Appendix A: Review of Mathematical Concepts

A.1 Matrix Theory and Linear Algebra

A.1.1 Matrix Algebra

A *matrix* is an array of elements arranged in rows and columns. For example, the matrix

$$A = \begin{bmatrix} 3 & 17 \\ 21 & 0 \\ 4 & 16 \end{bmatrix} \tag{A.1}$$

has three rows and two columns and is referred to as a 3×2 matrix. The symbol a_{ij} is used to denote the element in row i and column j, so that a general 3×2 matrix may be written

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix}. \tag{A.2}$$

A vector with n elements (or components) may be either an $n \times 1$ matrix, which is a *column vector*, or a $1 \times n$ matrix, which is a *row vector*. A matrix with the same number of rows and columns is called a *square matrix*.

The product of multiplication of a matrix A and a scalar c is defined as the matrix whose elements are ca_{ij}. The sum of two matrices A and B is defined only if they have the same number of rows and columns, in which case the elements of $A + B$ are $a_{ij} + b_{ij}$. The product of two matrices AB is defined only if B has the same number of rows as A has columns. In this case the elements of AB are obtained by adding the products across the rows of A and down the columns of B. For example, suppose

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \; B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}. \tag{A.3}$$

Then

$$AB = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix}. \tag{A.4}$$

In general, $AB \neq BA$. To see this, make up a pair of 2×2 matrices whose elements are the numbers 1–8 and compute the products.

The *identity matrix* is the matrix I that satisfies $AI = IA = A$ for any square matrix

A. The 3×3 identity matrix, for example, is

$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{A.5}$$

If A is a square matrix, then the *inverse* of A, A^{-1}, if it exists, is the matrix that satisfies $AA^{-1} = A^{-1}A = I$. If the inverse of a matrix does not exist, the matrix is said to be *singular.* To determine whether a matrix A is singular, one computes a quantity called the *determinant*, denoted det A (Noble and Daniel, 1977, p. 198). The matrix A is singular if and only if det A = 0. The *transpose* of the matrix A whose elements are a_{ij} is the matrix A' whose elements are a_{ji}. A matrix is *symmetric* if $A' = A$, i.e., if $a_{ij} = a_{ji}$ for all i and j.

A.1.2 Random Vectors and Matrices

A random vector or matrix is a vector or matrix that contains random variables. Let

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{bmatrix}. \tag{A.6}$$

be a random vector. Then the expected value $E\{Y\}$ is given by

$$E\{Y\} = \begin{bmatrix} E\{Y_1\} \\ E\{Y_2\} \\ \dots \\ E\{Y_n\} \end{bmatrix}, \tag{A.7}$$

and the variance-covariance matrix var$\{Y\}$ is given by

$$\begin{aligned} \text{var}\{Y\} &= E\{(Y - E\{Y\})(Y - E\{Y\})'\} \\ &= \begin{bmatrix} \text{var}\{Y_1\} & \text{cov}\{Y_1, Y_2\} & \dots & \text{cov}\{Y_1, Y_n\} \\ \text{cov}\{Y_2, Y_1\} & \text{var}\{Y_2\} & \dots & \text{cov}\{Y_2, Y_n\} \\ \dots & \dots & \dots & \dots \\ \text{cov}\{Y_n, Y_1\} & \text{cov}\{Y_n, Y_2\} & \dots & \text{var}\{Y_n\} \end{bmatrix}. \end{aligned} \tag{A.8}$$

Since $\text{cov}(Y_i, Y_j\} = \text{cov}\{Y_j, Y_i\}$, the matrix var$\{Y\}$ is symmetric.

Suppose $W = AY$ where Y is a random vector and A is a fixed matrix. Then the expected value and variance-covariance matrix of W are given by

$$\begin{aligned} E\{W\} &= E\{AY\} = AE\{Y\} \\ \text{var}\{W\} &= \text{var}\{AY\} = A\,\text{var}\{Y\}A'. \end{aligned} \tag{A.9}$$

These same relationships hold for the sample mean, sample variance, and sample covariance.

Now, suppose that instead of a single data field Y, we have a set of k data fields denoted Y_j, $i = 1,...,k$. Each data field is composed of n data records denoted Y_{ij}, $i = 1,...,n$. The set of all the data is represented by a matrix Y whose columns are the data fields Y_j and whose rows are the data records, that is,

$$Y = \begin{bmatrix} Y_{11} & Y_{12} & ... & Y_{1k} \\ Y_{21} & Y_{22} & ... & Y_{2k} \\ ... & ... & ... & ... \\ Y_{n1} & Y_{n2} & ... & Y_{nk} \end{bmatrix}. \tag{A.10}$$

We can write the sample variance-covariance matrix S_Y as

$$S_Y = \begin{bmatrix} \text{var}\{Y_1\} & \text{cov}\{Y_1, Y_2\} & ... & \text{cov}\{Y_1, Y_k\} \\ \text{cov}\{Y_2, Y_1\} & \text{var}\{Y_2\} & ... & \text{cov}\{Y_2, Y_k\} \\ ... & ... & ... & ... \\ \text{cov}\{Y_k, Y_1\} & \text{cov}\{Y_k, Y_2\} & ... & \text{var}\{Y_k\} \end{bmatrix}, \tag{A.11}$$

where the variances and covariances are now interpreted as sample statistics rather than population parameters. As an example, suppose that a data set consists of two data fields (say, mean annual rainfall and mean annual temperature) denoted by Y_1 and Y_2, and suppose that each quantity is measured at 30 locations. Suppose further that Y_1 is a normally distributed random variable and Y_2 is related to Y_1 according to the equation $Y_2 = 2Y_1 + 1.3 + \varepsilon$, where ε is normally distributed. The following code generates artificial data representing these data fields (please ignore the fact that the numerical values don't make any sense as measures of rainfall and temperature) and centers them so that their sample means are 0 (this is done for use in Section A.1.3). It then computes the sample variances and covariance and the covariance matrix in Equation A.11.

```
> set.seed(123)
> # Generate the data
> Y1 <- rnorm(30)
> Y2 <- 2 * Y1 + 1.3 * rnorm(30)
> # Center the variables
> Y1 <- Y1 - mean(Y1)
> Y2 <- Y2 - mean(Y2)
> var(Y1)
[1] 0.9624212
> var(Y2)
[1] 4.357569
> cov(Y1,Y2)
[1] 1.757146
> Y <- cbind(Y1,Y2)
> var(Y)
          Y1        Y2
Y1 0.9624212 1.757146
Y2 1.7571458 4.357569
```

A.1.3 Eigenvectors, Eigenvalues, and Projections

In a two-dimensional Cartesian coordinate system, every point in a region is described by a pair of coordinates that we can call x and y. Similarly, in a three-dimensional system we can add a third coordinate z. This notion of coordinates can be extended to general vectors through the use of *coordinate vectors*. In order to visualize this concept, we will illustrate it with the data fields Y_1 and Y_2 created in the previous subsection. Figure A.1a shows a scatterplot of the two data fields.

In order to avoid confusion, with too many symbols containing a Y, we will denote the i^{th} data record by P_i. For example, data record 11, denoted P_{11}, is shown as a black dot in Figure A.1.a. This record has components (to three decimal places) $Y_{11,1} = 1.271$, $Y_{11,2} = 1.407$. Also shown in the scatterplot are two unit vectors u_1 and u_2. These unit vectors have coordinates (1,0) and (0,1), respectively. If we think of the data record P_{11} as a vector in this coordinate plane (shown in Figure A.1a as the arrow whose tip is at $Y_{11,1} = 1.271$, $Y_{11,2} = 1.407$), then this vector can be written in terms of the unit vectors u_1 and u_2 as (Figure A.1a)

$$P_{11} = 1.271u_1 + 1.407u_2. \tag{A.12}$$

Thus, there are two ways to interpret a vector. The first is as an array with one column, as in Equation A.6, and the second is as an arrow in a *vector space* like that of Figure A.1a. In this example, the vector as an array has two elements and is written

$$P_{11} = \begin{bmatrix} 1.271 \\ 1.407 \end{bmatrix}, \tag{A.13}$$

while the vector as an arrow is written in the form of Equation A.12. In the interpretation of vectors as arrows, the sum of two vectors is the arrow obtained by placing the tail of one arrow in the summand on the head of the other. That the representation of P_{11} in the form

