UNIVERSITY OF CAPE TOWN
DEPARTMENT OF MATHEMATICAL STATISTICS

CONTRIBUTIONS TO THE THEORY OF GENERALIZED INVERSES, THE LINEAR MODEL AND OUTLIERS

by

Timothy Terence Dunne

A thesis prepared under the supervision of
Professor C G Troskie
in fulfilment of the requirements for the degree of
Doctor of Philosophy in Mathematical Statistics

A C K N O W L E D G E M E N T S

The convention, in Ph.D. theses, of tributes to those who assisted the aspiring candidate in the years before its production, is a congenial practice because it allows one to record permanently and directly the gratitude which otherwise is not always easily expressed.

In the first place, my happy thanks to Professor C.G. Troskie for his skilful participation, criticism and supervision. While leaving me to my own creative processes, he had a pertinent leading question to contribute at each separate phase of the research. He shared insights generously.

To Mrs M.I. Cousins for accurate typing of difficult notation, and a tolerant attitude to alterations, my gratitude. She knows that I could not have completed the thesis without her assistance.

My wife, Lucille, tolerated my preoccupation with the thoughts and theorems of others, and my own, for some years. Her special place in what I am and what I do is, I hope, nonetheless evident through the mists of concentration. She has presented me with a son, Rowan Christopher, as I write the last of these pages. Both of them will be receiving my full attention in the days ahead.

My parents, Pat and Edward, gave me in every aspect of my life, more than I could ever repay. I hope to emulate their example.

Colleagues and teachers have broadened my knowledge, shared their experience, and become trusted friends. While remembering all of them at the University of Cape Town, and the University of Natal, I must single out Professor G.D.L. Schreiner and Professor B.M. Nevin, to whom I have turned for help at several important stages of my life.

Dr D.M. Hawkins, of the National Research Institute for the Mathematical Sciences, Council for Scientific and Industrial Research, Pretoria, assisted with a visit to the Institute and personal discussions. Professor Adi Ben-Israel, while visiting the University of Cape Town, also discussed issues generously. Responsibility for statements and results nonetheless falls entirely upon myself.

Finally, with fondness and respect, I salute Professor A.A. Rayner, in this year of his retirement from the University of Natal. A fortunate set of circumstances brought me under his influence. Diligent teacher, prickly administrator and creative biometrician, he epitomises for me what academic excellence must be in its broadest terms. I hope his vacating the Chair will not mean that he vacates his desk. It is to him that this thesis is gratefully dedicated ....
ad jucundam senectutem.

A B S T R A C T

Column-space conditions are shown to be at the heart of a number of identities linking generalized inverses of rectangular matrices.

These identities give some new insights into reparametrizations of the general linear model, and into the imposition of constraints, when the variance-covariance structure is $\sigma^2.I$.

Hypothesis-test statistics for non-estimable functions are shown to give no further information than underlying estimable functions.

For an arbitrary variance-covariance structure the "sweep-out" method is generalized. The John and Draper model for outliers is extended, and distributional results established. Some diagnostic statistics for outlying or influential observations are considered. A Bayesian formulation of outliers in the general linear model is attempted.

CONTENTS

PREFACE

In accordance with the regulations for the Degree of Ph.D. from the University of Cape Town, the candidate presents a summary of the contents of the thesis indicating in what way they constitute a contribution to knowledge.

Chapter 1 is an introduction to the problem of estimation, and related issues. No new results are given.

Chapter 2 comprises some well-known results that are required for developments in subsequent chapters. They have been arranged so as to avoid unnecessary deviations in later proofs. However several new results are included. Theorems 2.5 and 2.6 and corollaries examine the construction of two-condition generalized inverses. Theorems 2.10, 2.11 and 2.20 through 2.22 show how column-space conditions underpin results previously associated with non-singularity of matrices. Theorems 2.13 and 2.15 and some related corollaries systematize some known special cases of g_1-inverses for partitioned and bordered matrices. An ancillary result on conditional distributions, and another on the predictive distribution of a set of future observations, are given in Section 2.2.

Chapter 3 summarizes issues concerned with estimation in the usual linear model. The statistical folk-lore associated with estimation is vast, but some insights and remarks concerning relative error (RE) are apparently unpublished. Some relatively unknown work of Rayner (1977) is included. Theorem 3.4 specifies a (trivial) uniqueness property associated with BLUE's. Results on reparametrization (Theorems 3.12 and 3.13) and on prior linear constraints (Theorems 3.16 and 3.17) are extensions of known theory.

Chapter 4 concerns partitioning of the sums of squares associated with parts of a linear model. Section 4.2 draws partly on the folk-lore but does not appear to have been presented before. Theorem 4.2 and Lemma 4.4 with corollaries explore hypothesis-testing under restrictions which involve non-estimable functions. A fairly extensive literature summary constitutes the remainder of the material.

Chapter 5 seeks to inter-relate four distinct approaches to the general linear model with arbitrary variance-covariance structure $\sigma^2.V$. Goldman and Zelen (1964) results are generalized as Lemma 5.3 and Theorems 5.5 through 5.7. A proof is given for a claim of Zyskind and Martin (1967). Properties of possible F-ratios, given by Rao (1971) are examined and some new insights provided, notably by Lemma 5.12 and Theorem 5.13. A corrected proof of Theorem 5.17 seeks to improve an apparent mis-statement in the Rao paper above.

Chapter 6 has Theorem 6.1 and corollaries generalize the John and Draper formulation of outliers in the general linear model, and provide a test-statistic which includes as special cases some important statistics in the literature. Diagnostic methods are recorded in Section 6.2 where the test-statistic is extended to principal component regression methods. A ridge regression extension fails. Section 6.3 explores an extension to the predictive distribution, and a model selection procedure which may assist in locating outliers. Results on predictive distributions are generalized for the case of appropriate natural conjugate priors in the multiple regression model with normality assumed.

For the convenience of the referees new results and extensions of old results are indicated by the appearance of the candidate's name behind the heading of the results, e.g. as in Theorem 2.5 on p.2.6. In all such cases, prior research is acknowledged in context, and it is not intended to suggest that a specified result is necessarily new in its entirety. Theorems and corollaries are, where possible, attributed to an original source. For some of such earlier results, alternative more convenient proofs have been given by the candidate, and these are indicated by a label after the heading of the proof, as in Corollary 2.4.2 on p.2.6. Any errors are therefore totally the responsibility of the candidate. An end-of-proof symbol □ has been liberally employed.

Possible areas of further research or extensions are sketched in the context and content of appropriate chapters.

Timothy T. Dunne
August 1982

CHAPTER 1

INTRODUCTION

Consider the usual well-known (linear) model

(1.1) $\underset{\sim}{y} = X\underline{\beta} + \underline{\varepsilon}$,

where $\underset{\sim}{y}$ is a known vector of n observed variate-values, X is an $n \times k$ matrix of known elements which are by assumption non-stochastic, $\underline{\beta}$ is an unknown vector of k parameters associated with the random variate $\underset{\sim}{y}$, and $\underline{\varepsilon}$ is an unknown and unobservable random variate. It is assumed that

(1.2) $E(\underline{\varepsilon}) = \underline{0}$,

where E indicates expected value or mean, and thus,

(1.3) $E(\underset{\sim}{y}) = X\underline{\beta}$.

The imposed function of the model is to mathematically describe the effects of concomitant variables on the induced stochastic variate of observation, viz. $\underset{\sim}{y}$. Clearly the description implied by the model is limited to a particular form, which is in the first place additive (and hence linear) over the assumed underlying random variate $\underline{\varepsilon}$, and secondly, governed by exactly k constants. It is taken that $k \lesssim n$. Were $k > n$, the model would be said to be "over-specified", in respect of parameters.

Essentially the model is an attempt to specify simultaneously the expected values of (up to n) various assumed subpopulations in the assumed population serving as the conceptual model or source of the n observations, and to do so *a priori*, i.e. before the observations are made. As such it attempts to impose an explanatory structure, of the type (1.3) described above, on the expected value $\underline{\mu}_y$ of the vector-variate $\underline{y}$, as though $\underline{\mu}_y$ were indeed to have such a form.

The known entries in the X matrix may arise from deliberate manipulation of the observation source, e.g. for presence or absence of particular influences in the observations, as in experimental designs. Whoever is responsible for such manipulations, and however they are constructed, the model presupposes an objective quantifiable response associated with each influence or influence level, i.e. $\underline{\beta}$, which results in a cumulated additive effect $X\underline{\beta}$ as the expected value.

More loosely, the X matrix entries may themselves result from observation, but be regarded as non-stochastic. In such cases the assumption that a constant vector $\underline{\beta}$ exists and will apply over the entire possible ranges of entry observations in X, can be misleading. The model is intended as a device to reasonably explain observations, in such a way that statistical approximation, or *estimation* of the unknown quantities in $\underline{\beta}$ (or some linear functions of those quantities) is achieved.

Estimates, once found, tend to be used predictively. Assumptions of constancy, and estimates associated with $\underline{\beta}$, must take into consideration the nature and range of X matrix entries. This problem is the problem of extrapolation (which is briefly discussed in a later chapter).

At this point nothing has been said of the statistical structure of $\underline{\varepsilon}$. The usual assumption is that $\underline{\varepsilon}$ is composed of n uncorrelated copies of a single random variate, whose mean is zero, and whose unknown variance is σ^2. The variance-covariance structure V of $\underline{\varepsilon}$ is therefore of the $(n \times n)$ matrix form:

(1.4) $V = \sigma^2.I$,

and is, on the assumptions of the model, also the variance structure of $\underline{y}$.

Theoretical attention has been given to other variance structures. Where $\underline{\varepsilon}$ is assumed to be multivariate normal in distribution, (1.4) implies the independence of the component variates in $\underline{\varepsilon}$. However multivariate normal distributions exist, without independence, e.g. the simplest case of autocorrelation, so that consideration of V non-singular and not satisfying (1.4) will provide one area of examination. The literature also includes a considerable body of theory for the V singular case, though it is difficult to describe a real situation in which a singular multivariate normal distribution may serve as an appropriate model. This cannot be a serious criticism of the pertinence or otherwise of investigating the singular V, because for the most part continuous distributions are themselves models without real world counterparts except in an approximate sense. Nonetheless, as simple examples, we may note that directional data-vectors $\underline{x}$ in circles, spheres and hyperspheres may be described in similar terms. We will therefore have recourse to examine the consequences of

(1.5) $\quad \text{var}(\underline{\varepsilon}) = \sigma^2 . V$

where σ^2 is an unknown scale parameter and V, $(n \times n)$ of arbitrary rank, describes the relative variance structure. It will furthermore become clear that artificial examples exist, and that it will be mathematically economical to treat certain problems in this artificial way.

