9

Selected Applications in Statistics

Data come in many forms. In the broad view, the term “data” embraces all representations of information or knowledge. There is no single structure that can efficiently contain all of these representations. Some data are in free-form text (for example, the Federalist Papers, which was the subject of a famous statistical analysis), other data are in a hierarchical structure (for example, political units and subunits), and still other data are encodings of methods or algorithms. (This broad view is entirely consistent with the concept of a “stored-program computer”; the program is the data.)

Several of the results in this chapter have already been presented in Chap. 8 or even in previous chapters, for example, the smoothing matrix H_λ that we discuss in Sect. 9.3.8 has already been encountered on page 364 in Chap. 8. The purpose of the apparent redundancy is to present the results from a different perspective. (None of the results are new; all are standard in the statistical literature.)

9.1 Structure in Data and Statistical Data Analysis

Data often have a logical structure as described in Sect. 8.1.1; that is, a two-dimensional array in which columns correspond to variables or measurable attributes and rows correspond to an observation on all attributes taken together. A matrix is obviously a convenient object for representing numeric data organized this way. An objective in analyzing data of this form is to uncover relationships among the variables, or to characterize the distribution of the sample over $\mathbb{R}^m$. Interesting relationships and patterns are called “structure” in the data. This is a different meaning from that of the word used in the phrase “logical structure” or in the phrase “data structure” used in computer science.

J.E. Gentle, *Matrix Algebra*, Springer Texts in Statistics,
DOI 10.1007/978-3-319-64867-5_9

Another type of pattern that may be of interest is a temporal pattern; that is, a set of relationships among the data and the time or the sequence in which the data were observed.

The objective of this chapter is to illustrate how some of the properties of matrices and vectors that were covered in previous chapters relate to statistical models and to data analysis procedures. The field of statistics is far too large for a single chapter on "applications" to cover more than just a small part of the area. Similarly, the topics covered previously are too extensive to give examples of applications of all of them.

A probability distribution is a specification of the stochastic structure of random variables, so we begin with a brief discussion of properties of multivariate probability distributions. The emphasis is on the multivariate normal distribution and distributions of linear and quadratic transformations of normal random variables. We then consider an important structure in multivariate data, a linear model. We discuss some of the computational methods used in analyzing the linear model. We then describe some computational method for identifying more general linear structure and patterns in multivariate data. Next we consider approximation of matrices in the absence of complete data. Finally, we discuss some models of stochastic processes. The special matrices discussed in Chap. 8 play an important role in this chapter.

9.2 Multivariate Probability Distributions

Most methods of statistical inference are based on assumptions about some underlying probability distribution of a random variable. In some cases these assumptions completely specify the form of the distribution, and in other cases, especially in nonparametric methods, the assumptions are more general. Many statistical methods in estimation and hypothesis testing rely on the properties of various transformations of a random variable.

In this section, we do not attempt to develop a theory of probability distribution; rather we assume some basic facts and then derive some important properties that depend on the matrix theory of the previous chapters.

9.2.1 Basic Definitions and Properties

One of the most useful descriptors of a random variable is its probability density function (PDF), or probability function. Various functionals of the PDF define standard properties of the random variable, such as the mean and variance, as we discussed in Sect. 4.5.3.

If X is a random variable over $\mathrm{I\!R}^d$ with PDF $p_X(\cdot)$ and $f(\cdot)$ is a measurable function (with respect to a dominating measure of $p_X(\cdot)$) from $\mathrm{I\!R}^d$ to $\mathrm{I\!R}^k$, the *expected value* of $f(X)$, which is in $\mathrm{I\!R}^k$ and is denoted by $\mathrm{E}(g(X))$, is defined by

$$\mathrm{E}(f(X)) = \int_{\mathrm{I\!R}^d} f(t) p_X(t)\,\mathrm{d}t.$$

The *mean* of X is the d-vector $\mathrm{E}(X)$, and the *variance* or *variance-covariance* of X, denoted by $\mathrm{V}(X)$, is the $d \times d$ matrix

$$\mathrm{V}(X) = \mathrm{E}\left((X - \mathrm{E}(X))\,(X - \mathrm{E}(X))^{\mathrm{T}}\right).$$

Given a random variable X, we are often interested in a random variable defined as a function of X, say $Y = g(X)$. To analyze properties of Y, we identify g^{-1}, which may involve another random variable. (For example, if $g(x) = x^2$ and the support of X is $\mathrm{I\!R}$, then $g^{-1}(Y) = (-1)^\alpha \sqrt{Y}$, where $\alpha = 1$ with probability $\Pr(X < 0)$ and $\alpha = 0$ otherwise.) Properties of Y can be evaluated using the Jacobian of $g^{-1}(\cdot)$, as in equation (4.12).

9.2.2 The Multivariate Normal Distribution

The most important multivariate distribution is the multivariate normal, which we denote as $\mathrm{N}_d(\mu, \Sigma)$ for d dimensions; that is, for a random d-vector. The PDF for the d-variate normal distribution, as we have discussed before, is

$$p_X(x) = (2\pi)^{-d/2}|\Sigma|^{-1/2}\mathrm{e}^{-(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)/2}, \tag{9.1}$$

where the normalizing constant is Aitken's integral given in equation (4.75). The multivariate normal distribution is a good model for a wide range of random phenomena.

9.2.3 Derived Distributions and Cochran's Theorem

If X is a random variable with distribution $\mathrm{N}_d(\mu, \Sigma)$, A is a $q \times d$ matrix with rank q (which implies $q \le d$), and $Y = AX$, then the straightforward change-of-variables technique yields the distribution of Y as $\mathrm{N}_d(A\mu,\, A\Sigma A^{\mathrm{T}})$.

Useful transformations of the random variable X with distribution $\mathrm{N}_d(\mu, \Sigma)$ are $Y_1 = \Sigma^{-1/2}X$ and $Y_2 = \Sigma_C^{-1}X$, where Σ_C is a Cholesky factor of Σ. In either case, the variance-covariance matrix of the transformed variate Y_1 or Y_2 is I_d.

Quadratic forms involving a Y that is distributed as $\mathrm{N}_d(\mu, I_d)$ have useful properties. For statistical inference it is important to know the distribution of these quadratic forms. The simplest quadratic form involves the identity matrix: $S_d = Y^{\mathrm{T}}Y$.

We can derive the PDF of S_d by beginning with $d = 1$ and using induction. If $d = 1$, for $t > 0$, we have

$$\Pr(S_1 \le t) = \Pr(Y \le \sqrt{t}) - \Pr(Y \le -\sqrt{t}),$$

where $Y \sim \mathrm{N}_1(\mu, 1)$, and so the PDF of S_1 is

$$p_{S_1}(t) = \frac{1}{2\sqrt{2\pi t}}\left(\mathrm{e}^{-(\sqrt{t}-\mu)^2/2} + \mathrm{e}^{-(-\sqrt{t}-\mu)^2/2}\right)$$

$$
\begin{aligned}
&= \frac{\mathrm{e}^{-\mu^2/2}\mathrm{e}^{-t/2}}{2\sqrt{2\pi t}}\left(\mathrm{e}^{\mu\sqrt{t}} + \mathrm{e}^{-\mu\sqrt{t}}\right) \\
&= \frac{\mathrm{e}^{-\mu^2/2}\mathrm{e}^{-t/2}}{2\sqrt{2\pi t}}\left(\sum_{j=0}^{\infty}\frac{(\mu\sqrt{t})^j}{j!} + \sum_{j=0}^{\infty}\frac{(-\mu\sqrt{t})^j}{j!}\right) \\
&= \frac{\mathrm{e}^{-\mu^2/2}\mathrm{e}^{-t/2}}{\sqrt{2t}}\sum_{j=0}^{\infty}\frac{(\mu^2 t)^j}{\sqrt{\pi}(2j)!} \\
&= \frac{\mathrm{e}^{-\mu^2/2}\mathrm{e}^{-t/2}}{\sqrt{2t}}\sum_{j=0}^{\infty}\frac{(\mu^2 t)^j}{j!\Gamma(j+1/2)2^{2j}},
\end{aligned}
$$

in which we use the fact that

$$
\Gamma(j+1/2) = \frac{\sqrt{\pi}(2j)!}{j!2^{2j}}
$$

(see page 595). This can now be written as

$$
p_{S_1}(t) = \mathrm{e}^{-\mu^2/2}\sum_{j=0}^{\infty}\frac{(\mu^2)^j}{j!2^j}\frac{1}{\Gamma(j+1/2)2^{j+1/2}}t^{j-1/2}\mathrm{e}^{-t/2}, \tag{9.2}
$$

in which we recognize the PDF of the central chi-squared distribution with $2j+1$ degrees of freedom,

$$
p_{\chi^2_{2j+1}}(t) = \frac{1}{\Gamma(j+1/2)2^{j+1/2}}t^{j-1/2}\mathrm{e}^{-t/2}. \tag{9.3}
$$

A similar manipulation for $d = 2$ (that is, for $Y \sim \mathrm{N}_2(\mu, 1)$, and maybe $d = 3$, or as far as you need to go) leads us to a general form for the PDF of the $\chi^2_d(\delta)$ random variable S_d:

$$
p_{S_d}(t) = \mathrm{e}^{-\mu^2/2}\sum_{j=0}^{\infty}\frac{(\mu^2/2)^j}{j!}p_{\chi^2_{2j+1}}(t). \tag{9.4}
$$

We can show that equation (9.4) holds for any d by induction. The distribution of S_d is called the noncentral chi-squared distribution with d degrees of freedom and noncentrality parameter $\delta = \mu^{\mathrm{T}}\mu$. We denote this distribution as $\chi^2_d(\delta)$.

The induction method above involves a special case of a more general fact: if X_i for $i = 1, \ldots, k$ are independently distributed as $\chi^2_{n_i}(\delta_i)$, then $\sum_i X_i$ is distributed as $\chi^2_n(\delta)$, where $n = \sum_i n_i$ and $\delta = \sum_i \delta_i$. (Compare this with the result for Wishart distributions in Exercise 4.12b on page 225.)

In applications of linear models, a quadratic form involving Y is often partitioned into a sum of quadratic forms. Assume that Y is distributed as $\mathrm{N}_d(\mu, I_d)$, and for $i = 1, \ldots k$, let A_i be a $d \times d$ symmetric matrix with rank

r_i such that $\sum_i A_i = I_d$. This yields a partition of the total sum of squares $Y^{\mathrm{T}}Y$ into k components:

$$Y^{\mathrm{T}}Y = Y^{\mathrm{T}}A_1Y + \cdots + Y^{\mathrm{T}}A_kY. \tag{9.5}$$

One of the most important results in the analysis of linear models states that the $Y^{\mathrm{T}}A_iY$ have independent noncentral chi-squared distributions $\chi^2_{r_i}(\delta_i)$ with $\delta_i = \mu^{\mathrm{T}}A_i\mu$ if and only if $\sum_i r_i = d$.

This is called Cochran's theorem. Beginning on page 355, we discussed a form of Cochran's theorem that applies to properties of idempotent matrices. Those results immediately imply the conclusion above.

9.3 Linear Models

Some of the most important applications of statistics involve the study of the relationship of one variable, often called a "response variable", to other variables. The response variable is usually modeled as a random variable, which we indicate by using a capital letter. A general model for the relationship of a variable, Y, to other variables, x (a vector), is

$$Y \approx f(x). \tag{9.6}$$

In this asymmetric model and others like it, we call Y the *dependent variable* and the elements of x the *independent variables.*

It is often reasonable to formulate the model with a *systematic component* expressing the relationship and an *additive random component* or "additive error". We write

$$Y = f(x) + E, \tag{9.7}$$

where E is a random variable with an expected value of 0; that is,

$$\mathrm{E}(E) = 0.$$

(Although this is by far the most common type of model used by data analysts, there are other ways of building a model that incorporates systematic and random components.) The zero expectation of the random error yields the relationship

$$\mathrm{E}(Y) = f(x),$$

although this expression is not equivalent to the additive error model above because the random component could just as well be multiplicative (with an expected value of 1) and the same value of $\mathrm{E}(Y)$ would result.

Because the functional form f of the relationship between Y and x may contain a *parameter*, we may write the model as

$$Y = f(x; \theta) + E. \tag{9.8}$$

A specific form of this model is

$$Y = \beta^{\mathrm{T}} x + E, \tag{9.9}$$

which expresses the systematic component as a linear combination of the xs using the vector parameter β.

A model is more than an equation; there may be associated statements about the distribution of the random variable or about the nature of f or x. We may assume β (or θ) is a fixed but unknown constant, or we may assume it is a realization of a random variable. Whatever additional assumptions we may make, there are some standard assumptions that go with the model. We assume that Y and x are *observable* and θ and E are *unobservable*.

Models such as these that express an asymmetric relationship between some variables ("dependent variables") and other variables ("independent variables") are called regression models. A model such as equation (9.9) is called a linear regression model. There are many useful variations of the model (9.6) that express other kinds of relationships between the response variable and the other variables.

Notation

In data analysis with regression models, we have a set of observations $\{y_i, x_i\}$ where x_i is an m-vector. One of the primary tasks is to determine a reasonable value of the parameter. That is, in the linear regression model, for example, we think of β as an unknown variable (rather than as a fixed constant or a realization of a random variable), and we want to find a value of it such that the model fits the observations well,

$$y_i = \beta^{\mathrm{T}} x_i + \epsilon_i, \tag{9.10}$$

where β and x_i are m-vectors. (In the expression (9.9), "E" is an uppercase epsilon. We attempt to use notation consistently; "E" represents a random variable, and "ϵ" represents a realization, though an unobservable one, of the random variable. We will not always follow this convention, however; sometimes it is convenient to use the language more loosely and to speak of ϵ_i as a random variable.) The meaning of the phrase "the model fits the observations well" may vary depending on other aspects of the model, in particular, on any assumptions about the distribution of the random component E. If we make assumptions about the distribution, we have a basis for statistical estimation of β; otherwise, we can define some purely mathematical criterion for "fitting well" and proceed to determine a value of β that optimizes that criterion.

For any choice of β, say b, we have $y_i = b^{\mathrm{T}} x_i + r_i$. The r_is are determined by the observations. An approach that does not depend on any assumptions about the distribution but can nevertheless yield optimal estimators under many distributions is to choose the estimator so as to minimize some measure of the set of r_is.

Given the observations $\{y_i, x_i\}$, we can represent the regression model and the data as

$$y = X\beta + \epsilon, \tag{9.11}$$

where X is the $n \times m$ matrix whose rows are the x_is and ϵ is the vector of deviations ("errors") of the observations from the functional model. Throughout the rest of this section, *we will assume that the number of rows of X (that is, the number of observations n) is greater than the number of columns of X (that is, the number of variables m).*

We will occasionally refer to submatrices of the basic data matrix X using notation developed in Chap. 3. For example, $X_{(i_1,\ldots,i_k)(j_1,\ldots,j_l)}$ refers to the $k \times l$ matrix formed by retaining only the $i_1, \ldots, i_k$ rows and the $j_1, \ldots, j_l$ columns of X, and $X_{-(i_1,\ldots,i_k)(j_1,\ldots,j_l)}$ refers to the matrix formed by deleting the $i_1, \ldots, i_k$ rows and the $j_1, \ldots, j_l$ columns of X. We also use the notation x_{i*} to refer to the i^{th} row of X (the row is a vector, a column vector), and x_{*j} to refer to the j^{th} column of X. See page 599 for a summary of this notation.

9.3.1 Fitting the Model

In a model for a given dataset as in equation (9.11), although the errors are no longer random variables (they are realizations of random variables), they are not observable. To fit the model, we replace the unknowns with variables: β with b and ϵ with r. This yields

$$y = Xb + r. \tag{9.12}$$

We then proceed by applying some criterion for fitting.

The criteria generally focus on the "residuals" $r = y - Xb$. Two general approaches to fitting are:

- Define a likelihood function of r based on an assumed distribution of E, and determine a value of b that maximizes that likelihood.
- Decide on an appropriate norm on r, and determine a value of b that minimizes that norm.

There are other possible approaches, and there are variations on these two approaches. For the first approach, it must be emphasized that r is not a realization of the random variable E. Our emphasis will be on the second approach, that is, on methods that minimize a norm on r.

9.3.1.1 Statistical Estimation

The statistical problem is to *estimate* β. (Notice the distinction between the phrases "to *estimate* β" and "to determine a value of β that minimizes …". The mechanical aspects of the two problems may be the same, of course.) The statistician uses the model and the given observations to explore relationships between the response and the regressors. Considering ϵ to be a realization of a random variable E (a vector) and assumptions about a distribution of the random variable ϵ allow us to make statistical inferences about a "true" β.

9.3.1.2 Ordinary Least Squares

The r vector contains the distances of the observations on y from the values of the variable y defined by the hyperplane $b^{\mathrm{T}}x$, measured *in the direction of the y axis*. The objective is to determine a value of b that minimizes some norm of r. The use of the L_2 norm is called "least squares". The estimate is the b that minimizes the dot product

$$(y - Xb)^{\mathrm{T}}(y - Xb) = \sum_{i=1}^{n}(y_i - x_{i*}^{\mathrm{T}}b)^2. \tag{9.13}$$

As we saw in Sect. 6.6 (where we used slightly different notation), using elementary calculus to determine the minimum of equation (9.13) yields the "normal equations"

$$X^{\mathrm{T}}X\widehat{\beta} = X^{\mathrm{T}}y. \tag{9.14}$$

9.3.1.3 Weighted Least Squares

The elements of the residual vector may be weighted differently. This is appropriate if, for instance, the variance of the residual depends on the value of x; that is, in the notation of equation (9.7), $\mathrm{V}(E) = g(x)$, where g is some function. If the function is known, we can address the problem almost identically as in the use of ordinary least squares, as we saw on page 295. Weighted least squares may also be appropriate if the observations in the sample are not independent. In this case also, if we know the variance-covariance structure, after a simple transformation, we can use ordinary least squares. If the function g or the variance-covariance structure must be estimated, the fitting problem is still straightforward, but formidable complications are introduced into other aspects of statistical inference. We discuss weighted least squares further in Sect. 9.3.6.

