PART I

The Analysis of Variance in the Case of Models with Fixed Effects and Independent Observations of Equal Variance

CHAPTER 1

Point Estimation

1.1. INTRODUCTION

The following rough definition of our subject may serve tentatively: The analysis of variance[1] is a statistical technique for analyzing measurements depending on several kinds of effects operating simultaneously, to decide which kinds of effects are important and to estimate the effects. The measurements or observations may be in an experimental science like genetics or a nonexperimental one like astronomy. A theory of analyzing measurements naturally has implications about how the experiment should be planned or the observations should be taken, i.e., *experimental design*. Historically, the present technique of analysis of variance has been developed mainly in connection with problems of agricultural experimentation.

An agricultural experiment of a relatively simple structure to which the analysis of variance would be applicable would be the following: In each of three localities four varieties of tomatoes are grown in tanks containing chemical solutions. Two different chemical solutions, which we shall call "treatments," are used, with different proportions of the chemicals. For each treatment in each locality there is a mixing tank from which the fluid is pumped to all the tanks on this treatment, connected "in parallel." We do not want a "series" connection, where the outflow from one tank is the inflow to another, because this would *confound* the effects of the varieties in these two tanks with the effects (if any) of order in the "series" connection. The tanks are arranged outdoors with the same orientation,

[1] The analysis of variance, as commonly understood and practiced today, has been developed chiefly by R. A. Fisher (1918, 1925, 1935), who introduced the terms *variance* and *analysis of variance* into statistics. The latter term would seem more appropriate for the random-effects models (Ch. 7), and these may constitute the path by which Fisher himself originally approached the subject. For some historical background see Scheffé (1956*b*). Names followed by dates in parentheses refer to the author index and bibliography at the end of the book.

so that the plants in one tank will not appreciably shade those in another, etc. For each treatment in the three localities the chemicals are renewed according to the same specifications. Each variety is grown in a separate tank, with the same number of plants in each. The *yield* of each tank is the weight of ripe tomatoes produced. (Later we shall speak about an observed yield and a "true" or expected yield.)

The yield from a tank may depend on the variety, the chemical treatment, and the locality. In particular, it will depend on *interactions* among these factors, a useful concept of the analysis of variance that we will develop later (sec. 4.1). The sort of questions for which our theory offers answers is the following: Are the varieties different in yield when averaged over the two treatments and three localities? Do the yields demonstrate differential effects of the varieties for different localities? How can we quantitatively express the differences with a given degree of confidence? Etc.

The developments in the remainder of this chapter and in Ch. 2 are of a generality that may somewhat dismay a reader expecting to find results whose usefulness is clearly visible to him. Such a reader may be encouraged to work through these two chapters by the following remarks: Beginning with Ch. 3, he will find that most of the material in the rest of the book is of a more obvious usefulness. The general developments of Chs. 1 and 2 furnish the foundation not only for obtaining these results, but also for carrying out the analysis of variance in cases he may encounter that do not fall under any of those treated specifically and in detail in the rest of the book.

1.2. MATHEMATICAL MODELS

Suppose that we have n observations or measurements. In the mathematical models employed in this book it is assumed that these observations are values taken on by n random variables[2] $y_1, y_2, \cdots, y_n$, which are constituted of linear combinations of p unknown quantities $\beta_1, \beta_2, \cdots, \beta_p$ plus errors $e_1, e_2, \cdots, e_n$,

$$(1.2.1) \qquad y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + \cdots + x_{pi}\beta_p + e_i \qquad (i = 1, 2, \cdots, n),$$

where[3] the $\{x_{ji}\}$ are known constant coefficients. (The reader unfamiliar with the brace notation $\{\ \}$ should read the footnote.[4]) The

[2] We will generally use the same symbols for random variables and for their observed values in this book. (Exceptions occur in secs. 2.10 and 9.3.)

[3] It might seem more natural to permute the subscripts on the x's in (1.2.1), but the present notation is standard. It would seem appropriate in situations where x_{ji} is the value assumed by an "independent" variable x_j in the ith observation; see sec. 6.1.

[4] The brace notation denotes the *set* of quantities indicated: In this case $\{x_{ji}\}$ means the set consisting of the np quantities x_{ji} with $j = 1, 2, \cdots, p$; $i = 1, 2, \cdots, n$.

$\{\beta_j\}$ are more or less idealized formulations of some aspects of interest to the investigator in the phenomena underlying the observations. The purpose of the analysis of variance is to make inferences about the $\{e_i\}$ and some of the $\{\beta_j\}$, the inferences to be valid regardless of the values of the other $\{\beta_j\}$, if any, which we may be more desirous of "eliminating" than "assessing."

A minimal assumption which is always made on the random variables $\{e_i\}$ is that their expected values are zero:

$$E(e_i) = 0 \qquad (i = 1, 2, \cdots, n). \tag{1.2.2}$$

We shall also usually assume that

$$E(e_i e_j) = \sigma^2 \delta_{ij}, \tag{1.2.3}$$

where σ^2 is an unknown constant and δ_{ij} is 0 or 1, according as $i \neq j$ or $i = j$, respectively. This is equivalent to saying that the random variables $\{e_i\}$ are uncorrelated (i.e., have zero coefficients of correlation) and have equal variance σ^2.

We may now make our definition in sec. 1.1 gradually more precise: The *analysis of variance* is a body of statistical methods of analyzing measurements assumed to be of the structure (1.2.1), where the coefficients $\{x_{ji}\}$ are integers usually 0 or 1. In order to clarify[5] this it is necessary to consider not only the values of the $\{x_{ji}\}$ but also their origin in the real situation being investigated: In the analysis of variance the $\{x_{ji}\}$ are the values of "counter variables" or "indicator variables" which refer to the presence or absence of the *effects* $\{\beta_j\}$ in the conditions under which the observations are taken: x_{ji} is the number of times β_j occurs in the ith observation, and this is usually[6] 0 or 1. If the $\{x_{ji}\}$ are values taken on in the observations not by counter variables but by continuous variables like t = time, T = temperature, t^2, e^{-t}, tT, t^0, etc. (these are called independent or concomitant variables, and the observations $\{y_i\}$ are then said to be on a dependent variable y; see sec. 6.1), we say we have a case of *regression analysis*. If there are some $\{x_{ji}\}$ of both kinds, we have an *analysis of covariance*. More natural and meaningful but equivalent definitions to distinguish among the three kinds of analysis, all of which fall under the general theory of Chs. 1 and 2, will be formulated in Ch. 6 after the reader has become accustomed to thinking in terms of the factors that are varied in an experiment or series of observations.

[5] These definitions and those of sec. 6.1 grew out of helpful discussions I had with Professor William Kruskal and Dr. Mervin Muller.

[6] For an example where some $x_{ji} = -1$ see Scheffé (1952), sec. 7; where some $x_{ji} = 2$, see Kempthorne (1952), sec. 6.8.

Up to now we have not specified the nature of the unknown effects $\{\beta_j\}$: They may be either unknown constants, which we then call parameters, or unobservable random variables subject to further assumptions about their distribution involving other unknown parameters. We shall call a model in which all the $\{\beta_j\}$ are unknown constants a *fixed-effects model.*[7] It often happens that one of the $\{\beta_j\}$ is a constant which occurs with every observation with coefficient 1 so that, for this j, $x_{ji} = 1$ for all i. We may call such a β_j an *additive constant* (in applications it is usually a "general mean" in some sense). A model in which all the $\{\beta_j\}$ are random variables, except possibly for one which is an additive constant, is called a *random-effects model.* Intermediate cases, where at least one β_j is a random variable and at least one is a constant not an additive constant, are called *mixed models.*

Examples: We wish to illustrate the notation now, but it is not convenient to use typical analysis-of-variance examples at this point because they would introduce other complications better postponed.

1. Consider the problem of fitting a polynomial of degree three, $y = a_0+a_1x+a_2x^2+a_3x^3$, to a set of observed values (x_i, y_i), $i = 1, \cdots, n$, assuming that y_i is a random variable, x_i is not, and the expected value of y_i is the ordinate on the cubic curve at $x = x_i$:

$$E(y_i) = a_0 + a_1x_i + a_2x_i^2 + a_3x_i^3.$$

We have in this case $p = 4$, $\beta_j = a_{j-1}$ $(j = 1, \cdots, 4)$. We note that the regression, in this case $a_0+a_1x+a_2x^2+a_3x^3$, need not be linear in the "independent" variable x, but only in the unknown parameters.

2. Another problem might be to fit a trigonometric polynomial to some periodic data with known period (which by change of the time scale we could make 2π):

$$\begin{aligned} E(y_i) = a_0 &+ a_1 \cos t_i + b_1 \sin t_i \\ &+ a_2 \cos 2t_i + b_2 \sin 2t_i \\ &+ a_3 \cos 3t_i + b_3 \sin 3t_i. \end{aligned}$$

Here the observation y_i is made at time t_i and the $\{\beta_j\}$ are the seven a's and b's.

These examples indicate that our models include a great variety of situations.

The development of the general theory in Chs. 1, 2, and 6 is greatly facilitated[8] by the use of vector and matrix algebra. The author hopes he has given a sufficient introduction to this in Apps. I and II. We define

[7] Fixed-effects models are also called Model I, and random-effects models, Model II, following Eisenhart (1947).

[8] See Preface.

the vectors (vectors and matrices will always be printed in **boldface** type)

$$\mathbf{y}^{n\times 1} = \begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix}, \qquad \boldsymbol{\beta}^{p\times 1} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \cdot \\ \beta_p \end{pmatrix}, \qquad \mathbf{e}^{n\times 1} = \begin{pmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ \cdot \\ e_n \end{pmatrix},$$

and the matrix

$$\mathbf{X}^{p\times n} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{pmatrix},$$

where superscripts $r\times s$ on a matrix indicate that the matrix has r rows and s columns. When there is no risk of ambiguity we drop the superscripts. The set of equations (1.2.1) then takes the simple form

$$\mathbf{y} = \mathbf{X}'\boldsymbol{\beta} + \mathbf{e}, \tag{1.2.4}$$

where $\mathbf{X}'$ denotes the transpose of $\mathbf{X}$.