In all cases we will take the symmetric matrix V to be known. Again this is a usual assumption, evident even in the case of (1.4). The theoretical intractability of distribution theory when V has itself to be estimated from the data renders serious difficulties for researchers. It may be that the examination of the V-assumed-known case will shed light on estimated-V or partially estimated-V problems. In any event a first approximation, of using an estimated V as though it were known, can serve as an initial development.

Chapter 2 presents a body of theory which underpins a full distributional analysis of the linear model under either of the variance-covariance structures defined by (1.4) or (1.5). Generalized inverse theory and its relationship to the linear model and regression have been extensively surveyed by Pringle and Rayner (1971) and Albert (1972). These texts draw heavily upon results of Penrose (1955), Bjerhammar

(1958), Bose (1959), Rohde (1964) and Chipman (1964, 1968). An overview is presented, and some of the results are generalized or their inter-relationships described, in Section 2.1. Some aspects of two-way partitioning of matrices are involved in the material of Chapter 2. Distributional results for linear and quadratic forms in multivariate normal distributions are summarised in Sections 2.2 and 2.3. In Section 2.4, a generalization of two-way partitioning yields a result that is useful in analysis of covariance.

The immediate interest after constructing a model for observations and recording the observations is that of estimating the unknown constants $\underline{\beta}$ and σ^2 on the basis of the available data information. This process is subject to induced stochastic variation, because of the intrusion of $\underline{\varepsilon}$ in the estimates that are eventually chosen. The first step is to decide what can in fact be estimated in the model proposed, on the basis of the observations $\underline{y}$ and the information implied by X. Definition of what is estimable, and properties of the corresponding estimators and estimates are the concern of Chapter 3.

Some criteria of estimability, defined by Bose (1944), are examined, including that of Milliken (1971). Golub and Styan (1973) examined ill-conditioned matrices X, and suggested for estimable parameters $\underline{\beta}$ methods of computation which may be reasonably expected to be stable. It will be argued that the condition number reflects variance and extrapolation effects. An alternative measure of computational accuracy, proposed by Longley (1967), involves checking whether or not

$$(1.6) \qquad X'\underline{\delta} = \underline{0} \qquad ,$$

where $\underline{\delta}$ is the vector of estimated deviations. This is generalized in Chapter 3 to obtain a scalar-valued index of the accuracy of an algorithm on a given matrix X, with the index independent of the observations $\underline{y}$. Many issues related to the stability of an algorithm are examined in Wampler (1970). The general finding was that programs using Householder transformations and Gram-Schmidt orthogonalization procedures appeared to be more accurate than elimination procedures.

When $\underline{\beta}$ is not estimable, the model is said to have non-full rank. An early paper of Rohde and Harvey (1965) suggested that methods for the full-rank case, and specifically the Doolittle method of finding estimates could be extended to non-full rank models. Other techniques involve reparametrization of the model or the imposition of linear restrictions on $\underline{\beta}$. The effects of such restrictions have been examined *inter alia* by Chipman (1964), Graybill (1961) and Pringle and Rayner (1971, 1973), who have examined conditions under which linear restrictions and reparametrizations coincide.

The term prior linear constraints is used to refer to restraints which may intrude into the space of estimable functions. These correspond to reduced models, and distributional results may be extended to cover such situations.

Criteria other than unbiasedness and minimum variance have been proposed by a number of authors. Some of these are examined in Section 3.7. In a sense then, Chapter 3 is concerned with anomalies, inadequacies or superfluous elements within the model itself.

Chapter 4 considers similar elements, but now judgments of a statistical nature are made, and these involve estimates of particular functions of the parameters. Essentially the model and the data are simultaneously considered. Methods such as hypothesis-testing are considered, with particular emphasis on the well-known underlying principle of partitioning the sums of squares associated with various components of the model. Besides the inappropriateness of a model *per se*, a source of evidence against the internal consistency of a set of observed variate-values $\underline{y}$ (with respect to a given model) may be the presence of erratic or anomalous data-values within the observations $\underline{y}$. This distinction is roughly that between having the right-hand side of (1.1) as misleading or false, and having one or more unrepresentative elements intruding into the left-hand side. Such elements are termed "outliers" and have served as a focus for much recent research (Hawkins, 1980; Barnett and Lewis, 1978). Those tests of hypotheses for such the presence of outliers have been formulated *inter alia* by Gentleman and Wilk (1975b) and John and Draper (1978), these issues are deferred to Chapter 6 in which they are examined under more general conditions than those of the model (1.1).

Specific types of hypotheses, involving orthogonal functions of the observations, or non-testable relations in the usual sense, are also examined. Partitioning of models and the consequent partitioning of sums of squares give rise to well-known tests for model reductions, either by dropping higher-order parameters or by ignoring the available covariates. Sections 4.2 to 4.5 deal with these issues.

Related to the idea of outliers is that of missing observations, and hence up-dating of regression by augmenting additional data. These are briefly described in Section 4.6.

It has been noted that the assumption of an arbitrary variance-covariance structure may be made as in (1.5). It then follows that all the linear model theory of Chapters 3 and 4 has to be generalized for the new assumption. Historically the origin of the concern for V possibly non-singular is described by Rao (1971) as arising out of research in 1954-1955 on anthropometric data obtained on families of Hiroshima and Nagasaki, reflecting the effects of atomic bomb radiation. In that case the design matrix X was non-orthogonal and had some rank deficiency. Assuming the usual variance-covariance structure $\sigma^2.I$, it was known that when X had full-column rank, the variance-covariance matrix of $\underline{\beta}$ could be obtained from Fisher's matrix $C = S^{-1}$. Rao required to estimate different contrasts with different precisions, and did not know which contrasts were of most interest to those who had commissioned his analysis. He was then led to the problem of finding a matrix C which sufficed for X non-full rank. His matrix C was singular, but generated the variances of all estimable functions of $\underline{\beta}$. The relationship of C to g_1-inverses was discovered only later by him.

Several formulations of the problem of finding estimates for $\underline{\beta}$ under arbitrary variance-covariance structure are summarized and interrelated in Chapter 5. It is interesting to note that all of the methods can be extended to a complete treatment of the least squares approach for the more general assumption here. We will also note in Chapter 5 that in general, not every consistent hypothesis on estimable functions is open to a test in the usual sense. Accordingly a notion of strong testability is defined, which appears to be a generalization of a notion of Roy and Roy (1960) recorded in Elston and Bush (1964). These matters occupy Sections 5.1 to 5.4. In the last two sections of Chapter 5 we examine some relations which may lead to both computational simplifications and a wider understanding of what estimation entails when the variance-covariance structure is singular.

Recent literature gives evidence of renewed interest in the problem of outliers. Hawkins (1980) has given an extensive survey of current theory and an overview of one hundred and fifty years of interest in the topic. Similarly Barnett and Lewis (1978) devote a lengthy chapter to outliers in designed experiments, regression and time series. The texts of these authors give some attention to outliers in the linear model, with the usual assumption of the variance matrix

(1.4) $\quad var(\underset{\sim}{\varepsilon}) = \sigma^2.I \quad .$

A regression formulation of the problem has been given by John and Draper (1978), and is generalized in Chapter 6. A number of results have special cases which are equivalent to well-known elements of theory, including properties of the statistics of Ellenberg (1973), Gentleman and Wilk (1975b) and several others. These will be examined within Section 6.1. It appears that some consideration of possible diagnostic methods for outliers in principal component analyses and in ridge regression should be attempted, and also, more generally, a Bayesian approach. In Sections 6.2 and 6.3 we present such approaches, and attempt to relate the development to the notion of influential observations as defined by Cook (1977). It is in this chapter that the notion of arbitrary variance-covariance structure yields interesting insights into the nature of least squares estimation in the general linear model.

CHAPTER 2

ANCILLARY ALGEBRAIC AND STATISTICAL THEOREMS

This chapter specifies a wide class of results and theorems that are of fundamental importance in the development of the theory of the general linear model. The presentation is intended to provide the reader with ready access to a body of theory which underpins the later chapters. Unless otherwise stated, the results are not new. Proofs are only provided for new work, or for special cases that yield an economy of proof in the material of subsequent chapters.

2.1 GENERALIZED INVERSE THEORY

Pringle and Rayner (1971, pp. 1-54) give a history of the development and major results. In keeping with their notation, we denote real matrices by ordinary capitals A, B,... . Underlined lower case letters denote vectors, and may appear with or without suffices. Scalars are denoted by lower case letters. The Greek alphabet is used where population parameters and related concepts are discussed, in keeping with the conventions of much of the literature.

For any matrix A of dimension $n \times k$, the Moore-Penrose (generalized) inverse, G, of A, must satisfy

(2.1)	(1)	$AGA = A$
	(2)	$GAG = G$
	(3)	$AG = (AG)'$
	(4)	$GA = (GA)'$

We note that, for conformability, G must be $k \times n$, and that the properties refer to matrices of four different orders.

<u>Theorem 2.1</u> (Moore, 1920; Penrose, 1955).

The equations (2.1) have a unique solution, denoted by A^g, for each matrix A.

It is however well-known that there exist classes of matrices satisfying some, but not necessarily all, of the properties in (2.1). Using the notation of Chipman (1968) we write

(2.2)	A^{g_1}	satisfying at least (2.1)	(1)
	A^{g_2}		(2)
	$A^{g_{12}}$		(1) and (2)
	$A^{g_{13}}$		(1) and (3)
	$A^{g_{14}}$		(1) and (4)
	$A^{g_{123}}$		(1), (2) and (3)
	$A^{g_{124}}$		(1), (2) and (4)
	$A^{g_{134}}$		(1), (3) and (4)

This list is not exhaustive, but comprises the cases of interest in this discussion. Further, (2.2) serves to emphasize the (trivial) inclusion relations existing between such matrix classes, e.g.

(2.3) $A^g \in \{A^{g_{134}}\} \subset \{A^{g_{13}}\} \subset \{A^{g_1}\}$

where A^{g_1} is clearly $k \times n$. Throughout this discussion the symbols in (2.2) will be used in two senses : either as the class itself, that is in the style of (2.3), or as an arbitrary or particular element of that class. This notational device is in keeping with presentations elsewhere, and where the dual use is confusing, an explanatory note or explicit form will be provided. We write

A^g is the g -inverse of A,

A^{g_1} is a g_1-inverse of A,

$A^{g_{13}}$ is a g_{13}-inverse of A, and so on.

<u>Theorem 2.2</u> (Rohde, 1964, p.34).

If P_1 and P_2 are non-singular matrices, then, given conformability,

$$[P_1 A P_2]^{g_1} \equiv P_2^{-1} A^{g_1} P_1^{-1} \quad . \tag{2.4}$$

<u>Corollary 2.2.1</u> (Bose, 1959).