9.3.1.4 Variations on the Criteria for Fitting

Rather than minimizing a norm of r, there are many other approaches we could use to fit the model to the data. Of course, just the choice of the norm yields different approaches. Some of these approaches may depend on distributional assumptions, which we will not consider here. The point that we want to emphasize here, with little additional comment, is that the standard approach to regression modeling is not the only one. We mentioned some of these other approaches and the computational methods of dealing with them in Sect. 6.7. Alternative criteria for fitting regression models are sometimes considered in the many textbooks and monographs on data analysis using a linear regression model. This is because the fits may be more "robust" or more resistant to the effects of various statistical distributions.

9.3.1.5 Regularized Fits

Some variations on the basic approach of minimizing residuals involve a kind of regularization that may take the form of an additive penalty on the objective function. Regularization often results in a shrinkage of the estimator toward 0. One of the most common types of shrinkage estimator is the ridge regression estimator, which for the model $y = X\beta + \epsilon$ is the solution of the modified normal equations $(X^{\mathrm{T}}X + \lambda I)\beta = X^{\mathrm{T}}y$. We discuss this further in Sect. 9.5.4.

9.3.1.6 Orthogonal Distances

Another approach is to define an optimal value of β as one that minimizes a norm of the distances of the observed values of y from the vector $X\beta$. This is sometimes called "orthogonal distance regression". The use of the L_2 norm on this vector is sometimes called "total least squares". This is a reasonable approach when it is assumed that the observations in X are realizations of some random variable; that is, an "errors-in-variables" model is appropriate. The model in equation (9.11) is modified to consist of two error terms: one for the errors in the variables and one for the error in the equation. The methods discussed in Sect. 6.7.3 can be used to fit a model using a criterion of minimum norm of orthogonal residuals. As we mentioned there, weighting of the orthogonal residuals can be easily accomplished in the usual way of handling weights on the different observations.

The weight matrix often is formed as an inverse of a variance-covariance matrix Σ; hence, the modification is to premultiply the matrix $[X|y]$ in equation (6.56) by the Cholesky factor Σ_{C}^{-1}. In the case of errors-in-variables, however, there may be another variance-covariance structure to account for. If the variance-covariance matrix of the columns of X (that is, the independent variables) together with y is T, then we handle the weighting for variances and covariances of the columns of X in the same way, except of course we postmultiply the matrix $[X|y]$ in equation (6.56) by T_{C}^{-1}. This matrix is $(m+1) \times (m+1)$; however, it may be appropriate to assume any error in y is already accounted for, and so the last row and column of T may be 0 except for the $(m+1, m+1)$ element, which would be 1. The appropriate model depends on the nature of the data, of course.

9.3.1.7 Collinearity

A major problem in regression analysis is collinearity (or "multicollinearity"), by which we mean a "near singularity" of the X matrix. This can be made more precise in terms of a condition number, as discussed in Sect. 6.1. Ill-conditioning may not only present computational problems, but also may result in an estimate with a very large variance.

9.3.2 Linear Models and Least Squares

The most common estimator of β is one that minimizes the L_2 norm of the vertical distances in equation (9.11); that is, the one that forms a least squares fit. This criterion leads to the normal equations (9.14), whose solution is

$$\widehat{\beta} = (X^{\mathrm{T}}X)^{-}X^{\mathrm{T}}y. \tag{9.15}$$

(As we have pointed out many times, we often write formulas that are not to be used for computing a result; this is the case here.) If X is of full rank, the generalized inverse in equation (9.15) is, of course, the inverse, and $\widehat{\beta}$ is the unique least squares estimator. If X is not of full rank, we generally use the Moore-Penrose inverse, $(X^{\mathrm{T}}X)^{+}$, in equation (9.15).

As we saw in equations (6.43) and (6.44), we also have

$$\widehat{\beta} = X^{+}y. \tag{9.16}$$

On page 293, we derived this least squares solution by use of the QR decomposition of X. In Exercises 6.5a and 6.5b we mentioned two other ways to derive this important expression.

Equation (9.16) indicates the appropriate way to compute $\widehat{\beta}$. As we have seen many times before, however, we often use an expression without computing the individual terms. Instead of computing X^{+} in equation (9.16) explicitly, we use either Householder or Givens transformations to obtain the orthogonal decomposition

$$X = QR,$$

or

$$X = QRU^{\mathrm{T}}$$

if X is not of full rank. As we have seen, the QR decomposition of X can be performed row-wise using Givens transformations. This is especially useful if the data are available only one observation at a time. The equation used for computing $\widehat{\beta}$ is

$$R\widehat{\beta} = Q^{\mathrm{T}}y, \tag{9.17}$$

which can be solved by back substitution in the triangular matrix R.

Because

$$X^{\mathrm{T}}X = R^{\mathrm{T}}R,$$

the quantities in $X^{\mathrm{T}}X$ or its inverse, which are useful for making inferences using the regression model, can be obtained from the QR decomposition.

If X is not of full rank, the expression (9.16) not only is a least squares solution but the one with minimum length (minimum Euclidean norm), as we saw in equations (6.44) and (6.45).

The vector $\widehat{y} = X\widehat{\beta}$ is the projection of the n-vector y onto a space of dimension equal to the (column) rank of X, which we denote by r_X. The vector of the model, $\mathrm{E}(Y) = X\beta$, is also in the r_X-dimensional space span(X). The projection matrix $I - X(X^{\mathrm{T}}X)^{+}X^{\mathrm{T}}$ projects y onto an $(n - r_X)$-dimensional residual space that is orthogonal to span(X). Figure 9.1 represents these subspaces and the vectors in them.

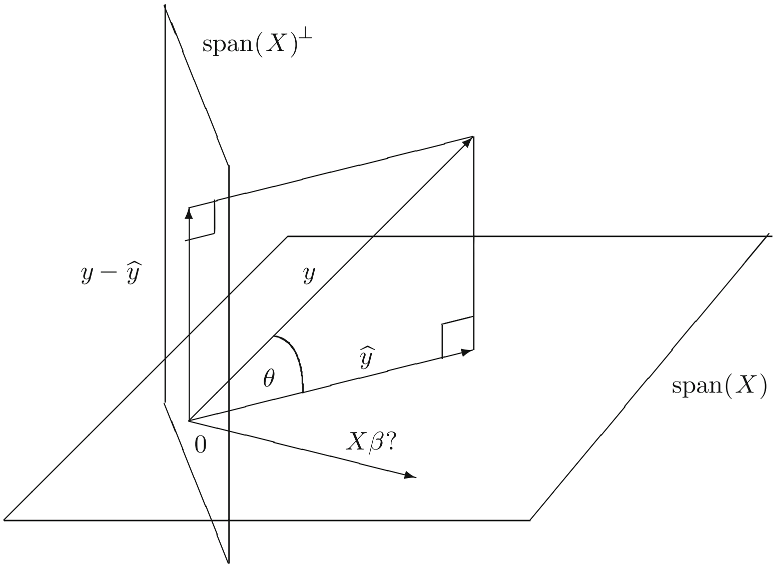

Figure 9.1. The linear least squares fit of y with X

Recall from page 291 that orthogonality of the residuals to span(X) is not only a property of a least squares solution, it actually characterizes a least squares solution; that is, if $\widehat{b}$ is such that $X^{\mathrm{T}}(y - X\widehat{b}) = 0$, then $\widehat{b}$ is a least squares solution.

In the $(r_X + 1)$-order vector space of the variables, the hyperplane defined by $\widehat{\beta}^{\mathrm{T}}x$ is the estimated model (assuming $\widehat{\beta} \neq 0$; otherwise, the space is of order r_X).

9.3.2.1 Degrees of Freedom

In general, the vector y can range freely over an n-dimensional space. We say the degrees of freedom of y, or the *total degrees of freedom*, is n. If we fix the mean of y, then the *adjusted total degrees of freedom* is $n - 1$.

The model $X\beta$ can range over a space with dimension equal to the (column) rank of X; that is, r_X. We say that the *model degrees of freedom* is r_X. Note that the space of $X\widehat{\beta}$ is the same as the space of $X\beta$.

Finally, the space orthogonal to $X\widehat{\beta}$ (that is, the space of the residuals $y - X\widehat{\beta}$) has dimension $n - r_X$. We say that the *residual (or error) degrees* *of freedom* is $n - r_X$. (Note that the error vector ϵ can range over an n-dimensional space, but because $\widehat{\beta}$ is a least squares fit, $y - X\widehat{\beta}$ can only range over an $(n - r_X)$-dimensional space.)

9.3.2.2 The Hat Matrix and Leverage

The projection matrix $H = X(X^{\mathrm{T}}X)^{+}X^{\mathrm{T}}$ is sometimes called the "hat matrix" because

$$\begin{aligned} \widehat{y} &= X\widehat{\beta} \\ &= X(X^{\mathrm{T}}X)^{+}X^{\mathrm{T}}y \\ &= Hy, \end{aligned} \tag{9.18}$$

that is, it projects y onto $\widehat{y}$ in the span of X. Notice that the hat matrix can be computed without knowledge of the observations in y.

The elements of H are useful in assessing the effect of the particular pattern of the regressors on the predicted values of the response. The extent to which a given point in the row space of X affects the regression fit is called its "leverage". The leverage of the i^{th} observation is

$$h_{ii} = x_{i*}^{\mathrm{T}}(X^{\mathrm{T}}X)^{+}x_{i*}. \tag{9.19}$$

This is just the partial derivative of $\hat{y}_i$ with respect to y_i (Exercise 9.2). A relatively large value of h_{ii} compared with the other diagonal elements of the hat matrix means that the i^{th} observed response, y_i, has a correspondingly relatively large effect on the regression fit.

9.3.3 Statistical Inference

Fitting a model by least squares or by minimizing some other norm of the residuals in the data might be a sensible thing to do without any concern for a probability distribution. "Least squares" per se is not a statistical criterion. Certain statistical criteria, such as maximum likelihood or minimum variance estimation among a certain class of unbiased estimators, however, lead to an estimator that is the solution to a least squares problem for specific probability distributions.

For statistical inference about the parameters of the model $y = X\beta + \epsilon$ in equation (9.11), we must add something to the model. As in statistical inference generally, we must identify the random variables and make some statements (assumptions) about their distribution. The simplest assumptions are that ϵ is a random variable and $\mathrm{E}(\epsilon) = 0$. Whether or not the matrix X is random, our interest is in making inference conditional on the observed values of X.

9.3.3.1 Estimability

One of the most important questions for statistical inference involves estimating or testing some linear combination of the elements of the parameter β; for example, we may wish to estimate $\beta_1 - \beta_2$ or to test the hypothesis that $\beta_1 - \beta_2 = c_1$ for some constant c_1. In general, we will consider the linear combination $l^{\mathrm{T}}\beta$. Whether or not it makes sense to estimate such a linear combination depends on whether there is a function of the observable random variable Y such that $g(\mathrm{E}(Y)) = l^{\mathrm{T}}\beta$.

We generally restrict our attention to linear functions of $\mathrm{E}(Y)$ and formally define a linear combination $l^{\mathrm{T}}\beta$ to be *linearly estimable* if there exists a vector t such that

$$t^{\mathrm{T}}\mathrm{E}(Y) = l^{\mathrm{T}}\beta \tag{9.20}$$

for any β.

It is clear that if X is of full column rank, $l^{\mathrm{T}}\beta$ is linearly estimable for any l or, more generally, $l^{\mathrm{T}}\beta$ is linearly estimable for any $l \in \mathrm{span}(X^{\mathrm{T}})$. (The t vector is just the normalized coefficients expressing l in terms of the columns of X.)

Estimability depends only on the simplest distributional assumption about the model; that is, that $\mathrm{E}(\epsilon) = 0$. Under this assumption, we see that the estimator $\widehat{\beta}$ based on the least squares fit of β is unbiased for the linearly estimable function $l^{\mathrm{T}}\beta$. Because $l \in \mathrm{span}(X^{\mathrm{T}}) = \mathrm{span}(X^{\mathrm{T}}X)$, we can write $l = X^{\mathrm{T}}X\tilde{t}$. Now, we have

$$\begin{aligned}\mathrm{E}(l^{\mathrm{T}}\widehat{\beta}) &= \mathrm{E}(l^{\mathrm{T}}(X^{\mathrm{T}}X)^{+}X^{\mathrm{T}}y)\\ &= \tilde{t}^{\mathrm{T}}X^{\mathrm{T}}X(X^{\mathrm{T}}X)^{+}X^{\mathrm{T}}X\beta\\ &= \tilde{t}^{\mathrm{T}}X^{\mathrm{T}}X\beta\\ &= l^{\mathrm{T}}\beta.\end{aligned} \tag{9.21}$$

Although we have been taking $\widehat{\beta}$ to be $(X^{\mathrm{T}}X)^{+}X^{\mathrm{T}}y$, the equations above follow for other least squares fits, $b = (X^{\mathrm{T}}X)^{-}X^{\mathrm{T}}y$, for any generalized inverse. In fact, the estimator of $l^{\mathrm{T}}\beta$ is invariant to the choice of the generalized inverse. This is because if $b = (X^{\mathrm{T}}X)^{-}X^{\mathrm{T}}y$, we have $X^{\mathrm{T}}Xb = X^{\mathrm{T}}y$, and so

$$l^{\mathrm{T}}\widehat{\beta} - l^{\mathrm{T}}b = \tilde{t}^{\mathrm{T}}X^{\mathrm{T}}X(\widehat{\beta} - b) = \tilde{t}^{\mathrm{T}}(X^{\mathrm{T}}y - X^{\mathrm{T}}y) = 0. \tag{9.22}$$

In the context of the linear model, we call an estimator of β a *linear estimator* if it can be expressed as Ay for some matrix A, and we call an estimator of $l^{\mathrm{T}}\beta$ a linear estimator if it can be expressed as $a^{\mathrm{T}}y$ for some vector a. It is clear that the least squares estimators $\widehat{\beta}$ and $l^{\mathrm{T}}\widehat{\beta}$ are linear estimators.

Other properties of the estimators depend on additional assumptions about the distribution of ϵ, and we will consider some of them below.

When X is not of full rank, we often are interested in an orthogonal basis for $\mathrm{span}(X^{\mathrm{T}})$. If X includes a column of 1s, the elements of any vector in the basis must sum to 0. Such vectors are called *contrasts*. The second and subsequent rows of the Helmert matrix (see Sect. 8.8.1 on page 381) are contrasts that are often of interest because of their regular patterns and their interpretability in applications involving the analysis of levels of factors in experiments.

9.3.3.2 Testability

We define a linear hypothesis $l^{\mathrm{T}}\beta = c_1$ as *testable* if $l^{\mathrm{T}}\beta$ is estimable. We generally restrict our attention to testable hypotheses.

It is often of interest to test multiple hypotheses concerning linear combinations of the elements of β. For the model (9.11), the *general linear hypothesis* is

$$\mathrm{H}_0: \; L^{\mathrm{T}}\beta = c,$$

where L is $m \times q$, of rank q, and such that $\mathrm{span}(L) \subseteq \mathrm{span}(X)$.

The test for a hypothesis depends on the distributions of the random variables in the model. If we assume that the elements of ϵ are i.i.d. normal with a mean of 0, then the general linear hypothesis is tested using an F statistic whose numerator is the difference in the residual sum of squares from fitting the model with the restriction $L^{\mathrm{T}}\beta = c$ and the residual sum of squares from fitting the unrestricted model. This reduced sum of squares is

$$(L^{\mathrm{T}}\widehat{\beta} - c)^{\mathrm{T}} \, (L^{\mathrm{T}}(X^{\mathrm{T}}X)^{*}L)^{-1} \, (L^{\mathrm{T}}\widehat{\beta} - c), \tag{9.23}$$

where $(X^{\mathrm{T}}X)^{*}$ is any g_2 inverse of $X^{\mathrm{T}}X$. This test is a likelihood ratio test. (See a text on linear models, such as Searle 1971, for more discussion on this testing problem.)

To compute the quantity in expression (9.23), first observe

$$L^{\mathrm{T}}(X^{\mathrm{T}}X)^{*}L = (X(X^{\mathrm{T}}X)^{*}L)^{\mathrm{T}} \, (X(X^{\mathrm{T}}X)^{*}L). \tag{9.24}$$

Now, if $X(X^{\mathrm{T}}X)^{*}L$, which has rank q, is decomposed as

$$X(X^{\mathrm{T}}X)^{*}L = P\begin{bmatrix} T \\ 0 \end{bmatrix},$$

where P is an $m \times m$ orthogonal matrix and T is a $q \times q$ upper triangular matrix, we can write the reduced sum of squares (9.23) as

$$(L^{\mathrm{T}}\widehat{\beta} - c)^{\mathrm{T}} \, (T^{\mathrm{T}}T)^{-1} \, (L^{\mathrm{T}}\widehat{\beta} - c)$$

or

$$\left(T^{-\mathrm{T}}(L^{\mathrm{T}}\widehat{\beta} - c)\right)^{\mathrm{T}} \, \left(T^{-\mathrm{T}}(L^{\mathrm{T}}\widehat{\beta} - c)\right)$$

or

$$v^{\mathrm{T}}v. \tag{9.25}$$

To compute v, we solve

$$T^{\mathrm{T}}v = L^{\mathrm{T}}\widehat{\beta} - c \tag{9.26}$$

for v, and the reduced sum of squares is then formed as $v^{\mathrm{T}}v$.

9.3.3.3 The Gauss-Markov Theorem

The Gauss-Markov theorem provides a restricted optimality property for estimators of estimable functions of β under the condition that $\mathrm{E}(\epsilon) = 0$ and $\mathrm{V}(\epsilon) = \sigma^2 I$; that is, in addition to the assumption of zero expectation, which we have used above, we also assume that the elements of ϵ have constant variance and that their covariances are zero. (We are not assuming independence or normality, as we did in order to develop tests of hypotheses.)

Given $y = X\beta + \epsilon$ and $\mathrm{E}(\epsilon) = 0$ and $\mathrm{V}(\epsilon) = \sigma^2 I$, the Gauss-Markov theorem states that $l^{\mathrm{T}}\widehat{\beta}$ is the unique *best linear unbiased estimator* (BLUE) of the estimable function $l^{\mathrm{T}}\beta$.