Matrix Random Variables

Definition: Given a matrix $\mathbf{V}^{r\times s}$ of jointly distributed random variables $\{v_{ij}\}$ with finite expectations,

$$\mathbf{V} = \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1s} \\ v_{21} & v_{22} & \cdots & v_{2s} \\ \cdot & \cdot & \cdots & \cdot \\ v_{r1} & v_{r2} & \cdots & v_{rs} \end{pmatrix},$$

we define the *expected value of the matrix* $\mathbf{V}$ to be the matrix

$$E(\mathbf{V}) = \begin{pmatrix} E(v_{11}) & E(v_{12}) & \cdots & E(v_{1s}) \\ E(v_{21}) & E(v_{22}) & \cdots & E(v_{2s}) \\ \cdot & \cdot & \cdots & \cdot \\ E(v_{r1}) & E(v_{r2}) & \cdots & E(v_{rs}) \end{pmatrix}. \tag{1.2.5}$$

This definition enables us to write the conditions (1.2.2) and (1.2.3) in the condensed matrix form

$$E(\mathbf{e}) = \mathbf{0}, \qquad E(\mathbf{ee}') = \sigma^2\mathbf{I}, \tag{1.2.6}$$

where $\mathbf{0}$ is the $n \times 1$ zero matrix and $\mathbf{I}$ is the $n \times n$ identity matrix.

Lemma: If $\mathbf{A}^{q\times r}$ and $\mathbf{B}^{s\times t}$ are matrices of constants and $\mathbf{V}^{r\times s}$ is a matrix of random variables we have the relation

$$E(\mathbf{AVB}) = \mathbf{A}\,E(\mathbf{V})\mathbf{B}. \tag{1.2.7}$$

Proof: In the proof only the linear operator property of the operator E on ordinary random variables is utilized, i.e., $E(ax+by) = a\,E(x)+b\,E(y)$, if a and b are constants, and x and y are random variables.

Covariance Matrices

Consider a vector $\mathbf{v} = (v_1, \cdots, v_n)'$ of jointly distributed random variables all having finite variance. We call the matrix

$$\mathbf{\Sigma}_v = (\mathrm{Cov}\,(v_i, v_j)), \tag{1.2.8}$$

whose i,j element is the covariance of v_i and v_j, the *covariance matrix* of $\mathbf{v}$. Write $\mu_i = E(v_i)$, so $\mathrm{Cov}\,(v_i, v_j) = E[(v_i-\mu_i)(v_j-\mu_j)]$. Then by (1.2.5) we may write

$$\mathbf{\Sigma}_v = E[(\mathbf{v}-\boldsymbol{\mu})(\mathbf{v}-\boldsymbol{\mu})'], \tag{1.2.9}$$

where $\boldsymbol{\mu} = E(\mathbf{v})$.

We shall make frequent use of the following property: For a linear transformation

$$\mathbf{w}^{m\times 1} = \mathbf{A}^{m\times n}\mathbf{v}^{n\times 1}$$

from n random variables $v_1, \cdots, v_n$ to m random variables $w_1, \cdots, w_m$, with matrix $\mathbf{A}$, the covariance matrix of $\mathbf{w}$ is given by

$$\mathbf{\Sigma}_w = \mathbf{A}\mathbf{\Sigma}_v\mathbf{A}'. \tag{1.2.10}$$

Proof: $\mathbf{\Sigma}_w = E([\mathbf{w} - E(\mathbf{w})][\mathbf{w} - E(\mathbf{w})]') = E(\mathbf{A}[\mathbf{v} - E(\mathbf{v})][\mathbf{v} - E(\mathbf{v})]'\mathbf{A}')$ $= \mathbf{A}\,E([\mathbf{v} - E(\mathbf{v})][\mathbf{v} - E(\mathbf{v})]')\mathbf{A}' = \mathbf{A}\mathbf{\Sigma}_v\mathbf{A}'$.

1.3. LEAST-SQUARES ESTIMATES AND NORMAL EQUATIONS

We use the symbol Ω throughout this book to denote a set of fundamental or underlying assumptions. Here we consider the following ones already introduced in sec. 1.2:

$$\Omega: \quad \mathbf{y}^{n\times 1} = \mathbf{X}'\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad E(\mathbf{e}) = \mathbf{0}, \qquad E(\mathbf{ee}') = \sigma^2\mathbf{I},$$

which may be written even more briefly as

$$\Omega: \quad E(\mathbf{y}) = \mathbf{X}'\boldsymbol{\beta}, \qquad \boldsymbol{\Sigma}_y = \sigma^2\mathbf{I}.$$

Suppose that $b_1, \cdots, b_p$ denote quantities which we might consider using as estimates of $\beta_1, \cdots, \beta_p$. The $\beta_1, \cdots, \beta_p$ are fixed unknown constants, whereas the $b_1, \cdots, b_p$ will be quantities that we vary freely in deciding which are the "best" values in some sense. For any $\mathbf{b} = (b_1, \cdots, b_p)'$ we form

$$\mathscr{S}(\mathbf{y}, \mathbf{b}) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ji} b_j \right)^2. \tag{1.3.1}$$

This might be interpreted as $\Sigma_1^n \hat{e}_i^2$, where $\hat{e}_i$ denotes the estimate of the error e_i in the observation y_i in (1.2.1) if $\boldsymbol{\beta}$ is estimated by $\mathbf{b}$. It can be regarded as a measure of how well the model with $\boldsymbol{\beta}$ estimated by $\mathbf{b}$ fits the observations; the smaller $\mathscr{S}$ is, the better the fit. In matrix notation we may write $\mathscr{S} = (\mathbf{y}-\mathbf{X}'\mathbf{b})'(\mathbf{y}-\mathbf{X}'\mathbf{b})$ or, if the length of a vector $\mathbf{v}$ is denoted by $\|\mathbf{v}\|$, as

$$\mathscr{S}(\mathbf{y}, \mathbf{b}) = \|\mathbf{y}-\mathbf{X}'\mathbf{b}\|^2. \tag{1.3.2}$$

Definition: A set of functions[9] of $\mathbf{y}$ (i.e., a set of *statistics*),

$$\hat{\beta}_1 = \hat{\beta}_1(\mathbf{y}), \cdots, \hat{\beta}_p = \hat{\beta}_p(\mathbf{y}),$$

such that the values $b_j = \hat{\beta}_j$ $(j = 1, \cdots, p)$ minimize $\mathscr{S}(\mathbf{y}, \mathbf{b})$, is called a set of *LS* (*least-squares*)[10] *estimates* of the $\{\beta_j\}$.

Normal Equations

We shall see that LS estimates always exist, but need not be unique. Later it will be seen that any set of LS estimates satisfies the conditions (the reader whose calculus is rusty should read the footnote[11])

$$\partial\mathscr{S}(\mathbf{y}, \mathbf{b})/\partial b_\nu = 0 \qquad (\nu = 1, \cdots, p).$$

[9] For the mathematically advanced reader we remark that here and elsewhere we mean *measurable* functions. If the $\{\hat{\beta}_j\}$ are unique they turn out to be linear functions of the $\{y_i\}$; if they are not unique there are infinitely many which are linear functions, and it might be convenient (for example in considering their covariance matrix) to restrict them to linear functions.

[10] The LS (least-squares) method of estimation was invented independently, and published in books on astronomical problems, by Gauss (1809) and Legendre (1806).

[11] The notation $\partial\mathscr{S}(\mathbf{y}, \mathbf{b})/\partial b_\nu$ denotes the partial derivative of $\mathscr{S}(\mathbf{y}, \mathbf{b})$ with respect to b_ν, meaning an ordinary derivative with respect to b_ν when the other $\{b_i\}$ are held constant. All the partial derivatives needed in this book can be written down at once by the following rule: We shall always want the partial derivative with respect to some variable θ of a function $\mathscr{S}$ which is a sum of squares of expressions each of which is linear in θ (i.e., of the form $A+B\theta$, where A and B do not depend on θ). This partial derivative is equal to twice the sum of products of those expressions (not their squares!) containing θ by the coefficients of θ in those expressions.

These give

$$\partial \mathscr{S}/\partial b_\nu = -2 \sum_{i=1}^{n} \Big(y_i - \sum_{j=1}^{p} x_{ji} b_j \Big) x_{\nu i} = 0 \qquad (\nu = 1, \cdots, p),$$

or

$$\sum_{i=1}^{n} \sum_{j=1}^{p} x_{\nu i} x_{ji} b_j = \sum_{i=1}^{n} x_{\nu i} y_i \qquad (\nu = 1, \cdots, p). \tag{1.3.3}$$

These equations may now be written in matrix form

$$\mathbf{X}\mathbf{X}'\mathbf{b} = \mathbf{X}\mathbf{y},$$

or, with $\mathbf{S} = \mathbf{X}\mathbf{X}'$,

$$\mathbf{S}\mathbf{b} = \mathbf{X}\mathbf{y}. \tag{1.3.4}$$

These are the *normal equations*. We use the symbol $\hat{\hat{\boldsymbol{\beta}}}$ for any solution $\mathbf{b}$ of them, reserving the symbol $\hat{\boldsymbol{\beta}}$ for LS estimates exclusively. However, we are going to show that every solution of the normal equations is a set of LS estimates, and every set of LS estimates satisfies the normal equations, so that thereafter it will be justified to write simply $\hat{\boldsymbol{\beta}}$ instead of $\hat{\hat{\boldsymbol{\beta}}}$. In setting up and solving the normal equations in practice, we do not distinguish in the notation between $\mathbf{b}$ and $\boldsymbol{\beta}$, setting $\partial \mathscr{S}(\mathbf{y}, \boldsymbol{\beta})/\partial \beta_\nu = 0$ $(\nu = 1, \cdots, p)$, solving for $\boldsymbol{\beta}$, and then denoting the solution $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$. We hope that it will not be confusing after this explanation.

Geometrical Interpretation

We are now going to prove the existence of the LS estimates and their equivalence to the solutions of the normal equations. For this purpose we use results from vector algebra which are derived in App. I.