If A has rank r, and P_1 and P_2 as above satisfy

$$P_1 A P_2 = N = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} \quad , \tag{2.5}$$

then G is a g_1-inverse of A if and only if

$$G = P_2 \begin{bmatrix} I_r & U \\ V & W \end{bmatrix} P_1 \quad , \tag{2.6}$$

for arbitrary U, V, and W of appropriate orders.

<u>Corollary 2.2.2</u> (Pringle and Rayner, 1971, p.14).

For G in (2.6) to be a g_{12}-inverse of A, we require the additional necessary and sufficient condition

(2.7) $W = VU$,

for property (2) of (2.1) to hold.

This type of approach does not admit a typification of conditions in U, V and W for g_{13}- and g_{14}-inverses. We may however generate such a typification by means of an alternative approach examined below, based upon solutions to linear equations.

<u>Corollary 2.2.3</u> (Pringle and Rayner, 1971, p.18).

$A^{g_1} \in \{A^{g_{12}}\}$ is equivalent to

(2.8) $r(A) = r(A^{g_1})$,

whereas, in general, $r(A^{g_1}) \geq r(A)$.

Inspection of (2.6) shows that it is always possible to find g_1- , g_{13}- , and g_{14}-inverses of maximum rank p, where $p = \min(k,n)$.

<u>Theorem 2.3</u> (Rao, 1967).

A necessary and sufficient condition for B to be in the column-space, $\mathcal{C}(A)$, of A, is

(2.9) $AA^{g_1}B = B$.

Likewise, for B in the row-space, $\mathcal{R}(A)$, of A,

(2.10) $BA^{g_1}A = B$.

Attention is drawn to the dual meanings of A^{g_1} in (2.9) and (2.10), as in the following.

Theorem 2.4 (Penrose, 1955).

A necessary and sufficient condition for the equations

(2.11) $AXB = H$

to have a solution for X, is consistency, i.e.

(2.12) $AA^{g_1}HB^{g_1}B = H$.

Then the general solution, with Z arbitrary but conformable, is

(2.13) $X = A^{g_1}HB^{g_1} + Z - A^{g_1}AZBB^{g_1}$.

Corollary 2.4.1 (Penrose, 1955).

Equation (2.13) may be taken, without loss of generality, as

(2.14) $X = A^{g}HB^{g} + W - A^{g}AWBB^{g}$.

Proof (Dunne): For arbitrary conformable W, (2.14) may be made to generate any particular instance of (2.13) simply by setting W equal to the required value. Thus (2.13) and (2.14) define identical sets of matrices X, as required. □

Corollary 2.4.2 (Bjerhammar, 1958; Pringle and Rayner, 1971).

The equations $AXA = A$ satisfy consistency trivially, and thus the set of g_1-inverses A^{g_1} of A may be typified as

(2.15) $X = A^{g_1}AA^{g_1} + Z - A^{g_1}AZAA^{g_1}$

with all choices of conformable Z and of g_1-inverses A^{g_1} arbitrary, or as

(2.16) $X = A^g + W - A^g AWAA^g$.

Proof (Dunne): In (2.15) take all $A^{g_1} = A^g$ and $Z = W$, thus generating (2.16) by noting $A^g AA^g = A^g$. Now in (2.16) set W equal to any required instance of (2.15), so that (2.16) reduces to (2.15) as required. □

Since (2.16) is an economical generator of all g_1-inverses, of A, we investigate the form to typify all g_{12}-, g_{13}- and g_{14}-inverses of A.

Theorem 2.5 (Dunne)

A necessary and sufficient condition for G to be a g_{12}-inverse of A is that, for arbitrary conformable W, and without loss of generality,

(2.17) $G = A^g + W - A^g AWAA^g + (I-A^g A)(WAW-W)(I-AA^g)$.

Proof: Pringle and Rayner (p.24) note that G is a g_{12}-inverse of A if and only if $G = G_1 AG_2$ for arbitrary choices G_1 and G_2 of g -inverses of A. That being so, the inclusion of g_{12}-inverses in the set of g_1-inverses implies that G is, without loss of generality, of the form $G_1 AG_1$ for some G_1 generated by (2.16). The result follows. □

Theorem 2.6 (Dunne)

A necessary and sufficient condition for G to be a g_{13}-inverse of A, is that W in (2.16) satisfy

(2.18) $AW = KA'$,

for some K. Likewise, for g_{14}-inverses, that W satisfy

$$(2.19)\quad WA = A'L$$

for some L.

Proof: Post-, or pre-multiply (2.18) by A. Then properties (3) and (4) of (2.1) yield

$$(2.20)\quad AX = A^{g'}A' - AW + A'A^{g'}WA^{g'}A' = A'X' \quad , \quad \text{and}$$

$$(2.21)\quad XA = A'A^{g'} - WA + A'A^{g'}WA^{g'}A' = A'X'$$

if and only if (2.18) and (2.19) hold, respectively. □

<u>Corollary 2.6.1</u> (Chipman, 1968, p.119).

The following equations hold over arbitrary choices of g_{13}- and g_{14}-inverses of A:

$$(2.22)\quad AA^{g_{13}} = AA^{g} \quad ,$$

$$(2.23)\quad A^{g_{14}}A = A^{g}A \quad .$$

Proof: (Dunne). For W as in Theorem 2.6, the terms involving W in (2.20) and (2.21) fall away, since

$$\begin{aligned}(2.24)\quad A(A^{g}AWAA^{g}) &= AWA^{g'}A' \\ &= KA'A^{g'}A' \\ &= KA' \\ &= AW \quad ,\end{aligned}$$

and similarly,

$$(2.25)\quad (A^{g}AWAA^{g})A = WA \quad .$$ □

There is nothing really profound about the typifications (2.16), (2.17), (2.18) and (2.19). They serve only to highlight the essential distinguishing characteristics of two-property g_1-inverses, and to generate all such inverses. Given any particular g_1-inverse, A^{g_1}, it can be generated by setting $W = \pm A^{g_1}$, or $W = (A^{g_1}-A^g)$, in (2.16), regardless of whether or not it possesses any of the other properties (2), (3) or (4) of (2.2). Similarly taking $W = A^{g_{12}}$ or $W = (A^{g_{12}}-A^g)$ in (2.17) yields $A^{g_{12}}$ itself. Particular inverses, such as g_{13}- and g_{14}-inverses, will be related to specific procedures in the analysis of the linear model, in the following chapter. It has transpired that many examples of g_{13}-inverses in the literature have in fact been g_{123}-inverses, and similarly g_{14}-inverses have satisfied property (2). Chipman (1968) implies that the additional property is assumed because such inverses often appear in forms such as (2.22) and (2.23), whence it may be convenient to focus attention on those $A^{g_{13}}$ with the same rank as A. This is not always the case.

Corollary 2.6.2 (Dunne)

The following equations, with W arbitrary, typify g_{13}- and g_{14}-inverses of A, respectively:

(2.26) $A^{g_{13}} = A^g + (I-A^gA)W$, and

(2.27) $A^{g_{14}} = A^g + W(I-AA^g)$.

Proof: For W as in Theorem 2.6 we have

(2.28) $A^gAWAA^g = A^g(KA')(A^g)'A'$

$= A^gKA'$

$= A^gAW$

or, similarly,

(2.29) $A^gAWAA^g = A'LA^g = WAA^g$.

Substitution in (2.16) completes the proof. □

Pringle and Rayner (p.28) showed a similar result for the special case of A of full row-rank.

Corollary 2.6.3 (Dunne)

For the g_{13}- and g_{14}-inverses of (2.26) and (2.27) to satisfy condition (2), and thus generate g_{123}- and g_{124}-inverses, a necessary and sufficient condition for each expression respectively is that

(2.30) $(I-A^gA)W = KA'$, and

(2.31) $W(I-AA^g) = A'L$.

Proof: Since

(2.32) $A^g = A^gAA^g$

$= A^gAA^gAA^g$

$= A'(A^g)'A^g(A^g)'A'$,

equations (2.30) and (2.31) are the respective equivalent conditions for

(2.33) $r(A^{g_{13}}) \leq r(A')$

and hence, (2.8) in Corollary 2.2.3. □

Similar results to those of Theorems 2.4 through 2.6, and the corollaries, may be obtained by using another characterization of g_1-inverses (Searle, 1971, p.25)

(2.34) $A^{g_1} = A^g + (I-A^gA)U + V(I-AA^g)$.

This is equivalent to (2.16), because the right-hand side of (2.34) may be obtained from (2.16) by taking W equal to the desired expression in U and V, and because the right-hand side of (2.16) is obtained from (2.34) by setting $U = WAA^g$ and $V = (W-U)$. In the alternative development based on (2.34), only the form of (2.17), the typification of g_{12}-inverses, alters substantially to

(2.35) $$A^{g_{12}} = A^g + (I-A^gA)U + V(I-AA^g) + (I-A^gA)(UAV-U-V)(I-AA^g)$$

The focus of attention is now turned to some square matrices, and their generalized inverses, which have application in linear model theory.

<u>Theorem 2.7</u> For arbitrary choice of g -inverses, the matrices AA^{g_1}, $A^{g_1}A$, $I_n-AA^{g_1}$, $I_k-A^{g_1}A$ are all idempotent, with ranks equal to $r(A)$, $r(A)$, $n-r(A)$ and $k-r(A)$ respectively.

Proof: Idempotency follows from (2.1) and (2.2) and the collection of terms. The rank results follow from the rank of products property in

(2.36) $r(A) = r(AA^{g_1}A) \leq r(AA^{g_1}) \leq r(A)$

and the rank-trace property

(2.37) $r(Q) = tr(Q)$

for idempotent matrices Q. □

The theorem is well-known and widely presented (e.g. Rao, 1973, pp.28 and 25). Since each idempotent matrix is easily shown to be one of its own g_{12}-inverses, from idempotency and (2.10), it is also clear that the matrices AA^{g_1} and $A^{g_1}A$, for arbitrary A and arbitrary choices of g_1-inverses, form the complete class of idempotent matrices of order $n \times n$ and $k \times k$ respectively. Projection matrices are precisely the symmetric idempotent matrices, and the classes of such matrices of the above orders are constituted by taking $AA^{g_{13}}$ and $A^{g_{14}}A$ respectively. By (2.22) and (2.23) these classes may be thought of as AA^g and A^gA respectively, for arbitrary A. The notion of idempotency will be called upon so as to apply (2.37) in later distribution theory. A related development is that of

Theorem 2.8 (Bose, 1959)

The matrix $A(A'A)^{g_1}A'$ is unique, idempotent and symmetric, with rank equal to the rank of A, and

$$(2.38) \quad A(A'A)^{g_1}A' = AA^g = (A')^gA' .$$

Proof: Uniqueness is a consequence of

$$(2.39) \quad R(A) = R(A'A) \quad , \qquad \text{and}$$

$$(2.40) \quad C(A') = C(A'A) \quad ,$$

which give, for some K and L, that

$$(2.41) \quad A(A'A)^{g_1}A' = K(A'A)(A'A)^{g_1}(A'A)L$$

$$= K(A'A)(A'A)^{g}(A'A)L$$

$$(2.42) \quad = A(A'A)^{g}A' \quad ,$$

together with

(2.43) $A'.A(A'A)^{g_1}A' = A'$, and

(2.44) $A(A'A)^{g_1}A'.A = A$.