"Linear" estimator in this context means a linear combination of y; that is, an estimator in the form $a^{\mathrm{T}}y$. It is clear that $l^{\mathrm{T}}\widehat{\beta}$ is linear, and we have already seen that it is unbiased for $l^{\mathrm{T}}\beta$. "Best" in this context means that its variance is no greater than any other estimator that fits the requirements. Hence, to prove the theorem, first let $a^{\mathrm{T}}y$ be any unbiased estimator of $l^{\mathrm{T}}\beta$, and write $l = X^{\mathrm{T}}X\tilde{t}$ as above. Because $a^{\mathrm{T}}y$ is unbiased for any β, as we saw above, it must be the case that $a^{\mathrm{T}}X = l^{\mathrm{T}}$. Recalling that $X^{\mathrm{T}}X\widehat{\beta} = X^{\mathrm{T}}y$, we have

$$\begin{aligned}
\mathrm{V}(a^{\mathrm{T}}y) &= \mathrm{V}(a^{\mathrm{T}}y - l^{\mathrm{T}}\widehat{\beta} + l^{\mathrm{T}}\widehat{\beta}) \\
&= \mathrm{V}(a^{\mathrm{T}}y - \tilde{t}^{\mathrm{T}}X^{\mathrm{T}}y + l^{\mathrm{T}}\widehat{\beta}) \\
&= \mathrm{V}(a^{\mathrm{T}}y - \tilde{t}^{\mathrm{T}}X^{\mathrm{T}}y) + \mathrm{V}(l^{\mathrm{T}}\widehat{\beta}) + 2\mathrm{Cov}(a^{\mathrm{T}}y - \tilde{t}^{\mathrm{T}}X^{\mathrm{T}}y,\ \tilde{t}^{\mathrm{T}}X^{\mathrm{T}}y).
\end{aligned}$$

Now, under the assumptions on the variance-covariance matrix of ϵ, which is also the (conditional, given X) variance-covariance matrix of y, we have

$$\begin{aligned}
\mathrm{Cov}(a^{\mathrm{T}}y - \tilde{t}^{\mathrm{T}}X^{\mathrm{T}}y,\ l^{\mathrm{T}}\widehat{\beta}) &= (a^{\mathrm{T}} - \tilde{t}^{\mathrm{T}}X^{\mathrm{T}})\sigma^2 I X\tilde{t} \\
&= (a^{\mathrm{T}}X - \tilde{t}^{\mathrm{T}}X^{\mathrm{T}}X)\sigma^2 I\tilde{t} \\
&= (l^{\mathrm{T}} - l^{\mathrm{T}})\sigma^2 I\tilde{t} \\
&= 0;
\end{aligned}$$

that is,

$$\mathrm{V}(a^{\mathrm{T}}y) = \mathrm{V}(a^{\mathrm{T}}y - \tilde{t}^{\mathrm{T}}X^{\mathrm{T}}y) + \mathrm{V}(l^{\mathrm{T}}\widehat{\beta}).$$

This implies that

$$\mathrm{V}(a^{\mathrm{T}}y) \geq \mathrm{V}(l^{\mathrm{T}}\widehat{\beta});$$

that is, $l^{\mathrm{T}}\widehat{\beta}$ has minimum variance among the linear unbiased estimators of $l^{\mathrm{T}}\beta$. To see that it is unique, we consider the case in which $\mathrm{V}(a^{\mathrm{T}}y) = \mathrm{V}(l^{\mathrm{T}}\widehat{\beta})$; that is, $\mathrm{V}(a^{\mathrm{T}}y - \tilde{t}^{\mathrm{T}}X^{\mathrm{T}}y) = 0$. For this variance to equal 0, it must be the case that $a^{\mathrm{T}} - \tilde{t}^{\mathrm{T}}X^{\mathrm{T}} = 0$ or $a^{\mathrm{T}}y = \tilde{t}^{\mathrm{T}}X^{\mathrm{T}}y = l^{\mathrm{T}}\widehat{\beta}$; that is, $l^{\mathrm{T}}\widehat{\beta}$ is the unique linear unbiased estimator that achieves the minimum variance.

If we assume further that $\epsilon \sim \mathrm{N}_n(0, \sigma^2 I)$, we can show that $l^{\mathrm{T}}\widehat{\beta}$ is the uniformly minimum variance unbiased estimator (UMVUE) for $l^{\mathrm{T}}\beta$. This is because $(X^{\mathrm{T}}y,\ (y - X\widehat{\beta})^{\mathrm{T}}(y - X\widehat{\beta}))$ is complete and sufficient for (β, σ^2). This line of reasoning also implies that $(y - X\widehat{\beta})^{\mathrm{T}}(y - X\widehat{\beta})/(n - r)$, where $r = \mathrm{rank}(X)$, is UMVUE for σ^2. We will not go through the details here. The interested reader is referred to a text on mathematical statistics, such as Shao (2003).

9.3.4 The Normal Equations and the Sweep Operator

The coefficient matrix in the normal equations, $X^{\mathrm{T}}X$, or the adjusted version $X_c^{\mathrm{T}}X_c$, where X_c is the centered matrix as in equation (8.64) on page 366, is often of interest for reasons other than just to compute the least squares estimators. The condition number of $X^{\mathrm{T}}X$ is the square of the condition number of X, however, and so any ill-conditioning is exacerbated by formation of the sums of squares and cross products matrix. The adjusted sums of squares and cross products matrix, $X_c^{\mathrm{T}}X_c$, tends to be better conditioned, so it is usually the one used in the normal equations, but of course the condition number of $X_c^{\mathrm{T}}X_c$ is the square of the condition number of X_c.

A useful matrix can be formed from the normal equations:

$$\begin{bmatrix} X^{\mathrm{T}}X & X^{\mathrm{T}}y \\ y^{\mathrm{T}}X & y^{\mathrm{T}}y \end{bmatrix}. \tag{9.27}$$

Applying m elementary operations on this matrix, we can get

$$\begin{bmatrix} (X^{\mathrm{T}}X)^{+} & X^{+}y \\ y^{\mathrm{T}}X^{+\mathrm{T}} & y^{\mathrm{T}}y - y^{\mathrm{T}}X(X^{\mathrm{T}}X)^{+}X^{\mathrm{T}}y \end{bmatrix}. \tag{9.28}$$

(If X is not of full rank, in order to get the Moore-Penrose inverse in this expression, the elementary operations must be applied in a fixed manner; otherwise, we get a different generalized inverse.)

The matrix in the upper left of the partition (9.28) is related to the estimated variance-covariance matrix of the particular solution of the normal equations, and it can be used to get an estimate of the variance-covariance matrix of estimates of any independent set of linearly estimable functions of

β. The vector in the upper right of the partition is the unique minimum-length solution to the normal equations, $\widehat{\beta}$. The scalar in the lower right partition, which is the Schur complement of the full inverse (see equations (3.190) and (3.214)), is the square of the residual norm. The squared residual norm provides an estimate of the variance of the errors in equation (9.11) after proper scaling.

The partitioning in expression (9.28) is the same that we encountered on page 363.

The elementary operations can be grouped into a larger operation, called the "sweep operation", which is performed for a given row. The sweep operation on row i, S_i, of the nonnegative definite matrix A to yield the matrix B, which we denote by

$$\mathrm{S}_i(A) = B,$$

is defined in Algorithm 9.1.

Algorithm 9.1 Sweep of the i^{th} row ’

1. If $a_{ii} = 0$, skip the following operations.
2. Set $b_{ii} = a_{ii}^{-1}$.
3. For $j \neq i$, set $b_{ij} = a_{ii}^{-1} a_{ij}$.
4. For $k \neq i$, set $b_{kj} = a_{kj} - a_{ki} a_{ii}^{-1} a_{ij}$. ∎

Skipping the operations if $a_{ii} = 0$ allows the sweep operator to handle non-full rank problems. The sweep operator is its own inverse:

$$\mathrm{S}_i(\mathrm{S}_i(A)) = A.$$

The sweep operator applied to the matrix (9.27) corresponds to adding or removing the i^{th} variable (column) of the X matrix to the regression equation.

9.3.5 Linear Least Squares Subject to Linear Equality Constraints

In the regression model (9.11), it may be known that β satisfies certain constraints, such as that all the elements be nonnegative. For constraints of the form $g(\beta) \in C$, where C is some m-dimensional space, we may estimate β by the *constrained least squares estimator*; that is, the vector $\widehat{\beta}_C$ that minimizes the dot product (9.13) among all b that satisfy $g(b) \in C$.

The nature of the constraints may or may not make drastic changes to the computational problem. (The constraints also change the statistical inference problem in various ways, but we do not address that here.) If the constraints are nonlinear, or if the constraints are inequality constraints (such as that all the elements be nonnegative), there is no general closed-form solution.

It is easy to handle linear equality constraints of the form

$$\begin{aligned} g(\beta) &= L\beta \\ &= c, \end{aligned}$$

where L is a $q \times m$ matrix of full rank. The solution is, analogous to equation (9.15),

$$\widehat{\beta}_C = (X^{\mathrm{T}}X)^{+}X^{\mathrm{T}}y + (X^{\mathrm{T}}X)^{+}L^{\mathrm{T}}(L(X^{\mathrm{T}}X)^{+}L^{\mathrm{T}})^{+}(c - L(X^{\mathrm{T}}X)^{+}X^{\mathrm{T}}y). \tag{9.29}$$

When X is of full rank, this result can be derived by using Lagrange multipliers and the derivative of the norm (9.13) (see Exercise 9.4 on page 452). When X is not of full rank, it is slightly more difficult to show this, but it is still true. (See a text on linear regression, such as Draper and Smith 1998).

The restricted least squares estimate, $\widehat{\beta}_C$, can be obtained (in the (1, 2) block) by performing $m + q$ sweep operations on the matrix,

$$\begin{bmatrix} X^{\mathrm{T}}X & X^{\mathrm{T}}y & L^{\mathrm{T}} \\ y^{\mathrm{T}}X & y^{\mathrm{T}}y & c^{\mathrm{T}} \\ L & c & 0 \end{bmatrix}, \tag{9.30}$$

analogous to matrix (9.27).

9.3.6 Weighted Least Squares

In fitting the regression model $y \approx X\beta$, it is often desirable to weight the observations differently, and so instead of minimizing equation (9.13), we minimize

$$\sum w_i(y_i - x_{i*}^{\mathrm{T}}b)^2,$$

where w_i represents a nonnegative weight to be applied to the i^{th} observation. One purpose of the weight may be to control the effect of a given observation on the overall fit. If a model of the form of equation (9.11),

$$y = X\beta + \epsilon,$$

is assumed, and ϵ is taken to be a random variable such that ϵ_i has variance σ_i^2, an appropriate value of w_i may be $1/\sigma_i^2$. (Statisticians almost always naturally assume that ϵ is a random variable. Although usually it is modeled this way, here we are allowing for more general interpretations and more general motives in fitting the model.)

The normal equations can be written as

$$\left(X^{\mathrm{T}}\mathrm{diag}((w_1, w_2, \ldots, w_n))X\right)\widehat{\beta} = X^{\mathrm{T}}\mathrm{diag}((w_1, w_2, \ldots, w_n))y.$$

More generally, we can consider W to be a weight matrix that is not necessarily diagonal. We have the same set of normal equations:

$$(X^{\mathrm{T}}WX)\widehat{\beta}_W = X^{\mathrm{T}}Wy. \tag{9.31}$$

When W is a diagonal matrix, the problem is called “weighted least squares”. Use of a nondiagonal W is also called weighted least squares but is sometimes called "generalized least squares". The weight matrix is symmetric and generally positive definite, or at least nonnegative definite. The weighted least squares estimator is

$$\widehat{\beta}_W = (X^{\mathrm{T}}WX)^{+}X^{\mathrm{T}}Wy.$$

As we have mentioned many times, an expression such as this is not necessarily a formula for computation. The matrix factorizations discussed above for the unweighted case can also be used for computing weighted least squares estimates.

In a model $y = X\beta + \epsilon$, where ϵ is taken to be a random variable with variance-covariance matrix $\varSigma$, the choice of W as $\varSigma^{-1}$ yields estimators with certain desirable statistical properties. (Because this is a natural choice for many models, statisticians sometimes choose the weighting matrix without fully considering the reasons for the choice.) As we pointed out on page 295, weighted least squares can be handled by premultiplication of both y and X by the Cholesky factor of the weight matrix. In the case of an assumed variance-covariance matrix $\varSigma$, we transform each side by $\varSigma_{\mathrm{C}}^{-1}$, where $\varSigma_{\mathrm{C}}$ is the Cholesky factor of $\varSigma$. The residuals whose squares are to be minimized are $\varSigma_{\mathrm{C}}^{-1}(y - Xb)$. Under the assumptions, the variance-covariance matrix of the residuals is I.

9.3.7 Updating Linear Regression Statistics

In Sect. 6.6.5 on page 295, we discussed the general problem of updating a least squares solution to an overdetermined system when either the number of equations (rows) or the number of variables (columns) is changed. In the linear regression problem these correspond to adding or deleting observations and adding or deleting terms in the linear model, respectively.

9.3.7.1 Adding More Variables

Suppose first that more variables are added, so the regression model is

$$y \approx \left[\, X \; X_{+} \,\right] \theta,$$

where X_{+} represents the observations on the additional variables. (We use θ to represent the parameter vector; because the model is different, it is not just β with some additional elements.)

If $X^{\mathrm{T}}X$ has been formed and the sweep operator is being used to perform the regression computations, it can be used easily to add or delete variables from the model, as we mentioned above. The Sherman-Morrison-Woodbury formulas (6.28) and (6.30) and the Hemes formula (6.31) (see page 288) can also be used to update the solution.

In regression analysis, one of the most important questions is the identification of independent variables from a set of potential explanatory variables that should be in the model. This aspect of the analysis involves adding and deleting variables. We discuss this further in Sect. 9.5.2.

9.3.7.2 Adding More Observations

If we have obtained more observations, the regression model is

$$\begin{bmatrix} y \\ y_+ \end{bmatrix} \approx \begin{bmatrix} X \\ X_+ \end{bmatrix} \beta,$$

where y_+ and X_+ represent the additional observations.

We first note some properties of the new $X^{\mathrm{T}}X$ matrix, although we will make direct use of the new X matrix, as usual. We see that

$$\begin{bmatrix} X \\ X_+ \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} X \\ X_+ \end{bmatrix} = X^{\mathrm{T}}X + X_+^{\mathrm{T}}X_+.$$

The relation of the inverse of $X^{\mathrm{T}}X + X_+^{\mathrm{T}}X_+$ to the inverse of $X^{\mathrm{T}}X$ can be seen in equation (3.177) on page 119, or in equation (3.185) for the vector corresponding to a single additional row.

If the QR decomposition of X is available, we simply augment it as in equation (6.47):

$$\begin{bmatrix} R & c_1 \\ 0 & c_2 \\ X_+ & y_+ \end{bmatrix} = \begin{bmatrix} Q^{\mathrm{T}} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} X & y \\ X_+ & y_+ \end{bmatrix}.$$

We now apply orthogonal transformations to this to zero out the last rows and produce

$$\begin{bmatrix} R_* & c_{1*} \\ 0 & c_{2*} \end{bmatrix},$$

where R_* is an $m \times m$ upper triangular matrix and c_{1*} is an m-vector as before, but c_{2*} is an $(n - m + k)$-vector. We then have an equation of the form (9.17) and we use back substitution to solve it.

9.3.7.3 Adding More Observations Using Weights

Another way of approaching the problem of adding or deleting observations is by viewing the problem as weighted least squares. In this approach, we also have more general results for updating regression statistics. Following Escobar and Moser (1993), we can consider two weighted least squares problems: one with weight matrix W and one with weight matrix V. Suppose we have the solutions $\widehat{\beta}_W$ and $\widehat{\beta}_V$. Now let

$$\Delta = V - W,$$

and use the subscript $*$ on any matrix or vector to denote the subarray that corresponds only to the nonnull rows of Δ. The symbol Δ_*, for example, is the square subarray of Δ consisting of all of the nonzero rows and columns of Δ, and X_* is the subarray of X consisting of all the columns of X and only the rows of X that correspond to Δ_*. From the normal equations (9.31) using W and V, and with the solutions $\widehat{\beta}_W$ and $\widehat{\beta}_V$ plugged in, we have

$$(X^{\mathrm{T}}WX)\widehat{\beta}_V + (X^{\mathrm{T}}\Delta X)\widehat{\beta}_V = X^{\mathrm{T}}Wy + X^{\mathrm{T}}\Delta y,$$

and so

$$\widehat{\beta}_V - \widehat{\beta}_W = (X^{\mathrm{T}}WX)^+X_*^{\mathrm{T}}\Delta_*(y - X\widehat{\beta}_V)_*.$$

This gives

$$(y - X\widehat{\beta}_V)_* = (I + X(X^{\mathrm{T}}WX)^+X_*^{\mathrm{T}}\Delta_*)^+(y - X\widehat{\beta}_W)_*,$$

and finally

$$\widehat{\beta}_V = \widehat{\beta}_W + (X^{\mathrm{T}}WX)^+X_*^{\mathrm{T}}\Delta_*\left(I + X_*(X^{\mathrm{T}}WX)^+X_*^{\mathrm{T}}\Delta_*\right)^+(y - X\widehat{\beta}_W)_*.$$

If Δ_* can be written as $\pm GG^{\mathrm{T}}$, using this equation and the equations (3.176) on page 119 (which also apply to pseudoinverses), we have

$$\widehat{\beta}_V = \widehat{\beta}_W \pm (X^{\mathrm{T}}WX)^+X_*^{\mathrm{T}}G(I \pm G^{\mathrm{T}}X_*(X^{\mathrm{T}}WX)^+X_*^{\mathrm{T}}G)^+G^{\mathrm{T}}(y - X\widehat{\beta}_W)_*. \tag{9.32}$$

The sign of GG^{T} is positive when observations are added and negative when they are deleted.