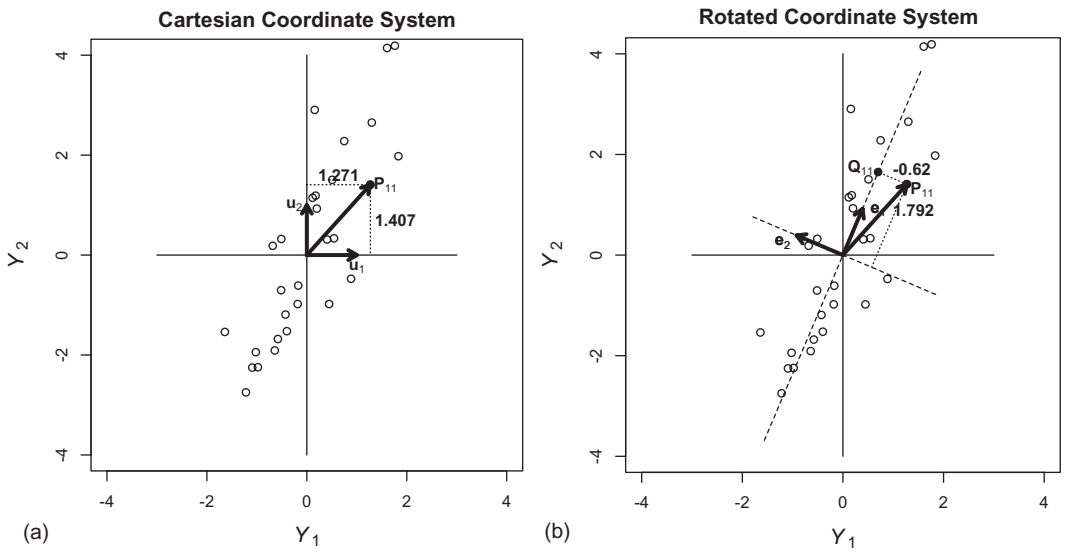

FIGURE A.1
(a) A vector expressed in Cartesian coordinates; (b) the same vector in a rotated coordinate system.

of Equation A.12 is equivalent to the representation in the form of Equation A.13 follows from the fact that we can write

$$u_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, u_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \tag{A.14}$$

in which case Equation A.13 is obtained from Equation A.12 by a combination of multiplication by a scalar and matrix addition. The dual interpretation of a vector as an array or as an arrow in a vector space does not depend on the number of elements in the vector. If the vector as an array has more than three elements, then we can no longer visualize the corresponding vector space in which the vector as an arrow lives, but as an abstract mathematical entity it is equally valid.

We can rotate the unit vectors u_1 and u_2 as shown in Figure A.1b and define the vector P_{11} (or any other vector) in terms of a new rotated coordinate system, which is denoted e_1 and e_2 in the figure. We will continue to work in two dimensions; but all of our results generalize readily to three or more dimensions. Suppose we wish to represent the vector P_{11} in terms of the rotated coordinate system composed of e_1 and e_2 as $P_{11} = a_1e_1 + a_2e_2$. It turns out that if the coordinate vectors e_1 and e_2 are obtained by rotating through an angle θ counterclockwise from the original coordinate vectors u_1 and u_2, then the values of a_1 and a_2 are obtained by multiplying the array in Equation A.13 by a *rotation matrix* that is expressed as (Noble and Daniel, 1977, p. 282).

$$W = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}, \tag{A.15}$$

For example, the coordinate vectors e_1 and e_2 in Figure A.1b are obtained from u_1 and u_2 by rotating through an angle of approximately $\theta = 67°$ (the reason for choosing this particular value for θ will be explained later). We have $\cos\theta = 0.3906338$ and $\sin\theta = 0.9205462$. Therefore, the vector P_{11} is expressed as follows:

```
> W
           [,1]       [,2]
[1,]  0.3906338  0.9205462
[2,] -0.9205462  0.3906338
> print(P11 <- Y[11,])
      Y1       Y2
1.271186 1.407412
> W %*% P11
           [,1]
[1,]  1.7921559
[2,] -0.6204022
```

Thus, $P_{11} = 1.792e_1 - 0.62e_2$. This is shown in Figure A.1b.

Because the vector P_{11} represents a data record in a sample composed of the data fields Y_1 and Y_2, there is a particular rotation that has special significance. This rotation has to do with the sample variance-covariance matrix. The variance-covariance matrix S_Y of the sample in this example is defined in Equation A.11 and is computed in the code given at the end of Section A.1.2. Like all variance-covariance matrices, it is symmetric. Suppose we

wish to compute two new variables, denoted C_1 and C_2, that are linear combinations of Y_1 and Y_2 but are uncorrelated. That is, we would like to determine coefficients a_{ij} such that

$$\begin{aligned} C_{i1} &= a_{11}Y_{i1} + a_{12}Y_{i2} \\ C_{i2} &= a_{21}Y_{i1} + a_{22}Y_{i2}, \; i = 1, 2, \ldots, n, \end{aligned} \tag{A.16}$$

so that the covariance between the vectors C_1 and C_2 is zero. In matrix theoretical terms, this says that the variance-covariance matrix S_C of the matrix C whose columns are these two new variables is a diagonal matrix. If we consider any data record $P_i = [Y_{i1}\, Y_{i2}]'$ (remember that the prime denotes transpose, and that P_i is a column vector) as a vector in the array sense, then we can write Equation A.16 as

$$C_i = AP_i, \; i = 1, \ldots, n \tag{A.17}$$

where A is a 2×2 matrix and C_i is a vector. For example, if for P_i we use the value P_{11} from our artificial data set, then we could write

$$\begin{bmatrix} C_{11,1} \\ C_{11,2} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} Y_{11,1} \\ Y_{11,2} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} 1.271 \\ 1.407 \end{bmatrix}. \tag{A.18}$$

One has to be a little careful of the subscripts here, because they have different uses. The subscript 11 of P_{11} and the $Y_{11,i}$ indicates that this is the eleventh data record. The subscripts 1 and 2 of $Y_{11,1}$ and $Y_{11,2}$ indicate that these are the first and second values of the eleventh data record (for example, they might represent the centered values of mean annual rainfall and mean annual temperature of the eleventh data record). The subscripts i and j of the elements a_{ij} represent the position of the element in the matrix A.

Returning to our problem, we want to determine a matrix A such that the variance-covariance matrix S_C of the data set expressed in the new variables C_1 and C_2 given by Equation A.16, or equivalently by Equation A.17, is a diagonal matrix. It turns out (Noble and Daniel, 1977, p. 272) that if a matrix A satisfying Equation A.17 exists such that the covariance matrix S_C is a diagonal matrix, then the matrices S_C, S_Y, and A are related by the equation

$$S_C = A' S_Y A. \tag{A.19}$$

It further turns out that because the variance-covariance matrix S_Y is symmetric, the matrix A satisfying Equation A.19 is guaranteed to exist. It further turns out that the matrix A satisfies

$$A^{-1} = A', \tag{A.20}$$

that is, its inverse is equal to its transpose. The properties represented in Equations A.19 and A.20 are a consequence of the symmetry of S_Y, and since all variance-covariance matrices are symmetric, these properties hold for any data set, no matter how many variables it contains.

The diagonal elements of S_C in Equation A.19 are called the *eigenvalues* of S_Y, and the columns of A are called the *eigenvectors*. *Eigen* is a German word that can be translated as "particular." "Eigenvalue" and "eigenvector" are hybrid German-English words that are

usually rendered in English as "characteristic value" and "characteristic vector." In general, if there is a nonsingular matrix A such that the matrices B and D satisfy

$$D = A^{-1}BA, \tag{A.21}$$

then B and D are said to be *similar*. If B is any matrix and D is a diagonal matrix, then the diagonal elements of D are the eigenvalues of B. Similarity is an equivalence relation, so any two matrices that are similar to each other have the same eigenvalues (Noble and Daniel, 1977, p. 279).

Continuing with the example begun earlier in this section, the R function `eigen()` computes the eigenvalues and eigenvectors of a matrix.

```
> print(SY <- var(P))
          Y1       Y2
Y1 0.9624212 1.757146
Y2 1.7571458 4.357569
> print(eigen.SY <- eigen(SY))
$values
[1] 5.1032143 0.2167763
$vectors
           [,1]       [,2]
[1,] 0.3906338 -0.9205462
[2,] 0.9205462  0.3906338
> A <- eigen.SY$vectors
> print (Sc <- t(A) %*% SY %*% A)
              [,1]          [,2]
[1,]  5.103214e+00 -1.053845e-16
[2,] -7.859110e-17  2.167763e-01
```

Within the roundoff error of the computation, the matrix S_C is a diagonal matrix whose elements are the eigenvalues of S_Y.