In turn, either of these last two equations implies

(2.45) $r(A) \leq r(A(A'A)^{g_1}A') \leq r(A)$

by repeated use of the rank of products inequality. Further, setting B as in

(2.46) $B = (A'A)^{g}A'$,

it is easily verified that (2.38) holds, because

(2.47) $B = A^{g}$. □

Corollary 2.8.1 (Rohde, 1964, p.14)

A matrix G is a g_{123}-inverse of A if and only if it is of the form

(2.48) $G = (A'A)^{g_1}A'$

Similarly, for g_{124}-inverses of A,

(2.49) $G = A'(AA')^{g_1}$.

Corollary 2.8.2 (Bjerhammar, 1958)

For arbitrary g_1-inverses,

(2.50) $A^{g} = A'(AA')^{g_1}A(A'A)^{g_1}A'$.

Corollary 2.8.3 (Pringle and Rayner, 1971, p.32)

If L is $q \times k$ and $S = A'A$, an equivalent condition to $L \subset \mathcal{R}(S)$, or $L \subset \mathcal{R}(A)$ is that

(2.51) $LS^{g_1}L' = LS^{g}L'$, and

(2.52) $r(LS^{g_1}L') = r(L)$.

Furthermore, if $Q = LS^{g_1}L'$, then Q and $R = L'Q^{g_1}L$ are unique, symmetric, positive semi-definite, and R has rank equal to the rank of L.

Proof: Using $L = KS$ for some S, we have

$$(2.53)\quad LS^{g_1}L' = KSS^{g_1}SK' = KSK' = KSS^{g}SK' = LS^{g}L' \quad .$$

But, from (2.53), we also have

$$(2.54)\quad r(Q) = r(LS^{g}L') = r(KSK') = r(KS) = r(L),$$

and thus $L' \subset R(Q)$. Applying the results (2.51) and (2.52) to these matrices, we have the uniqueness of

$$(2.55)\quad L'Q^{g_1}L = L'Q^{g}L \quad ,$$

whose symmetry and positive definiteness follow that the corresponding properties of Q, in (2.53). □

Using the above theorem and its corollaries we now examine and generalize some results which have been crucial to a development of a complete theory of the linear model.

<u>Theorem 2.9</u> (Pringle and Rayner, 1971, p.32)

For S and L as in Corollary 2.8.3,

$$(2.56)\quad (S+L'L)^{g_1} = S^{g_1} - S^{g_1}L'(I+LS^{g_1}L')^{-1}LS^{g_1} \quad ,$$

with all choices of S^{g_1} arbitrary.

Proof: From Theorem 2.3, and $L \subset R(S)$, (2.57) $SS^{g_1}(S+L'L) = (S+L'L)$.

Also, collection of terms yields

(2.58) $(S+L'L)[S^{g_1}-S^{g_1}L'(I+LS^{g_1}L')^{-1}LS^{g_1}]$

$= SS^{g_1} + L'LS^{g_1} - L'(I+LS^{g}L')(I+LS^{g}L')^{-1}LS^{g_1}$

$= SS^{g_1} + L'LS^{g_1} - L'LS^{g_1}$,

with initial choices of g_1-inverses arbitrary. Post-multiplication by $(S+L'L)$, and (2.51) with (2.57), give the result. □

Note that in view of (2.51) the third g_1-inverse on the right-hand side of (2.56) is always arbitrary. It remains an open question whether or not (2.56) determines the entire class of g_1-inverses. However, by Theorem 2.2, if it can be shown that

(2.59) $P_1 = I-L'(I+LS^{g_1}L')^{-1}LS^{g_1}$

can be taken as non-singular for suitably chosen g_1-inverses of S, then for $P_2 = I$, (2.56) is an equivalence condition on the class of g_1-inverses of $(S+L'L)$. For such a matrix P_1,

(2.60) $P_1(S+L'L) = S$,

(2.61) $[(S+L'L)^{g_1}P_1^{-1}] \equiv S^{g}$, and

(2.62) $(S+L'L)^{g_1} \equiv S^{g_1}P_1$.

We will return to this question after the next result, which is a generalization of a theorem of Chipman (1964).

<u>Theorem 2.10</u> (Dunne)

Let A and H satisfy the row-space condition

$$(2.63)\quad \mathcal{R}(A) \cap \mathcal{R}(H) = \{0\} \quad .$$

Then, for all such H,

$$(2.64)\quad A(A'A+H'H)^{g_1}H' = 0 \quad ,$$

$$(2.65)\quad (A'A+H'H)^{g_1} \in \{(A'A)^{g_1}\} \cap \{(H'H)^{g_1}\} \quad , \qquad \text{and}$$

$$(2.66)\quad (A'A+H'H)^{g_1}B' = B^{g_{123}}$$

for $B = A$ and $B = H$.

Proof: Let A be $n\times k$ of rank r and H be $q\times k$ of rank p. Without loss of generality we assume the first r or p rows of A and H are bases for their row-spaces, that is, for some N, M,

$$(2.67)\quad A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} I_r \\ N \end{bmatrix} A_1 \quad , \qquad \text{and}$$

$$(2.68)\quad H = \begin{bmatrix} H_1 \\ H_2 \end{bmatrix} = \begin{bmatrix} I_p \\ M \end{bmatrix} H_1 \quad .$$

In view of (2.63), we may write

$$(2.69)\quad r\begin{bmatrix} A_1 \\ H_1 \end{bmatrix} = r\begin{bmatrix} A \\ H \end{bmatrix} = r(A'A+H'H) \quad ,$$

so that for

$$(2.70)\quad Q = \begin{bmatrix} A_1 \\ H_1 \end{bmatrix} \quad ,$$

we investigate the existence of a right inverse R, with

$$(2.71)\quad QR = I_{r+p} \quad .$$

Pringle and Rayner (1971, p.27) have shown that consistency of (2.71) requires that, for all g_1-inverses of Q,

$$(2.72) \quad QQ^{g_1} = I_{r+p} \quad .$$

Using the full row-rank property of Q and Theorem 2.2 we may write

$$(2.73) \quad P_1QP_2 = [I_{r+p} : 0] \quad ,$$

$$(2.74) \quad [I_{r+p} : 0]^{g_1} \equiv \begin{bmatrix} I_{r+p} \\ V \end{bmatrix}$$

for arbitrary V, and

$$(2.75) \quad Q^{g_1} \equiv P_2\begin{bmatrix} I_{r+p} \\ V \end{bmatrix}P_1 \quad .$$

Every form of (2.75) is an admissible R, with

$$(2.76) \quad QQ^{g_1} = P_1^{-1}[I_{r+p} : 0]P_2^{-1}P_2\begin{bmatrix} I_{r+p} \\ V \end{bmatrix}P_1$$

$$= I_{r+p} \quad .$$

Now partition R conformably with Q so that

$$(2.77) \quad QR = \begin{bmatrix} A_1 \\ H_1 \end{bmatrix}[X : Y] = \begin{bmatrix} I_r & 0 \\ 0 & I_p \end{bmatrix} \quad .$$

This implies

$$(2.78) \quad AY = \begin{bmatrix} I_r \\ N \end{bmatrix}A_1Y = 0 \quad , \qquad \text{and}$$

$$(2.79) \quad HY = \begin{bmatrix} I_p \\ M \end{bmatrix}H_1Y = \begin{bmatrix} I_p \\ M \end{bmatrix} \quad ,$$

so that

(2.80) $(A'A+H'H)Y = H'HY = H_1'[I_p : M']\begin{bmatrix} I_p \\ M \end{bmatrix}$.

Thus, setting

(2.81) $W = (A'A+H'H) = [A' : H']\begin{bmatrix} A \\ H \end{bmatrix}$

in (2.80), and noting $\mathcal{R}(A) \subset \mathcal{R}(W) = \mathcal{R}\left(\begin{bmatrix} A \\ H \end{bmatrix}\right)$

we have, from (2.78) and (2.80), that

(2.82) $$\begin{aligned} 0 &= AY \\ &= AW^{g_1}WY \\ &= AW^{g_1}H'\begin{bmatrix} I_p \\ M \end{bmatrix} \end{aligned}$$.

Post-multiplication by $[I_p+M'M]^{-1}.[I_p : M']$ yields the desired result (2.64), which is equivalent to

(2.83) $(A'A)W^{g_1}(H'H) = 0$.

Applying (2.83) in both

(2.84) $$\begin{aligned} A'A &= A'AW^{g_1}W \\ &= A'AW^{g_1}(A'A+H'H) \\ &= A'AW^{g_1}A'A \end{aligned}$$, and

(2.85) $$\begin{aligned} H'H &= H'HW^{g_1}W \\ &= H'HW^{g_1}(A'A+H'H) \\ &= H'HW^{g_1}H'H \end{aligned}$$,

proves (2.65). In view of (2.65) and Corollary 2.8.1, (2.66) is a special case of (2.48). □

The foregoing proof follows the method that Chipman (1964) derived for the special case of (2.63) in which $R(A) \cup R(H)$ exhausts R^k, the k-dimensional Euclidean (row-) space, and in which W is nonsingular, with W^{-1} replacing the arbitary W^{g_1} throughout the theorem. Using this extended result we are in a position to finalize the discussion on equations (2.59) to (2.62) inclusive, and strengthen Theorem 2.9, as

Theorem 2.11 (Dunne)

For L $q \times k$ and $S = A'A$ with $L \subset R(S)$, then

(2.86) $(S+L'L)^{g_1} \equiv S^{g_1} - S^{g_1}L'(I+LS^{g_1}L')^{-1}LS^{g_1}$

with all choices of S^{g_1} arbitrary.