In the n-dimensional space V_n we introduce the vector of means $\boldsymbol{\eta} = E(\mathbf{y})$; so under Ω

$$\boldsymbol{\eta}^{n \times 1} = \mathbf{X}'\boldsymbol{\beta}, \tag{1.3.5}$$

which we also write as (this is obvious by the interpretation of matrix multiplication above (II.7) of App. II)

$$\boldsymbol{\eta} = \beta_1 \boldsymbol{\xi}_1 + \beta_2 \boldsymbol{\xi}_2 + \cdots + \beta_p \boldsymbol{\xi}_p,$$

where the vector $\boldsymbol{\xi}_j^{n \times 1}$ is the jth column of $\mathbf{X}'$.

Let r denote the rank of $\mathbf{X}$, and V_r the r-dimensional vector space spanned (see App. I) by the vectors $\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_p$. Then a vector $\mathbf{z}^{n \times 1}$ lies in V_r if and only if there exist coefficients $b_1, \cdots, b_p$ such that $\mathbf{z} = b_1 \boldsymbol{\xi}_1 + \cdots + b_p \boldsymbol{\xi}_p$. In particular, $\boldsymbol{\eta} \in V_r$ under Ω.

Let $\mathbf{z} = \mathbf{X}'\mathbf{b}$, where we think of varying $\mathbf{b}$. Then from Theorem 2 of App. I it follows that $\mathscr{S}(\mathbf{y}, \mathbf{b}) = \|\mathbf{y} - \mathbf{z}\|^2$ has a minimum which is attained

when and only when $\mathbf{z}$ is the vector $\hat{\boldsymbol{\eta}}$ defined to be the projection of $\mathbf{y}$ on V_p. Since $\hat{\boldsymbol{\eta}}$ is a vector in V_p, it can be written as a linear combination of $\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_p$; i.e., there exist $b_1, \cdots, b_p$ such that

$$\hat{\boldsymbol{\eta}} = b_1\boldsymbol{\xi}_1 + \cdots + b_p\boldsymbol{\xi}_p; \tag{1.3.6}$$

here $\hat{\boldsymbol{\eta}}$ is unique but the $\{b_j\}$ in general are not. Since $\hat{\boldsymbol{\eta}}$ is a function of $\mathbf{y}$ only, and not of unknown parameters, the $\{b_j\}$ in (1.3.6) may also be taken to be functions of $\mathbf{y}$ only, and they are then LS estimates whose existence we have now demonstrated. Furthermore, any $\{b_1, \cdots, b_p\}$ which are functions of $\mathbf{y}$ only will be a set of LS estimates if and only if $\mathbf{X}'\mathbf{b} = \hat{\boldsymbol{\eta}}$, which is true if and only if each of the following statements holds (each statement is true if and only if the following one is true; the symbol $\perp$ denotes "is orthogonal to"),

$$\begin{aligned}
&\mathbf{X}'\mathbf{b} = \hat{\boldsymbol{\eta}},\\
&\mathbf{y} - \mathbf{X}'\mathbf{b} \perp V_p,\\
&\mathbf{y} - \mathbf{X}'\mathbf{b} \perp \boldsymbol{\xi}_j \qquad (j = 1, \cdots, p),\\
&\boldsymbol{\xi}_j'(\mathbf{y} - \mathbf{X}'\mathbf{b}) = 0 \qquad (j = 1, \cdots, p),\\
&\mathbf{X}(\mathbf{y} - \mathbf{X}'\mathbf{b}) = \mathbf{0},
\end{aligned} \tag{1.3.7}$$

$$\mathbf{X}\mathbf{X}'\mathbf{b} = \mathbf{X}\mathbf{y}. \tag{1.3.8}$$

Here (1.3.7) follows from Lemma 8 of App. I, and (1.3.8) states that $\mathbf{b}$ satisfies the normal equations. We have now proved that LS estimates $\hat{\beta}_1, \cdots, \hat{\beta}_p$ always exist, that any set of LS estimates satisfies the normal equations, and that any solution $\hat{\hat{\beta}}_1, \cdots, \hat{\hat{\beta}}_p$ of the normal equations, which is a function of $\mathbf{y}$ only, is a set of LS estimates.

Thus, there is no reason for using the symbol $\hat{\hat{\boldsymbol{\beta}}}$ any more, and $\hat{\boldsymbol{\beta}}$ will denote a solution of the normal equations as well as a set of LS estimates.

The situation can be easily visualized as in Fig. 1.3.1.

Notation: We use the symbol $\mathscr{S}_\Omega$ for the minimum value of $\mathscr{S}(\mathbf{y}, \mathbf{b})$,

$$\mathscr{S}_\Omega = \mathscr{S}(\mathbf{y}, \hat{\boldsymbol{\beta}}), \tag{1.3.9}$$

where $\hat{\boldsymbol{\beta}}$ is any set of LS estimates, or any solution of the normal equations. We shall call $\mathscr{S}_\Omega$ the *error sum of squares*, because in sec. 1.6 we shall see that it provides an estimate of the error variance σ^2. Although $\hat{\boldsymbol{\beta}}$ is not in general unique, $\mathscr{S}(\mathbf{y}, \hat{\boldsymbol{\beta}})$ is. A very useful expression for $\mathscr{S}_\Omega$ is

$$\mathscr{S}_\Omega = \sum_{i=1}^{n} y_i^2 - \sum_{\nu=1}^{p} \hat{\beta}_\nu r_\nu, \tag{1.3.10}$$

where $\{\hat{\beta}_1, \cdots, \hat{\beta}_p\}$ is any set of LS estimates, and r_ν is the right member

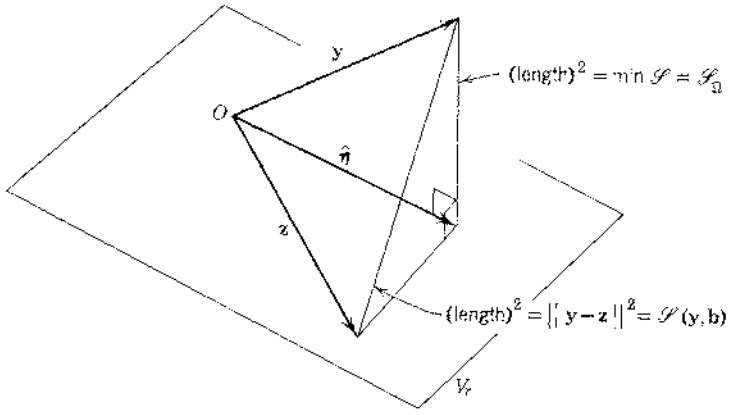

FIG. 1.3.1

of the rth normal equation (1.3.3). This expression may be derived as follows:

$$\mathscr{S}_\Omega = (\mathbf{y}-\mathbf{X}'\hat{\boldsymbol{\beta}})'(\mathbf{y}-\mathbf{X}'\hat{\boldsymbol{\beta}}) = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'(\mathbf{X}\mathbf{y}) + \hat{\boldsymbol{\beta}}'(\mathbf{X}\mathbf{X}'\hat{\boldsymbol{\beta}}-\mathbf{X}\mathbf{y}),$$

where we have used the fact that $\mathbf{y}'\mathbf{X}'\hat{\boldsymbol{\beta}}$ equals its transpose because it is a 1×1 matrix. This reduces to (1.3.10) because $\hat{\boldsymbol{\beta}}$ satisfies the normal equations $\mathbf{X}\mathbf{X}'\boldsymbol{\beta} = \mathbf{X}\mathbf{y}$.

Case where $\hat{\boldsymbol{\beta}}$ is Unique

The case where the $p\times n$ matrix $\mathbf{X}$ is of rank p is often called the *case of maximal rank*, or the *case of full rank*, because usually $p < n$. If rank $\mathbf{X} = p$ then (1.3.4) has a unique solution (and only then). In Theorem 7 of App. II we prove that rank $\mathbf{S}$ = rank $\mathbf{X}$, and hence in this case $\mathbf{S}$ is nonsingular. Thus $\mathbf{S}^{-1}$ exists, and the solution is given uniquely by

$$\hat{\boldsymbol{\beta}} = \mathbf{S}^{-1}\mathbf{X}\mathbf{y}. \tag{1.3.11}$$

Applying (1.2.10), we then obtain for the covariance matrix of $\hat{\boldsymbol{\beta}}$

$$\boldsymbol{\Sigma}_{\hat{\beta}} = (\mathbf{S}^{-1}\mathbf{X})\,\boldsymbol{\Sigma}_y(\mathbf{S}^{-1}\mathbf{X})'.$$

But $\mathbf{S}^{-1}$ is symmetric since $\mathbf{S}$ is, hence

$$\boldsymbol{\Sigma}_{\hat{\beta}} = \sigma^2\mathbf{S}^{-1}\mathbf{X}\mathbf{X}'\mathbf{S}^{-1},$$

and so finally

$$\boldsymbol{\Sigma}_{\hat{\beta}} = \sigma^2\mathbf{S}^{-1}. \tag{1.3.12}$$

Remarks: The case in which rank $\mathbf{X} = p$ sometimes occurs in practice. (It occurs usually in regression theory but not usually in the analysis of variance.) One does not then need the matrix $\mathbf{S}^{-1}$ in order to solve the normal equations. However, it is generally a good idea to compute $\mathbf{S}^{-1}$ first and then $\hat{\boldsymbol{\beta}}$ from (1.3.11) since it is almost always desirable subsequently to get also the covariance matrix $\boldsymbol{\Sigma}_{\hat{\beta}}$. The lack of uniqueness of the LS estimates $\{\hat{\beta}_j\}$ in the case where rank $\mathbf{X} < p$ is related to a similar nonuniqueness of the parameter values $\{\beta_j\}$; this is discussed further at the end of sec. 1.4. In connection with the result (1.3.10) involving the right members of the normal equations, and the result (1.3.12) if we think of $\mathbf{S}$ denoting the matrix of coefficients of the left members of the normal equations, it is of course essential that the normal equations be in exactly the form (1.3.3), where the νth equation is obtained by dividing by -2 the equation $\partial \mathscr{S} / \partial \beta_\nu = 0$ and transposing to the right member the known term resulting from the differentiation.