Equation (9.32) is particularly simple in the case where W and V are identity matrices (of different sizes, of course). Suppose that we have obtained more observations in y_+ and X_+. (In the following, the reader must be careful to distinguish "+" as a subscript to represent more data and "+" as a superscript with its usual meaning of a Moore-Penrose inverse.) Suppose we already have the least squares solution for $y \approx X\beta$, say $\widehat{\beta}_W$. Now $\widehat{\beta}_W$ is the weighted least squares solution to the model with the additional data and with weight matrix

$$W = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}.$$

We now seek the solution to the same system with weight matrix V, which is a larger identity matrix. From equation (9.32), the solution is

$$\widehat{\beta} = \widehat{\beta}_W + (X^{\mathrm{T}}X)^+X_+^{\mathrm{T}}(I + X_+(X^{\mathrm{T}}X)^+X_+^{\mathrm{T}})^+(y - X\widehat{\beta}_W)_*. \tag{9.33}$$

9.3.8 Linear Smoothing

The interesting reasons for doing regression analysis are to understand relationships and to predict a value of the dependent value given a value of the independent variable. As a side benefit, a model with a smooth equation $f(x)$ "smoothes" the observed responses; that is, the elements in $\hat{y} = \widehat{f(x)}$ exhibit less variation than the elements in y. Of course, the important fact for our purposes is that $\|y - \hat{y}\|$ is smaller than $\|y\|$ or $\|y - \bar{y}\|$.

The use of the hat matrix emphasizes the smoothing perspective as a projection of the original y:

$$\hat{y} = Hy.$$

The concept of a smoothing matrix was discussed in Sect. 8.6.2. From this perspective, using H, we project y onto a vector in $\mathrm{span}(H)$, and that vector has a smaller variation than y; that is, H has *smoothed* y. It does not matter what the specific values in the vector y are so long as they are associated with the same values of the independent variables.

We can extend this idea to a general $n \times n$ *smoothing matrix* H_λ:

$$\tilde{y} = H_\lambda y.$$

The smoothing matrix depends only on the kind and extent of smoothing to be performed and on the observed values of the independent variables. The extent of the smoothing may be indicated by the indexing parameter λ. Once the smoothing matrix is obtained, it does not matter how the independent variables are related to the model.

In Sect. 6.7.2, we discussed regularized solutions of overdetermined systems of equations, which in the present case is equivalent to solving

$$\min_b \left((y - Xb)^{\mathrm{T}}(y - Xb) + \lambda b^{\mathrm{T}} b\right).$$

The solution of this yields the smoothing matrix

$$S_\lambda = X(X^{\mathrm{T}}X + \lambda I)^{-1}X^{\mathrm{T}},$$

as we have seen on page 364. This has the effect of shrinking the $\widehat{y}$ toward 0. (In regression analysis, this is called "ridge regression".)

We discuss ridge regression and general shrinkage estimation in Sect. 9.5.4. Loader (2012) provides additional background and discusses more general issues in smoothing.

9.3.9 Multivariate Linear Models

A simple modification of the model (9.10), $y_i = \beta^{\mathrm{T}} x_i + \epsilon_i$, on page 404, extends the scalar responses to vector responses; that is, y_i is a vector, and of course, the vector of parameters β must be replaced by a matrix. Let d be the order of y_i. Similarly, ϵ_i is a d-vector.

This is a "multivariate" linear model, meaning among other things, that the error term has a multivariate distribution (it is not a set of i.i.d. scalars).

A major difference in the multivariate linear model arises from the structure of the vector ϵ_i. It may be appropriate to assume that the ϵ_is are independent from one observation to another, but it is not likely that the individual elements within an ϵ_i vector are independent from each other or even that they have zero correlations. A reasonable assumption to complete the model is that the vectors ϵ_is are independently and identically distributed with mean 0 and variance-covariance matrix Σ. It might be reasonable also to assume that they have a normal distribution.

In statistical applications in which univariate responses are modeled, instead of the model for a single observation, we are more likely to write the model for a set of observations on y and x in the form of equation (9.11), $y = X\beta + \epsilon$, in which y and ϵ are d-vectors, X is a matrix in which the rows correspond to the individual x_i. Extending this form to the multivariate model, we write

$$Y = XB + E, \tag{9.34}$$

where now Y is an $n \times d$ matrix, X is an $n \times m$ matrix as before, B is an $m \times d$ matrix and E is an $n \times d$ matrix. Under the assumptions on the distribution of the vector ϵ_i above, and including the assumption of normality, E in (9.34) has a matrix normal distribution (see expression (4.78) on page 221):

$$E \sim \mathrm{N}_{n,d}(0, I, \Sigma), \tag{9.35}$$

or in the form of (4.79),

$$\mathrm{vec}\left(E^{\mathrm{T}}\right) \sim \mathrm{N}_{dn}\left(0, \mathrm{diag}(\Sigma, \ldots, \Sigma)\right).$$

Note that the variance-covariance matrix in this distribution has Kronecker structure, since $\mathrm{diag}(\Sigma, \ldots, \Sigma) = I \otimes \Sigma$ (see also equation (3.102)).

9.3.9.1 Fitting the Model

Fitting multivariate linear models is done in the same way as fitting univariate linear models. The most common criterion for fitting is least squares, which as we have pointed out before is the same as a maximum likelihood criterion if the errors are identically and independently normally distributed (which follows from the identity matrix I in expression (9.35)). This is the same fitting problem that we considered in Sect. 9.3.1 in this chapter or, earlier in Sect. 6.6 on page 289.

In an approach similar to the development in Sect. 6.6, for a given choice of B, say $\widetilde{B}$, we have, corresponding to equation (6.33),

$$X\widetilde{B} = Y - R, \tag{9.36}$$

where R is an $n \times d$ matrix of residuals.

A least squares solution $\widehat{B}$ is one that minimizes the sum of squares of the residuals (or, equivalently, the square root of the sum of the squares, that is, $\|R\|_{\mathrm{F}}$). Hence, we have the optimization problem

$$\min_{\widetilde{B}} \left\| Y - X\widetilde{B} \right\|_{\mathrm{F}}. \tag{9.37}$$

As in Sect. 6.6, we rewrite the square of this norm, using equation (3.291) from page 168, as

$$\mathrm{tr}\left((Y - X\widetilde{B})^{\mathrm{T}}(Y - X\widetilde{B}) \right). \tag{9.38}$$

This is similar to equation (6.35), which, as before, we differentiate and set equal to zero, getting the normal equations in the vector $\widehat{B}$,

$$X^{\mathrm{T}}X\widehat{B} = X^{\mathrm{T}}Y. \tag{9.39}$$

(Exercise.)

We note that the columns of the matrices in these equations are each the same as the univariate normal equations (6.36):

$$X^{\mathrm{T}}X[\widehat{B}_{*1}, \ldots, \widehat{B}_{*d}] = X^{\mathrm{T}}[Y_{*1}, \ldots, Y_{*d}].$$

9.3.9.2 Partitioning the Sum of Squares

On page 363, we discussed the partitioning of the sum of squares of an observed vector of data, $y^{\mathrm{T}}y$. We did this in the context of the Gramian of the partitioned matrix $[X\, y]$. In the multivariate case, first of all, instead of the sum of squares $y^{\mathrm{T}}y$, we have the matrix of sums of squares and cross products, $Y^{\mathrm{T}}Y$. We now consider the Gramian matrix $[X\, Y]^{\mathrm{T}}[X\, Y]$, and partition it as in expression (9.27),

$$\begin{bmatrix} X^{\mathrm{T}}X & X^{\mathrm{T}}Y \\ Y^{\mathrm{T}}X & Y^{\mathrm{T}}Y \end{bmatrix}. \tag{9.40}$$

From this we can get

$$\begin{bmatrix} (X^{\mathrm{T}}X)^{+} & X^{+}Y \\ Y^{\mathrm{T}}X^{+\mathrm{T}} & Y^{\mathrm{T}}Y - Y^{\mathrm{T}}X(X^{\mathrm{T}}X)^{+}X^{\mathrm{T}}Y \end{bmatrix}. \tag{9.41}$$

Note that the term in the lower right side in this partitioning is the Schur complement of $X^{\mathrm{T}}X$ in $[X\, Y]^{\mathrm{T}}\, [X\, Y]$ (see equation (3.191) on page 122). This matrix of residual sums of squares and cross products provides a maximum likelihood estimator of Σ:

$$\widehat{\Sigma} = \left(Y^{\mathrm{T}}Y - Y^{\mathrm{T}}X(X^{\mathrm{T}}X)^{+}X^{\mathrm{T}}Y \right)/n. \tag{9.42}$$

If the normalizing factor is $1/(n-d)$ instead of $1/n$, the estimator is unbiased.

This partitioning breaks the matrix of total sums of squares and cross products into a sum of a matrix of sums of squares and cross products due to the fitted relationship between Y and X and a matrix of residual sums of squares and cross products. The analysis of these two matrices of sums of squares and cross products is one of the most fundamental and important techniques in multivariate statistics.

The matrix in the upper left of the partition (9.41) can be used to get an estimate of the variance-covariance matrix of estimates of any independent set of linearly estimable functions of B. The matrix in the upper right of the partition is the solution to the normal equations, $\widehat{B}$. The matrix in the lower right partition, which is the Schur complement of the full inverse (see equations (3.190) and (3.214)), is the matrix of sums of squares and cross products of the residuals. With proper scaling, it provides an estimate of the variance-covariance Σ of each row of E in equation (9.34).

9.3.9.3 Statistical Inference

Statistical inference for multivariate linear models is similar to what is described in Sect. 9.3.3 with some obvious changes and extensions. First order properties of distributions of the analogous statistics are almost the same. Second order properties (variances), however, are rather different. The solution of the normal equations $\widehat{B}$ has a matrix normal distribution with expectation B. The scaled matrix of sums of squares and cross products of the residuals, call it $\widehat{\Sigma}$, has a Wishart distribution with parameter Σ.

The basic null hypothesis of interest that the distribution of Y is not dependent on X is essentially the same as in the univariate case. The F-test in the corresponding univariate case, which is the ratio of two independent chi-squared random variables, has an analogue in a comparison of two matrices of sums of squares and cross products. In the multivariate case, the basis for statistical inference is $\widehat{\Sigma}$, and it can be used in various ways. The relevant fact is that $\widehat{\Sigma} \sim \mathrm{W}_d(\Sigma, n-d)$, that is, it has a Wishart distribution with variance-covariance matrix Σ and parameters d and $n-d$ (see Exercise 4.12 on page 224).

In hypothesis testing, depending on the null hypothesis, there are other matrices that have Wishart distributions. There are various scalar transformations of Wishart matrices whose distributions are known (or which have been approximated). One of the most common ones, and which even has some of the flavor of an F statistic, is Wilk's Λ,

$$\Lambda = \frac{\det\left(\widehat{\Sigma}\right)}{\det\left(\widehat{\Sigma}_0\right)}, \tag{9.43}$$

where $\widehat{\Sigma}_0$ is a scaled Wishart matrix yielding a maximum of the likelihood under a null hypothesis. Other related test statistics involving $\widehat{\Sigma}$ are Pillai's trace, the Lawley-Hotelling trace (and Hotelling's T^2), and Roy's maximum root. We will not discuss these here, and the interested reader is referred to a text on multivariate analysis.

In multivariate analysis, there are other properties of the model that are subject to statistical inference. For example, we may wish to estimate or test the rank of the coefficient matrix, B. Even in the case of a single multivariate random variable, we may wish to test whether the variance-covariance matrix is of full rank. If it is not, there are fixed relationships among the elements of the random variable, and the distribution is said to be singular. We will discuss the problem of testing the rank of a matrix briefly in Sect. 9.5.5, beginning on page 433, but for more discussion on issues of statistical inference, we again refer the reader to a text on multivariate statistical inference.

9.4 Principal Components

The analysis of multivariate data involves various linear transformations that help in understanding the relationships among the features that the data represent. The second moments of the data are used to accommodate the differences in the scales of the individual variables and the covariances among pairs of variables.

If X is the matrix containing the data stored in the usual way, a useful statistic is the sums of squares and cross products matrix, $X^{\mathrm{T}}X$, or the "adjusted" squares and cross products matrix, $X_{\mathrm{c}}^{\mathrm{T}}X_{\mathrm{c}}$, where X_{c} is the centered matrix formed by subtracting from each element of X the mean of the column containing that element. The sample variance-covariance matrix, as in equation (8.67), is the Gramian matrix

$$S_X = \frac{1}{n-1} X_{\mathrm{c}}^{\mathrm{T}} X_{\mathrm{c}}, \tag{9.44}$$

where n is the number of observations (the number of rows in X).

In data analysis, the sample variance-covariance matrix S_X in equation (9.44) plays an important role. In more formal statistical inference, it is a consistent estimator of the population variance-covariance matrix (if it is positive definite), and under assumptions of independent sampling from a normal distribution, it has a known distribution. It also has important numerical properties; it is symmetric and positive definite (or, at least, nonnegative definite; see Sect. 8.6). Other estimates of the variance-covariance matrix or the correlation matrix of the underlying distribution may not be positive definite, however, and in Sect. 9.5.6 and Exercise 9.15 we describe possible ways of adjusting a matrix to be positive definite.

9.4.1 Principal Components of a Random Vector

It is often of interest to transform a given random vector into a vector whose elements are independent. We may also be interested in which of those elements of the transformed random vector have the largest variances. The transformed vector may be more useful in making inferences about the population. In more informal data analysis, it may allow use of smaller observational vectors without much loss in information.

Stating this more formally, if Y is a random d-vector with variance-covariance matrix Σ, we seek a transformation matrix A such that $\widetilde{Y} = AY$ has a diagonal variance-covariance matrix. We are additionally interested in a transformation $a^{\mathrm{T}}Y$ that has maximal variance for a given $\|a\|$.

Because the variance of $a^{\mathrm{T}}Y$ is $\mathrm{V}(a^{\mathrm{T}}Y) = a^{\mathrm{T}}\Sigma a$, we have already obtained the solution in equation (3.265). The vector a is the eigenvector corresponding to the maximum eigenvalue of Σ, and if a is normalized, the variance of $a^{\mathrm{T}}Y$ is the maximum eigenvalue.

Because Σ is symmetric, it is orthogonally diagonalizable and the properties discussed in Sect. 3.8.10 on page 153 not only provide the transformation immediately but also indicate which elements of $\widetilde{Y}$ have the largest variances. We write the orthogonal diagonalization of Σ as (see equation (3.252))

$$\Sigma = \Gamma \Lambda \Gamma^{\mathrm{T}}, \tag{9.45}$$

where $\Gamma\Gamma^{\mathrm{T}} = \Gamma^{\mathrm{T}}\Gamma = I$, and Λ is diagonal with elements $\lambda_1 \geq \cdots \geq \lambda_m \geq 0$ (because a variance-covariance matrix is nonnegative definite). Choosing the transformation as

$$\widetilde{Y} = \Gamma^{\mathrm{T}}Y, \tag{9.46}$$

we have $\mathrm{V}(\widetilde{Y}) = \Lambda$; that is, the i^{th} element of $\widetilde{Y}$ has variance λ_i, and

$$\mathrm{Cov}(\widetilde{Y}_i, \widetilde{Y}_j) = 0 \quad \text{if } i \neq j.$$

The elements of $\widetilde{Y}$ are called the *principal components* of Y. The *first principal component*, $\widetilde{Y}_1$, which is the signed magnitude of the projection of Y in the direction of the eigenvector corresponding to the maximum eigenvalue, has the maximum variance of any of the elements of $\widetilde{Y}$, and $\mathrm{V}(\widetilde{Y}_1) = \lambda_1$. (It is, of course, possible that the maximum eigenvalue is not simple. In that case, there is no one-dimensional first principal component. If m_1 is the multiplicity of λ_1, all one-dimensional projections within the m_1-dimensional eigenspace corresponding to λ_1 have the same variance, and m_1 projections can be chosen as mutually independent.)

The second and third principal components, and so on, are likewise determined directly from the spectral decomposition.

9.4.2 Principal Components of Data

The same ideas of principal components in probability models carry over to observational data. Given an $n \times d$ data matrix X, we seek a transformation as above that will yield the linear combination of the columns that has maximum sample variance, and other linear combinations that are independent. This means that we work with the centered matrix X_{c} (equation (8.64)) and the variance-covariance matrix S_X, as above, or the centered and scaled matrix X_{cs} (equation (8.65)) and the correlation matrix R_X (equation (8.69)). See Section 3.3 in Jolliffe (2002) for discussions of the differences in using the centered but not scaled matrix and using the centered and scaled matrix.

In the following, we will use S_X, which plays a role similar to Σ for the random variable. (This role could be stated more formally in terms of statistical estimation. Additionally, the scaling may require more careful consideration. The issue of scaling naturally arises from the arbitrariness of units of measurement in data. Random variables discussed in Sect. 9.4.1 have no units of measurement.)

In data analysis, we seek a normalized transformation vector a to apply to any centered observation x_c, so that the sample variance of $a^{\mathrm{T}}x_c$, that is,

$$a^{\mathrm{T}}S_X a, \tag{9.47}$$

is maximized.

From equation (3.265) or the spectral decomposition equation (3.256), we know that the solution to this maximization problem is the eigenvector, v_1, corresponding to the largest eigenvalue, c_1, of S_X, and the value of the expression (9.47); that is, $v_1^{\mathrm{T}}S_Xv_1$ at the maximum is the largest eigenvalue. In applications, this vector is used to transform the rows of X_{c} into scalars. If we think of a generic row of X_{c} as the vector x, we call $v_1^{\mathrm{T}}x$ the *first principal component* of x. There is some ambiguity about the precise meaning of "principal component". The definition just given is a scalar; that is, a combination of values of a vector of variables. This is consistent with the definition that arises in the population model in Sect. 9.4.1. Sometimes, however, the eigenvector v_1 itself is referred to as the first principal component. More often, the vector $X_{\mathrm{c}}v_1$ of linear combinations of the columns of X_{c} is called the first principal component. We will often use the term in this latter sense.

If the largest eigenvalue, c_1, is of algebraic multiplicity $m_1 > 1$, we have seen that we can choose m_1 orthogonal eigenvectors that correspond to c_1 (because S_X, being symmetric, is simple). Any one of these vectors may be called a first principal component of X.

The second and third principal components, and so on, are likewise determined directly from the nonzero eigenvalues in the spectral decomposition of S_X. Because the eigenvectors are orthogonal (or can be chosen to be), the principal components have the property

$$z_i^{\mathrm{T}}S_Xz_j = z_i^{\mathrm{T}}z_j = 0, \quad \text{for } i \neq j.$$

The full set of principal components of X_{c}, analogous to equation (9.46) except that here the random vectors correspond to the rows in X_{c}, is

$$Z = X_{\mathrm{c}}V, \tag{9.48}$$

where V has r_X columns. (As before, r_X is the rank of X.). Figure 9.2 shows two principal components, z_1 and z_2, formed from the data represented in x_1 and x_2.