Proof: As stated in the discussion, the crucial element of the proof is the existence of a non-singular choice of P_1 in (2.59). Since $R(S) = R(S+L'L)$ we may find H such that $R(H)$ and $R(S)$ are virtually disjoint, and with the spaces $R(S+H'H)$ and $R(S+L'L+H'H)$ identically equal to R^k. This implies that the matrices $(S+H'H)$ and $(S+L'L+H'H)$ are non-singular. Taking P_1 as in (2.59), with $W = (S+H'H)$ and $S^{g_1} = W^{-1}$, yields

(2.87) $$\begin{aligned} P_1(S+L'L+H'H) &= P_1(S+L'L) + P_1H'H \\ &= (S+L'L-L'L) + (H'H-0) \\ &= S + H'H \end{aligned}$$

from (2.60) and (2.83). The rank inequality for products now implies that $r(P_1) \geq r(S+H'H)$, and thus that P_1 is non-singular as required. □

Theorem 2.11 depends essentially on the equivalence of the row- and column-spaces of S, which is guaranteed for S of the stated form, viz. positive semidefinite matrices. Such matrices occur throughout the linear model theory, and often enough in partitioned form. Suppose

$$(2.88)\quad S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}$$

is a partitioning with S_{11} and S_{22} square matrices. We investigate methods of finding g_1-inverses of S in terms of the constituent submatrices. This may be attempted through defining a conformable matrix T similarly partitioned, and investigating the relations on the constituent submatrices of T given by

$$(2.89)\quad STS = S \quad .$$

However some simplification is achieved, if we apply Theorem 2.2. Rohde (1964) noted that the solution obtained by using a matrix B defined as

$$(2.90)\quad B = P_1SP_2 = \begin{bmatrix} I & -S_{12}S_{22}^{g_1} \\ 0 & I \end{bmatrix} S \begin{bmatrix} I & 0 \\ -S_{22}^{g_1}S_{21} & I \end{bmatrix}$$

$$= \begin{bmatrix} I & -S_{12}G_1 \\ 0 & I \end{bmatrix} S \begin{bmatrix} I & 0 \\ -G_2S_{21} & I \end{bmatrix}$$

$$= \begin{bmatrix} Q & 0 \\ 0 & S_{22} \end{bmatrix}$$

where all choices of g_1-inverse are arbitrary, and

$$(2.91)\quad Q = S_{11} - S_{12}S_{22}^{g_1}S_{21} = S_{11} - S_{12}S_{22}^{g}S_{21} \quad .$$

Though Pringle and Rayner (1971, pp.53-54) draw attention to a wider class of g_1-inverses of B, it is clear that we may take

$$(2.92)\quad B^{g_1} = \begin{bmatrix} Q^{g_1} & 0 \\ 0 & S_{22}^{g_1} \end{bmatrix} = \begin{bmatrix} A & 0 \\ 0 & G_3 \end{bmatrix}$$

and, by applying Theorem 2.2, obtain

Theorem 2.12 (Rohde, 1964)

$$(2.93)\quad S^{g_1} = \begin{bmatrix} A & -AS_{12}G_1 \\ -G_2S_{21}A & G_3+G_2S_{21}AS_{12}G_1 \end{bmatrix}$$ □

Inspection of this matrix indicates firstly that the choice of A is identical in the four submatrices, and secondly, notwithstanding the form and derivation of (2.93), all choices of the g_1-inverses, G_i, of S_{22} are arbitrary. We may therefore write

$$(2.94)\quad S^{g_1} = \begin{bmatrix} A & -AS_{12}G_1 \\ -G_2S_{21}A & G_3+G_4S_{21}AS_{12}G_5 \end{bmatrix} .$$

Corollary 2.12.1 (Dunne)

For S^{g_1} as in (2.94) to be a g_{12}-inverse of S, we require that

$$(2.95)\quad A = Q^{g_{12}}$$

$$(2.96)\quad G_3 = S_{22}^{g_{12}}$$

and that (2.94) reduce to the form (2.93) by either having arbitrariness restricted to

$$(2.97)\quad G_1 = G_5 \quad ,$$ and

(2.98) $G_2 = G_4$,

or

(2.99) $S_{12}G_1 = S_{12}G_5 = S_{12}S_{22}^g$, and

(2.100) $G_2S_{21} = G_4S_{21} = S_{22}^gS_{21}$.

The latter conditions amount to having G_1 and G_5 an arbitrary pair of g_{13}-inverses of S_{22}, and similarly that G_2 and G_4 are arbitrary g_{14}-inverses of S_{22}.

Proof: A simple extension of Theorem 2.2 to g_{12}-inverses, implies that

(2.101) $$S^{g_{12}} = \begin{bmatrix} A & -AS_{12}G \\ -G_2S_{21}A & G_3+G_2S_{21}AS_{12}G_1 \end{bmatrix}$$

as in (2.93), but with (2.95) and (2.96) additionally, in which case (2.99) and (2.100) are sufficient in (2.94).

A similar result has also been established by Rayner, in unpublished lecture notes, and serves to correct misleading expressions for partitioned inverses in Pringle and Rayner (1971, p.46). As noted elsewhere, Theorem 2.2 and its corollaries do not lead to typifications of g_{13}- and g_{14}-inverses. However we may use similar results of Rohde (1964), Pringle and Rayner (1971) and Zelen and Federer (1965) and state, more precisely, the following

<u>Theorem 2.13</u> (Dunne)

The necessary and sufficient conditions for g_1-inverses of S as in (2.93) to satisfy the requirements for g_{13}-inverses, are that (2.99) hold, (2.102) $r(S) = r(S_{11}) + r(S_{22})$,

(2.103) $G_3 = S_{22}^{g_{13}}$, and

(2.104) $A = Q^{g_{13}}$.

Similarly, for g_{14}-inverses, the conditions are that (2.100) hold with (2.102), and

(2.105) $G_3 = S_{22}^{g_{14}}$, and

(2.106) $A = Q^{g_{14}}$.

Proof: Consider, from (2.93),

(2.107) $SS^{g_1} = \begin{bmatrix} QA & S_{12}G_3 - QAS_{12}G_1 \\ 0 & S_{22}G_3 \end{bmatrix}$.

Symmetry of QA and $S_{22}G_3$ implies and is implied by (2.103) and (2.104). The upper-right submatrix reduces to zero if and only if

(2.108) $(I-QA)S_{12} = 0$,

from (2.99). In turn, this is equivalent to

(2.109) $(I-QA)S_{11} = (I-QA)(Q+S_{12}S_{22}^{g}S_{21}) = 0$

and thus also

(2.110) $S_{11} \subset C(Q) \subset C(S_{11})$.

Noting that (2.90) implies

(2.111) $r(S) = r(Q) + r(S_{22})$,

it follows from (2.110) that

(2.112) $r(Q) = r(S_{11})$,

proving the result. A similar series of arguments applies in the g_{14}-inverse development. □

In the special case where S has a natural form of partitioning

$$(2.113) \quad S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} = \begin{bmatrix} X_1' \\ X_2' \end{bmatrix} [X_1 : X_2] \quad ,$$

the relation (2.112) is equivalent to

$$(2.114) \quad r[(I-X_2S_{22}^{g}X_2')X_1] = r(X_1)$$

which in turn gives

$$(2.115) \quad \mathcal{C}(X_2) \cap \mathcal{C}(X_1) = \{\underline{0}\} \quad .$$

Clearly a parallel development in which Q of (2.92) is replaced by $S_{11}^{g_1}$, and $S_{22}^{g_1}$ replaced by a g_1-inverse of $S_{22}-S_{21}S_{11}^{g}S_{12}$, leads to conditions similar to those of Theorem 2.13. The rank or column-space condition (2.102) or (2.115) remains unchanged for the parallel case.

The effect of the condition is to limit the cases in which the diagonalization approach leads directly to g_{13}- and g_{14}-inverses of partitioned matrices that are symmetric positive semi-definite. These include the g-inverse itself. However in practice the use of arbitrary g_1-inverses of S will be sufficient for the applications required in these chapters.

The so-called bordered matrices of the form

$$(2.116) \quad M = \begin{bmatrix} S & L' \\ L & 0 \end{bmatrix}$$

are not in general positive semidefinite matrices. Pringle and Rayner have shown that setting

$$(2.117)\quad P_1 = \begin{bmatrix} I & L' \\ 0 & I \end{bmatrix} \quad ,$$

$$(2.118)\quad P_2 = \begin{bmatrix} I & 0 \\ -LK^{g_1} & I \end{bmatrix} \quad , \quad \text{and}$$

$$(2.119)\quad P_3 = \begin{bmatrix} I & -K^{g_1}L' \\ 0 & I \end{bmatrix} \quad ,$$

yields

$$(2.120)\quad P_2P_1MP_3 = \begin{bmatrix} K & 0 \\ 0 & -R \end{bmatrix}$$

where the positive semi-definite matrices K and R are defined by

$$(2.121)\quad K = S + L'L \quad , \quad \text{and}$$

$$(2.122)\quad R = LK^{g_1}L' \quad .$$

However, though (2.120) and the non-singularity of P_1, P_2 and P_3 confirm the absence of a positive semi-definiteness property, these conditions allow applications of Theorem 2.2 to obtain

Theorem 2.14 (Pringle and Rayner, 1971, pp.48-51)

For S any $k\times k$ positive semi-definite matrix and L any $q\times k$ matrix, then M as in (2.116) admits a g_1-inverse of the form

$$(2.123)\quad \begin{bmatrix} S & L' \\ L & 0 \end{bmatrix}^{g_1} = \begin{bmatrix} K^{g_1}-K^{g_1}L'R^{g_1}LK^{g_1} & K^{g_1}L'R^{g_1} \\ R^{g_1}LK^{g_1} & R^{g_1}R-R^{g_1} \end{bmatrix}$$

Proof: The result follows immediately from

$$(2.124) \quad M^{g_1} = P_3 \begin{bmatrix} K^{g_1} & 0 \\ 0 & -R^{g_1} \end{bmatrix} P_2 P_1$$

where K and R are taken as in (2.120) and (2.121). □

<u>Corollary 2.14.1</u> (Pringle and Rayner, 1971)

For $L \subset R(S)$, then we may take

$$(2.125) \quad R = LS^{g_1}L'$$

whence

$$(2.126) \quad \begin{bmatrix} S & L' \\ L & 0 \end{bmatrix}^{g_1} = \begin{bmatrix} S^{g_1}-S^{g_1}L'R^{g_1}LS^{g_1} & S^{g_1}L'R^{g_1} \\ R^{g_1}LS^{g_1} & -R^{g_1} \end{bmatrix} .$$

Proof: Let $P_1 = I$ in (2.117), and $K = S$ in (2.121) so that (2.125) holds. Under these conditions (2.120) reduces to

$$(2.127) \quad \begin{bmatrix} I & 0 \\ -LS^{g_1} & I \end{bmatrix} M \begin{bmatrix} I & -S^{g_1}L' \\ 0 & I \end{bmatrix} = \begin{bmatrix} S & 0 \\ 0 & -R \end{bmatrix} .$$

The remainder of the proof follows easily. □

We note in passing that if L has full row-rank q, then under the conditions of the corollary R is non-singular and we take $R^{g_1} = R^{-1} = (LS^{g_1}L')^{-1}$. Further if S itself were also non-singular, the expression for M^{g_1} reduces to the regular inverse M^{-1}, as presented by Plackett (1960, p.67).