1.4. ESTIMABLE FUNCTIONS. THE GAUSS–MARKOFF THEOREM

The useful concept of *estimable functions*[12] is formulated in the following two definitions.

Definition: A *parametric function* is defined to be a linear function of the unknown parameters $\{\beta_1, \cdots, \beta_p\}$ with known constant coefficients $\{c_1, \cdots, c_p\}$,

$$\psi = \sum_{j=1}^{p} c_j \beta_j. \tag{1.4.1}$$

We introduce the vector $\mathbf{c}^{p \times 1} = (c_1, \cdots, c_p)'$; then we can write $\psi = \mathbf{c}'\boldsymbol{\beta}$.

Definition: A parametric function ψ is called an *estimable function* if it has an unbiased linear estimate, in other words, if there exists a vector of constant coefficients $\mathbf{a}^{n \times 1}$ such that

$$E(\mathbf{a}'\mathbf{y}) = \psi, \tag{1.4.2}$$

identically in $\boldsymbol{\beta}$ (i.e., no matter what the true values of the unknown parameters $\{\beta_j\}$).

Theorem 1: The parametric function $\psi = \mathbf{c}'\boldsymbol{\beta}$ is estimable if and only if $\mathbf{c}'$ is a linear combination of the rows of $\mathbf{X}'$, i.e., if and only if there exists a vector $\mathbf{a}^{n \times 1}$ such that

$$\mathbf{c}' = \mathbf{a}'\mathbf{X}'.$$

[12] Due to R. C. Bose (1944).

Proof: $\psi = \mathbf{c}'\boldsymbol{\beta}$ is estimable if and only if there exists $\mathbf{a}^{n\times 1}$ such that (1.4.2) is satisfied. But $E(\mathbf{a}'\mathbf{y}) = \mathbf{a}'E(\mathbf{y}) = \mathbf{a}'\mathbf{X}'\boldsymbol{\beta}$, and the condition $\mathbf{a}'\mathbf{X}'\boldsymbol{\beta} = \mathbf{c}'\boldsymbol{\beta}$ is satisfied identically in $\boldsymbol{\beta}$ if and only if $\mathbf{a}'\mathbf{X}' = \mathbf{c}'$.

We note that in nonmatrix notation the totality of estimable functions is $\{\sum_{i=1}^{n} a_i\eta_i\}$ where $\eta_i = E(y_i) = \sum_{j=1}^{p} x_{ij}\beta_j$, and $\{a_1, \cdots, a_n\}$ is an arbitrary set of n known constants.

For the proof[13] of the main theorem of this section we shall use the

Lemma: If $\psi = \mathbf{c}'\boldsymbol{\beta}$ is estimable, and if V_r is the space spanned by the columns of $\mathbf{X}'$, there exists a unique linear unbiased estimate of ψ, say $\mathbf{a}^{*\prime}\mathbf{y}$, with $\mathbf{a}^* \in V_r$. If $\mathbf{a}'\mathbf{y}$ is any unbiased linear estimate of ψ, then $\mathbf{a}^*$ is the projection of $\mathbf{a}$ on V_r.

Proof: Since ψ is estimable there exists an $\mathbf{a}^{n\times 1}$ for which $E(\mathbf{a}'\mathbf{y}) = \psi$. Let $\mathbf{a} = \mathbf{a}^* + \mathbf{b}$, where $\mathbf{a}^* \in V_r$, $\mathbf{b} \perp V_r$. Then

$$\psi = E(\mathbf{a}'\mathbf{y}) = E(\mathbf{a}^{*\prime}\mathbf{y}) + E(b'\mathbf{y}) = E(\mathbf{a}^{*\prime}\mathbf{y}),$$

since $E(\mathbf{b}'\mathbf{y}) = \mathbf{b}'\mathbf{X}'\boldsymbol{\beta}$ and $\mathbf{b}'\mathbf{X}' = \mathbf{0}$ by orthogonality of $\mathbf{b}$ to the columns of $\mathbf{X}'$. Thus $\mathbf{a}^{*\prime}\mathbf{y}$ is an unbiased linear estimate of ψ with $\mathbf{a}^* \in V_r$. Suppose that the same is true of $\boldsymbol{\alpha}'\mathbf{y}$. Then we have identically in $\boldsymbol{\beta}$

$$\mathbf{0} = E(\mathbf{a}^{*\prime}\mathbf{y}) - E(\boldsymbol{\alpha}'\mathbf{y}) = (\mathbf{a}^* - \boldsymbol{\alpha})'\mathbf{X}'\boldsymbol{\beta},$$

so $(\mathbf{a}^* - \boldsymbol{\alpha})'\mathbf{X}' = \mathbf{0}$. Thus $\mathbf{a}^* - \boldsymbol{\alpha} \perp V_r$ and $\in V_r$, and hence $= \mathbf{0}$. This proves the uniqueness of $\mathbf{a}^{*\prime}\mathbf{y}$. The earlier part of the proof shows that for *any* unbiased estimate $\mathbf{a}'\mathbf{y}$, $\mathbf{a}^*$ is the projection of $\mathbf{a}$ on V_r.

Theorem 2 (Gauss–Markoff Theorem): Under the assumptions Ω: $E(\mathbf{y}) = \mathbf{X}'\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_y = \sigma^2\mathbf{I}$, every estimable function $\psi = \mathbf{c}'\boldsymbol{\beta}$ has a unique unbiased linear estimate $\hat{\psi}$ which has minimum variance in the class of all unbiased linear estimates. The estimate $\hat{\psi}$ may be obtained from $\psi = \sum_{j=1}^{p} c_j\beta_j$ by replacing the $\{\beta_j\}$ by any set of LS estimates $\{\hat{\beta}_1, \cdots, \hat{\beta}_p\}$.

Proof: Let $\mathbf{a}^{*\prime}\mathbf{y}$ be the unbiased linear estimate of ψ with $\mathbf{a}^* \in V_r$, whose existence and uniqueness is given by the lemma, and let $\mathbf{a}'\mathbf{y}$ be any unbiased linear estimate of ψ. Then $\mathbf{a}^*$ is the projection of $\mathbf{a}$ on V_r, by the lemma, and

$$\|\mathbf{a}\|^2 = \|\mathbf{a}^*\|^2 + \|\mathbf{a} - \mathbf{a}^*\|^2.$$

By (1.2.10) with $m = 1$,

$$\begin{aligned}\operatorname{Var}(\mathbf{a}'\mathbf{y}) &= \mathbf{a}'\boldsymbol{\Sigma}_y\mathbf{a} = \sigma^2\|\mathbf{a}\|^2 = \sigma^2\|\mathbf{a}^*\|^2 + \sigma^2\|\mathbf{a} - \mathbf{a}^*\|^2 \\ &= \operatorname{Var}(\mathbf{a}^{*\prime}\mathbf{y}) + \sigma^2\|\mathbf{a} - \mathbf{a}^*\|^2.\end{aligned}$$

Thus $\operatorname{Var}(\mathbf{a}'\mathbf{y}) \geqq \operatorname{Var}(\mathbf{a}^{*\prime}\mathbf{y})$ with equality only if $\mathbf{a} = \mathbf{a}^*$. Hence $\mathbf{a}^{*\prime}\mathbf{y}$ is the unique unbiased linear estimate of ψ with minimum variance.

[13] This method of proof of the Gauss–Markoff theorem was suggested to me by Professor Werner Gautschi.

It remains to prove that $\mathbf{a}^{*\prime}\mathbf{y} = \mathbf{c}'\hat{\boldsymbol{\beta}}$. Now $\mathbf{a}^{*\prime}(\mathbf{y}-\hat{\boldsymbol{\eta}}) = 0$, where $\hat{\boldsymbol{\eta}} = \mathbf{X}'\hat{\boldsymbol{\beta}}$ is the projection of $\mathbf{y}$ on V_r, since $\mathbf{a}^* \in V_r$ and $\mathbf{y}-\hat{\boldsymbol{\eta}} \perp V_r$. Also $\mathbf{c}' = \mathbf{a}^{*\prime}\mathbf{X}'$ since $\mathbf{c}'\boldsymbol{\beta} = E(\mathbf{a}^*\mathbf{y}) = \mathbf{a}^*\mathbf{X}'\boldsymbol{\beta}$ identically in $\boldsymbol{\beta}$. Hence $\mathbf{a}^{*\prime}\mathbf{y} = \mathbf{a}^{*\prime}\hat{\boldsymbol{\eta}} = \mathbf{a}^{*\prime}\mathbf{X}'\hat{\boldsymbol{\beta}} = \mathbf{c}'\hat{\boldsymbol{\beta}}$.

Definition: For any estimable function ψ, its unique minimum-variance unbiased linear estimate $\hat{\psi}$, whose existence and structure are given by Theorem 2, will be called the *LS estimate of* ψ.

We have previously employed the terminology of "LS estimates" for the LS estimates $\{\hat{\beta}_j\}$ of the $\{\beta_j\}$. It might be extended to calling $\Sigma_1^p c_j\hat{\beta}_j$ the LS estimate of any linear function $\Sigma_1^p c_j\beta_j$, if the $\{\hat{\beta}_j\}$ are any set of LS estimates of the $\{\beta_j\}$. It would then follow that the LS estimate of $\Sigma_1^p c_j\beta_j$ is unique if and only if $\Sigma_1^p c_j\beta_j$ is estimable. However, we shall be interested in LS estimates only of estimable functions and of the $\{\beta_j\}$.

Corollary 1: If $\{\psi_1, \cdots, \psi_q\}$ are estimable functions every linear combination $\psi = \Sigma_1^q h_i\psi_i$ is estimable and its LS estimate $\hat{\psi}$ is $\Sigma_1^q h_i\hat{\psi}_i$, where $\hat{\psi}_i$ is the LS estimate of ψ_i.