9.4.2.1 Principal Components Directly from the Data Matrix

Formation of the S_X matrix emphasizes the role that the sample covariances play in principal component analysis. However, there is no reason to form a matrix such as $X_c^{\mathrm{T}}X_c$, and indeed we may introduce significant rounding errors by doing so. (Recall our previous discussions of the condition numbers of $X^{\mathrm{T}}X$ and X.)

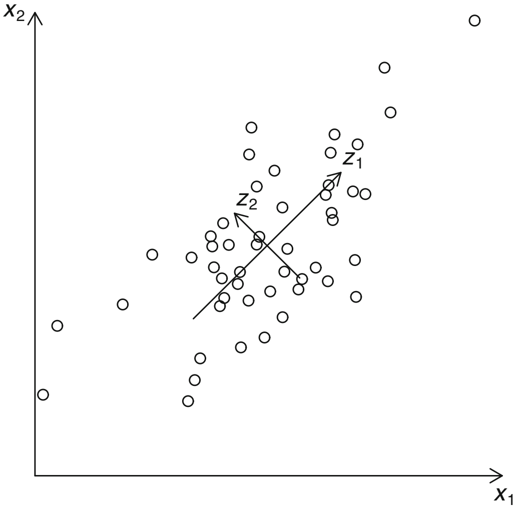

Figure 9.2. Principal components

The singular value decomposition of the $n \times m$ matrix X_c yields the square roots of the eigenvalues of $X_c^{\mathrm{T}}X_c$ and the same eigenvectors. (The eigenvalues of $X_c^{\mathrm{T}}X_c$ are $(n-1)$ times the eigenvalues of S_X.) We will assume that there are more observations than variables (that is, that $n > m$). In the SVD of the centered data matrix $X_c = UAV^{\mathrm{T}}$, U is an $n \times r_X$ matrix with orthogonal columns, V is an $m \times r_X$ matrix whose first r_X columns are orthogonal and the rest are 0, and A is an $r_X \times r_X$ diagonal matrix whose entries are the nonnegative singular values of $X - \overline{X}$. (As before, r_X is the column rank of X.)

The spectral decomposition in terms of the singular values and outer products of the columns of the factor matrices is

$$X_c = \sum_{i}^{r_X} \sigma_i u_i v_i^{\mathrm{T}}. \tag{9.49}$$

The vectors u_i are the same as the eigenvectors of S_X.

9.4.2.2 Dimension Reduction

If the columns of a data matrix X are viewed as variables or features that are measured for each of several observational units, which correspond to rows in the data matrix, an objective in principal components analysis may be to determine some small number of linear combinations of the columns of X that contain almost as much information as the full set of columns. (Here we are not using "information" in a precise sense; in a general sense, it means having similar statistical properties.) Instead of a space of dimension equal to the (column) rank of X (that is, r_X), we seek a subspace of span(X) with rank less than r_X that approximates the full space (in some sense). As we discussed on page 176, the best approximation in terms of the usual norm (the Frobenius norm) of X_{c} by a matrix of rank p is

$$\widetilde{X}_p = \sum_{i}^{p} \sigma_i u_i v_i^{\mathrm{T}} \tag{9.50}$$

for some $p < \min(n, m)$.

Principal components analysis is often used for "dimension reduction" by using the first few principal components in place of the original data. There are various ways of choosing the number of principal components (that is, p in equation (9.50)). There are also other approaches to dimension reduction. A general reference on this topic is Mizuta (2012).

9.5 Condition of Models and Data

In Sect. 6.1, we describe the concept of "condition" of a matrix for certain kinds of computations. In Sect. 6.1.3, we discuss how a large condition number may indicate the level of numerical accuracy in the solution of a system of linear equations, and on page 292 we extend this discussion to overdetermined systems such as those encountered in regression analysis. (We return to the topic of condition in Sect. 11.2 with even more emphasis on the numerical computations.) The condition of the X matrices has implications for the accuracy we can expect in the numerical computations for regression analysis.

There are other connections between the condition of the data and statistical analysis that go beyond just the purely computational issues. Analysis involves more than just computations. Ill-conditioned data also make interpretation of relationships difficult because we may be concerned with both conditional and marginal relationships. In ill-conditioned data, the relationships between any two variables may be quite different depending on whether or not the relationships are conditioned on relationships with other variables in the dataset.

9.5.1 Ill-Conditioning in Statistical Applications

We have described ill-conditioning heuristically as a situation in which small changes in the input data may result in large changes in the solution. Ill-conditioning in statistical modeling is often the result of high correlations among the independent variables. When such correlations exist, the computations may be subject to severe rounding error. This was a problem in using computer software many years ago, as Longley (1967) pointed out. When there are large correlations among the independent variables, the model itself must be examined, as Beaton, Rubin, and Barone (1976) emphasize in reviewing the analysis performed by Longley. Although the work of Beaton, Rubin, and Barone was criticized for not paying proper respect to high-accuracy computations, ultimately it is the utility of the fitted model that counts, not the accuracy of the computations.

Large correlations are reflected in the condition number of the X matrix. A large condition number may indicate the possibility of harmful numerical errors. Some of the techniques for assessing the accuracy of a computed result may be useful. In particular, the analyst may try the suggestion of Mullet and Murray (1971) to regress $y + dx_j$ on $x_1, \ldots, x_m$, and compare the results with the results obtained from just using y.

Other types of ill-conditioning may be more subtle. Large variations in the leverages may be the cause of ill-conditioning.

Often, numerical problems in regression computations indicate that the linear model may not be entirely satisfactory for the phenomenon being studied. Ill-conditioning in statistical data analysis often means that the approach or the model is not appropriate.

9.5.2 Variable Selection

Starting with a model such as equation (9.9),

$$Y = \beta^{\mathrm{T}} x + E,$$

we are ignoring the most fundamental problem in data analysis: which variables *are really related* to Y, and *how are they related*?

We often begin with the premise that a linear relationship is at least a good approximation locally; that is, with restricted ranges of the variables. This leaves us with one of the most important tasks in linear regression analysis: selection of the variables to include in the model. There are many statistical issues that must be taken into consideration. We will not discuss these issues here; rather we refer the reader to a comprehensive text on regression analysis, such as Draper and Smith (1998), or to a text specifically on this topic, such as Miller (2002). Some aspects of the statistical analysis involve tests of linear hypotheses, such as discussed in Sect. 9.3.3. There is a major difference, however; those tests were based on knowledge of the *correct* model. The basic problem in variable selection is that we do not know the correct model. Most reasonable procedures to determine the correct model yield biased statistics. Some people attempt to circumvent this problem by recasting the problem in terms of a "full" model; that is, one that includes all independent variables that the data analyst has looked at. (Looking at a variable and then making a decision to exclude that variable from the model can bias further analyses.)

We generally approach the variable selection problem by writing the model with the data as

$$y = X_{\mathrm{i}}\beta_{\mathrm{i}} + X_{\mathrm{o}}\beta_{\mathrm{o}} + \epsilon, \tag{9.51}$$

where X_{i} and X_{o} are matrices that form some permutation of the columns of X, $X_{\mathrm{i}}|X_{\mathrm{o}} = X$, and β_{i} and β_{o} are vectors consisting of corresponding elements from β. (The i and o are "in" and "out".) We then consider the model

$$y = X_{\mathrm{i}}\beta_{\mathrm{i}} + \epsilon_{\mathrm{i}}. \tag{9.52}$$

It is interesting to note that the least squares estimate of β_{i} in the model (9.52) is the same as the least squares estimate in the model

$$\hat{y}_{\mathrm{io}} = X_{\mathrm{i}}\beta_{\mathrm{i}} + \epsilon_{\mathrm{i}},$$

where $\hat{y}_{\mathrm{io}}$ is the vector of predicted values obtained by fitting the full model (9.51). An interpretation of this fact is that fitting the model (9.52) that includes only a subset of the variables is the same as using that subset to *approximate* the predictions of the full model. The fact itself can be seen from the normal equations associated with these two models. We have

$$X_{\mathrm{i}}^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}} = X_{\mathrm{i}}^{\mathrm{T}}. \tag{9.53}$$

This follows from the fact that $X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ is a projection matrix, and X_{i} consists of a set of columns of X (see Sect. 8.5 and Exercise 9.12 on page 455).

As mentioned above, there are many difficult statistical issues in the variable selection problem. The exact methods of statistical inference generally do not apply (because they are based on a model, and we are trying to choose a model). In variable selection, as in any statistical analysis that involves the choice of a model, the effect of the given dataset may be greater than warranted, resulting in overfitting. One way of dealing with this kind of problem is to use part of the dataset for fitting and part for validation of the fit. There are many variations on exactly how to do this, but in general, "cross validation" is an important part of any analysis that involves building a model.

The computations involved in variable selection are the same as those discussed in Sects. 9.3.3 and 9.3.7.

9.5.3 Principal Components Regression

A somewhat different approach to the problem of variable selection involves selecting some linear combinations of all of the variables. The first p principal components of X cover the space of span(X) optimally (in some sense), and so these linear combinations themselves may be considered as the "best" variables to include in a regression model. If V_p is the first p columns from V in the full set of principal components of X, equation (9.48), we use the regression model

$$y \approx Z_p \gamma, \tag{9.54}$$

where

$$Z_p = X V_p. \tag{9.55}$$

This is the idea of principal components regression.

In principal components regression, even if $p < m$ (which is the case, of course; otherwise principal components regression would make no sense), all of the original variables are included in the model. Any linear combination forming a principal component may include all of the original variables. The weighting on the original variables tends to be such that the coefficients of the original variables that have extreme values in the ordinary least squares regression are attenuated in the principal components regression using only the first p principal components.

The principal components do not involve y, so it may not be obvious that a model using only a set of principal components selected without reference to y would yield a useful regression model. Indeed, sometimes important independent variables do not get sufficient weight in principal components regression.

9.5.4 Shrinkage Estimation

As mentioned in the previous section, instead of selecting specific independent variables to include in the regression model, we may take the approach of shrinking the coefficient estimates toward zero. This of course has the effect of introducing a bias into the estimates (in the case of a true model being used), but in the process of reducing the inherent instability due to collinearity in the independent variables, it may also reduce the mean squared error of linear combinations of the coefficient estimates. This is one approach to the problem of overfitting.

The shrinkage can also be accomplished by a regularization of the fitting criterion. If the fitting criterion is minimization of a norm of the residuals, we add a norm of the coefficient estimates to minimize

$$\|r(b)\|_f + \lambda \|b\|_b, \tag{9.56}$$

where λ is a tuning parameter that allows control over the relative weight given to the two components of the objective function. This regularization is also related to the variable selection problem by the association of superfluous variables with the individual elements of the optimal b that are close to zero.

9.5.4.1 Ridge Regression

If the fitting criterion is least squares, we may also choose an L_2 norm on b, and we have the fitting problem

$$\min_b \left((y - Xb)^{\mathrm{T}}(y - Xb) + \lambda b^{\mathrm{T}} b\right). \tag{9.57}$$

This is called Tikhonov regularization (from A. N. Tikhonov), and it is by far the most commonly used regularization. This minimization problem yields the modified normal equations

$$(X^{\mathrm{T}}X + \lambda I)b = X^{\mathrm{T}}y, \tag{9.58}$$

obtained by adding λI to the sums of squares and cross products matrix. This is the ridge regression we discussed on page 364, and as we saw in Sect. 6.1, the addition of this positive definite matrix has the effect of reducing numerical ill-conditioning.

Interestingly, these normal equations correspond to a least squares approximation for

$$\begin{pmatrix} y \\ 0 \end{pmatrix} \approx \begin{bmatrix} X \\ \sqrt{\lambda} I \end{bmatrix} \beta. \tag{9.59}$$

(See Exercise 9.11.) The shrinkage toward 0 is evident in this formulation. Because of this, we say the “effective” degrees of freedom of a ridge regression model decreases with increasing λ. In equation (8.61), we formally defined the *effective model degrees of freedom* of any linear fit

$$\widehat{y} = S_\lambda y$$

as

$$\mathrm{tr}(S_\lambda),$$

and we saw in equation (8.62) that indeed it does decrease with increasing λ.

Even if all variables are left in the model, the ridge regression approach may alleviate some of the deleterious effects of collinearity in the independent variables.

9.5.4.2 Lasso Regression

The norm for the regularization in expression (9.56) does not have to be the same as the norm applied to the model residuals. An alternative fitting criterion, for example, is to use an L_1 norm,

$$\min_b (y - Xb)^{\mathrm{T}}(y - Xb) + \lambda \|b\|_1.$$

Rather than strictly minimizing this expression, we can formulate a constrained optimization problem

$$\min_{\|b\|_1<t} (y - Xb)^{\mathrm{T}}(y - Xb), \tag{9.60}$$

for some tuning constant t. The solution of this quadratic programming problem yields a b with some elements identically 0, depending on t. As t decreases, more elements of the optimal b are identically 0, and thus this is an effective method for variable selection. The use of expression (9.60) is called lasso regression. ("Lasso" stands for "least absolute shrinkage and selection operator".)

Lasso regression is computationally expensive if several values of t are explored. Efron et al. (2004) propose "least angle regression" (LAR), the steps of which effectively yield the entire lasso regularization path.

9.5.5 Statistical Inference about the Rank of a Matrix

An interesting problem in numerical linear algebra is to approximate the rank of a given matrix. A related problem in statistical inference is to estimate or to test an hypothesis concerning the rank of an unknown matrix. For example, in the multivariate regression model discussed in Sect. 9.3.9 (beginning on page 420), we may wish to test whether the coefficient matrix B is of full rank.

In statistical inference, we use observed data to make inferences about a model, but we do not "estimate" or "test an hypothesis" concerning the rank of a given matrix of data.

9.5.5.1 Numerical Approximation and Statistical Inference

The rank of a matrix is not a continuous function of the elements of the matrix. It is often difficult to compute the rank of a given matrix; hence, we often seek to approximate the rank. We alluded to the problem of approximating the rank of a matrix on page 252, and indicated that a QR factorization of the given matrix might be an appropriate approach to the problem. (In Sect. 11.4, we discuss the rank-revealing QR (or LU) method for approximating the rank of a matrix.)

The SVD can also be used to approximate the rank of a given $n \times m$ matrix. The approximation would be based on a decision that either the rank is $\min(n, m)$, or that the rank is r because $d_i = 0$ for $i > r$ in the decomposition UDV^{T} given in equation (3.276) on page 161.

Although we sometimes refer to the problem as one of "estimating the rank of a matrix", "estimation" in the numerical-analytical sense refers to "approximation", rather than to statistical estimation. This is an important distinction that is often lost. Estimation and testing in a statistical sense do not apply to a given entity; these methods of inference apply to properties of a random variable. We use observed realizations of the random variable to make inferences about unobserved properties or parameters that describe the distribution of the random variable.

A statistical test is a decision rule for rejection of an hypothesis about which empirical evidence is available. The empirical evidence consists of observations on some random variable, and the hypothesis is a statement about the distribution of the random variable. In simple cases of hypothesis testing, the distribution is assumed to be characterized by a parameter, and the hypothesis merely specifies the value of that parameter. The statistical test is based on the distribution of the underlying random variable if the hypothesis is true.

9.5.5.2 Statistical Tests of the Rank of a Class of Matrices

Most common statistical tests involve hypotheses concerning a scalar parameter. We have encountered two examples that involve tests of hypotheses concerning matrix parameters. One involved tests of the variance-covariance matrix Σ in a multivariate distribution (Exercise 4.12 on page 224), and the other was for tests of the coefficient matrix B in multivariate linear regression (see page 423). The tests of the variance-covariance matrix are based on a Wishart matrix W, but for a specific hypothesis, the test statistic is a chi-squared statistic. In the multivariate linear regression testing problem, the least-squares estimator of the coefficient matrix, which has a matrix normal distribution, is used to form two matrices that have independent Wishart distributions. The hypotheses of interest are that certain elements of the coefficient matrix are zero, and the test statistics involve functions of the Wishart matrices, such as Wilk's Λ, which is the ratio of the determinants.

In multivariate linear regression, given n observations on the vectors y and x, we use the model for the data given in equation (9.34), on page 421,

$$Y = XB + E,$$

where Y is an $n \times d$ matrix and X is an $n \times m$ matrix of observations, B is an $m \times d$ unknown matrix, E is an $n \times d$ matrix of n unobserved realizations of a d-variate random variable. The canonical problem in statistical applications is to test whether $B = 0$, that is, whether there is any linear relationship between y and x. A related but less encompassing question is whether B is of full rank. (If $B = 0$, its rank is zero.) Testing whether B is of full rank is similar to the familiar univariate statistical problem of testing if some elements of β in the model $y = x^{\mathrm{T}}\beta + \epsilon$ are zero. In the multivariate case, this is sometimes referred to as the "reduced rank regression" problem. The null hypothesis of interest is

$$H_0 : \ \mathrm{rank}(B) \leq \min(m, d) - 1.$$

One approach is to test sequentially the null hypotheses $H_{0_i} : \ \mathrm{rank}(B) = i$ for $i = 1, \ldots \min(m, d) - 1$.

The other problem referred to above is, given n d-vectors $y_1, \ldots y_n$ assumed to be independent realizations of random vectors distributed as $\mathrm{N}_d(\mu, \Sigma)$, to test whether Σ is of full rank (that is, whether the multivariate normal distribution is singular; see page 219).

In other applications, in vector stochastic processes, the matrix of interest is one that specifies the relationship of one time series to another. In such applications the issue of stationarity is important, and may be one of the reasons for performing the rank test.

The appropriate statistical models in these settings are different, and the forms of the models in the different applications affect the distributions of any test statistics. The nature of the subject of the hypothesis, that is, the rank of a matrix, poses some difficulty. Much of the statistical theory on hypothesis testing involves an open parameter space over a dense set of reals, but of course the rank is an integer. Because of this, even if for no other reason, we would not expect to be able to work out an exact distribution of any estimator or test statistic for the rank. At best we would seek an estimator or test statistic for which we could derive, or at least approximate, an asymptotic distribution.

Problems of testing the rank of a matrix have been addressed in the statistical literature for some time; see, for example, Anderson (1951), Gill and Lewbel (1992), and Cragg and Donald (1996). They have also been discussed frequently in econometric applications; see, for example, Robin and Smith (2000) and Kleibergen and Paap (2006). In some of the literature, it is not clear whether or not the authors are describing a test for the rank of a given matrix, which, as pointed out above, is not a statistical procedure, even if a "test statistic" and a "null probability distribution" are involved.