In view of Theorem 2.10, it is possible to extend two corollaries of Pringle and Rayner (1971, pp.51 and 52), to the following

<u>Theorem 2.15</u> (Dunne)

If L is a $q \times k$ matrix such that

(2.128) $R(L) \cap R(S) = \{\underline{0}\}$,

then we may write

(2.129) $$M^{g_1} = \begin{bmatrix} K^{g_1}SK^{g_1} & K^{g_1}L' \\ LK^{g_1} & 0 \end{bmatrix} .$$

Proof: By Theorem 2.10, we have that K^{g_1} is a g_1-inverse of K, S and $L'L$, and that

(2.130) $SK^{g_1}L' = 0.$

Now, either by examining the effect of (2.130) in the pre- and post-multiplication of (2.129) by M, or by observing that

(2.131) $R = LK^{g_1}L' = L(L'L)^{g_1}L'$

is idempotent and that we may take

(2.132) $R^{g_1} = R = R^2$

to reduce (2.123) to the form (2.129), the proof is complete. □

Special cases of the above theorem when

(2.133) $q > r(L) = k - r(S)$, or

(2.134) $q = r(L) = k - r(S)$

give the Pringle and Rayner corollaries:

(2.135) $$M^{g} = \begin{bmatrix} K^{-1}SK^{-1} & K^{-1}L' \\ LK^{-1} & 0 \end{bmatrix} ,$$ and

$$(2.136)\quad M^{-1} = \begin{bmatrix} K^{-1}SK^{-1} & K^{-1}L' \\ LK^{-1} & 0 \end{bmatrix}$$

respectively. Moreover, if

$$(2.137)\quad r(L) \leq k - r(S) \quad ,$$ then

$$(2.138)\quad M^{g} = \begin{bmatrix} K^{g}SK^{g} & K^{g}L' \\ LK^{g} & 0 \end{bmatrix}$$

under the condition (2.128). Since (2.138) yields

$$(2.139)\quad MM^{g} = \begin{bmatrix} KK^{g} & 0 \\ 0 & LK^{g}L' \end{bmatrix} \quad ,$$

the well-known relation

$$(2.140)\quad r(M) = r(MM^{g}) = r(M^{g}) \quad ,$$

implies that the necessary and sufficient condition for (2.138) to yield M^{-1} is precisely that both K and $LK^{g}L'$ are non-singular, or equivalently that (2.134) holds. As Goldman and Zelen (1964) have implicitly observed, (2.134) with the additional condition of orthogonality of the row-spaces of S and L, i.e.

$$(2.141)\quad SL' = 0 \quad ,$$ yields

$$(2.142)\quad M^{-1} = \begin{bmatrix} S^{g} & L^{g} \\ (L^{g})' & 0 \end{bmatrix} \quad .$$

However (2.141) may hold even if (2.134) does not. In that case, since (2.141) is equivalent to

$$(2.143)\quad S^{g}L' = 0 = SL^{g} \quad ,$$

from simple pre- and post-multiplication of (2.141) by $(S^{g})^{2}$ and $(LL')^{g}$ respectively, we may write the more general relation

$$(2.144)\quad M^g = \begin{bmatrix} S^g & L^g \\ (L^g)' & 0 \end{bmatrix} .$$

Clearly (2.143) does not require L to have full row-rank, so that the number of rows of L, viz. q, is not necessarily determined in any way by the maximum-rank quantity $k-r(S)$ of (2.137).

At this point in the development of the theory of the linear model, the fore-going results amount mainly to a classification of preceding work. The most important consequences for computation have been long known and applied, and discussion of computational methods is to be found, *inter alia,* in Rohde and Harvey (1965), Pringle and Rayner (1971, pp.65-69), and Golub and Kahan (1964), where g_1-inverses (with various possible additional properties) are described with the aid of Doolittle, Cholesky or Householder transformations. We return to applications in Chapter 3.

2.2 SINGULAR MULTIVARIATE NORMAL DISTRIBUTIONS

The following results will be applied in a Bayesian approach to outlier detection in a later chapter. Rohde (1968) states that where a normal multivariate $\underline{y} \sim N(\underline{\mu}, V)$ has V singular, the absence of a density function which is absolutely continuous with respect to Lebesque measure, renders as pointless the generalization of the density function from the non-singular V case. Any such density is of necessity a density only on the hyperspace described by

(2.145) $H_2\underline{y} = H_2\underline{\mu}$,

where H_2 spans the zero eigen-space of V. The values of the variable $\underline{y}$ must satisfy the relation (2.145) with probability one. Nonetheless, it is both useful and constructive to examine the marginal and conditional distributions associated with the singular case of V. A partial development is presented in Pringle and Rayner (1971, pp.70-72).

<u>Theorem 2.16</u> (Harris and Helvig, 1966)

Let $\underline{y} \sim N(\underline{\mu}, V)$ be partitioned conformably as

(2.146) $$\begin{bmatrix}\underline{y}_1 \\ \underline{y}_2\end{bmatrix} \sim N\left(\begin{bmatrix}\underline{\mu}_1 \\ \underline{\mu}_2\end{bmatrix}, \begin{bmatrix}V_{11} & V_{12} \\ V_{21} & V_{22}\end{bmatrix}\right) .$$

Then there exists C such that the conditional distribution of $\underline{y}_1$, given $\underline{y}_2 = \underline{k}_2$, is given by

(2.147) $$\underline{y}_1\Big|_{\underline{y}_2 = \underline{k}_2} \sim N(\underline{\mu}_1 + C(\underline{k}_2 - \underline{\mu}_2), V_{11} - CV_{22}C')$$

In fact, for arbitrary choice of g_1-inverse,

(2.148) $$C = V_{12}V_{22}^{g_1} ,$$

uniquely determines the mean and variance in (2.147).

Proof: Consider

(2.149) $$U = \begin{bmatrix} I & -C \\ 0 & I \end{bmatrix}$$

so that we define by linear transformation, a normal variate

(2.150) $\underline{z} = \begin{bmatrix} \underline{z}_1 \\ \underline{z}_2 \end{bmatrix} = U\underline{y} = \begin{bmatrix} \underline{y}_1 - C\underline{y}_2 \\ \underline{y}_2 \end{bmatrix}$

with

(2.151) $\underline{z} \sim N\left(\begin{bmatrix} \underline{\mu}_1 - C\underline{\mu}_2 \\ \underline{\mu}_2 \end{bmatrix}, \begin{bmatrix} V_{11} - CV_{22}C' & 0 \\ 0 & V_{22} \end{bmatrix}\right)$

if and only if C satisfies (2.148). In such a case $\underline{z}_1$ is independent of $\underline{z}_2$, and both are multivariate normal. Thus $\underline{z}_1 \Big| \underline{z}_2 = \underline{k}_2$ has the same distribution as $\underline{z}_1$.

Equivalently

(2.152) $\underline{y}_1 - V_{12}V_{22}^{g_1}\underline{y}_2 \Big|_{\underline{y}_2 = \underline{k}_2}$

is independent of $\underline{y}_2$, with mean

(2.153) $\underline{\mu}_1 - C\underline{\mu}_2 = \underline{\mu}_1 - V_{12}V_{22}^{g_1}\underline{\mu}_2$,

whence

(2.154) $E\left(\underline{y}_1 \Big|_{\underline{y}_2 = \underline{k}_2}\right) = \underline{\mu}_1 - C\underline{\mu}_2 + V_{12}V_{22}^{g_1}\underline{\mu}_2$
$= \underline{\mu}_1 + V_{12}V_{22}^{g_1}(\underline{y}_2 - \underline{\mu}_2)$

and

(2.155) $\mathrm{Var}\left(\underline{y}_1 \Big|_{\underline{y}_2 = \underline{k}_2}\right) = V_{11} - V_{12}V_{22}^{g_1}V_{22}V_{22}^{g_1}V_{21}$.

Clearly (2.155) is unique, and the well-known result that $\underline{x} \sim N(\underline{\mu}, V)$ implies $(\underline{x} - \underline{\mu}) \in C(V)$ with probability one, renders $(\underline{y}_2 - \underline{\mu}_2) \in C(V_{22})$ and hence the uniqueness of (2.154). The normality of the conditional variate is trivial. □

We now generalize an unpublished result of Pringle (1976).

Corollary 2.16.1 (Dunne)

Let $\underline{y} \sim N(\underline{\mu}, V)$, and suppose $R\underline{y} = \underline{c}$ is known. Then the conditional distribution of $\underline{y}\big|_{R\underline{y} = \underline{c}}$ is

$$(2.156) \quad N(\underline{\mu}+VR'(RVR')^{g_1}(\underline{c}-R\underline{\mu}),\ V-VR'(RVR')^{g_1}RV) \ .$$

Proof: We assume consistency of the given condition.

Consider

$$(2.157) \quad \underline{z} = \begin{bmatrix} I \\ R \end{bmatrix} \underline{y} = \begin{bmatrix} \underline{z}_1 \\ \underline{z}_2 \end{bmatrix} \ ,$$

with

$$(2.158) \quad \underline{z} = \begin{bmatrix} \underline{z}_1 \\ \underline{z}_2 \end{bmatrix} \sim N\left(\begin{bmatrix} I \\ R \end{bmatrix}\underline{\mu} \ , \begin{bmatrix} V & RV \\ VR' & RVR' \end{bmatrix}\right) \ .$$

Applying Theorem 2.16, to (2.158) the required result follows directly. □

The uniqueness over all g_1-inverses follows from

$$(2.159) \quad \underline{c} - R\underline{\mu} = R(\underline{y}-\underline{\mu}) \in C(RV)$$

and from the uniqueness of $VR'(RVR')^{g_1}RV$. It is the corollary and the uniqueness that allow an approach equivalent of that of Chipman (1968) to establish a more general result than that of Chipman:

Theorem 2.17 (Dunne)

Let $\underline{y} = X\underline{\beta} + \underline{\varepsilon}$ where $\underline{\varepsilon} \sim N(\underline{0}, V)$. Let $\underline{y}$ be observed, and assume a prior for $\underline{\beta}$, namely $\underline{\beta} \sim N(\underline{\beta}_0, A)$, where V, A and $\underline{\beta}_0$ are known, and the variance-covariance matrices have arbitrary rank. We may define a posterior for

$\underline{\beta}$ as the conditional distribution of $\underline{\beta}\big|_{\underline{y}}$, and write

(2.160) $\underline{\beta}\big|_{\underline{y}} \sim N(\underline{\beta}_0+AX'(V+XAX')^{g_1}(\underline{y}-X\underline{\beta}_0),\ A-AX'(V+XAX')^{g_1}XA)$.