Proof: Since $\Sigma_1^q h_i\hat{\psi}_i$ is an unbiased linear estimate of ψ, ψ is estimable. Suppose that $\psi_i = \Sigma_{j=1}^p c_{ij}\beta_j$. Then $\psi = \Sigma_j(\Sigma_i h_i c_{ij})\beta_j$. Applying Theorem 2 to both ψ_i and ψ, we find their LS estimates to be $\hat{\psi}_i = \Sigma_j c_{ij}\hat{\beta}_j$ and $\hat{\psi} = \Sigma_j(\Sigma_i h_i c_{ij})\hat{\beta}_j$, where the $\{\hat{\beta}_j\}$ are any set of LS estimates of the $\{\beta_j\}$. Hence $\hat{\psi} = \Sigma_i h_i\hat{\psi}_i$.

Side Conditions on the Parameters and Estimates

If rank $\mathbf{X} < p$ then we have seen that the LS estimates $\{\hat{\beta}_1, \cdots, \hat{\beta}_p\}$ are not unique, since they are any set $\{b_1, \cdots, b_p\}$ of statistics satisfying

$$b_1\boldsymbol{\xi}_1 + \cdots + b_p\boldsymbol{\xi}_p = \hat{\boldsymbol{\eta}}, \tag{1.4.3}$$

where $\boldsymbol{\xi}_j$ is the jth column of $\mathbf{X}'$, and $\hat{\boldsymbol{\eta}}$ is the projection of $\mathbf{y}$ on V_r, the space spanned by the $\{\boldsymbol{\xi}_j\}$. A similar indeterminacy affects the parameters $\{\beta_1, \cdots, \beta_p\}$ through the relation

$$\beta_1\boldsymbol{\xi}_1 + \cdots + \beta_p\boldsymbol{\xi}_p = \boldsymbol{\eta}, \tag{1.4.4}$$

in the sense that different sets of values for the $\{\beta_j\}$ will give the same $\boldsymbol{\eta}$ and hence the same vector of observations $\mathbf{y} = \boldsymbol{\eta}+\mathbf{e}$. We note however that if $\mathbf{c}'\boldsymbol{\beta}$ is any estimable function it has the same value regardless of which $\boldsymbol{\beta}$ is used in (1.4.4), since by Theorem 1 there exists a constant vector $\mathbf{a}$ such that $\mathbf{c}' = \mathbf{a}'\mathbf{X}'$, and thus $\mathbf{c}'\boldsymbol{\beta} = \mathbf{a}'\boldsymbol{\eta}$ depends only on $\boldsymbol{\eta}$. If it is desired to eliminate these indeterminacies two courses are open:

(i). Consider a "reduced" problem with only r parameters $\{\beta_j\}$. This can be achieved by choosing r linearly independent vectors from the set $\{\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_p\}$ as a basis for V_r, as in the proof of Lemma 2 of App. I, and

keeping only the r corresponding $\{\beta_j\}$. This gives us a new $n \times r$ matrix of coefficients instead of the old $\mathbf{X}'$ and the resulting "reduced" problem is a case of maximal rank, i.e., the rank of the new $\mathbf{X}$ equals the new number of $\{\beta_j\}$.

(ii). Put suitable side conditions on the p parameters $\{\beta_j\}$ and their estimates. Thus, we would achieve the same result as in (i) if we agreed that for the $p-r$ parameters $\{\beta_j\}$ there discarded we always take $\beta_j = 0$ and $\hat{\beta}_j = 0$. In most analysis-of-variance situations it is convenient to add linear restrictions of a more general form than this to produce the desired uniqueness. We therefore consider subjecting the $\{\beta_j\}$ to t ($t \geqq p-r$) linear restrictions $\mathbf{H}'\boldsymbol{\beta} = \mathbf{0}$, where $\mathbf{H}'$ is a $t \times p$ matrix of known constants. It will usually be almost obvious that the restrictions adopted in practice make the $\{\beta_j\}$ unique in the sense that for every possible set $\{\beta_j\}$ in the original problem there will exist a unique set $\{\tilde{\beta}_j\}$ satisfying

$$\mathbf{X}'\boldsymbol{\beta} = \mathbf{X}'\tilde{\boldsymbol{\beta}} \quad \text{and} \quad \mathbf{H}'\tilde{\boldsymbol{\beta}} = \mathbf{0}. \tag{1.4.5}$$

The first of these conditions says the $\{\tilde{\beta}_j\}$ give the same $\boldsymbol{\eta} = \mathbf{X}'\boldsymbol{\beta}$ as the $\{\beta_j\}$. The two conditions (1.4.5) will then make the $\{\tilde{\beta}_j\}$ uniquely determined functions of the $\{\beta_j\}$. We will prove below that these are estimable functions in the original problem, so that then *every* parametric function $\mathbf{c}'\tilde{\boldsymbol{\beta}}$ in the new problem is an estimable function in the old problem. We shall also show that there is then a unique set of LS estimates $\{\hat{\beta}_j\}$ which satisfy the side conditions $\mathbf{H}'\hat{\boldsymbol{\beta}} = \mathbf{0}$, i.e., a unique solution of the normal equations which satisfies the side conditions. In later parts of the book, when applying this theory of making the parameters and their estimates unique by subjecting them to appropriate side conditions, we shall omit the tildes ($\sim$) from the $\{\tilde{\beta}_j\}$, but it will clarify the derivation of the theory to keep the distinction in the notation for the present. (If the non-mathematically inclined reader is willing to accept all this without proof he may skip the rest of this section.)

We will see in a moment that we shall have to consider the trivial estimable function $\mathbf{c}'\boldsymbol{\beta}$ all of whose coefficients are zero, which we shall write $\mathbf{0}'\boldsymbol{\beta}$. It is clear that, aside from $\mathbf{0}'\boldsymbol{\beta}$, an estimable function $\mathbf{c}'\boldsymbol{\beta}$ can take on every possible value k for suitable choice of $\boldsymbol{\beta}$ (for some ν such that $c_\nu \neq 0$ take $\beta_j = k\delta_{\nu j}/c_\nu$). Denote the t rows of $\mathbf{H}'$ by $\mathbf{h}_1', \cdots, \mathbf{h}_t'$. Evidently we cannot let any $\mathbf{h}_i'\boldsymbol{\beta}$ be estimable, unless it is $\mathbf{0}'\boldsymbol{\beta}$, since we are going to add the restriction $\mathbf{H}'\tilde{\boldsymbol{\beta}} = \mathbf{0}$, or all $\mathbf{h}_i'\tilde{\boldsymbol{\beta}} = 0$, and, if $\mathbf{h}_i'\boldsymbol{\beta}$ is estimable, $\mathbf{h}_i'\tilde{\boldsymbol{\beta}} = \mathbf{h}_i'\boldsymbol{\beta}$ by the remark made after (1.4.4), and can take on every value k. More generally, no linear combination of the $\{\mathbf{h}_i'\boldsymbol{\beta}\}$ must be an estimable function except $\mathbf{0}'\boldsymbol{\beta}$ (it is quite possible that $\mathbf{0}'\boldsymbol{\beta}$ may be a linear combination of the $\{\mathbf{h}_i'\boldsymbol{\beta}\}$ with coefficients not all 0, since we permit the $\{\mathbf{h}_i\}$ to be linearly dependent). By Theorem 1 we may state this as

follows: No linear combination of the rows of $\mathbf{H}'$ except[14] $\mathbf{0}'$ can be a linear combination of the rows of $\mathbf{X}'$. On the other hand if the solution $\tilde{\boldsymbol{\beta}}$ of (1.4.5) is to be unique, the rank of the composite $(n+t)\times p$ matrix

$$\mathbf{G}' = \begin{pmatrix} \mathbf{X}' \\ \mathbf{H}' \end{pmatrix} \tag{1.4.6}$$

must be p: For, by use of partitioned matrices (see end of App. II), (1.4.5) may be written as $\mathbf{G}'\tilde{\boldsymbol{\beta}} = \boldsymbol{\eta}^*$, where

$$\boldsymbol{\eta}^* = \begin{pmatrix} \mathbf{X}'\boldsymbol{\beta} \\ \mathbf{0} \end{pmatrix} \tag{1.4.7}$$

is a vector with $n+t$ components, or

$$\tilde{\beta}_1\mathbf{g}_1 + \tilde{\beta}_2\mathbf{g}_2 + \cdots + \tilde{\beta}_p\mathbf{g}_p = \boldsymbol{\eta}^*, \tag{1.4.8}$$

where $\mathbf{g}_j$ is the jth column of $\mathbf{G}'$. By Lemma 3 of App. I, the coefficients $\{\tilde{\beta}_j\}$ in (1.4.8) will be unique if and only if the $\{\mathbf{g}_j\}$ are linearly independent, i.e., rank $\mathbf{G}' = p$.

That the two necessary conditions we have found on the matrix $\mathbf{H}'$ are also sufficient for our purpose will follow as Corollary 2 to the following theorem. The theorem is stated as a purely algebraic result. In its statistical interpretation, condition (b) of the theorem is equivalent to (b′): No linear combination of the rows of $\mathbf{H}'\boldsymbol{\beta}$ (i.e., no linear combination of the parametric functions we equate to zero in the side conditions) is an estimable function except $\mathbf{0}'\boldsymbol{\beta}$.

Theorem 3: Suppose that $\mathbf{X}'$ is $n\times p$, $\mathbf{H}'$ is $t\times p$, rank $\mathbf{X}' = r$ $(p > r, t \geq p - r)$, and V_r is the space spanned by the columns of $\mathbf{X}'$. Then the system

$$\mathbf{X}'\mathbf{b} = \mathbf{z}, \qquad \mathbf{H}'\mathbf{b} = \mathbf{0} \tag{1.4.9}$$

has a unique solution $\mathbf{b}^{p\times 1}$ for every $\mathbf{z}^{n\times 1} \in V_r$ if and only if the following two conditions are satisfied: (*a*) The rank of the composite matrix

$$\mathbf{G}' = \begin{pmatrix} \mathbf{X}' \\ \mathbf{H}' \end{pmatrix}$$

is p. (*b*) No linear combination of the rows of $\mathbf{H}'$ is a linear combination of the rows of $\mathbf{X}'$ except $\mathbf{0}'$.

Proof: Most of the proof consists of showing that there exists a solution $\mathbf{b}$ for every $\mathbf{z} \in V_r$ if and only if condition (*b*) is satisfied. It follows by the argument stated above in connection with (1.4.8) that if a solution $\mathbf{b}$ exists it is unique if and only if (*a*) is satisfied.