The graphical interpretation shown in Figure A.1b demonstrates the connection of eigenvalues and eigenvectors to the rotation of coordinate systems discussed above. Applying the matrix multiplication in Equation A.19 is equivalent to rotating the coordinate system from the solid axes to the dashed axes, so that the unit vectors in this new coordinate system are the eigenvectors e_1 and e_2. The matrix A in Equation A.19 is the same as the rotation matrix W in Equation A.15 and corresponds to a rotation of 67°. The following shows the first column of the rotation matrix in that equation.

```
> W <- A
> acos(W[1,1]) * 180 / pi
[1] 67.00606
> asin(W[2,1]) * 180 / pi
[1] -67.00606
```

The values of data record 11 in this new coordinate system are given by equation (A.16).

```
> print(P11 <- Y[11,])
      Y1       Y2
1.271186 1.407412
> print(C <- A %*% P11)
```

```
           [,1]
[1,]  1.7921559
[2,] -0.6204022
```

This is displayed in Figure A.1b, where the vector P_{11} is written in terms of the unit vectors e_1 and e_2 as $P_{11} = 1.792e_1 - 0.620e_2$.

Figure A.1b appears to indicate that the axis provided by the eigenvector e_1 has the property that among all rotations of the axes, it is the linear combination of the two variables having the maximum variance. This is in fact the case. The variable C_1 has the maximum variance among the possible rotations. Moreover, it turns out that the first eigenvalue, whose value to three decimal places is 4.859, is the variance of C_1. The variable C_2 is uncorrelated with C_1, and its variance is given by the second eigenvalue, which is 0.135. The sum of these variances is equal to the sum of the variances of the original variables Y_1 and Y_2 as shown here.

```
> with(eigen.SY, values[1] + values[2])
[1] 5.319991
> with(Y, var(Y1) + var(Y2))
[1] 5.319991
```

Now suppose we want to study a simplified version of the data set $\{Y_1, Y_2\}$ that consists of only one variable. It makes sense to use a variable that lies along the e_1 axis in Figure A.1b, since this would incorporate the greatest amount of variation (i.e., of information) from the original pair of variables Y_1 and Y_2. Consider the point P_{11} above. The point on the line of the e_1 axis that lies closest to P_{11} is the point Q_{11} shown in FigureA.1b. This point is called the *orthogonal projection*, or simply the *projection* of P_{11} onto e_1.

A.2 Linear Regression

A.2.1 Basic Regression Theory

The objective of regression is to predict the value of a random variable Y (the *response variable*) given one or more variables (which may or may not be random) denoted $X_1, X_2, ..., X_{p-1}$. The X_j are variously called *predictors*, *carriers*, and *explanatory variables*. The reason we write $p-1$ instead of just p will become evident later. Since the objective of the analysis in this book is generally to gain an increased understanding of process rather than to do prediction, we will use the last term of the three above for the X_j. Each explanatory variable is a vector that may take on more than one value X_{ij}, *viz.*, $X_{i1}, X_{i2}, ...X_{i,p-1}$, $i = 1, ..., n$. Each explanatory variable represents a data field, and each value represents one record (i.e., one measurement) of that data field.

We begin by discussing the case in which there is only one explanatory variable, denoted X, taking on values $X_1, X_2, ..., X_n$. This case is called *simple linear regression*. The explanatory variable X may be either a random variable, or it may take on fixed, predetermined values. The latter case is called the *Gaussian case*, and when this case applies, X is called a

mathematical variable. We will illustrate the difference between a mathematical variable and a random variable using three examples.

Example 1. An agronomist conducts an experiment in which fertilizer is applied to corn at the rate of 0, 50, 100, and 150 kg ha^{-1}. The explanatory variable X is fertilizer rate and the response Y is corn yield.

Example 2. The same agronomist interviews 20 farmers and asks them how much fertilizer they apply and what their yield is. X and Y play the same roles. Although X is not "controlled" by the experimenter, it is selected by the farmer.

Example 3. Our indefatigable agronomist randomly selects 20 corn fields and measures both the total accumulated precipitation during the growing season and the corn yield in each field. Here the explanatory variable X is annual precipitation and the response Y again is the yield.

Example 1 is clearly a Gaussian case, in which X is a mathematical variable. Example 3 is clearly a case in which X is a random variable. In example 2, X may look like a random variable, but actually it is a mathematical variable. Although the agronomist did not select the fertilization rates, the farmers did select them, and therefore they are not random.

It is easiest to develop regression theory for the Gaussian case. We assume a relationship between X and Y of the form

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,\ i = 1, \ldots, n \tag{A.22}$$

where the β_i are parameters, the X_i are known constants and the ε_i are random variables (the "error" terms). The assumptions made in classical simple linear regression are

1. The ε_i are uncorrelated
2. $E\{\varepsilon_i\} = 0$
3. $\text{var}\{\varepsilon_i\} = \sigma^2$ is constant for all i
4. The ε_i are normally distributed.

Assumption 4 is not necessary for the results derived in this section to be valid. It will be needed in the later sections, when we discuss hypothesis testing.

Given a set of points (X_i, Y_i), $i = 1, \ldots, n$, we want to fit a straight line

$$\hat{Y}_i = b_0 + b_1 X_i \tag{A.23}$$

such that the quantity $Q(b_0, b_1) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ is minimized. The quantities b_0 and b_1 called the *regression coefficients*, are the estimators of β_0 and β_1, respectively. Differentiating Q with respect to the b_i and setting the derivatives equal to zero yields, after some algebra (Kutner et al., 2005, p. 17), the *normal equations*,

$$\begin{aligned} nb_0 + b_1 \Sigma X_i &= \Sigma Y_i \\ b_0 \Sigma X_i + b_1 \Sigma X_i^2 &= \Sigma X_i Y \end{aligned} \tag{A.24}$$

These equations can then be solved for b_0 and b_1 to yield

$$b_1 = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}$$
$$b_0 = \bar{Y} - b_1\bar{X} = \frac{1}{n}(\Sigma Y_i - b_1 \Sigma X_i) \tag{A.25}$$

An estimator b_i of the parameter β_i is *unbiased* if $E\{b_i\} = \beta_i$. The Gauss-Markov theorem (Kutner et al., 2005, p. 18) states that the estimator $\hat{Y}_i$ of Equation A.23, where b_0 and b_1 satisfy Equation A.25, is the best linear unbiased estimator (or BLUE), in the sense that it has minimum variance among all unbiased linear estimators. The derivation of Equation A.24 assumes that X is a mathematical variable (the Gaussian case). Suppose X is a random variable. If both X and Y are normally distributed (the so-called *bivariate normal* case), then it turns out (Sprent, 1969, p. 8) that one may write $E\{Y \mid X = X_i\} = \mu_Y + \frac{\rho\sigma_Y}{\sigma_X}(X_i - \mu_X)$. As a result, some students have the impression that the linear regression model with random X is only valid for bivariate normal data. That is not true (Dawes and Corrigan, 1974). Theil (1971, p. 102) provides a lucid discussion of linear regression when X is a random variable. The results are summarized by Kutner et al. (2005, p. 83) as follows. The statistical results associated with the regression model discussed here are mathematically valid for random X and Y if the following two conditions are met:

1. The conditional distributions of the Y_i, given X_i, are normal and independent with mean $\beta_0 + \beta_1 X_i$ and constant variance σ^2.
2. The X_i are independent random variables whose probability distribution does not depend on β_0, β_1, or σ^2.

The exploitation of the relationship between regression and correlation does require bivariate normal data. Of course, to say that a regression model is mathematically valid does not mean that it is ecologically appropriate. Theil (1971, p. 155) gives an excellent short discussion of some of the issues associated with interpretation of regression models, and Venables (2000) provides an excellent discussion in more detail.