Proof: It is reasonable to assume $\underline{\beta}$ and $\underline{\varepsilon}$ are uncorrelated, and thus

(2.161) $\begin{bmatrix}\underline{\beta}\\ \underline{\varepsilon}\end{bmatrix} \sim N\left(\begin{bmatrix}\underline{\beta}_0\\ \underline{0}\end{bmatrix}, \begin{bmatrix}A & 0\\ 0 & V\end{bmatrix}\right)$.

In Corollary 2.16.1 take $R = [X : I]$, and write

(2.162) $W = V + XAX'$,

so that we have

(2.163) $\begin{bmatrix}\underline{\beta}\\ \underline{\varepsilon}\end{bmatrix}\Bigg|_{\underline{y}} \sim N(\underline{\mu}_c, V_c)$, where

(2.164) $\underline{\mu}_c = \begin{bmatrix}\underline{\beta}_0+AX'W^{g_1}(\underline{y}-X\underline{\beta}_0)\\ VW^{g_1}(\underline{y}-X\underline{\beta}_0)\end{bmatrix}$, and

(2.165) $V_c = \begin{bmatrix}A-AX'W^{g_1}XA & -AX'W^{g_1}V\\ -VW^{g_1}XA & V-VW^{g_1}V\end{bmatrix}$.

Taking the marginal distribution for $\underline{\beta}\big|_{\underline{y}}$ in (2.163) gives the result. □

<u>Corollary 2.17.1</u> (Dunne)

The predictive distribution for the expected value of the observations $\tilde{\underline{y}} = X\underline{\beta}\big|_{\underline{y}}$ is given by

(2.166) $\tilde{\underline{y}} \sim N(\tilde{\underline{\mu}}, \tilde{V})$, where

$$(2.167)\quad \tilde{\underline{\mu}} = X\underline{\beta}_0 + XAX'W^{g_1}(\underline{y}-X\underline{\beta}_0)$$

$$= WW^{g_1}\underline{y} - VW^{g_1}(\underline{y}-X\underline{\beta}_0)$$

$$= XAX'W^{g_1}\underline{y} - VW^{g_1}X\underline{\beta}_0 \qquad \text{and}$$

$$(2.168)\quad \tilde{V} = XAX' - XAX'W^{g_1}XAX'$$

$$= (W-V) - (W-V)W^{g_1}(W-V)$$

$$= V - VW^{g_1}V \quad . \qquad \square$$

In Theorem 2.17 and its corollary all g_1-inverses are arbitrary, though all the expressions involved are uniquely determined. Rao (1971) has a similar result.

2.3 CONDITIONS FOR CHI-SQUAREDNESS AND INDEPENDENCE

The history of the derivation of necessary and sufficient conditions for quadratic forms to be

(i) chi-squared in distribution,

and (ii) independently distributed,

has been surveyed by Scarowsky (1973) and Rayner (1974). The most general form of these results was given by Khatri (1963), and will be presented here, with a general outline of what are largely his methods of proof. The introduction of matrix notation in the approach to this problem is due to Cochran (1934) in a seminal paper on the distribution of quadratic forms, who also first utilized the moment generating function (m.g.f.) of a quadratic form in normal variates. It transpired that the development was subject to a number of errors of varying degrees of severity, but there will be no analysis of those pitfalls in this thesis.

Theorem 2.18 (Khatri, 1962, 1963)

If $\underline{x}$ is $N(\underline{\mu}, V)$, a set of necessary and sufficient conditions for $\underline{x}'Q\underline{x}$ to follow a non-central $\chi^2_r(\lambda)$ distribution is that

(2.169) $VQVQV = VQV$

(2.170) $VQVQ\underline{\mu} = VQ\underline{\mu}$

(2.171) $\underline{\mu}'QVQ\underline{\mu} = \underline{\mu}'Q\underline{\mu}$

in which case the degrees of freedom and the non-centrality parameter are respectively given by

(2.172) $r = r(VQV) = tr(QV)$, and

(2.173) $\lambda = \underline{\mu}'Q\underline{\mu}$.

Proof: Without loss of generality, take Q as symmetric.

The m.g.f. of the quadratic form $q = \underline{x}'Q\underline{x}$ is given by

(2.174) $M(q,t) = E(\exp t\,\underline{x}'Q\underline{x})$

$= E(\exp t(\underline{\mu}+K\underline{y})'Q(\underline{\mu}+K\underline{y}))$,

with probability one, where $V = KK'$ and $\underline{y}$ is $N(\underline{0}, I)$, by Pringle's theorem (Pringle and Rayner, 1971, p.76). After an appropriate completion of the square and the application of

(2.175) $\int_{-\infty}^{\infty} \exp -\tfrac{1}{2}(\underline{w}-\underline{\beta})'S^{-1}(\underline{w}-\underline{\beta})d\underline{w} = (2\pi)^{r/2}|S|^{\frac{1}{2}}$

we may obtain

(2.176) $M(q,t) = |I-2tK'QK|^{-\frac{1}{2}}\, e(t)$, where

(2.177) $e(t) = \exp\{t\underline{\mu}'[I+2tQK(I-2tK'QK)^{-1}K']Q\underline{\mu}\}$.

From (2.59) and the theory of determinants these reduce to

(2.178) $M(q,t) = |I-2tQV|^{-\frac{1}{2}}\exp\{t\underline{\mu}'(I-2tQV)^{-1}Q\underline{\mu}\}$.

The m.g.f. of a noncentral $\chi^2_r(\lambda)$ variate, z say, is

(2.179) $M(z,t) = (1-2t)^{-r/2}.\exp(\lambda t(1-2t)^{-1})$.

The equality of these m.g.f.'s in the variable t implies equality in corresponding pairs of rational analytic functions, and by expanding these functions as infinite series in t, the necessary and sufficient conditions for the equality of coefficients for all powers of t transpire to be precisely (2.169) through to (2.173). □

Theorem 2.19 (Khatri, 1962, 1963; Rayner 1963)

The necessary and sufficient conditions for two quadratic forms $q_i = \underline{x}'Q_i\underline{x}$ $(i = 1,2)$ in normal variables $\underline{x} \sim N(\underline{\mu}, V)$, to be independently distributed are that

(2.180) $VQ_1VQ_2V = 0$

(2.181) $VQ_1VQ_2\underline{\mu} = \underline{0}$

(2.182) $VQ_2VQ_1\underline{\mu} = \underline{0}$

(2.183) $\underline{\mu}'Q_1VQ_2\underline{\mu} = 0$.

Equivalently, due to Scarowsky (1973)

(2.184) $\begin{bmatrix} V \\ \underline{\mu}' \end{bmatrix} Q_1VQ_2[V\ \underline{\mu}] = 0$.

Proof: We examine the conditions under which the joint m.g.f. $M(q_1,q_2;\ t_1,t_2)$ of q_1 and q_2 factorizes as the product of the marginal m.g.f.'s. Writing

$$(2.185)\quad Q^* = t_1Q_1 + t_2Q_2 = Q_1^* + Q_2^*$$

we have

$$(2.186)\quad M(q_1,q_2;\ t_1,t_2) = |I-2Q^*V|^{-\frac{1}{2}}\exp\{\underline{\mu}'(I-2Q^*V)^{-1}Q^*\underline{\mu}\}$$

and require the expression to factorise as

$$(2.187)\quad M(q_1,q_2;\ t_1,t_2) = \Pi\ |I-2Q_i^*V|^{-\frac{1}{2}}\exp\{\underline{\mu}'(I-2Q_i^*V)^{-1}Q_i^*\underline{\mu}\}$$

again equating the corresponding rational analytical functions and expanding the exponent terms as infinite series in t_1 and t_2, the mixed terms $t_1^i t_2^j$ fall away if and only if (2.180) and (2.183) hold with

$$(2.188)\quad \underline{\mu}'Q_2VQ_1VQ_1VQ_2\underline{\mu} = 0 \quad , \qquad \text{and}$$

$$(2.189)\quad \underline{\mu}'Q_1VQ_2VQ_2VQ_1\underline{\mu} = 0 \quad ,$$

which are equivalent to (2.181) and (2.182) respectively. □

It should be noted that both theorems hold regardless of the choice of variance-covariance structure V, and the symmetry or otherwise of matrices Q, Q^*, Q_i and Q_i^* in the quadratic forms. In particular Theorem 2.19 does not require the forms to follow $\chi_r^2(\lambda)$ distributions.

A plethora of special case results can be tabulated, but we will not do so here and will note simplifying conditions when these results are applied.

Mitra (1968) has investigated the solutions Q to equations of the form

(2.190) $VQVQV = VQV$

for given V. By taking Q as hermitian and V as positive semi-definite, some equivalent conditions to those of Theorem 2.18 are noted, and interesting special cases tabulated. Mitra notes that under the condition (2.190), V is easily verified to be a g_1-inverse of $QVQVQ$. In fact it is then, trivially, a g_1-inverse of QVQ.

However Q is also easily verified as a g_1-inverse of VQV, and the condition (2.180) for independence, writing

(2.191) $Q = Q_1 + Q_2$

for positive semi-definite Q, amounts to

(2.192) $VQ_1V(Q-Q_1)V = 0$.

If $\underline{x}'Q_1\underline{x}$ is $\chi^2_r(\lambda)$ in distribution, independence requires that Q be a g_1-inverse of VQ_1V as well as of VQV, and in that case it is a g_1-inverse of VQ_2V. These remarks augment those of Mitra in discussing the conditions, but do not appear to have any advantages over the conventional methods of verifying the properties.

All these formulations suffer from the involuted nature of any attempt to describe a matrix in terms of g_1-inverses of matrix products that contain the matrix in question.

Two important corollaries to Theorems 2.18 and 2.19 determine the conditions under which quadratic polynomials

(2.193) $\underline{x}'Q\underline{x} + \underline{m}'\underline{x} + d$, and bilinear forms

(2.194) $\underline{x}_1'A\underline{x}_2$

follow $\chi_r^2(\lambda)$ distributions, and the conditions under which pairs of such forms are independently distributed.

Corollary 2.18.1 (Khatri, 1963)

If $\underline{x}$ is $N(\underline{\mu}, V)$, then necessary and sufficient conditions to the quadratic polynomial $\underline{x}'Q\underline{x} + \underline{m}'\underline{x} + d$ to have $\chi_r^2(\lambda)$ distribution are

(2.195) $VQVQV = VQV$

(2.196) $V(Q\underline{\mu}+\tfrac{1}{2}\underline{m}) = VQV(Q\underline{\mu}+\tfrac{1}{2}\underline{m})$

(2.197) $\underline{\mu}'Q\underline{\mu} + \underline{m}'\underline{\mu} + d = (Q\underline{\mu}+\tfrac{1}{2}\underline{m})'V(Q\underline{\mu}+\tfrac{1}{2}\underline{m})$

and in that case

(2.198) $r = r(VQV) = tr(QV)$, and

(2.199) $\lambda = \underline{\mu}'Q\underline{\mu} + \underline{m}'\underline{\mu} + d$.