[14] We write $\mathbf{0}'$ because we are thinking here of $\mathbf{0}$ as the zero vector; however it would be perfectly correct to write $\mathbf{0}$ instead for the $t\times 1$ zero matrix.

Write (1.4.9) as $\mathbf{G}'\mathbf{b} = \mathbf{z}^*$, where $\mathbf{z}^*$ is the vector with $n+t$ components

$$\mathbf{z}^* = \begin{pmatrix} \mathbf{z}^{n\times 1} \\ \mathbf{0}^{t\times 1} \end{pmatrix},$$

or

$$b_1\mathbf{g}_1 + b_2\mathbf{g}_2 + \cdots + b_p\mathbf{g}_p = \mathbf{z}^*,$$

where $\mathbf{g}_j$ is the jth column of $\mathbf{G}'$. Then a solution $\mathbf{b}$ is seen to exist if and only if $\mathbf{z}^* \in W$, where W is the space of vectors of $n+t$ components spanned by the $\{\mathbf{g}_j\}$. By Theorem 3 of App. I, $\mathbf{z}^* \in W$ if and only if $\mathbf{u}'\mathbf{z}^* = 0$ for every $\mathbf{u} \perp W$. If we partition $\mathbf{u}$,

$$\mathbf{u} = \begin{pmatrix} \mathbf{v}^{n\times 1} \\ \mathbf{w}^{t\times 1} \end{pmatrix},$$

so that

$$\mathbf{u}'\mathbf{z}^* = (\mathbf{v}', \mathbf{w}')\begin{pmatrix} \mathbf{z} \\ \mathbf{0} \end{pmatrix} = \mathbf{v}'\mathbf{z},$$

we see that $\mathbf{u}'\mathbf{z}^* = 0$ if and only if $\mathbf{v}'\mathbf{z} = 0$. Also $\mathbf{u} \perp W$ if and only if $\mathbf{u}$ is orthogonal to the columns of $\mathbf{G}'$ which span W; hence if and only if $\mathbf{u}'\mathbf{G}' = \mathbf{0}'$,

$$(\mathbf{v}', \mathbf{w}')\begin{pmatrix} \mathbf{X}' \\ \mathbf{H}' \end{pmatrix} = \mathbf{0}',$$

or

$$\mathbf{v}'\mathbf{X}' + \mathbf{w}'\mathbf{H}' = \mathbf{0}'. \tag{1.4.10}$$

We now have that a solution $\mathbf{b}$ exists for a given $\mathbf{z} \in V_r$ if and only if $\mathbf{v}'\mathbf{z} = 0$ for every $\mathbf{v}^{n\times 1}$ and $\mathbf{w}^{t\times 1}$ satisfying (1.4.10).

Suppose first that condition (*b*) is satisfied and suppose that $\mathbf{v}$ and $\mathbf{w}$ satisfy (1.4.10). Then $\mathbf{v}'\mathbf{X}' = -\mathbf{w}'\mathbf{H}'$ is a linear combination of the rows of $\mathbf{X}'$ and also of the rows of $\mathbf{H}'$ and hence must be $\mathbf{0}'$ by (*b*). Then $\mathbf{v}'\mathbf{X}' = \mathbf{0}'$ implies that $\mathbf{v}$ is orthogonal to the columns of $\mathbf{X}'$, thus $\mathbf{v} \perp V_r$; therefore $\mathbf{v} \perp \mathbf{z}$ for every $\mathbf{z} \in V_r$, i.e., $\mathbf{v}'\mathbf{z} = 0$, and hence there exists a solution $\mathbf{b}$ for every $\mathbf{z} \in V_r$. Suppose next that (*b*) is not satisfied, so that there exists a linear combination of the rows of $\mathbf{H}'$, say $-\mathbf{w}'\mathbf{H}'$ which is a linear combination of the rows of $\mathbf{X}'$, say $\mathbf{v}'\mathbf{X}'$, and is not $\mathbf{0}'$: $\mathbf{v}'\mathbf{X}' = -\mathbf{w}'\mathbf{H}' = \boldsymbol{\lambda}'$, say, where $\boldsymbol{\lambda}^{p\times 1} \neq \mathbf{0}$. Now take $\mathbf{z} = \mathbf{X}'\boldsymbol{\lambda}$, so $\mathbf{z} \in V_r$. Then $\mathbf{v}'\mathbf{z} = \mathbf{v}'\mathbf{X}'\boldsymbol{\lambda} = \boldsymbol{\lambda}'\boldsymbol{\lambda} \neq 0$, while $\mathbf{v}$ and $\mathbf{w}$ satisfy (1.4.10). Thus for this $\mathbf{z} \in V_r$ there is no solution $\mathbf{b}$.

By taking $\mathbf{b} = \tilde{\boldsymbol{\beta}}$ and $\mathbf{z} = \mathbf{X}'\boldsymbol{\beta}$ in Theorem 3 we get

Corollary 2: The system (1.4.5) has a unique solution $\hat{\boldsymbol{\beta}}$ for every $\boldsymbol{\beta}$ if and only if conditions (*a*) and (*b*) of Theorem 3 are satisfied.

Recalling that any $\{b_1, \cdots, b_p\}$ that satisfy (1.4.3) and are functions of

$\mathbf{y}$ only[15] constitute a set of LS estimates, and taking $\mathbf{z} = \hat{\boldsymbol{\eta}}$ in Theorem 3, we get

Corollary 3: If conditions (*a*) and (*b*) of Theorem 3 are satisfied, there exists a unique set of LS estimates $\{\hat{\beta}_1, \cdots, \hat{\beta}_p\}$ (i.e., a unique solution of the normal equations) for which $\mathbf{H}'\hat{\boldsymbol{\beta}} = \mathbf{0}$.

This says that we may subject the LS estimates to the same side conditions as the parameters. Finally we need the following result, which implies that every linear combination of the parameters $\{\tilde{\beta}_i\}$, the $\{\tilde{\beta}_j\}$ being subject to the side conditions, is estimable:

Theorem 4: If conditions (*a*) and (*b*) of Theorem 3 are satisfied, so that the $\{\tilde{\beta}_j\}$ are functions of the $\{\beta_j\}$ determined uniquely by (1.4.5), then the $\{\tilde{\beta}_j\}$ are estimable functions.

Proof: We will obtain an explicit formula for the $\{\tilde{\beta}_j\}$ in terms of the $\{\beta_j\}$: For any $\boldsymbol{\beta}$ let $\tilde{\boldsymbol{\beta}}$ be the unique solution of (1.4.5), so

$$\mathbf{G}'\tilde{\boldsymbol{\beta}} = \begin{pmatrix} \mathbf{X}'\boldsymbol{\beta} \\ \mathbf{0} \end{pmatrix}.$$

Multiply by $\mathbf{G}$ on the left to get

$$\mathbf{G}\mathbf{G}'\tilde{\boldsymbol{\beta}} = (\mathbf{X}, \mathbf{H}) \begin{pmatrix} \mathbf{X}'\boldsymbol{\beta} \\ \mathbf{0} \end{pmatrix} = \mathbf{X}\mathbf{X}'\boldsymbol{\beta}.$$

Now by Theorem 7 of App. II the rank of the $p \times p$ matrix $\mathbf{G}\mathbf{G}'$ is equal to rank $\mathbf{G} = p$; so $\mathbf{G}\mathbf{G}'$ has an inverse, and thus $\tilde{\boldsymbol{\beta}} = (\mathbf{G}\mathbf{G}')^{-1}\mathbf{X}\mathbf{X}'\boldsymbol{\beta}$, or

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}\mathbf{X}' + \mathbf{H}\mathbf{H}')^{-1}\mathbf{X}\mathbf{X}'\boldsymbol{\beta},$$

the promised formula. Since $E(\mathbf{y}) = \mathbf{X}'\boldsymbol{\beta}$, $\tilde{\boldsymbol{\beta}}$ has the unbiased estimate $(\mathbf{X}\mathbf{X}' + \mathbf{H}\mathbf{H}')^{-1}\mathbf{X}\mathbf{y}$.

1.5. REDUCTION OF THE CASE WHERE THE OBSERVATIONS HAVE KNOWN CORRELATIONS AND KNOWN RATIOS OF VARIANCES

We consider now the case where the covariance matrix $\boldsymbol{\Sigma}_y$ of the observations $\{y_i\}$ is not of the form $\sigma^2\mathbf{I}$ but $\boldsymbol{\Sigma}_y$ is known except for a scalar factor, i.e., $\boldsymbol{\Sigma}_y = \theta\mathbf{B}$, where θ is an unknown positive constant and $\mathbf{B}^{n\times n}$ is a known constant matrix; $\mathbf{B}$ is necessarily symmetric and positive indefinite, and we shall assume furthermore that it is nonsingular (see App. V). This is equivalent to knowing the correlation coefficients of all pairs of observations y_i and the ratios of their variances.

[15] That these $\{b_i\}$ are functions of $\mathbf{y}$ only, and in fact *linear* functions, follows from their being the unique solution of the linear system $\mathbf{X}\mathbf{X}'\mathbf{b} = \mathbf{X}\mathbf{y}$, $\mathbf{H}'\mathbf{b} = \mathbf{0}$.