My purpose in this section is not to review the various approaches to statistical inference about the rank of matrices or to discuss the "best" tests under various scenarios, but rather to describe one test in order to give the flavor of the approaches.

9.5.5.3 Statistical Tests of the Rank Based on an LDU Factorization

Gill and Lewbel (1992) and Cragg and Donald (1996) describe tests of the rank of a matrix that use factors from an LDU factorization. For an $m \times d$ matrix Θ, the tests are of the null hypothesis $H_0 : \;\mathrm{rank}(\Theta) = r$, where $r < \min(m, d)$. (There are various ways an alternative hypothesis could be phrased, but we will not specify one here.)

We first decompose the unknown matrix Θ as in equation (5.32), using permutation matrices so that the diagonal elements of D are nonincreasing in absolute value: $E_{(\pi_1)} \Theta E_{(\pi_2)} = LDU$.

The $m \times d$ matrix Θ (with $m \geq d$ without loss of generality) can be decomposed as

$$\begin{aligned} E_{(\pi_1)}\Theta E_{(\pi_2)} &= LDU \\ &= \begin{bmatrix} L_{11} & 0 & 0 \\ L_{21} & L_{22} & 0 \\ L_{31} & L_{32} & I_{m-d} \end{bmatrix} \begin{bmatrix} D_1 & 0 & 0 \\ 0 & D_2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} \\ 0 & U_{22} \\ 0 & 0 \end{bmatrix}, \end{aligned} \tag{9.61}$$

where the unknown matrices L_{11}, U_{11}, and D_1 are $r \times r$, and the elements of the diagonal submatrices D_1 and D_2 are arranged in nonincreasing order. If the rank of Θ is r, $D_2 = 0$, but no diagonal element of D_1 is 0.

For a statistical test of the rank of Θ, we need to identify an observable random variable (random vector) whose distribution depends on Θ. To proceed, we take a sample of realizations of this random variable. We let $\widehat{\Theta}$ be an estimate of Θ based on n such realizations, and assume the central limit property,

$$\sqrt{k}\,\mathrm{vec}(\widehat{\Theta} - \Theta) \to_{\mathrm{d}} \mathrm{N}(0, V), \tag{9.62}$$

where V is $nm \times nm$ and positive definite. (For example, if B is the coefficient matrix in the multivariate linear regression model (9.34) and $\widehat{B}$ is the least-squares estimator from expression (9.37), then $\widehat{B}$ and B have this property. If Σ is the variance-covariance matrix in the multivariate normal distribution, expression (4.74), then we use the sample variance-covariance matrix, equation (8.67), page 367; but in this case, the analogous asymptotic distribution relating $\widehat{\Sigma}$ and Σ is a Wishart distribution.)

Now if $D_2 = 0$ (that is, if Θ has rank r) and $\widehat{\Theta}$ is decomposed in the same way as Θ in equation (9.61), then

$$\sqrt{k}\,\mathrm{diag}(\widehat{D}_2) \to_{\mathrm{d}} \mathrm{N}(0, W)$$

for some positive definite matrix W, and the quantity

$$n\widehat{d}_2^{\mathrm{T}} W^{-1} \widehat{d}_2, \tag{9.63}$$

where

$$\widehat{d}_2 = \mathrm{diag}(\widehat{D}_2),$$

has an asymptotic chi-squared distribution with $(m - r)$ degrees of freedom. If a consistent and independent estimator of W, say $\widehat{W}$, is used in place of W in the expression (9.63), this would be a test statistic for the hypothesis that the rank of Θ is r. (Note that W is $m - r \times m - r$.)

Gill and Lewbel (1992) derive a consistent estimator to use in expression (9.63) as a test statistic. Following their derivation, first let $\widehat{V}$ be a consistent estimator of V. (It would typically be a sample variance-covariance matrix.) Then

$$\left(\widehat{Q}^{\mathrm{T}} \otimes \widehat{P}\right) \widehat{V} \left(\widehat{Q} \otimes \widehat{P}^{\mathrm{T}}\right)$$

is a consistent estimator of the variance-covariance of $\mathrm{vec}(\widehat{P}(\widehat{\Theta} - \Theta)\widehat{Q})$. Next, define the matrices

$$\widehat{H} = \left[-\widehat{L}_{22}^{-1}\widehat{L}_{21}\widehat{L}_{11}^{-1} \;\middle|\; \widehat{L}_{22}^{-1} \;\middle|\; 0 \right],$$

$$\widehat{K} = \begin{bmatrix} -\widehat{U}_{11}^{-1}\widehat{U}_{12}\widehat{U}_{22}^{-1} \\ \widehat{U}_{22}^{-1} \end{bmatrix},$$

and T such that

$$\mathrm{vec}(\widehat{D}_2) = T\widehat{d}_2.$$

The matrix T is $(m-r)^2 \times (m-r)$, consisting of a stack of square matrices with 0s in all positions except for a 1 in one diagonal element. The matrix is orthogonal; that is,

$$T^{\mathrm{T}}T = I_{m-r}.$$

The matrix

$$(\widehat{K} \otimes \widehat{H}^{\mathrm{T}})T$$

transforms $\mathrm{vec}(\widehat{P}(\widehat{\Theta} - \Theta)\widehat{Q})$ into $\widehat{d}_2$; hence the variance-covariance estimator, $(\widehat{Q}^{\mathrm{T}} \otimes \widehat{P})\widehat{V}(\widehat{Q} \otimes \widehat{P}^{\mathrm{T}})$, is adjusted by this matrix. The estimator $\widehat{W}$ therefore is given by

$$\widehat{W} = T^{\mathrm{T}}(\widehat{K}^{\mathrm{T}} \otimes \widehat{H})(\widehat{Q}^{\mathrm{T}} \otimes \widehat{P})\widehat{V}(\widehat{Q} \otimes \widehat{P}^{\mathrm{T}})(\widehat{K} \otimes \widehat{H}^{\mathrm{T}})T.$$

The test statistic is

$$n\widehat{d}_2^{\mathrm{T}}\widehat{W}^{-1}\widehat{d}_2, \tag{9.64}$$

with an approximate chi-squared distribution with $(m-r)$ degrees of freedom.

Cragg and Donald (1996), however, have pointed out that the indeterminacy of the LDU decomposition casts doubts on the central limiting distribution in (9.62). Kleibergen and Paap (2006) proposed a related test for a certain class of matrices based on the SVD. Because the SVD is unique (within the limitations mentioned on page 163), it does not suffer from the indeterminacy.

9.5.6 Incomplete Data

Missing values in a dataset can not only result in ill-conditioned problems but can cause some matrix statistics to lack their standard properties, such as covariance or correlation matrices formed from the available data not being positive definite.

In the standard flat data file represented in Fig. 8.1, where a row holds data from a given observation and a column represents a specific variable or feature, it is often the case that some values are missing for some observation/variable combination. This can occur for various reasons, such as a failure of a measuring device, refusal to answer a question in a survey, or an indeterminate or infinite value for a derived variable (for example, a coefficient of variation when the mean is 0). This causes problems for our standard storage of data in a matrix. The values for some cells are not available.

The need to make provisions for missing data is one of the important differences between statistical numerical processing and ordinary numerical analysis. First of all, we need a method for representing a “not available” (NA) value, and then we need a mechanism for avoiding computations with this NA value. There are various ways of doing this, including the use of special computer numbers (see pages 464 and 475).

The layout of the data may be of the form

$$X = \begin{bmatrix} \mathtt{X} & \mathtt{X} & \mathtt{NA} \\ \mathtt{X} & \mathtt{NA} & \mathtt{NA} \\ \mathtt{X} & \mathtt{NA} & \mathtt{X} \\ \mathtt{X} & \mathtt{X} & \mathtt{X} \end{bmatrix}. \tag{9.65}$$

In the data matrix of equation (9.65), all rows could be used for summary statistics relating to the first variable, but only two rows could be used for summary statistics relating to the second and third variables. For summary statistics such as the mean or variance for any one variable, it would seem to make sense to use all of the available data.

The picture is not so clear, however, for statistics on two variables, such as the covariance. If all observations that contain data on both variables are used for computing the covariance, then the covariance matrix may not be positive definite. If the correlation matrix is computed using covariances computed in this way but variances computed on all of the data, some off-diagonal elements may be larger than 1. If the correlation matrix is computed using covariances from all available pairs and variances computed only from the data in complete pairs (that is, the variances used in computing correlations involving a given variable are different for different variables), then no off-diagonal element can be larger than 1, but the correlation matrix may not be nonnegative definite.

An alternative, of course, is to use only data in records that are complete. This is called “casewise deletion”, whereas use of all available data for bivariate statistics is called “pairwise deletion”. One must be very careful in computing bivariate statistics from data with missing values; see Exercise 9.14 (and a solution on page 612).

Estimated or approximate variance-covariance or correlation matrices that are not positive definite can arise in other ways in applications. For example, the data analyst may have an estimate of the correlation matrix that was not based on a single sample.

Various approaches to handling an approximate correlation matrix that is not positive definite have been considered. Devlin et al. (1975) describe a method of shrinking the given R toward a chosen positive definite matrix, R_1, which may be an estimator of a correlation matrix computed in other ways (perhaps a robust estimator) or may just be chosen arbitrarily; for example, R_1 may just be the identity matrix. The method is to choose the largest value α in $[0, 1]$ such that the matrix

$$\widetilde{R} = \alpha R + (1 - \alpha) R_1 \tag{9.66}$$

is positive definite. This optimization problem can be solved iteratively starting with $\alpha = 1$ and decreasing α in small steps while checking whether $\widetilde{R}$ is positive definite. (The checks may require several computations.) A related method is to use a modified Cholesky decomposition. If the symmetric matrix S is not positive definite, a diagonal matrix D can be determined so that $S + D$ is positive definite. Eskow and Schnabel (1991), for example, describe one way to determine D with values near zero and to compute a Cholesky decomposition of $S + D$.

Devlin, Gnanadesikan, and Kettenring (1975) also describe nonlinear shrinking methods in which all of the off-diagonal elements r_{ij} are replaced iteratively, beginning with $r_{ij}^{(0)} = r_{ij}$ and proceeding with

$$r_{ij}^{(k)} = \begin{cases} f^{-1}\left(f\left(r_{ij}^{(k-1)}\right) + \delta\right) & \text{if } r_{ij}^{(k-1)} < -f^{-1}(\delta) \\ 0 & \text{if } \left|r_{ij}^{(k-1)}\right| \le f^{-1}(\delta) \\ f^{-1}\left(f\left(r_{ij}^{(k-1)}\right) - \delta\right) & \text{if } r_{ij}^{(k-1)} > f^{-1}(\delta) \end{cases} \tag{9.67}$$

for some invertible positive-valued function f and some small positive constant δ (for example, 0.05). The function f may be chosen in various ways; one suggested function is the hyperbolic tangent, which makes f^{-1} Fisher's variance-stabilizing function for a correlation coefficient; see Exercise 9.19b.

Rousseeuw and Molenberghs (1993) suggest a method in which some approximate correlation matrices can be adjusted to a nearby correlation matrix, where closeness is determined by the Frobenius norm. Their method applies to pseudo-correlation matrices. Recall that any symmetric nonnegative definite matrix with ones on the diagonal is a correlation matrix. A *pseudo-correlation matrix* is a symmetric matrix R with positive diagonal elements (but not necessarily 1s) and such that $r_{ij}^2 \le r_{ii} r_{jj}$. (This is inequality (8.12), which is a necessary but not sufficient condition for the matrix to be nonnegative definite.)

The method of Rousseeuw and Molenberghs adjusts an $m \times m$ pseudo-correlation matrix R to the closest correlation matrix $\widetilde{R}$, where closeness is determined by the Frobenius norm; that is, we seek $\widetilde{R}$ such that

$$\|R - \widetilde{R}\|_{\mathrm{F}} \tag{9.68}$$

is minimum over all choices of $\widetilde{R}$ that are correlation matrices (that is, matrices with 1s on the diagonal that are positive definite). The solution to this optimization problem is not as easy as the solution to the problem we consider on page 176 of finding the best approximate matrix of a given rank. Rousseeuw and Molenberghs describe a computational method for finding $\widetilde{R}$ to minimize expression (9.68). A correlation matrix $\widetilde{R}$ can be formed as a Gramian matrix formed from a matrix U whose columns, $u_1, \ldots, u_m$, are normalized vectors, where

$$\tilde{r}_{ij} = u_i^{\mathrm{T}} u_j.$$

If we choose the vector u_i so that only the first i elements are nonzero, then they form the Cholesky factor elements of $\widetilde{R}$ with nonnegative diagonal elements,

$$\widetilde{R} = U^{\mathrm{T}} U,$$

and each u_i can be completely represented in $\mathrm{I\!R}^i$. We can associate the $m(m-1)/2$ unknown elements of U with the angles in their spherical coordinates. In u_i, the j^{th} element is 0 if $j > i$ and otherwise is

$$\sin(\theta_{i1}) \cdots \sin(\theta_{i,i-j}) \cos(\theta_{i,i-j+1}),$$

where $\theta_{i1}, \ldots, \theta_{i,i-j}, \theta_{i,i-j+1}$ are the unknown angles that are the variables in the optimization problem for the Frobenius norm (9.68). The problem now is to solve

$$\min \sum_{i=1}^{m} \sum_{j=1}^{i} (r_{ij} - \sin(\theta_{i1}) \cdots \sin(\theta_{i,i-j}) \cos(\theta_{i,i-j+1}))^2. \tag{9.69}$$

This optimization problem is well-behaved and can be solved by steepest descent (see page 201). Rousseeuw and Molenberghs (1993) also mention that a weighted least squares problem in place of equation (9.69) may be more appropriate if the elements of the pseudo-correlation matrix R result from different numbers of observations.

In Exercise 9.15, we describe another way of converting an approximate correlation matrix that is not positive definite into a correlation matrix by iteratively replacing negative eigenvalues with positive ones.

9.6 Optimal Design

When an experiment is designed to explore the effects of some variables (usually called "factors") on another variable, the settings of the factors (independent variables) should be determined so as to yield a maximum amount of information from a given number of observations. The basic problem is to determine from a set of candidates the best rows for the data matrix X. For example, if there are six factors and each can be set at three different levels, there is a total of $3^6 = 729$ combinations of settings. In many cases, because of the expense in conducting the experiment, only a relatively small number of runs can be made. If, in the case of the 729 possible combinations, only 30 or so runs can be made, the scientist must choose the subset of combinations that will be most informative. A row in X may contain more elements than just the number of factors (because of interactions), but the factor settings completely determine the row.

We may quantify the information in terms of variances of the estimators. If we assume a linear relationship expressed by

$$y = \beta_0 1 + X\beta + \epsilon$$

and make certain assumptions about the probability distribution of the residuals, the variance-covariance matrix of estimable linear functions of the least squares solution (9.15) is formed from

$$(X^{\mathrm{T}}X)^{-}\sigma^2.$$

(The assumptions are that the residuals are independently distributed with a constant variance, σ^2. We will not dwell on the statistical properties here, however.) If the emphasis is on estimation of β, then X should be of full rank. In the following, we assume X is of full rank; that is, that $(X^{\mathrm{T}}X)^{-1}$ exists.

An objective is to minimize the variances of estimators of linear combinations of the elements of β. We may identify three types of relevant measures of the variance of the estimator $\widehat{\beta}$: the average variance of the elements of $\widehat{\beta}$, the maximum variance of any elements, and the "generalized variance" of the vector $\widehat{\beta}$. The property of the design resulting from maximizing the information by reducing these measures of variance is called, respectively, A-optimality, E-optimality, and D-optimality. They are achieved when X is chosen as follows:

- A-optimality: minimize $\mathrm{tr}((X^{\mathrm{T}}X)^{-1})$.
- E-optimality: minimize $\rho((X^{\mathrm{T}}X)^{-1})$.
- D-optimality: minimize $\det((X^{\mathrm{T}}X)^{-1})$.

Using the properties of eigenvalues and determinants that we discussed in Chap. 3, we see that E-optimality is achieved by maximizing $\rho(X^{\mathrm{T}}X)$ and D-optimality is achieved by maximizing $\det(X^{\mathrm{T}}X)$.

9.6.1 D-Optimal Designs

The D-optimal criterion is probably used most often. If the residuals have a normal distribution (and the other distributional assumptions are satisfied), the D-optimal design results in the smallest volume of confidence ellipsoids for β. (See Titterington 1975; Nguyen and Miller 1992; and Atkinson and Donev 1992. Identification of the D-optimal design is related to determination of a minimum-volume ellipsoid for multivariate data.) The computations required for the D-optimal criterion are the simplest, and this may be another reason it is used often.

To construct an optimal X with a given number of rows, n, from a set of N potential rows, one usually begins with an initial choice of rows, perhaps random, and then determines the effect on the determinant by exchanging a selected row with a different row from the set of potential rows. If the matrix X has n rows and the row vector x^{T} is appended, the determinant of interest is

$$\det(X^{\mathrm{T}}X + xx^{\mathrm{T}})$$

or its inverse. Using the relationship $\det(AB) = \det(A)\det(B)$, it is easy to see that

$$\det(X^{\mathrm{T}}X + xx^{\mathrm{T}}) = \det(X^{\mathrm{T}}X)(1 + x^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x). \tag{9.70}$$

Now, if a row x_+^{T} is exchanged for the row x_-^{T}, the effect on the determinant is given by

$$\begin{aligned}\det(X^{\mathrm{T}}X + x_+x_+^{\mathrm{T}} - x_-x_-^{\mathrm{T}}) = \det(X^{\mathrm{T}}X)\ \times& \\ &\Big(1 + x_+^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_+ - \\ &x_-^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_-(1 + x_+^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_+)\ + \\ &(x_+^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_-)^2\Big)\end{aligned} \tag{9.71}$$

(see Exercise 9.8).