Given a regression model of the form (Equation A.23), we define the *error* e_i by

$$e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i). \tag{A.26}$$

The *error sum of squares, SSE*, the *regression sum of squares, SSR*, and the *total sum of squares, SST* are then given by

$$SSE = \Sigma(Y_i - \hat{Y}_i)^2$$
$$SSR = \Sigma(\hat{Y}_i - \bar{Y})^2 \tag{A.27}$$
$$SST = \Sigma(Y_i - \bar{Y})^2.$$

These sums of squares satisfy an *orthogonality relationship* (Kutner et al., 2005, p. 65)

$$SST = SSE + SSR. \tag{A.28}$$

The *coefficient of determination*, which is generally denoted r^2 for simple linear regression, is defined by

$$r^2 = \frac{SSR}{SST}. \tag{A.29}$$

From Equations A.27 and A.29, it follows that $0 \le r^2 \le 1$ and a high value of r^2 is considered to imply a "good fit," that is, a high level of linear association between X and Y. The *mean squared error*, *MSE*, is defined for simple linear regression by

$$MSE = \frac{SSE}{n-2}. \tag{A.30}$$

It turns out (Kutner et al., 2005, p. 68) that the *MSE* is an unbiased estimator of the error variance σ^2, that is,

$$E\{MSE\} = \sigma^2. \tag{A.31}$$

A.2.2 The Matrix Representation of Linear Regression

The equations characterizing linear regression can be expressed very cleanly in matrix notation. Let

$$Y = \begin{bmatrix} Y_1 \\ \dots \\ Y_n \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, X = \begin{bmatrix} 1 & X_1 \\ \dots & \dots \\ 1 & X_n \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_n \end{bmatrix}. \tag{A.32}$$

The matrix X is called the *design matrix*. With the definitions of Equation A.32, Equation A.22 becomes

$$Y = X\beta + \varepsilon, \tag{A.33}$$

and Equations A.25, the normal equations, become

$$X'Xb = X'Y. \tag{A.34}$$

Therefore, assuming $(X'X)^{-1}$ exists, Equations A.25 can be written

$$b = (X'X)^{-1}X'Y. \tag{A.35}$$

Finally, if we let $\hat{Y} = Xb$, then we have

$$\hat{Y} = Xb = X(X'X)^{-1}X'Y \equiv HY. \tag{A.36}$$

Because it puts a hat on Y, the matrix H is called the *hat matrix*.

Now that we have matrix notation, it is a very simple manner to extend the theory to cover multiple regression. Consider the multiple regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_{p-1} X_{i,p-1} + \varepsilon_i,\ i = 1, \dots, n \tag{A.37}$$

If we replace Equation A.32 with

$$Y = \begin{bmatrix} Y_1 \\ \dots \\ Y_n \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_{p-1} \end{bmatrix}, X = \begin{bmatrix} 1 & X_{11} & \dots & X_{1,p-1} \\ 1 & X_{21} & \dots & X_{2,p-1} \\ \dots & \dots & \dots & \dots \\ 1 & X_{n1} & \dots & X_{n,p-1} \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_n \end{bmatrix}, \tag{A.38}$$

then all of the matrix regression equations still hold without modification. By enumerating the index of the explanatory variables from 1 to $p-1$, we have p regression coefficients, counting β_0. In the artificial rainfall-temperature data set caricatured in Section A.1, the data from the point P_{11} would be entered into the design matrix X as $X_{11,1}$ and $X_{11,2}$. It is important to add parenthetically that the term "linear regression" means that the regression is linear in the regression coefficients β_i, not necessarily in the explanatory variables X_i. Any of the explanatory variables in Equation A.37 may consist of powers or products of other explanatory variables.

Although the regression model represented by Equations A.32 through A.37 provides the best linear unbiased estimates b_i, given the data, of the regression coefficients β_i in Equation A.33, there may be problems with the data that cause it to be a poor representation of reality. Some of these problems are discussed in Chapter 8. One problem, which we discuss in Chapter 6, occurs when either one or more of the explanatory variables X or the response variable Y take on extreme values. A number of statistical measures have been devised to identify data records that have extreme values, and these are discussed in the next subsection.

A.2.3 Regression Diagnostics

One can always compute regression coefficients using Equation A.35, but the quality of the regression model as a representation of reality relies on the quality of the data as a representation of that same reality. Data are never perfect, however, and therefore some measures are required that provide an indication of whether the departures from perfection of a particular data set are sufficient to seriously reduce the usefulness of the regression model.

Although it is a bit of an oversimplification, we will in this subsection refer to a data record (X_i, Y_i) as an *outlier* if it has an extreme Y value, as a *leverage point* if it has an extreme X value (even if the corresponding Y value is consistent with the regression), and as an *influence point* if it is either an outlier or a leverage point. Outliers are discussed more fully in Section 6.2.1. Diagnostic statistics that identify influence points are called *influence measures.* One of the simplest influence measures is the set of diagonals of the hat matrix H. From Equation A.37, $H = X(X'X)^{-1}X'$, so it only depends on X and not on Y. Thus, it cannot detect outliers (unusual Y values), but it can detect leverage points (unusual X values). It turns out (Kutner et al., 2005, p. 392) that the i^{th} diagonal h_{ii} of the hat matrix provides a measure of the influence of X_i on the regression. It further turns out that the h_{ii} satisfy $\Sigma h_{ii} = p$, so a "typical" value of h_{ii} should be roughly equal to p/n. The R function `influence.measures()` identifies data records with an h_{ii} value greater than $3p/n$ as influential.

In addition to the hat matrix, the matrix $C = (X'X)^{-1}$ is also useful in regression diagnostics. This stems from the fact that the variance-covariance matrix of the vector b of regression coefficients is given by $\text{var}\{b\} = \sigma^2 C$ (Kutner et al., 2005, p. 227). The matrix C is used in two diagnostic statistics, each of which is computed using a similar procedure. This is

to delete on a record-by-record basis each data record, compute a statistic using the data set with the record deleted, and compare this to the same statistic computed with the full data set. The notation for a quantity computed with record i deleted is a subscript with the i in parentheses. Thus, for example $X_{(i)}$ denotes the design matrix X obtained by deleting data record i.

One diagnostic statistic computed from the matrix C is the *covariance ratio*, denoted *COVRATIO* (Belsley et al., p. 22). This is defined as

$$COVRATIO_i = \frac{MSE \det[(X'X)^{-1}]}{MSE_{(i)} \det[(X'_{(i)}X_{(i)})^{-1}]}, \tag{A.39}$$

where $MSE_{(i)}$ is the value of *MSE* computed when the i^{th} data record is deleted. As with other diagnostic measures, it turns out that the covariance ratio can be computed using a much simpler formula than that of the definition (Belsley et al., p. 22). The R function `influence.measures()` identifies data records with a covariance ratio value greater than $3p/n$ as influential.

The *DFBETAS* diagnostic statistic also makes use of the C matrix. It measures the influence of each data record on each regression coefficient (Kutner et al., 2005, p. 404). Let b_k denote the k^{th} regression coefficient, $k = 0, \ldots, p-1$, and let $b_{k(i)}$ denote the value of b_k when the i^{th} data record is deleted. Let c_{kk} be the k^{th} diagonal element of the matrix $(X'X)^{-1}$. Then the variance of b_k is $\text{var}\{b_k\} = \sigma^2 c_{kk}$ (Kutner et al. 2005, p. 404). This motivates the following definition for the *DFBETAS* statistic for the effect of data record i on regression coefficient b_k:

$$DFBETAS_{k(i)} = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)} c_{kk}}}, \tag{A.40}$$

The R function `influence.measures()` identifies data records with a *DFBETAS* value greater than 1 as influential.

The *DFFITS* statistic measures the effect that data record i has on the fitted value $\hat{Y}_i$. It turns out that the variance of $\hat{Y}_i$ is $\text{var}\{\hat{Y}_i\} = \sigma^2 h_{ii}$, where h_{ii} is the i^{th} diagonal of the hat matrix. This motivates the definition of $DFFITS_i$ as

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}}, \tag{A.41}$$

where, as usual $\hat{Y}_{i(i)}$ is the value of $\hat{Y}_i$ computed when the i^{th} data record is deleted. The R function `influence.measures()` identifies data records with a *DFFITS* whose absolute value is greater than $3\sqrt{p/(n-p)}$ as influential.

The *DFFITS* statistic measures the effect of data record i on the value of $\hat{Y}_i$. By contrast, the Cook's distance measures the effect of data record i on the entire regression model. The Cook's distance is defined as (Kutner et al., 2005, p. 402)

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_{ji} - \hat{Y}_{j(i)})^2}{pMSE}. \tag{A.42}$$

The R function `influence.measures()` identifies as influential data records with a Cook's distance for which the percentile of the distribution of the F percentile with p and $n-p$ degrees of freedom has a value greater than 0.5.