Our underlying assumptions are now

$$\Omega: \quad E(\mathbf{y}) = \mathbf{X}'\boldsymbol{\beta}, \qquad \boldsymbol{\Sigma}_y = \theta\mathbf{B}, \qquad |\mathbf{B}| \neq 0, \qquad \text{rank } \mathbf{X}' = r. \tag{1.5.1}$$

This case may be reduced to that previously considered, where $\boldsymbol{\Sigma}_y = \sigma^2\mathbf{I}$, by appealing to Lemma 11′ and the discussion following it in App. II, which says there exists a nonsingular $\mathbf{P}^{n\times n}$ such that $\mathbf{P}'\mathbf{B}\mathbf{P} = \mathbf{I}$. Let $\tilde{\mathbf{y}} = \mathbf{P}'\mathbf{y}$. Then

$$E(\tilde{\mathbf{y}}) = \mathbf{P}'\,E(\mathbf{y}) = \mathbf{P}'\mathbf{X}'\boldsymbol{\beta} = \tilde{\mathbf{X}}'\boldsymbol{\beta},$$

where $\tilde{\mathbf{X}}' = \mathbf{P}'\mathbf{X}'$; so rank $\tilde{\mathbf{X}}'$ = rank $\mathbf{X}'$ = r, and

$$\boldsymbol{\Sigma}_{\tilde{y}} = \mathbf{P}'\boldsymbol{\Sigma}_y\mathbf{P} = \theta\mathbf{P}'\mathbf{B}\mathbf{P} = \sigma^2\mathbf{I},$$

where $\sigma^2 = \theta$. We may thus write (1.5.1) as

$$\Omega: \quad E(\tilde{\mathbf{y}}) = \tilde{\mathbf{X}}'\boldsymbol{\beta}, \qquad \boldsymbol{\Sigma}_{\tilde{y}} = \sigma^2\mathbf{I}, \qquad \text{rank } \tilde{\mathbf{X}}' = r,$$

which is the case previously considered.

In applications the transformed "observations" $\{\tilde{y}_i\}$ are tedious to calculate and one usually prefers to work with the actual observations $\{y_i\}$. The LS estimates of the parameters $\{\beta_j\}$ may then be found by minimizing the following sum of squares involving the $\{y_i\}$ and $\{\beta_j\}$

$$\mathscr{S}(\mathbf{y}, \boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}'\boldsymbol{\beta})'\,\mathbf{B}^{-1}(\mathbf{y} - \mathbf{X}'\boldsymbol{\beta}). \tag{1.5.2}$$

To see this we note that in the transformed problem, which falls under our previous theory, the $\{\hat{\beta}_j\}$ are found by minimizing

$$\tilde{\mathscr{S}}(\tilde{\mathbf{y}}, \boldsymbol{\beta}) = (\tilde{\mathbf{y}} - \tilde{\mathbf{X}}'\boldsymbol{\beta})'(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}'\boldsymbol{\beta}). \tag{1.5.3}$$

Now $\tilde{\mathbf{y}} - \tilde{\mathbf{X}}'\boldsymbol{\beta} = \mathbf{P}'(\mathbf{y} - \mathbf{X}'\boldsymbol{\beta})$, and substituting this in (1.5.3) and using $\mathbf{P}\mathbf{P}' = \mathbf{B}^{-1}$, we get that $\tilde{\mathscr{S}}(\tilde{\mathbf{y}}, \boldsymbol{\beta})$ equals the $\mathscr{S}(\mathbf{y}, \boldsymbol{\beta})$ defined by (1.5.2).

Besides the $\{\beta_j\}$ the model (1.5.1) contains the unknown parameter θ. In sec. 1.6 it will be shown that an unbiased estimate of σ^2 is $\tilde{\mathscr{S}}(\tilde{\mathbf{y}}, \hat{\boldsymbol{\beta}})/(n-r)$, where $\hat{\boldsymbol{\beta}}$ is any set of LS estimates. It follows that an unbiased estimate of the parameter θ is $\mathscr{S}(\mathbf{y}, \hat{\boldsymbol{\beta}})/(n-r)$, where $\mathscr{S}(\mathbf{y}, \hat{\boldsymbol{\beta}})$ is formed by replacing $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$ in (1.5.2).

The sum of squares (1.5.2) which is minimized to calculate the LS estimates may be called the "weighted sum of squares": In the particular case where the observations are uncorrelated, $\mathbf{B}$ is a diagonal matrix, and if we then write the ith diagonal element of $\mathbf{B}$ as w_i^{-1}, the $\{w_i\}$ are inversely proportional to the variances of the observations $\{y_i\}$, and (1.5.2) becomes

$$\mathscr{S}(\mathbf{y}, \boldsymbol{\beta}) = \sum_i w_i\left(y_i - \sum_j x_{ji}\beta_j\right)^2.$$

The case $\boldsymbol{\Sigma}_y = \sigma^2\mathbf{I}$ is the special case where the weights $\{w_i\}$ are all equal.

Sometimes in applications we may have some doubt about the correct weights, and we may then find some comfort in the fact that the method of least squares used with incorrect weights still leads to unbiased estimates; however, our calculations of the variances[16] of the estimates will be invalidated by incorrect weights. More generally, it is true that the use of any positive definite matrix $\mathbf{B}$ whatever (not just a correct one of the form $\theta^{-1}\boldsymbol{\Sigma}_y$) in (1.5.2) leads to unbiased estimates of estimable functions if the LS estimates of the $\{\beta_i\}$ are calculated by minimizing (1.5.2). We shall prove this only for the case where $\mathbf{X}'$ is of rank p:

Let $\mathbf{P}$ be defined as above for the $\mathbf{B}$ actually used, and again transform to $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{X}}'$ as above. In the transformed problem the normal equations are $\tilde{\mathbf{X}}\tilde{\mathbf{X}}'\boldsymbol{\beta} = \tilde{\mathbf{X}}\tilde{\mathbf{y}}$, and the solution, which we will denote by $\hat{\boldsymbol{\beta}}^*$, is

$$\hat{\boldsymbol{\beta}}^* = (\tilde{\mathbf{X}}\tilde{\mathbf{X}}')^{-1}\tilde{\mathbf{X}}\tilde{\mathbf{y}}.$$

But this solution will be the same as that found from minimizing (1.5.2). Since

$$\hat{\boldsymbol{\beta}}^* = (\tilde{\mathbf{X}}\tilde{\mathbf{X}}')^{-1}\tilde{\mathbf{X}}\mathbf{P}'\mathbf{y},$$

therefore

$$E(\hat{\boldsymbol{\beta}}^*) = (\tilde{\mathbf{X}}\tilde{\mathbf{X}}')^{-1}\tilde{\mathbf{X}}\mathbf{P}'\mathbf{X}'\boldsymbol{\beta}.$$

Substituting $\mathbf{P}'\mathbf{X}' = \tilde{\mathbf{X}}'$ in this expression, we get $E(\hat{\boldsymbol{\beta}}^*) = \boldsymbol{\beta}$.

1.6. THE CANONICAL FORM OF THE UNDERLYING ASSUMPTIONS Ω. THE MEAN SQUARE FOR ERROR

Let us introduce in the sample space V_n of the observation vector $\mathbf{y}^{n\times 1}$ the orthonormal basis $\{\boldsymbol{\rho}_1, \boldsymbol{\rho}_2, \cdots, \boldsymbol{\rho}_n\}$, where $\boldsymbol{\rho}_i = (\delta_{i1}, \delta_{i2}, \cdots, \delta_{in})'$ (this is the basis R of the example after Theorem 1 of App. I), so $\mathbf{y} = \sum_1^n y_i\boldsymbol{\rho}_i$. Let us also introduce an orthonormal basis $\{\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_r\}$ for V_r, the space spanned by the columns of $\mathbf{X}'$, and complete it to an orthonormal basis $\{\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_r, \boldsymbol{\alpha}_{r+1}, \cdots, \boldsymbol{\alpha}_n\}$ for V_n; this is always possible (Lemmas 6 and 7 of App. I). Write

$$\mathbf{y} = \sum_{i=1}^{n} z_i\boldsymbol{\alpha}_i, \tag{1.6.1}$$

where $\{z_i\}$ are the coordinates of $\mathbf{y}$ relative to the new basis, and hence $z_i = \boldsymbol{\alpha}_i'\mathbf{y}$, as we see by multiplying (1.6.1) by $\boldsymbol{\alpha}_i'$. This relation between the coordinates $\{z_i\}$ and $\{y_i\}$ may be written $\mathbf{z} = \mathbf{P}\mathbf{y}$, where $\mathbf{P}^{n\times n}$ is the orthogonal matrix whose ith row is $\boldsymbol{\alpha}_i'$. Let $\zeta_i = E(z_i)$, so $\zeta_i = E(\boldsymbol{\alpha}_i'\mathbf{y}) = \boldsymbol{\alpha}_i'\boldsymbol{\eta}$. It follows that for all values of the parameters, $\zeta_i = 0$ for $i > r$

[16] Bounds on the bias of the estimated covariance matrix are derived for some cases in Watson (1955).

since $\boldsymbol{\eta} \in V_r \perp \boldsymbol{\alpha}_i$ for $i > r$. Furthermore we have for the covariance matrix of the transformed "observations" $\{z_i\}$,

$$\boldsymbol{\Sigma}_z = \mathbf{P}\boldsymbol{\Sigma}_y\mathbf{P}' = \sigma^2\mathbf{P}\mathbf{P}' = \sigma^2\mathbf{I}.$$

We have now shown that by a suitable orthogonal transformation (not depending on unknown parameters) we can always reduce the Ω-assumptions to the *canonical form*

$$\Omega: \begin{cases} \mathbf{z} = (z_1, \cdots, z_n)', \\ E(z_i) = \zeta_i \qquad (i = 1, \cdots, r), \\ E(z_i) = 0 \qquad (i = r+1, \cdots, n), \\ \boldsymbol{\Sigma}_z = \sigma^2\mathbf{I}, \end{cases}$$

where $\zeta_1, \cdots, \zeta_r$, and σ^2 are unknown parameters, and the $\{z_i\}$ are a known transformation of the observations.