Following Miller and Nguyen (1994), writing $X^{\mathrm{T}}X$ as $R^{\mathrm{T}}R$ from the QR decomposition of X, and introducing z_+ and z_- as

$$Rz_+ = x_+$$

and

$$Rz_- = x_-,$$

we have the right-hand side of equation (9.71):

$$z_+^{\mathrm{T}}z_+ - z_-^{\mathrm{T}}z_-(1 + z_+^{\mathrm{T}}z_+) + (z_-^{\mathrm{T}}z_+)^2. \tag{9.72}$$

Even though there are $n(N-n)$ possible pairs (x_+, x_-) to consider for exchanging, various quantities in (9.72) need be computed only once. The corresponding (z_+, z_-) are obtained by back substitution using the triangular matrix R. Miller and Nguyen use the Cauchy-Schwarz inequality (2.26) (page 24) to show that the quantity (9.72) can be no larger than

$$z_+^{\mathrm{T}}z_+ - z_-^{\mathrm{T}}z_-; \tag{9.73}$$

hence, when considering a pair (x_+, x_-) for exchanging, if the quantity (9.73) is smaller than the largest value of (9.72) found so far, then the full computation of (9.72) can be skipped. Miller and Nguyen also suggest not allowing the last point added to the design to be considered for removal in the next iteration and not allowing the last point removed to be added in the next iteration.

The procedure begins with an initial selection of design points, yielding the $n \times m$ matrix $X^{(0)}$ that is of full rank. At the k^{th} step, each row of $X^{(k)}$ is considered for exchange with a candidate point, subject to the restrictions mentioned above. Equations (9.72) and (9.73) are used to determine the best exchange. If no point is found to improve the determinant, the process terminates. Otherwise, when the optimal exchange is determined, $R^{(k+1)}$ is formed using the updating methods discussed in the previous sections. (The programs of Gentleman 1974, referred to in Sect. 6.6.5 can be used.)

9.7 Multivariate Random Number Generation

The need to simulate realizations of random variables arises often in statistical applications, both in the development of statistical theory and in applied data analysis. In this section, we will illustrate only a couple of problems in multivariate random number generation. These make use of some of the properties we have discussed previously.

Most methods for random number generation assume an underlying source of realizations of a uniform $(0, 1)$ random variable. If U is a uniform $(0, 1)$ random variable, and F is the cumulative distribution function of a continuous random variable, then the random variable

$$X = F^{-1}(U)$$

has the cumulative distribution function F. (If the support of X is finite, $F^{-1}(0)$ and $F^{-1}(1)$ are interpreted as the limits of the support.) This same idea, the basis of the so-called inverse CDF method, can also be applied to discrete random variables.

9.7.1 The Multivariate Normal Distribution

If Z has a multivariate normal distribution with the identity as variance-covariance matrix, then for a given positive definite matrix Σ, both

$$Y_1 = \Sigma^{1/2} Z \tag{9.74}$$

and

$$Y_2 = \Sigma_C Z, \tag{9.75}$$

where Σ_C is a Cholesky factor of Σ, have a multivariate normal distribution with variance-covariance matrix Σ (see page 401). The mean of Y_1 is $\Sigma^{1/2}\mu$, where μ is the mean of Z, and the mean of Y_1 is $\Sigma_C\mu$. If Z has 0 mean, then the distributions are identical, that is, $Y_1 \overset{\text{d}}{=} Y_2$.

This leads to a very simple method for generating a multivariate normal random d-vector: generate into a d-vector z d independent $\text{N}_1(0, 1)$. Then form a vector from the desired distribution by the transformation in equation (9.74) or (9.75) together with the addition of a mean vector if necessary.

9.7.2 Random Correlation Matrices

Occasionally we wish to generate random numbers but do not wish to specify the distribution fully. We may want a “random” matrix, but we do not know an exact distribution that we wish to simulate. (There are only a few “standard” distributions of matrices. The Wishart distribution and the Haar distribution (page 222) are the only two common ones. We can also, of course, specify the distributions of the individual elements.)

We may want to simulate random correlation matrices. Although we do not have a specific distribution, we may want to specify some characteristics, such as the eigenvalues. (All of the eigenvalues of a correlation matrix, not just the largest and smallest, determine the condition of data matrices that are realizations of random variables with the given correlation matrix.)

Any nonnegative definite (symmetric) matrix with 1s on the diagonal is a correlation matrix. A correlation matrix is diagonalizable, so if the eigenvalues are $c_1, \ldots, c_d$, we can represent the matrix as

$$V \mathrm{diag}((c_1, \ldots, c_d))V^{\mathrm{T}}$$

for an orthogonal matrix V. (For a $d \times d$ correlation matrix, we have $\sum c_i = d$; see page 368.) Generating a random correlation matrix with given eigenvalues becomes a problem of generating the random orthogonal eigenvectors and then forming the matrix V from them. (Recall from page 153 that the eigenvectors of a symmetric matrix can be chosen to be orthogonal.) In the following, we let $C = \mathrm{diag}((c_1, \ldots, c_d))$ and begin with $E = I$ (the $d \times d$ identity) and $k = 1$. The method makes use of deflation in step 6 (see page 310). The underlying randomness is that of a normal distribution.

Algorithm 9.2 Random correlation matrices with given eigenvalues

```
1. Generate a $d$-vector $w$ of i.i.d. standard normal deviates, form $x = Ew$,
   and compute $a = x^{\mathrm{T}}(I - C)x$.
2. Generate a $d$-vector $z$ of i.i.d. standard normal deviates, form $y = Ez$,
   and compute $b = x^{\mathrm{T}}(I - C)y$, $c = y^{\mathrm{T}}(I - C)y$, and $e^2 = b^2 - ac$.
3. If $e^2 < 0$, then go to step 2.
4. Choose a random sign, $s = -1$ or $s = 1$. Set $r = \frac{b + se}{a}x - y$.
5. Choose another random sign, $s = -1$ or $s = 1$, and set $v_k = \frac{sr}{(r^{\mathrm{T}}r)^{\frac{1}{2}}}$.
6. Set $E = E - v_k v_k^{\mathrm{T}}$, and set $k = k + 1$.
7. If $k < d$, then go to step 1.
8. Generate a $d$-vector $w$ of i.i.d. standard normal deviates, form $x = Ew$,
   and set $v_d = \frac{x}{(x^{\mathrm{T}}x)^{\frac{1}{2}}}$.
9. Construct the matrix $V$ using the vectors $v_k$ as its rows. Deliver $VCV^{\mathrm{T}}$
   as the random correlation matrix. ∎
```

9.8 Stochastic Processes

Many stochastic processes are modeled by a "state vector" and rules for updating the state vector through a sequence of discrete steps. At time t, the elements of the state vector x_t are values of various characteristics of the system. A model for the stochastic process is a probabilistic prescription for x_{t_a} in terms of x_{t_b}, where $t_a > t_b$; that is, given observations on the state vector prior to some point in time, the model gives probabilities for, or predicts values of, the state vector at later times.

A stochastic process is distinguished in terms of the countability of the space of states, $\mathcal{X}$, and the index of the state (that is, the parameter space, $\mathcal{T}$); either may or may not be countable. If the parameter space is continuous, the process is called a *diffusion process*. If the parameter space is countable, we usually consider it to consist of the nonnegative integers.

If the properties of a stochastic process do not depend on the index, the process is said to be *stationary*. If the properties also do not depend on any initial state, the process is said to be *time homogeneous* or *homogeneous with respect to the parameter space*. (We usually refer to such processes simply as "homogeneous".)

9.8.1 Markov Chains

The *Markov* (or Markovian) *property* in a stochastic process is the condition in which the current state does not depend on any states prior to the immediately previous state; that is, the process is *memoryless*. If the transitions occur at discrete intervals, the Markov property is the condition where the probability distribution of the state at time $t + 1$ depends only on the state at time t.

In what follows, we will briefly consider some Markov processes in which both the set of states is countable and the transitions occur at discrete intervals (discrete times). Such a process is called a *Markov chain*. (Some authors' use of the term "Markov chain" allows the state space to be continuous, and others' allows time to be continuous; here we are not defining the term. We will be concerned with only a subclass of Markov chains, whichever way they are defined. The models for this subclass are easily formulated in terms of vectors and matrices.)

If the state space is countable, it is equivalent to $\mathcal{X} = \{1, 2, \ldots\}$. If X is a random variable from some sample space to $\mathcal{X}$, and

$$\pi_i = \Pr(X = i),$$

then the vector π defines a distribution of X on $\mathcal{X}$. (A vector of nonnegative numbers that sum to 1 is a *distribution*.)

Formally, we define a Markov chain (of random variables) $X_0, X_1, \ldots$ in terms of an initial distribution π and a conditional distribution for X_{t+1} given X_t. Let X_0 have distribution π, and given $X_t = i$, let X_{t+1} have distribution

$(p_{ij}; j \in \mathcal{X})$; that is, p_{ij} is the probability of a transition from state i at time t to state j at time $t + 1$. Let

$$P = (p_{ij}).$$

This square matrix is called the *transition matrix* of the chain. It is clear that P is a stochastic matrix (it is nonnegative and the elements in any row sum to 1), and hence $\rho(P) = \|P\|_\infty = 1$, and $(1, 1)$ is an eigenpair of P (see page 379).

If P does not depend on the time (and our notation indicates that we are assuming this), the Markov chain is *stationary*.

The initial distribution π and the transition matrix P characterize the chain, which we sometimes denote as $\mathit{Markov}(\pi, P)$.

If the set of states is countably infinite, the vectors and matrices have infinite order; that is, they have "infinite dimension". (Note that this use of "dimension" is different from our standard definition that is based on linear independence.)

We denote the distribution at time t by $\pi^{(t)}$ and hence often write the initial distribution as $\pi^{(0)}$. A distribution at time t can be expressed in terms of π and P if we extend the definition of (Cayley) matrix multiplication in equation (3.43) in the obvious way to handle any countable number of elements so that PP or P^2 is the matrix defined by

$$(P^2)_{ij} = \sum_{k \in \mathcal{X}} p_{ik} p_{kj}.$$

We see immediately that

$$\pi^{(t)} = (P^t)^{\mathrm{T}} \pi^{(0)}. \tag{9.76}$$

Because of equation (9.76), P^t is often called the *t-step transition matrix*. (The somewhat awkward notation with the transpose results from the historical convention in Markov chain theory of expressing distributions as "row vectors".)

9.8.1.1 Properties of Markov Chains

The transition matrix determines various relationships among the states of a Markov chain. State j is said to be *accessible* from state i if it can be reached from state i in a finite number of steps. This is equivalent to $(P^t)_{ij} > 0$ for some t. If state j is accessible from state i and state i is accessible from state j, states j and i are said to *communicate*. Communication is clearly an equivalence relation. (A binary relation $\sim$ is an *equivalence relation* over some set S if for $x, y, z \in S$, (1) $x \sim x$, (2) $x \sim y \Rightarrow y \sim x$, and (3) $x \sim y \wedge y \sim z \Rightarrow x \sim z$; that is, it is reflexive, symmetric, and transitive.) The set of all states that communicate with each other is an *equivalence class*. States belonging to different equivalence classes do not communicate, although a state in one class may be accessible from a state in a different class.

Identification and analysis of states that communicate can be done by the reduction of the transition matrix in the manner discussed on page 375 and illustrated in equation (8.76), in which by permutations of the rows and columns, square 0 submatrices are formed. If the transition matrix is irreducible, that is, if no such 0 submatrices can be formed, then all states in a Markov chain are in a single equivalence class. In that case the chain is said to be *irreducible.* Irreducible matrices are discussed in Sect. 8.7.3, beginning on page 375, and the implication (8.77) in that section provides a simple characterization of irreducibility. Reducibility of Markov chains is also clearly related to the reducibility in graphs that we discussed in Sect. 8.1.2. (In graphs, the connectivity matrix is similar to the transition matrix in Markov chains.)

If the transition matrix is primitive (that is, it is irreducible and its eigenvalue with maximum modulus has algebraic multiplicity of 1, see page 377), then the Markov chain is said to be *primitive.*

Primitivity and irreducibility are important concepts in analysis of Markov chains because they imply interesting limiting behavior of the chains.

9.8.1.2 Limiting Behavior of Markov Chains

The limiting behavior of the Markov chain is of interest. This of course can be analyzed in terms of $\lim_{t\to\infty} P^t$. Whether or not this limit exists depends on the properties of P. If P is primitive and irreducible, we can make use of the results in Sect. 8.7.3. In particular, because 1 is an eigenvalue and the vector 1 is the eigenvector associated with 1, from equation (8.79), we have

$$\lim_{t\to\infty} P^t = 1\pi_s^{\mathrm{T}}, \tag{9.77}$$

where π_s is the Perron vector of P^{T}.

This also gives us the *limiting distribution* for an irreducible, primitive Markov chain,

$$\lim_{t\to\infty} \pi^{(t)} = \pi_s.$$

The Perron vector has the property $\pi_s = P^{\mathrm{T}}\pi_s$ of course, so this distribution is the *invariant distribution* of the chain. This invariance is a necessary condition for most uses of Markov chains in Monte Carlo methods for generating posterior distributions in Bayesian statistical analysis. These methods are called Markov chain Monte Carlo (MCMC) methods, and are widely used in Bayesian analyses.

There are many other interesting properties of Markov chains that follow from various properties of nonnegative matrices that we discuss in Sect. 8.7, but rather than continuing the discussion here, we refer the interested reader to a text on Markov chains, such as Meyn and Tweedie (2009).

9.8.2 Markovian Population Models

A simple but useful model for population growth measured at discrete points in time, $t, t+1, \ldots$, is constructed as follows. We identify k age groupings for the members of the population; we determine the number of members in each age group at time t, calling this $p^{(t)}$,

$$p^{(t)} = \left(p_1^{(t)}, \ldots, p_k^{(t)}\right);$$

determine the reproductive rate in each age group, calling this α,

$$\alpha = (\alpha_1, \ldots, \alpha_k);$$

and determine the survival rate in each of the first $k-1$ age groups, calling this σ,

$$\sigma = (\sigma_1, \ldots, \sigma_{k-1}).$$

It is assumed that the reproductive rate and the survival rate are constant in time. (There are interesting statistical estimation problems here that are described in standard texts in demography or in animal population models.) The survival rate σ_i is the proportion of members in age group i at time t who survive to age group $i+1$. (It is assumed that the members in the last age group do not survive from time t to time $t+1$.) The total size of the population at time t is $N^{(t)} = 1^{\mathrm{T}} p^{(t)}$. (The use of the capital letter N for a scalar variable is consistent with the notation used in the study of finite populations.)

If the population in each age group is relatively large, then given the sizes of the population age groups at time t, the approximate sizes at time $t+1$ are given by

$$p^{(t+1)} = Ap^{(t)}, \tag{9.78}$$

where A is a Leslie matrix as in equation (8.85),

$$A = \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_{m-1} & \alpha_m \\ \sigma_1 & 0 & \cdots & 0 & 0 \\ 0 & \sigma_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{m-1} & 0 \end{bmatrix}, \tag{9.79}$$

where $0 \le \alpha_i$ and $0 \le \sigma_i \le 1$.

The Leslie population model can be useful in studying various species of plants or animals. The parameters in the model determine the vitality of the species. For biological realism, at least one α_i and all σ_i must be positive. This model provides a simple approach to the study and simulation of population dynamics. The model depends critically on the eigenvalues of A.

As we have seen (Exercise 8.10), the Leslie matrix has a single unique positive eigenvalue. If that positive eigenvalue is strictly greater in modulus than any other eigenvalue, then given some initial population size, $p^{(0)}$, the model yields a few damping oscillations and then an exponential growth,

$$p^{(t_0+t)} = p^{(t_0)} \mathrm{e}^{rt}, \tag{9.80}$$

where r is the *rate constant*. The vector $p^{(t_0)}$ (or any scalar multiple) is called the *stable age distribution*. (You are asked to show this in Exercise 9.22a.) If 1 is an eigenvalue and all other eigenvalues are strictly less than 1 in modulus, then the population eventually becomes constant; that is, there is a *stable population*. (You are asked to show this in Exercise 9.22b.)

The survival rates and reproductive rates constitute an age-dependent *life table*, which is widely used in studying population growth. The age groups in life tables for higher-order animals are often defined in years, and the parameters often are defined only for females. The first age group is generally age 0, and so $\alpha_1 = 0$. The *net reproductive rate*, r_0, is the average number of (female) offspring born to a given (female) member of the population over the lifetime of that member; that is,

$$r_0 = \sum_{i=2}^{m} \alpha_i \sigma_{i-1}. \tag{9.81}$$

The *average generation time*, T, is given by

$$T = \sum_{i=2}^{m} i \alpha_i \sigma_{i-1} / r_0. \tag{9.82}$$

The net reproductive rate, average generation time, and exponential growth rate constant are related by

$$r = \log(r_0)/T. \tag{9.83}$$

(You are asked to show this in Exercise 9.22c.)

Because the process being modeled is continuous in time and this model is discrete, there are certain averaging approximations that must be made. There are various refinements of this basic model to account for continuous time. There are also refinements to allow for time-varying parameters and for the intervention of exogenous events. Of course, from a statistical perspective, the most interesting questions involve the estimation of the parameters. See Cullen (1985), for example, for further discussions of this modeling problem.

Various starting age distributions can be used in this model to study the population dynamics.

9.8.3 Autoregressive Processes

An interesting type of stochastic process is the p^{th}-order autoregressive time series, defined by the stochastic difference equation

$$x_t = \phi_0 + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + e_t, \tag{9.84}$$

where $\phi_p \neq 0$, the e_t have constant mean of 0 and constant variance of $\sigma^2 > 0$, and any two different e_s and e_t have zero correlation. This model is called an autoregressive model of order p or $AR(p)$.

Comments on the model and notation:

> I have implied that e_t is considered to be a random variable, even though it is not written in upper case, and of course, if e_t is a random variable then x_t is also, even though it is not written in upper case either; likewise, I will sometimes consider x_{t-1} and the other x_{t-j} to be random variables.
>
> Stationarity (that is, constancy over time) is an issue. In the simple model above all of the simple parameters are constant; however, unless certain conditions are met, the moments of x_t can grow without bounds. (This is related to the "unit roots" mentioned below. Some authors require that some other conditions be satisfied in an $AR(p)$ model so that moments of the process do not grow without bounds.
>
> Also, many authors omit the ϕ_0 term in the $AR(p)$ model.