To illustrate the calculation influence measures, we will generate two normally distributed random variables according to the following code.

```
> set.seed(123)
> X1 <- rnorm(30)
> X2 <- 2 * X1 + 1.3 * rnorm(30)
```

Figure A.2a shows a plot of Y vs. X together with the least squares regression line. Applying the function `influence.measures()` to this regression yields the following.

```
> YX.lm <- lm(Y ~ X)
> im.lm <- influence.measures(YX.lm)
> im.lm$is.inf
   dfb.1_ dfb.X dffit cov.r cook.d hat
1   FALSE FALSE FALSE FALSE FALSE FALSE
2   FALSE FALSE FALSE FALSE FALSE FALSE
3   FALSE FALSE FALSE FALSE FALSE FALSE
4   FALSE FALSE FALSE FALSE FALSE FALSE
5   FALSE FALSE FALSE FALSE FALSE FALSE
6   FALSE FALSE FALSE FALSE FALSE FALSE
7   FALSE FALSE FALSE FALSE FALSE FALSE
8   FALSE FALSE FALSE FALSE FALSE FALSE
9   FALSE FALSE FALSE FALSE FALSE FALSE
10  FALSE FALSE FALSE FALSE FALSE FALSE
11  FALSE FALSE FALSE FALSE FALSE FALSE
12  FALSE FALSE FALSE FALSE FALSE FALSE
13  FALSE FALSE FALSE FALSE FALSE FALSE
14  FALSE FALSE FALSE FALSE FALSE FALSE
15  FALSE FALSE FALSE FALSE FALSE FALSE
16  FALSE FALSE FALSE FALSE FALSE FALSE
17  FALSE FALSE FALSE FALSE FALSE FALSE
18  FALSE FALSE FALSE TRUE FALSE FALSE
19  FALSE FALSE FALSE FALSE FALSE FALSE
20  FALSE FALSE FALSE FALSE FALSE FALSE
```

Point 18, which lies very close to the regression line, is identified as a potential influence point by the covariance ratio, which indicates that deleting this point results in a noticeable change in the variance-covariance matrix of the regression coefficients. This (presumably) false positive shows that not every data record identified as suspicious should be automatically treated as an influence point.

As a first example of an influence point, we create an outlier, but one whose X value is close to $\bar{X}$. Here is the code.

```
> set.seed(123)
> X <- rnorm(20)
> Y <- X + 0.5 * rnorm(20)
> Y[2] <- Y[2] + 2
> YX.lm <- lm(Y ~ X)
> im.lm <- influence.measures(YX.lm)
> im.lm$is.inf
   dfb.1_ dfb.X dffit cov.r cook.d hat
1   FALSE FALSE FALSE FALSE FALSE FALSE
2    TRUE FALSE TRUE TRUE FALSE FALSE
```

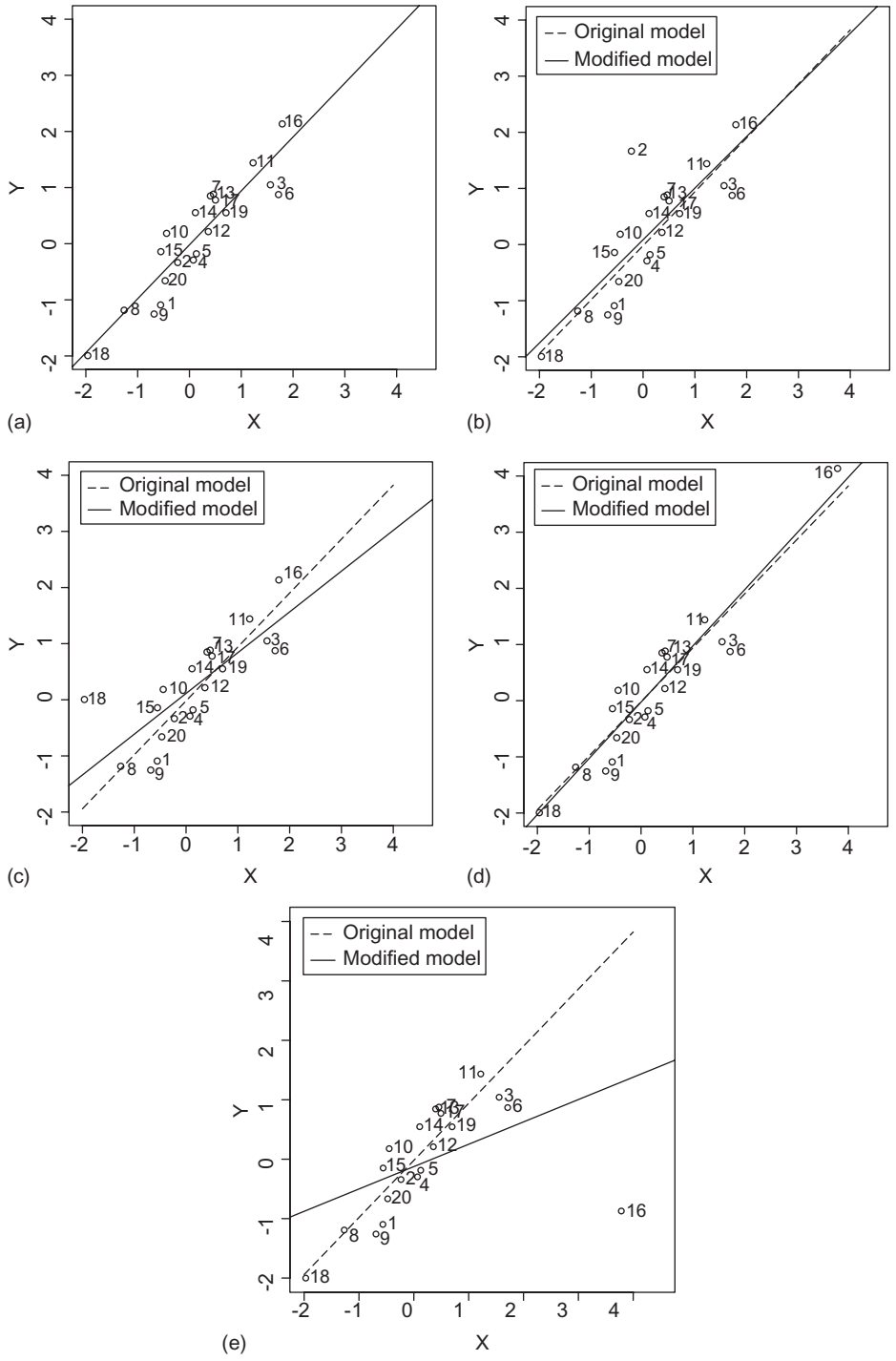

FIGURE A.2
A series of regression lines showing the effects of influence points: (a) no influence points; (b) an outlier near the median of the explanatory variables; (c) an influence point; (d) a point that is extreme in both the explanatory and response variables but consistent with the regression; (e) an outlier that is at an extreme value of the explanatory variables.

Figure A.2b shows the effect of this change: point 2 is clearly an extreme Y value, but the effect on the regression line is small. When a Y value corresponding to a more extreme X value is changed by the same amount, the effect is more dramatic (Figure A.2c).

```
> set.seed(123)
> X <- rnorm(20)
> Y <- X + 0.5 * rnorm(20)
> Y[18] <- Y[18] + 2
> YX.lm <- lm(Y ~ X)
> im.lm <- influence.measures(YX.lm)
> im.lm$is.inf
   dfb.1_ dfb.X dffit cov.r cook.d hat
   *    *    * DELETED    *    *    *
18 TRUE TRUE TRUE TRUE TRUE FALSE
```

This illustrates the reason for the term "leverage." As with a mechanical system, a point farther out from the middle exerts more "leverage" than a point closer to the middle. This is further illustrated in Figure A.2d, which is generated by the following code:

```
> set.seed(123)
> X <- rnorm(20)
> X[16] <- X[16] + 2
> Y <- X + 0.5 * rnorm(20)
> YX.lm <- lm(Y ~ X)
> im.lm <- influence.measures(YX.lm)
> im.lm$is.inf
   dfb.1_ dfb.X dffit cov.r cook.d hat
   *    *    * DELETED    *    *    *
16 FALSE TRUE TRUE TRUE FALSE TRUE
```

