and $SSR = 14.892$, so that $R^2 = 0.801$. This is larger than the $R^2 = 0.485$ in Chapter 2 (see Exhibit 2.6, p. 86). In Section 3.4 we will discuss a method to test whether this is a significant increase in the model fit. Panel 3 in Exhibit 3.5 also reports the standard error of the regression $s = \sqrt{SSR/(474 - 3)} = 0.178$ and the standard error of b_2 0.0039.

(vi) Orthogonality of residuals and explanatory variables

Panel 4 in Exhibit 3.5 shows the result of regressing the least squares residuals on the variables x_1, x_2, and x_3. This gives an $R^2 = 0$, which is in accordance with the property that the residuals are uncorrelated with the explanatory variables in the sense that $X'e = 0$ (see Exhibits 3.2 and 3.4).