The most important properties of this process arise from the autocorrelations, $\mathrm{Cor}(x_s, x_t)$. If these autocorrelations depend on s and t only through $|s-t|$ and if for given $h = |s-t|$ the autocorrelation,

$$\rho_h = \mathrm{Cor}(x_{t+h}, x_t),$$

is constant, the autoregressive process has some simple, but useful properties.

The model (9.84) is a little more complicated than it appears. This is because the specification of x_t is conditional on $x_{t-1}, \ldots, x_{t-p}$. Presumably, also, x_{t-1} is dependent on $x_{t-2}, \ldots, x_{t-p-1}$, and so on. There are no marginal (unconditional) properties of the xs that are specified in the model; that is, we have not specified a starting point.

The stationarity of the e_t (constant mean and constant variance) does not imply that the x_t are stationary. We can make the model more specific by adding a condition of stationarity on the x_t. Let us assume that the x_t have constant and finite means and variances; that is, the $\{x_t\}$ process is (weakly) stationary.

To continue the analysis, consider the $AR(1)$ model. If the x_t have constant means and variances, then

$$\mathrm{E}(x_t) = \frac{\phi_0}{1-\phi_1} \tag{9.85}$$

and

$$\mathrm{V}(x_t) = \frac{\sigma^2}{1-\phi_1^2}. \tag{9.86}$$

The first equation indicates that we cannot have $\phi_1 = 1$ and the second equation makes sense only if $|\phi_1| < 1$. For $AR(p)$ models in general, similar situations can occur. The denominators in the expressions for the mean and variance involve p-degree polynomials , similar to the first degree polynomials in the denominators of equations (9.85) and (9.86).

We call $f(z) = 1 - \phi_1 z^1 - \cdots - \phi_p z^p$ the *associated polynomial.* If $f(z) = 0$, we have situations similar to a 0 in the denominator of equation (9.86). If a root of $f(z) = 0$ is 1, the expression for a variance is infinite (which we see immediately from equation (9.86) for the $AR(1)$ model). This situation is called a "unit root". If some root is greater than 1, we have an expression for a variance that is negative. Hence, in order for the model to make sense in all respects, all roots of of the associated polynomial must be less than 1 in modulus. (Note some roots can contain imaginary components.)

Although many of the mechanical manipulations in the analysis of the model may be unaffected by unit roots, they have serious implications for the interpretation of the model.

9.8.3.1 Relation of the Autocorrelations to the Autoregressive Coefficients

From the model (9.84) we can see that $\rho_h = 0$ for $h > p$, and $\rho_1, \ldots \rho_p$ are determined by $\phi_1, \ldots \phi_p$ by the relationship

$$R\phi = \rho, \tag{9.87}$$

where ϕ and ρ are the p-vectors of the ϕ_is ($i \neq 0$) and the ρ_is, and R is the Toeplitz matrix (see Sect. 8.8.4)

$$R = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{p-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{p-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{p-3} \\ \vdots & & & \ddots & \vdots \\ \rho_{p-1} & \rho_{p-2} & \rho_{p-3} & \cdots & 1 \end{bmatrix}.$$

This system of equations is called the Yule-Walker equations. Notice the relationship of the Yule-Walker equations to the unit root problem mentioned above. For example, for $p = 1$, we have $\phi_1 = \rho_1$. In order for to be a correlation, it must be the case that $|\phi_1| \leq 1$.

For a given set of ρs, possibly estimated from some observations on the time series, Algorithm 9.3 can be used to solve the system (9.87).

Algorithm 9.3 Solution of the Yule-Walker system (9.87)

1. Set $k = 0$; $\phi_1^{(k)} = -\rho_1$; $b^{(k)} = 1$; and $a^{(k)} = -\rho_1$.
2. Set $k = k + 1$.
3. Set $b^{(k)} = \left(1 - \left(a^{(k-1)}\right)^2\right) b^{(k-1)}$.

```
4. Set $a^{(k)} = -\left(\rho_{k+1} + \sum_{i=1}^{k} \rho_{k+1-i}\phi_1^{(k-1)}\right) / b^{(k)}$.
5. For $i = 1, 2, \ldots, k$
     set $y_i = \phi_i^{(k-1)} + a^{(k)}\phi_{k+1-i}^{(k-1)}$.
6. For $i = 1, 2, \ldots, k$
     set $\phi_i^{(k)} = y_i$.
7. Set $\phi_{k+1}^{(k)} = a^{(k)}$.
8. If $k < p - 1$, go to step 1; otherwise terminate.    ■
```

This algorithm is $\mathrm{O}(p)$ (see Golub and Van Loan 1996).

The Yule-Walker equations arise in many places in the analysis of stochastic processes. Multivariate versions of the equations are used for a vector time series (see Fuller 1995; for example).

Exercises

9.1. Let X be an $n \times m$ matrix with $n > m$ and with entries sampled independently from a continuous distribution (of a real-valued random variable). What is the probability that $X^{\mathrm{T}}X$ is positive definite?

9.2. From equation (9.18), we have $\hat{y}_i = y^{\mathrm{T}}X(X^{\mathrm{T}}X)^{+}x_{i*}$. Show that h_{ii} in equation (9.19) is $\partial\hat{y}_i/\partial y_i$.

9.3. Formally prove from the definition that the sweep operator is its own inverse.

9.4. Consider the regression model

$$y = X\beta + \epsilon \tag{9.88}$$

subject to the linear equality constraints

$$L\beta = c, \tag{9.89}$$

and assume that X is of full column rank.

a) Let λ be the vector of Lagrange multipliers. Form

$$(b^{\mathrm{T}}L^{\mathrm{T}} - c^{\mathrm{T}})\lambda$$

and

$$(y - Xb)^{\mathrm{T}}(y - Xb) + (b^{\mathrm{T}}L^{\mathrm{T}} - c^{\mathrm{T}})\lambda.$$

Now differentiate these two expressions with respect to λ and b, respectively, set the derivatives equal to zero, and solve to obtain

$$\begin{aligned}\widehat{\beta}_C &= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y - \frac{1}{2}(X^{\mathrm{T}}X)^{-1}L^{\mathrm{T}}\widehat{\lambda}_C \\ &= \widehat{\beta} - \frac{1}{2}(X^{\mathrm{T}}X)^{-1}L^{\mathrm{T}}\widehat{\lambda}_C\end{aligned}$$

and

$$\widehat{\lambda}_C = -2(L(X^{\mathrm{T}}X)^{-1}L^{\mathrm{T}})^{-1}(c - L\widehat{\beta}).$$

Now combine and simplify these expressions to obtain expression (9.29) (on page 416).

b) Prove that the stationary point obtained in Exercise 9.4a actually minimizes the residual sum of squares subject to the equality constraints.
Hint: First express the residual sum of squares as

$$(y - X\widehat{\beta})^{\mathrm{T}}(y - X\widehat{\beta}) + (\widehat{\beta} - b)^{\mathrm{T}}X^{\mathrm{T}}X(\widehat{\beta} - b),$$

and show that is equal to

$$(y-X\widehat{\beta})^{\mathrm{T}}(y-X\widehat{\beta})+(\widehat{\beta}-\widehat{\beta}_C)^{\mathrm{T}}X^{\mathrm{T}}X(\widehat{\beta}-\widehat{\beta}_C)+(\widehat{\beta}_C-b)^{\mathrm{T}}X^{\mathrm{T}}X(\widehat{\beta}_C-b),$$

which is minimized when $b = \widehat{\beta}_C$.

c) Show that sweep operations applied to the matrix (9.30) on page 416 yield the restricted least squares estimate in the (1,2) block.

d) For the weighting matrix W, derive the expression, analogous to equation (9.29), for the generalized or weighted least squares estimator for β in equation (9.88) subject to the equality constraints (9.89).

9.5. Derive a formula similar to equation (9.33) to update $\widehat{\beta}$ due to the deletion of the i^{th} observation.

9.6. When data are used to fit a model such as $y = X\beta + \epsilon$, a large leverage of an observation is generally undesirable. If an observation with large leverage just happens not to fit the "true" model well, it will cause $\widehat{\beta}$ to be farther from β than a similar observation with smaller leverage.

a) Use artificial data to study influence. There are two main aspects to consider in choosing the data: the pattern of X and the values of the residuals in ϵ. The true values of β are not too important, so β can be chosen as 1. Use 20 observations. First, use just one independent variable ($y_i = \beta_0 + \beta_1 x_i + \epsilon_i$). Generate 20 x_is more or less equally spaced between 0 and 10, generate 20 ϵ_is, and form the corresponding y_is. Fit the model, and plot the data and the model. Now, set $x_{20} = 20$, set ϵ_{20} to various values, form the y_i's and fit the model for each value. Notice the influence of x_{20}.
Now, do similar studies with three independent variables. (Do not plot the data, but perform the computations and observe the effect.) Carefully write up a clear description of your study with tables and plots.

b) Heuristically, the leverage of a point arises from the distance from the point to a fulcrum. In the case of a linear regression model, the measure of the distance of observation i is

$$\Delta(x_i, X1/n) = \|x_i, X1/n\|.$$

(This is not the same quantity from the hat matrix that is defined as the leverage on page 410, but it should be clear that the influence of a point for which $\Delta(x_i, X1/n)$ is large is greater than that of a point for which the quantity is small.) It may be possible to overcome some of the undesirable effects of differential leverage by using weighted least squares to fit the model. The weight w_i would be a decreasing function of $\Delta(x_i, X1/n)$.

Now, using datasets similar to those used in the previous part of this exercise, study the use of various weighting schemes to control the influence. Weight functions that may be interesting to try include

$$w_i = e^{-\Delta(x_i, X1/n)}$$

and

$$w_i = \max(w_{\max}, \|\Delta(x_i, X1/n)\|^{-p})$$

for some $w_{\max}$ and some $p > 0$. (Use your imagination!)

Carefully write up a clear description of your study with tables and plots.

c) Now repeat Exercise 9.6b except use a decreasing function of the leverage, h_{ii} from the hat matrix in equation (9.18) instead of the function $\Delta(x_i, X1/n)$.

Carefully write up a clear description of this study, and compare it with the results from Exercise 9.6b.

9.7. By differentiating expression (9.38), derive the normal equations (9.39) for the multivariate linear model.

9.8. Formally prove the relationship expressed in equation (9.71) on page 442.

Hint: Use equation (9.70) twice.

9.9. On page 211, we used Lagrange multipliers to determine the normalized vector x that maximized $x^{\mathrm{T}}Ax$. If A is S_X, this is the first principal component. We also know the principal components from the spectral decomposition. We could also find them by sequential solutions of Lagrangians. After finding the first principal component, we would seek the linear combination z such that $X_{\mathrm{c}}z$ has maximum variance among all normalized z that are orthogonal to the space spanned by the first principal component; that is, that are $X_{\mathrm{c}}^{\mathrm{T}}X_{\mathrm{c}}$-*conjugate* to the first principal component (see equation (3.93) on page 94). If V_1 is the matrix whose columns are the eigenvectors associated with the largest eigenvector, this is equivalent to finding z so as to maximize $z^{\mathrm{T}}Sz$ subject to $V_1^{\mathrm{T}}z = 0$. Using the method of Lagrange multipliers as in equation (4.54), we form the Lagrangian corresponding to equation (4.56) as

$$z^{\mathrm{T}}Sz - \lambda(z^{\mathrm{T}}z - 1) - \phi V_1^{\mathrm{T}}z,$$

where λ is the Lagrange multiplier associated with the normalization requirement $z^{\mathrm{T}}z = 1$, and ϕ is the Lagrange multiplier associated with the orthogonality requirement. Solve this for the second principal component, and show that it is the same as the eigenvector corresponding to the second-largest eigenvalue.

9.10. Obtain the "Longley data". (It is a dataset in R, and it is also available from `statlib`.) Each observation is for a year from 1947 to 1962 and consists of the number of people employed, five other economic variables, and the year itself. Longley (1967) fitted the number of people employed to a linear combination of the other variables, including the year.

a) Use a regression program to obtain the fit.

b) Now consider the year variable. The other variables are measured (estimated) at various times of the year, so replace the year variable with a "midyear" variable (i.e., add $\frac{1}{2}$ to each year). Redo the regression. How do your estimates compare?

c) Compute the L_2 condition number of the matrix of independent variables. Now add a ridge regression diagonal matrix, as in the matrix (9.90), and compute the condition number of the resulting matrix. How do the two condition numbers compare?

9.11. Consider the least squares regression estimator (9.15) for full rank $n \times m$ matrix X $(n > m)$:

$$\widehat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y.$$

a) Compare this with the ridge estimator

$$\widehat{\beta}_{R(d)} = (X^{\mathrm{T}}X + dI_m)^{-1}X^{\mathrm{T}}y$$

for $d \geq 0$. Show that

$$\|\widehat{\beta}_{R(d)}\| \leq \|\widehat{\beta}\|.$$

b) Show that $\widehat{\beta}_{R(d)}$ is the least squares solution to the regression model similar to $y = X\beta + \epsilon$ except with some additional artificial data; that is, y is replaced with

$$\begin{pmatrix} y \\ 0 \end{pmatrix},$$

where 0 is an m-vector of 0s, and X is replaced with

$$\begin{bmatrix} X \\ dI_m \end{bmatrix}. \tag{9.90}$$

Now explain why $\widehat{\beta}_{R(d)}$ is shorter than $\widehat{\beta}$.

9.12. Use the Schur decomposition (equation (3.190), page 122) of the inverse of $(X^{\mathrm{T}}X)$ to prove equation (9.53).

9.13. Given the matrix

$$A = \begin{bmatrix} 2 & 1 & 3 \\ 1 & 2 & 3 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix},$$

assume the random 3×2 matrix X is such that

$$\mathrm{vec}(X - A)$$

has a $\mathrm{N}(0, V)$ distribution, where V is block diagonal with the matrix

$$\begin{bmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 2 \end{bmatrix}$$

along the diagonal. Generate ten realizations of X matrices, and use them to test that the rank of A is 2. Use the test statistic (9.64) on page 437.

9.14. Construct a 9×2 matrix X with some missing values, such that S_X computed using all available data for the covariance or correlation matrix is not nonnegative definite.

9.15. Consider an $m \times m$, symmetric nonsingular matrix, R, with 1s on the diagonal and with all off-diagonal elements less than 1 in absolute value. If this matrix is positive definite, it is a correlation matrix. Suppose, however, that some of the eigenvalues are negative. Iman and Davenport (1982) describe a method of adjusting the matrix to a "near-by" matrix that is positive definite. (See Ronald L. Iman and James M. Davenport, 1982, *An Iterative Algorithm to Produce a Positive Definite Correlation Matrix from an "Approximate Correlation Matrix"*, Sandia Report SAND81-1376, Sandia National Laboratories, Albuquerque, New Mexico.) For their method, they assumed the eigenvalues are unique, but this is not necessary in the algorithm.

Before beginning the algorithm, choose a small positive quantity, ϵ, to use in the adjustments, set $k = 0$, and set $R^{(k)} = R$.

1. Compute the eigenvalues of $R^{(k)}$,

$$c_1 \geq c_2 \geq \ldots \geq c_m,$$

and let p be the number of eigenvalues that are negative. If $p = 0$, stop. Otherwise, set

$$c_i^* = \begin{cases} \epsilon & \text{if } c_i < \epsilon \\ c_i & \text{otherwise} \end{cases} \quad \text{for } i = p_1, \ldots, m - p, \tag{9.91}$$

where $p_1 = \max(1, m - 2p)$.

2. Let
$$\sum_i c_i v_i v_i^{\mathrm{T}}$$
be the spectral decomposition of R (equation (3.256), page 154), and form the matrix R^*:
$$R^* = \sum_{i=1}^{p_1} c_i v_i v_i^{\mathrm{T}} + \sum_{i=p_1+1}^{m-p} c_i^* v_i v_i^{\mathrm{T}} + \sum_{i=m-p+1}^{m} \epsilon v_i v_i^{\mathrm{T}}.$$
3. Form $R^{(k)}$ from R^* by setting all diagonal elements to 1.
4. Set $k = k+1$, and go to step 1. (The algorithm iterates on k until $p = 0$.)

Write a program to implement this adjustment algorithm. Write your program to accept any size matrix and a user-chosen value for ϵ. Test your program on the correlation matrix from Exercise 9.14.

9.16. Consider some variations of the method in Exercise 9.15. For example, do not make the adjustments as in equation (9.91), or make different ones. Consider different adjustments of R^*; for example, adjust any off-diagonal elements that are greater than 1 in absolute value.
Compare the performance of the variations.

9.17. Investigate the convergence of the method in Exercise 9.15. Note that there are several ways the method could converge.

9.18. Suppose the method in Exercise 9.15 converges to a positive definite matrix $R^{(n)}$. Prove that all off-diagonal elements of $R^{(n)}$ are less than 1 in absolute value. (This is true for any positive definite matrix with 1s on the diagonal.)

9.19. Shrinkage adjustments of approximate correlation matrices.

a) Write a program to implement the linear shrinkage adjustment of equation (9.66). Test your program on the correlation matrix from Exercise 9.14.

b) Write a program to implement the nonlinear shrinkage adjustment of equation (9.67). Let $\delta = 0.05$ and
$$f(x) = \tanh(x).$$
Test your program on the correlation matrix from Exercise 9.14.

c) Write a program to implement the scaling adjustment of equation (9.68). Recall that this method applies to an approximate correlation matrix that is a pseudo-correlation matrix. Test your program on the correlation matrix from Exercise 9.14.

9.20. Show that the matrices generated in Algorithm 9.2 are correlation matrices. (They are clearly nonnegative definite, but how do we know that they have 1s on the diagonal?)

9.21. Consider a two-state Markov chain with transition matrix

$$P = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}$$

for $0 < \alpha < 1$ and $0 < \beta < 1$. Does an invariant distribution exist, and if so what is it?

9.22. Recall from Exercise 8.10 that a Leslie matrix has a single unique positive eigenvalue.

a) What are the conditions on a Leslie matrix A that allow a stable age distribution? Prove your assertion.
Hint: Review the development of the power method in equations (7.9) and (7.10).

b) What are the conditions on a Leslie matrix A that allow a stable population, that is, for some x_t, $x_{t+1} = x_t$?

c) Derive equation (9.83). (Recall that there are approximations that result from the use of a discrete model of a continuous process.)

9.23. Derive equations (9.85) and (9.86) under the stationarity assumptions for the model (9.84).

9.24. Derive the Yule-Walker equations (9.87) for the model (9.84).