Point 16 is consistent with the regression model (the adjustment to the X value is made *before* the Y value is computed), so this point does not change the regression line much, although it is identified as an influence point. The alteration created by a data record that is both an outlier and a leverage point has the greatest impact of all of the changes (Figure A.2e).

```
> set.seed(123)
> X <- rnorm(20)
> Y <- X + 0.5 * rnorm(20)
> X[16] <- X[16] + 2
> Y[16] <- Y[16] - 3
> YX.lm <- lm(Y ~ X)
> im.lm <- influence.measures(YX.lm)
> im.lm$is.inf
   dfb.1_ dfb.X dffit cov.r cook.d hat
   *    *    *   DELETD    *    *    *
16 FALSE TRUE TRUE TRUE TRUE TRUE
```

A.2.4 Regression and Causality

Finally, it is important to note that a high level of linear association between X and Y does not imply a causal relationship between X and Y. This should be evident if the roles of X and Y in Example 2 in Section A.2.1 above are reversed. One would not expect that the

fact that a strong linear relationship between the explanatory variable corn yield X and annual rainfall Y implies that the corn yield causes the rainfall. One of the more amusing examples of the misuse of correlation and regression to infer causality is the Theory of the Stork. There are a number of data sets, among them those of Box et al. (1978. p.8) and Höfer et al. (2004), that show a high level of linear association between the stork population in a given area and the birthrate in that area.

A.3 Nested Models and the General Linear Test

The *general linear test* is a means of testing a given null hypothesis involving the ordinary least squares regression model (Kutner et al. 2005, p. 72; Searle, 1971, p. 110). Suppose for example that, given a model of the form of Equation A.22, one wishes to test the null hypothesis $H_0 : \beta_1 = 0$ against the alternative hypothesis $H_a : \beta_1 \neq 0$. One regards Equation A.22 as the *full* model, denoted by the subscript F, and develops a model called the *restricted* model, denoted by the subscript R, in which the null hypothesis is satisfied. In our example the restricted model is

$$Y_i = \beta_0 + \varepsilon_i. \tag{A.43}$$

The restricted model of Equation A.43 is said to be *nested* in the full model of Equation A.22 because the model of Equation A.43 can be derived from the full model of Equation A.22 by restricting one or more of the parameters of the full model to certain values. In this particular example, the value of β_1 is restricted to $\beta_1 = 0$.

To carry out the general linear test, one computes the sums of squares of errors for both models and then computes the statistic

$$G = \frac{SSE_R - SSE_F}{df_R - df_F} \div \frac{SSE_F}{df_f}, \tag{A.44}$$

where SSE_i and df_i are the sums of squares of errors and the degrees of freedom of model i, $i = R, F$. If the null hypothesis is true, then the statistic G has an F distribution with $df_r - df_F$ and df_F degrees of freedom, and thus one can compute a p value based on this distribution. However, the statistic G has an F distribution only if the Y_i are independent (Kutner et al. 2005, p. 699). A consequence of the nonzero covariance structure of spatially autocorrelated errors is therefore that the general linear test on the sums of squares can no longer be used to test H_0. An alternative means of carrying out this test, the maximum likelihood method, remains valid when the response variables are not independent. This will be discussed in Section A.5.

A.4 The Method of Lagrange Multipliers

The Lagrange multiplier method is used to solve constrained optimization problems, that is, problems in which the objective is to find the maximum or minimum value of a scalar function $f(Y_1, Y_2, ..., Y_n)$ where the x_i are also required to satisfy some constraint of

the form $g(Y_1, Y_2, ..., Y_n) = 0$. For simplicity, we will restrict the explanation to the case of two independent variables Y_1 and Y_2. The extension to the more general case is straightforward. We will only deal with the problem of finding *necessary* conditions for the maximum or minimum, and, to keep the explanation short, we will refer to this as "solving" the problem. The basic idea of the Lagrange multiplier is as follows (Sokolnikoff and Redheffer, 1966, p. 342). Suppose $f(Y_1, Y_2)$ is a scalar function of two variables whose derivatives in both variables exist everywhere, and suppose one wishes to find the values of Y_1 and Y_2 that maximize this function. A necessary (but not sufficient) condition that the values of Y_1 and Y_2 occur at a maximum of f is that they satisfy the equations

$$\begin{aligned} \frac{\partial f}{\partial Y_1} &= 0 \\ \frac{\partial f}{\partial Y_2} &= 0. \end{aligned} \tag{A.45}$$

Now suppose that we impose the condition that Y_1 and Y_2 maximize the function $f(Y_1, Y_2)$ subject to a constraint of the form $g(Y_1, Y_2) = 0$. To solve this, one can define a new function $F(Y_1, Y_2, \psi) = f(Y_1, Y_2) + \psi\, g(Y_1, Y_2)$. The term ψ is called a *Lagrange multiplier*. Since we require $g(Y_1, Y_2) = 0$, any values Y_1 and Y_2 that maximize $f(Y_1, Y_2)$ must also maximize $F(Y_1, Y_2, \psi)$. But $F(Y_1, Y_2, \psi)$ does not have any constraints on it; instead, it has the extra variable ψ. Therefore, the problem of finding necessary condition for a maximum of $f(Y_1, Y_2)$ subject to the constraint $g(Y_1, Y_2) = 0$ is equivalent to solving the unconstrained problem in three variables

$$\begin{aligned} \frac{\partial F}{\partial Y_1} &= \frac{\partial f}{\partial Y_1} + \psi \frac{\partial g}{\partial Y_1} = 0 \\ \frac{\partial F}{\partial Y_2} &= \frac{\partial f}{\partial Y_2} + \psi \frac{\partial g}{\partial Y_2} = 0 \\ \frac{\partial F}{\partial \psi} &= g(Y_1, Y_2) = 0. \end{aligned} \tag{A.46}$$

Note that the constraint itself is recovered in the last equation. The problem of finding necessary conditions for a minimum of f subject to a constraint is solved in the same way since Equation A.46 is also a necessary constraint for a minimum. Thus, Lagrange multipliers can be used to convert a constrained optimization problem, which generally cannot be solved by simply setting the derivatives to zero, into an unconstrained optimization problem in one extra variable.

A.5 The Maximum Likelihood Method

A.5.1 The Likelihood Function

The method of maximum likelihood was worked out by R.A. Fisher over a period of years between 1912 and 1922 (Aldrich, 1997). Suppose there are a set of observations $Y_1, \ldots Y_n$, each having the same probability density $f(Y, \theta)$, where θ is a parameter or vector of parameters

such as the expected value μ and/or variance σ^2. The joint probability distribution, when considered as a function of θ, is called the *likelihood function*, and written as

$$L(\theta|Y) = f(\theta, Y_1, ..., Y_n). \tag{A.47}$$

The values of the observations $Y_1, ..., Y_n$ are considered to be parameters of the function L.

As a very simple example, we will work out the likelihood function for the problem of estimating the mean and variance of a data set, so that the model is $Y_i = \mu + \varepsilon_i$, where the ε_i are independent, identically distributed random variables satisfying $\varepsilon_i \sim N(0, \sigma^2)$. Thus, in this case $\theta = (\mu, \sigma^2)$. At the very beginning, we run into a problem. This is that we wish to compute the likelihood function $L(\theta|Y)$, expressed in terms of the random variable Y, but the random variable whose distribution we know is ε, the error term. We therefore need to transform the variables of the probability distribution f in Equation A.47 from the ε variables to the Y variables. In the present case this is easy. The mean μ is fixed, and the ε_i are the only random variables in the model. Therefore, the Y_i have the same distribution as the ε_i, and the Y_i are also independent and normally distributed with mean μ and variance σ^2. The fact that the Y_i are independent permits us to write the likelihood as a product of normal density functions, but if they were not independent (as they are not in the spatial autocorrelation case), then in order to write the likelihood as a product we would have to transform the variables from the Y_i into the ε_i.