Since we do not actually use the canonical form in analyzing data, we will never need to calculate the transformation matrix $\mathbf{P}$ explicitly (although it could be done by calculating its rows $\{\boldsymbol{\alpha}_i'\}$ by the Schmidt process of Lemma 6 of App. I). However, the canonical form is very useful for the derivation of distribution theory, for example:

An Unbiased Estimate of σ^2

The error sum of squares $\mathscr{S}_\Omega$ introduced at the end of sec. 1.3, namely

$$\mathscr{S}_\Omega = \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ji}\hat{\beta}_j\right)^2, \tag{1.6.2}$$

where $\{\hat{\beta}_j\}$ is any set of LS estimates, may be written $\mathscr{S}_\Omega = \|\mathbf{y}-\hat{\boldsymbol{\eta}}\|^2$, where $\hat{\boldsymbol{\eta}}$ is the projection of $\mathbf{y}$ on V_r. But $\mathbf{y} = \Sigma_1^n z_i\boldsymbol{\alpha}_i$, and $\hat{\boldsymbol{\eta}} = \Sigma_1^r z_i\boldsymbol{\alpha}_i$, where $\{\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_n\}$ is the above basis for the canonical form, and so $\mathscr{S}_\Omega = \|\Sigma_{r+1}^n z_i\boldsymbol{\alpha}_i\|^2$, or

$$\mathscr{S}_\Omega = \sum_{i=r+1}^{n} z_i^2. \tag{1.6.3}$$

Now for $i > r$, $E(z_i) = 0$, which implies that $E(z_i^2) = \mathrm{Var}(z_i) = \sigma^2$. Hence from (1.6.3), $E(\mathscr{S}_\Omega) = (n-r)\sigma^2$. If we define

$$s^2 = \mathscr{S}_\Omega/(n-r), \tag{1.6.4}$$

we have $E(s^2) = \sigma^2$, that is, s^2 is an unbiased estimate of σ^2. The quantity s^2 is called the *mean square for error* (later written also as MS_e) and it is said to have *$n-r$ degrees of freedom*. In general the number of degrees of freedom of a quadratic form in the observations is defined to be its rank (i.e., the rank of the symmetric matrix of the quadratic form), and we see from (1.6.3) that the rank of $\mathscr{S}_\Omega$ is $n-r$.

This result for estimating σ^2 is a supplement to the Gauss–Markoff theorem of great practical importance, since in applications we want some idea of the accuracy of our unbiased point estimates. If $\psi = \mathbf{c}'\boldsymbol{\beta}$ is an estimable function, then by the theorem there exists a unique linear combination of the observations, $\hat{\psi} = \mathbf{a}^{*\prime}\mathbf{y}$ which is the optimum estimate of ψ. Then the variance of the estimate $\hat{\psi}$ is $\sigma_{\hat{\psi}}^2 = \mathbf{a}^{*\prime}\mathbf{a}^*\sigma^2$, and this may then be estimated by $\hat{\sigma}_{\hat{\psi}}^2 = \mathbf{a}^{*\prime}\mathbf{a}^* s^2$. This estimate of the variance is evidently unbiased, and has been shown to have other optimum properties.[17] The expected value of the mean square for error s^2 in the case where the observations $\{y_i\}$ have unequal variances $\{\sigma_i^2\}$, but s^2 is calculated in the above way as though the $\{\sigma_i^2\}$ were equal, is given by the rule at the beginning of sec. 10.4.

Estimation and Error Spaces

Consider the set of all *linear forms* $\Sigma_1^n a_i y_i = \mathbf{a}'\mathbf{y}$ in the observations. The coefficients $\{a_i\}$ are assumed to be known constants (i.e., they do not depend on unknown parameters): we may call $\mathbf{a}$ the *coefficient vector* of the linear form $\mathbf{a}'\mathbf{y}$. We see there is a one-to-one correspondence between the totality of linear forms $\mathbf{a}'\mathbf{y}$ and the totality of vectors $\mathbf{a} \in V_n$, and that addition of linear forms or multiplication of a linear form by a constant corresponds to the same operation on the coefficient vectors. It is convenient to speak of *spaces of linear forms spanned by a given set of linear forms, independence of linear forms, orthogonality of forms and of spaces*, etc., the terms being defined by use of the corresponding properties of the coefficient vectors of the forms.

The canonical variables $\{z_1, \cdots, z_n\}$ are linear forms in the observations $\{y_i\}$, and they may be used to define two interesting orthogonal spaces of linear forms, namely the space spanned by $\{z_1, \cdots, z_r\}$, called the *estimation space*, and that spanned by $\{z_{r+1}, \cdots, z_n\}$, called the *error space*.[18] Since $z_i = \boldsymbol{\alpha}_i'\mathbf{y}$, we see that the forms $\{z_1, \cdots, z_n\}$ constitute an orthonormal basis for the n-dimensional space of forms (because their coefficient vectors constitute an orthonormal basis for V_n), and so the two spaces are orthogonal.

The reason for calling the latter the error space is that the error sum of squares $\mathscr{S}_\Omega$ involves only the set $\{z_{r+1}, \cdots, z_n\}$. It is easily shown that a linear form $\mathbf{a}'\mathbf{y}$ is in the error space if and only if its expected value is identically zero in the parameters: The relation $\mathbf{z} = \mathbf{P}\mathbf{y}$ may be inverted, $\mathbf{y} = \mathbf{P}'\mathbf{z}$ since $\mathbf{P}'\mathbf{P} = \mathbf{I}$, hence $\mathbf{a}'\mathbf{y} = \mathbf{b}'\mathbf{z}$, where $\mathbf{b} = \mathbf{P}\mathbf{a}$, and so $E(\mathbf{a}'\mathbf{y}) = \mathbf{b}'\boldsymbol{\zeta} = \Sigma_1^r b_i\zeta_i = 0$ if and only if $b_1 = b_2 = \cdots = b_r = 0$, i.e., if and only if $\mathbf{a}'\mathbf{y} = \Sigma_{r+1}^n b_i z_i$. The former space is called the estimation space

[17] By P. L. Hsu (1938*b*).

[18] By R. C. Bose (1944).

because if ψ is any estimable function and $\hat{\psi}$ is its LS estimate then the linear form $\hat{\psi}$ is a linear combination of $\{z_1, \cdots, z_r\}$ only, i.e., $\hat{\psi}$ is in the estimation space. To see this note that the columns of $\mathbf{P}'$ are $\{\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_n\}$, the orthonormal basis for V_n used in deriving the canonical form. If ψ is estimable, by the Gauss–Markoff theorem its LS estimate $\hat{\psi}$ is of the form $\mathbf{a}^{*\prime}\mathbf{y}$ with $\mathbf{a}^* \in V_r$, i.e., $\mathbf{a}^* \perp \boldsymbol{\alpha}_j$ for $j > r$. Now $\hat{\psi} = \mathbf{a}^{*\prime}\mathbf{y} = \mathbf{c}'\mathbf{z}$, where $\mathbf{c}' = \mathbf{a}^{*\prime}\mathbf{P}'$ is a row matrix whose jth element is $c_j = \mathbf{a}^{*\prime}\boldsymbol{\alpha}_j$, so $c_j = 0$ for $j > r$. Hence $\hat{\psi} = \Sigma_1^r c_j z_j$.

Although the linear forms $\{z_1, \cdots, z_n\}$ depend on the choice of the basis $\{\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_n\}$, it is clear that the estimation and error spaces do not, since the first is the space of all $\hat{\psi}$, and the second is the space of all $\mathbf{a}'\mathbf{y}$ for which $E(\mathbf{a}'\mathbf{y}) = 0$.

PROBLEMS

1.1. First- and second-degree polynomials are fitted by LS to n points (x_i, y_i), $i = 1, \cdots, n$. Let ω and Ω denote the assumptions*

$$\omega: \quad y_i = \alpha + \beta x_i + e_i, \qquad E(e_i) = 0, \qquad E(e_i e_{i'}) = \sigma^2 \delta_{ii'},$$
$$\Omega: \quad y_i = \alpha + \beta x_i + \gamma x_i^2 + e_i, \qquad E(e_i) = 0, \qquad E(e_i e_{i'}) = \sigma^2 \delta_{ii'}.$$

Find by differentiation the normal equations for the estimates of α and β under ω, and of α, β, and γ under Ω. Solve the former explicitly and indicate the solution of the latter by using determinants. Save the results of Problems 1.1, 1.2, and 1.3 for later use in Ch. 2.

1.2. In Problem 1.1 find the variances and covariance of the estimates of α and β under ω. Show that if we write $\delta + \beta(x_i - \bar{x})$ in place of $\alpha + \beta x_i$ in ω, then under ω, $\hat{\delta} = \bar{y}$ and $\mathrm{Cov}(\hat{\delta}, \hat{\beta}) = 0$.

1.3. In Problem 1.1 express $\mathrm{Var}(\hat{\gamma})$ under Ω by using determinants.

1.4. Prove the following lemma: If $\mathbf{y} = (y_1, \cdots, y_n)'$, $E(\mathbf{y}) = \boldsymbol{\eta}$, $\mathbf{e} = \mathbf{y} - \boldsymbol{\eta}$, and $Q(\mathbf{y})$ is a quadratic form in $\mathbf{y}$, then $E(Q(\mathbf{y})) = Q(\boldsymbol{\eta}) + E(Q(\mathbf{e}))$. Note that $Q(\boldsymbol{\eta})$ may be evaluated by replacing the $\{y_i\}$ by their expectations in $Q(\mathbf{y})$, and that $E(Q(\mathbf{e}))$ is the value of $E(Q(\mathbf{y}))$ when $\boldsymbol{\eta} = 0$.

1.5. Prove the following result, of importance in the theory of the design of experiments: Under Ω: $E(\mathbf{y}) = \Sigma_1^p \beta_j \boldsymbol{\xi}_j$ and $\boldsymbol{\Sigma}_y = \sigma^2 \mathbf{I}$, if $\boldsymbol{\xi}_\nu = \boldsymbol{\xi}_\nu^* + \boldsymbol{\xi}_\nu^\perp$, where $\boldsymbol{\xi}_\nu^*$ is the projection of $\boldsymbol{\xi}_\nu$ on the space spanned by the other $\{\boldsymbol{\xi}_j\}$, and if $\boldsymbol{\xi}_\nu^\perp \neq \mathbf{0}$, then β_ν is estimable, and the variance of its LS estimate $\hat{\beta}_\nu$ is $|\boldsymbol{\xi}_\nu^\perp|^{-2}\sigma^2$. [*Hint:* Assume $\nu = 1$ and take the vector $\boldsymbol{\alpha}_1$ of the canonical form of sec. 1.6 in the direction of $\boldsymbol{\xi}_1^\perp$.]

* It is convenient to denote the underlying assumptions by ω and Ω rather than Ω_1 and Ω_2 for later use in Ch. 2.