The probability density function for the random variable ε_i is

$$f(\varepsilon_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\varepsilon_i^2 / 2\sigma^2). \tag{A.48}$$

Since the ε_i are independent, the joint density function $f(\varepsilon_1, ..., \varepsilon_n)$ is given by

$$f(\varepsilon) = \prod_{i=1}^{n} f(\varepsilon_i, 0, \sigma^2). \tag{A.49}$$

where $f(\varepsilon_i, 0, \sigma^2)$ is the normal density function with mean 0 and variance σ^2. To express the likelihood in terms of Y, we substitute $\varepsilon_i = Y_i - \mu$ to get

$$\begin{aligned} L(\mu, \sigma^2 \mid Y) &= f(Y \mid \mu, \sigma^2) \\ &= \prod_{i=1}^{n} f(Y_i, \mu, \sigma^2) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp(-\Sigma(Y_i - \mu)^2 / 2\sigma^2). \end{aligned} \tag{A.50}$$

The fundamental principle of the maximum likelihood method is that estimates of the parameters of the model can be obtained by maximizing the likelihood function (Equation A.50). The likelihood function itself has a complex form. However, in the case of the normally distributed random variables in Equation A.50, the task of finding the maximum is simplified by maximizing the *log* likelihood, defined as

$$l(\theta \mid Y) = \log L(\theta \mid Y). \tag{A.51}$$

Since the logarithm is a monotonic function, l and L will attain their maxima at the same value of θ.

We now derive the log likelihood estimates for the parameters of our model $Y_i = \mu + \varepsilon_i$. The log likelihood function of Equation A.51 is (Theil, 1971, p. 90)

$$l(\mu, \sigma^2 \mid Y) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mu - \alpha_i)^2, \tag{A.52}$$

and therefore

$$\begin{aligned} \frac{\partial l}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \mu) \\ \frac{\partial l}{\partial \sigma^2} &= -\frac{n/2}{\sigma^2} + \frac{\Sigma(Y_i - \mu)}{2\sigma^4}. \end{aligned} \tag{A.53}$$

Solving the first equation for μ yields $\mu = \bar{Y}$. Substituting this into the second equation yields

$$\sigma^2_{ML} = \frac{1}{n}\sum(Y_i - \bar{Y})^2 \tag{A.54}$$

Equation A.54 illustrates an important fact, (Kutner, 2005, p. 34), namely, that the maximum likelihood estimator of σ^2 is biased. Kendall and Stuart (1979, p. 90) show that the least squares estimate of σ^2 is the unbiased estimate

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \tag{A.55}$$

The fact that the maximum likelihood method yields a biased estimate for σ^2 is a problem that will occasionally have to be dealt with. The great advantage of the method is that we can apply it in situations where the method of least squares breaks down. One of these situations occurs in the mixed-model formulation discussed in Chapter 12, in which the values of the response variable Y_{ij} are correlated. A second situation where the method can be used is the spatial regression case discussed in Chapter 13, in which the error terms ε_{ij} are autocorrelated.

A.5.2 The Likelihood Ratio Test

In the hypothesis tests involving the mixed model discussed in Chapter 12, in which the response variables Y_{ij} are correlated, the general linear test statistic defined by Equation A.44 does not have an F distribution, and so the test cannot be easily applied. Therefore, a different test is required, and the likelihood ratio test is frequently used. Consider first the case where the parameter θ under consideration is a scalar rather than a vector (e.g., suppose $\theta = \mu$). Suppose the null hypothesis is $H_0 : \theta = \theta_0$ and the alternative is $H_a : \theta \neq \theta_0$. It turns out that under very general conditions the likelihood function $L(\theta|Y)$ is asymptotically

normally distributed, that is, as the number n of observations increases the distribution of $L(\theta \mid Y)$ approaches normality (Theil 1971, p. 393). Let $\hat{\theta}$ be the maximum likelihood estimator of θ based on the observed data. For example, in the case $\theta = \mu$ the maximum likelihood estimator is the sample mean $\bar{Y}$. The likelihood ratio statistic

$$G \equiv -2\log(L(\theta)/L(\hat{\theta})) \tag{A.56}$$

is asymptotically distributed as χ_1^2, a chi square with one degree of freedom. Thus, the likelihood ratio test of the null hypothesis $H_0 : \theta = \theta_0$ consists of calculating G and comparing it with the percentiles of the chi-square probability distribution. This is an example of a likelihood ratio test. The test is only valid asymptotically, and McCulloch et al. (2008, p. 33) give an example in which the distribution of G is quite far from chi square for relatively moderate values of the sample size n. Nevertheless, the likelihood ratio test is very widely used.

More generally, suppose θ is a vector of parameters and we wish to compare a full model with a restricted model. For example, consider the ANCOVA model discussed in Chapter 12. In this case the full model is given by

$$Y_{ij} = \mu + \alpha_i + \beta(X_{ij} - \bar{X}_i) + \varepsilon_{ij}, \tag{A.57}$$

where the α_i are random effects and μ and β are fixed effects, and the restricted model is given by restricting $\beta = 0$ to obtain

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \tag{A.58}$$

Note that the restricted model is nested in the full model, since it is obtained by restricting a parameter of the full model. Suppose in general that there are k_F parameters in the full model and k_R in the restricted model ($k_F > k_R$). In our example, $k_F = 3$ and $k_R = 2$. Let L_F be the estimated value of the maximum likelihood of the full model, and let L_R of the restricted model. Then asymptotically in n the quantity

$$G = -2\log(L_F / L_R) = 2(l_F - l_R) \tag{A.59}$$

has a chi-square distribution with $(k_F - k_R)$ degrees of freedom (Kutner et al., 2005, p. 580; Theil, 1971, p. 396). In our case, $l_F = l(\mu, \alpha, \beta, \sigma^2 \mid Y)$, $l_R = l(\mu, \alpha, \sigma^2 \mid Y)$, and $k_F - k_R = 1$.

The likelihood ratio test is implemented by computing the maxima of the log likelihood statistics l_F and l_R and subtracting them. The computation is carried out by determining the values θ_i * of θ that maximize $l_i(\theta|z)$ for $i = R$ and F. If θ is a vector of k parameters (e.g., if $\theta = [\mu, \alpha_1, \ldots, \alpha_n, \beta, \sigma^2]'$, then $k = n + 3$), then to maximize $l_i(\theta \mid z)$ one must solve the equations

$$\frac{\partial l_i}{\partial \theta_j} = 0,\ i = R,\ S,\ \ j = 1, \ldots, k \tag{A.60}$$

In simple cases, these equations can be solved analytically. In more complicated cases, however, an approximate solution must be computed numerically.

One type of algorithm for the solution of problems such as these is called *hill climbing*, because if the values of the function $l_i(\theta|z)$ are considered as a surface or "hill" in the space of the variables $\theta_1, \ldots, \theta_k$, then the solution $\partial l_i / \partial \theta_j = 0$ is at the top of the hill (Figure A.3). Since the computer can only evaluate the function at specific values of $\theta_1, \ldots, \theta_k$, finding the solution is like using an altimeter to climb a hill in a thick fog. You know how high you

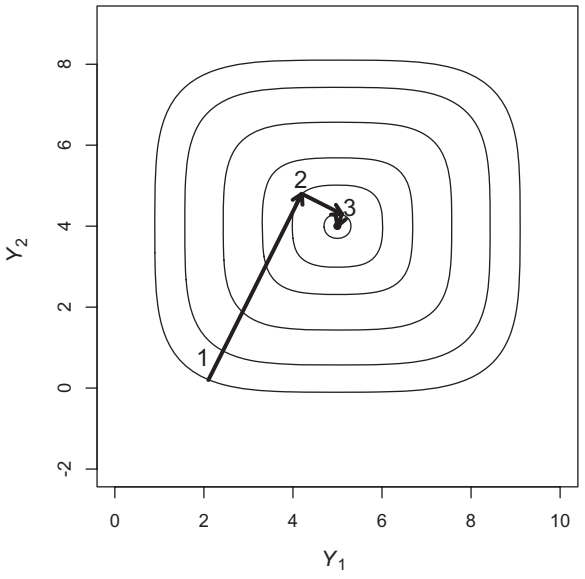

FIGURE A.3
An illustration of the Newton-Raphson hill climbing procedure.

are, but you can't see the quickest way to the top. The most common solution algorithm for a hill climbing problem is some form of the *Newton-Raphson* algorithm (Press et al., 1986, p. 269). To see how algorithms of this form work, imagine that you are at point 1 on the side of the likelihood hill in parameter space as shown in Figure A.3. You are in a thick fog so that you cannot see anything, but you are equipped with a compass and an altimeter. A good procedure to get to the top of the hill would be to take a small step in four perpendicular directions and based on this to estimate the direction of steepest ascent of the hill. This direction is called the *gradient*. Suppose you estimate that the direction of steepest ascent is along the dotted line passing through point 1. You move along this line, using your compass to maintain direction, counting paces and stopping every few paces to read your altimeter. Ultimately you will reach a crest, and the next altimeter reading will be less than the last. At this point, you stop, compute the gradient direction from this point, and repeat the process. As you get closer and closer to the top of the hill, the distance you travel before you reach a point where you change direction will decrease, until (after the third iteration in the figure) it becomes very small. When this distance is smaller than some predetermined value, you decide that you have reached the top of the hill and stop. If the likelihood function has the simple form shown in Figure A.3, then this iterative process is sure to converge, but if the function has a more complex form, then the algorithm may "get lost" and fail to converge. This occasionally happens when one attempts to compute the maximum likelihood. In addition, the computation of the gradient can also run into numerical difficulties.

A.5.3 Application of Maximum Likelihood to Linear Regression

The application of the theory of maximum likelihood to linear regression is a straightforward extension of the procedure discussed in Sections A.5.1 and A.5.2. We develop it here for

the simple linear regression model (Equation A.22) based on the discussion in Kutner et al. (2005, p. 27). By analogy with Equation A.50, the likelihood function for the simple regression model, assuming independent, normally distributed errors with constant variance, is

$$L(\beta_0, \beta_1, \sigma^2 \mid Y) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp(-\sum_{i=1}^{n} \Sigma(Y_i - \beta_0 - \beta_1 X_i)^2 / 2\sigma^2). \tag{A.61}$$

The log likelihood is therefore

$$l(\beta_0, \beta_1, \sigma^2 \mid Y) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2. \tag{A.62}$$

Taking partial derivatives with respect to the variables gives

$$\begin{aligned} \frac{\partial l}{\partial \beta_0} &= \frac{1}{\sigma^2} \sum_{1}^{n} (Y_i - \beta_0 - \beta_1 X_i) = 0 \\ \frac{\partial l}{\partial \beta_1} &= \frac{1}{\sigma^2} \sum_{1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i) = 0 \\ \frac{\partial l}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 = 0. \end{aligned} \tag{A.63}$$

The solution to these equations is

$$\begin{aligned} \sum_{1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) &= 0 \\ \sum_{1}^{n} X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) &= 0 \\ \sum_{1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 &= n\hat{\sigma}^2. \end{aligned} \tag{A.64}$$

The first two of these equations are the same as the normal Equations A.25 and indicate that the values $\hat{\beta}_i$ obtained by the method of maximum likelihood are equal to those obtained by least squares. The third equation indicates that the maximum likelihood estimate of the error variance is again biased, $\hat{\sigma}^2 = (n-1)s^2/n$.

The testing of a null hypothesis such as $H_0 : \beta_1 = 0$ via the likelihood ratio test is a straightforward extension of the method developed in Section A.5.2.

A.5.4 Maximum Likelihood and Restricted Maximum Likelihood

There is one further complication that must be discussed: the use of the *restricted maximum likelihood* (REML) method vs. the ordinary maximum likelihood (ML) method. As discussed above (Equation A.54), the maximum likelihood estimate of the variance of the simple mean estimation problem is biased, and indeed is $(n-1)/n$ times the unbiased

estimate and therefore always biased downward. This tendency of the maximum likelihood method to underestimate the variance persists in the application to the mixed model. The REML is an alternative method (Pinhiero and Bates, 2000, p. 75) that provides better variance estimates, and also has some computational advantages. The REML works by splitting the solution of the model into two stages. In the first stage, the fixed terms, for example, the terms μ and β in the full model of Equation 12.6, and the term μ of the restricted model of Equation 12.5, are in effect removed from the model by assuming that they have a particularly simple probability distribution and integrating them over this distribution, which removes them from the calculation. The random effects estimates are then computed by solving the maximum likelihood equations under this simplifying assumption. In the second stage, these estimates of the random effects parameters are plugged back into the model and used to compute the estimates of the fixed effects parameters. While the REML method generally computes improved variance estimates, it cannot be used in a likelihood ratio test of a fixed-effect term. This is because the full and restricted models are computed using different variance structures and are therefore not comparable (Pinhiero and Bates, 2000, p. 76). Thus, we cannot use the REML to test the null hypothesis $\beta = 0$ in the model of Equation 12.6, because it is a fixed effect and must instead use the full maximum likelihood method.

A.6 Change of Variables of a Probability Density

In maximizing the likelihood function, moving from Equation A.48 to Equation A.50 requires a change of variables. Viewed as a probability density function f of the random variables ε_i rather than a likelihood function L of the parameters of the model, f depends on n variables ε_i. In order to get some insight for how to transform this to a function of the Y_i, we consider a very simple case, in which there is only one variable Y, and thus in Equation A.48 ε is a scalar. However, we will consider a more general form for the relation between Y and ε. The following argument is taken from Johnston (1972, p. 374). Assume X is a mathematical variable and let

$$Y = X\beta + g(\varepsilon). \tag{A.65}$$

Suppose ε has a probability density function $\varphi(\varepsilon)$. Our task is to express the likelihood function L in terms of Y rather than ε while maintaining its properties as a probability density function. Suppose g is monotonic in some region (i.e., some range of values of ε, so that g^{-1} exists in that region, and let $h = g^{-1}$, so that $\varepsilon = h(Y)$. Suppose that the relation between Y and ε in this region is shown in Figure A.4. Whenever ε lies in the region $\Delta\varepsilon$ shown in the figure, Y will lie in the corresponding region ΔY. Thus, we can write the following probability equation:

$$\Pr\{Y \text{ lies in } \Delta\Psi\} = \Pr\{\varepsilon \text{ lies in } \Delta\varepsilon\}. \tag{A.66}$$

We have

$$\psi(Y)\Delta Y = \phi(\varepsilon)\Delta\varepsilon. \tag{A.67}$$

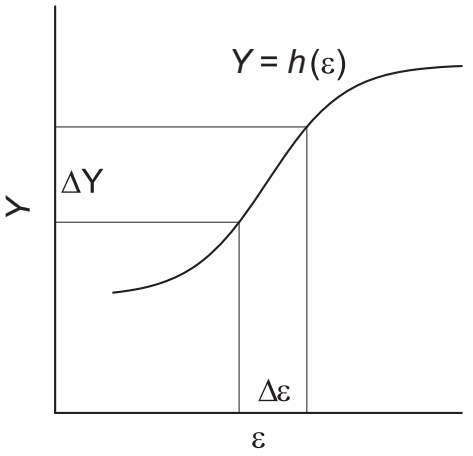

FIGURE A.4
Functional relationship between Y and ε.

where $\psi(Y)$ is the probability density expressed in terms of Y. Therefore,

$$\psi(Y) = \phi(\varepsilon)\frac{\Delta\varepsilon}{\Delta Y}. \tag{A.68}$$

Taking the limit for an infinitesimally small Δ leads to

$$\psi(Y) = \phi(\varepsilon)\frac{d\varepsilon}{dY} = \phi(h(Y))h'(Y). \tag{A.69}$$

where $h' = dh / dY$. In the figure, $h(Y)$ is an increasing function, so that h' is positive. If h' were negative, the same argument would hold except that, since the probability density must be positive, we would have $\psi(Y) = -\varphi(h(Y))h'(Y)$. Since we always multiply ϕ by a positive value, we can write the general equation

$$\psi(Y) = \varphi(h(Y))\left|h'(Y)\right|. \tag{A.70}$$

Now suppose ε and Y are vector functions with n components each. In this case, the formula is very similar (Hoel, 1971, p. 378). Suppose again that $\varepsilon = h(Y)$ locally, where now $h(Y)$ is an n dimensional function of the n dimensional vector Y. Then Equation A.70 becomes

$$\psi(\eta) = \varphi(h(Y))\left|J\right|, \tag{A.71}$$

where J is called the *Jacobian determinant* and is given by

$$J = \det\begin{bmatrix} \dfrac{\partial h_1}{\partial Y_1} & \dfrac{\partial h_1}{\partial Y_2} & \dfrac{\partial h_1}{\partial Y_n} \\ \dfrac{\partial h_2}{\partial Y_1} & \dfrac{\partial h_2}{\partial Y_2} & \dfrac{\partial h_2}{\partial Y_n} \\ \dfrac{\partial h_1}{\partial Y_n} & \dfrac{\partial h_2}{\partial Y_n} & \dfrac{\partial h_n}{\partial Y_n} \end{bmatrix}. \tag{A.72}$$