Proof: Follows directly from the theorem by using the notational device

(2.200) $$\underline{x}_*'Q_*\underline{x}_* = (\underline{x}':1)\begin{bmatrix} Q & \tfrac{1}{2}\underline{m} \\ \tfrac{1}{2}\underline{m}' & d \end{bmatrix}\begin{pmatrix} \underline{x} \\ 1 \end{pmatrix}$$
$$= \underline{x}'Q\underline{x} + \underline{m}'\underline{x} + d$$

where $\underline{x}_*$ is $N(\underline{\mu}_*, V_*)$ for

(2.201) $$\underline{\mu}_* = \begin{pmatrix} \underline{\mu} \\ 1 \end{pmatrix} ,$$ and

(2.202) $$V_* = \begin{bmatrix} V & \underline{0} \\ \underline{0}' & 0 \end{bmatrix} .$$

Note further that writing

$$(2.203)\quad \underline{x} = \begin{pmatrix} \underline{x}_1 \\ \underline{x}_2 \end{pmatrix}$$

we obtain the conditions for the bilinear form

$$(2.204)\quad \underline{x}_1 A \underline{x}_2 = \underline{x}' \begin{bmatrix} 0 & \frac{1}{2}A \\ \frac{1}{2}A' & 0 \end{bmatrix} \underline{x}$$

in terms of the quadratic polynomial.-

Corollary 2.19.1

The necessary and sufficient conditions for the variables $L\underline{x}$ and $\underline{x}'Q\underline{x} + \underline{m}'\underline{x} + d$ to be independently distributed are

$$(2.205)\quad LVQV = 0$$

$$(2.206)\quad LVQ\underline{\mu} + \tfrac{1}{2}LV\underline{m} = \underline{0} \quad ,$$

where $\underline{x}$ is $N(\underline{\mu}, V)$.

Proof: The result is a simple extension of the conditions of independence, to the quadratic forms $\underline{x}_*'Q_*\underline{x}_*$ as in (2.200), and the individual rows $\underline{\ell}_i'\underline{x}$ of $L\underline{x}$, written as

$$(2.207)\quad \underline{\ell}_i'\underline{x} = \underline{x}_*' \begin{bmatrix} 0 & \frac{1}{2}\underline{\ell}_i \\ \frac{1}{2}\underline{\ell}_i' & 0 \end{bmatrix} \underline{x}_* \quad .$$

Two of the four conditions (2.180) through to (2.183) are void in this special case. □

It is clearly possible using the $\underline{x}_*$ notation to extend Corollary 2.19.1 to the case of two quadratic polynomials $\underline{x}_*'Q_i\underline{x}_*$ $(i = 1,2)$.

2.4 RESULTS ON EXTENDED PARTITIONED MATRICES

This section provides a proof for a conjecture made by Linhart and Zucchini (1981) in private correspondence with the author. The result is of interest in the discussion of covariance analysis.

Theorem 2.20 (Dunne)

Let A_1, A_2, A_3 by $(n\times p_1)$, $(n\times p_2)$ and $(n\times p_3)$ matrices of full column-ranks p_1, p_2 and p_3 respectively. Let

$$(2.208)\quad C = \begin{bmatrix} A_1' \\ A_2' \\ A_3' \end{bmatrix} [A_1 A_2 A_3] = \begin{bmatrix} A_1'A_1 & A_1'A_2 & A_1'A_3 \\ A_2'A_1 & A_2'A_2 & A_2'A_3 \\ A_3'A_1 & A_3'A_2 & A_3'A_3 \end{bmatrix}$$

$$= \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{bmatrix}$$

and suppose C is non-singular, so that we may partition C^{-1} as

$$(2.209)\quad C^{-1} = \begin{bmatrix} C^{11} & C^{12} & C^{13} \\ C^{21} & C^{22} & C^{23} \\ C^{31} & C^{32} & C^{33} \end{bmatrix}$$

Let V_1, V_2 and V_3 be defined by

$$(2.210)\quad V_i = \begin{bmatrix} C_{jj} & C_{kj} \\ C_{jk} & C_{kk} \end{bmatrix} \quad \text{for} \quad \{i,j,k\} = \{1,2,3;\ i\neq j \lesssim k \neq i\}$$

with corresponding inverses partitioned as

(2.211) $$V_i^{-1} = \begin{bmatrix} V_i^{jj} & V_i^{jk} \\ V_i^{kj} & V_i^{kk} \end{bmatrix} = \begin{bmatrix} U_i \\ W_i \end{bmatrix} \quad .$$

Finally let

(2.212) $$C_i = [C_{ij} \quad C_{ik}] \quad .$$

Then we have

(2.213) $$\begin{bmatrix} U_i C_i' C^{ii} C_i U_i' & C_i' C^{ii} C_i \\ W_i C_i' C^{ii} C_i U_i' & W_i C_i' C^{ii} C_i W_i' \end{bmatrix} = \begin{bmatrix} C^{jj} & C^{jk} \\ C^{kj} & C^{kk} \end{bmatrix} - \begin{bmatrix} V_i^{jj} & V_i^{jk} \\ V_i^{kj} & V_i^{kk} \end{bmatrix}$$

Proof: By Theorems 2.12 and 2.13, and the fact that

(2.214) $$CC^{-1} = \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} = C^{-1}C$$

we have, for example, that

(2.215) $$C^{11}C_{13} + C^{12}C_{23} + C^{13}C_{33} = 0 \quad ,$$ and

(2.216) $$C^{31}C_{13} + C^{32}C_{23} + C^{33}C_{33} = I$$

and consequently

(2.217) $$-C^{13}C_{33} = C^{11}C_{13} + C^{12}C_{23} \quad ,$$ and

(2.218) $$I - C^{33}C_{33} = C^{31}C_{13} + C^{32}C_{21} \quad .$$

Similarly the Schur Identity

(2.219) $$B_{11}^{-1} = A_{11} - A_{12}A_{22}^{-1}A_{21} \quad ,$$ in

$$(2.220)\quad \begin{bmatrix} A_{11}A_{12} \\ A_{21}A_{22} \end{bmatrix}\begin{bmatrix} B_{11}B_{12} \\ B_{21}B_{22} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$$

yields

$$(2.221)\quad V_3^{-1} = \begin{bmatrix} C^{11}C^{12} \\ C^{21}C^{22} \end{bmatrix} - \begin{bmatrix} C^{13} \\ C^{23} \end{bmatrix}(C^{33})^{-1}[C^{31}C^{32}] \quad .$$

But from (2.215) and (2.216) we have

$$(2.222)\quad V_3^{-1}\begin{bmatrix} C_{13} \\ C_{23} \end{bmatrix} = \begin{bmatrix} -C^{13}C_{33} \\ -C^{23}C_{33} \end{bmatrix} - \begin{bmatrix} C^{13} \\ C^{23} \end{bmatrix}(C^{33})^{-1}(I-C^{33}C_{33})$$

$$= -\begin{bmatrix} C^{13} \\ C^{23} \end{bmatrix}(C^{33})^{-1}$$

and in turn, using symmetry of C and (2.212)

$$(2.223)\quad V_3^{-1}C_3'C^{33}C_3V_3^{-1} = +\begin{bmatrix} C^{13} \\ C^{23} \end{bmatrix}(C^{33})^{-1}[C^{31}C^{32}] \quad .$$

Now (2.221) implies

$$(2.224)\quad V_3^{-1}C_3'C^{33}C_3V_3^{-1} = \begin{bmatrix} C^{11}C^{12} \\ C^{21}C^{22} \end{bmatrix} - \begin{bmatrix} V_3^{11}V_3^{12} \\ V_3^{21}V_3^{22} \end{bmatrix}$$

and the left hand side may be written as

$$(2.225)\quad V_3^{-1}C_3'C^{33}C_3V_3^{-1} = \begin{bmatrix} U_3 \\ W_3 \end{bmatrix}C_3'C^{33}C_3[U_3'W_3']$$

by (2.211) and (2.213). This proves one case of the result, and the other cases follow similarly. □

<u>Corollary 2.20.1</u> (Linhart-Zucchini conjecture)

Under the stated conditions

(2.226) $\quad tr(C^{11}-V_3^{11}) = tr(C_3U_3'U_3C_3'C^{33})$

Proof: From the theorem, the equality of the leading submatrices in (2.225) implies

(2.227) $\quad C^{11}-V_3^{11} = U_3C_3'C^{33}C_3U_3'$

and the trace result follows directly since

(2.228) $\quad tr(AB) = tr(BA) \quad .$ □

It is however not the full rank of C which is the crux of the proof. The result depends crucially on the assumption that the column-spaces $\mathcal{C}(A_1)$, $\mathcal{C}(A_2)$ and $\mathcal{C}(A_3)$ are virtually disjoint. Consequently the theorem may be generalized. We will need

Lemma 2.21 (Dunne)

If the matrix

(2.229) $\quad X = [X_1 : X_2]$

has $\mathcal{C}(X_1)$ and $\mathcal{C}(X_2)$ virtually disjoint, i.e.

(2.230) $\quad r(X) = r(X_1) + r(X_2)$

then for $S = X'X$ we have

(2.231) $\quad SS^g = \begin{bmatrix} S_{11}S_{11}^g & 0 \\ 0 & S_{22}S_{22}^g \end{bmatrix} \quad ,$ where

(2.232) $\quad S_{ii}S_{ii}^g = (X_i'X_i)(X_i'X_i)^g \quad .$

Proof: From Theorems 2.12 and 2.13, we have, by multiplication and collection of terms

$$(2.233)\quad SS^g = \begin{bmatrix} QQ^g & 0 \\ 0 & S_{22}S_{22}^g \end{bmatrix}$$

where $Q = S_{11}-S_{12}S_{22}^gS_{21}$. Under the conditions of the theorem

$$(2.234)\quad r(Q) = r(X_1)$$

and QQ^g is the unique symmetric projection onto $C(X_1)$. Thus

$$(2.235)\quad QQ^g = S_{11}S_{11}^g$$

and the lemma follows. □

<u>Theorem 2.22</u> (Dunne)

Let A_1, A_2, A_3 be $n\times p_i$ matrices of rank p_i for $i = 1,2,3$. Let

$$(2.236)\quad r[A_1 : A_2 : A_3] = p_1 + p_2 + p_3 \quad .$$

Let C be as in (2.192) ,

$$C = [A_1 : A_2 : A_3]'[A_1 : A_2 : A_3]$$

and let B the g-inverse of C be conformably partitioned as

$$(2.237)\quad C^g = B = \begin{bmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & B_{23} \\ B_{31} & B_{32} & B_{33} \end{bmatrix}$$

Let V_1 V_2 V_3 be defined as in (2.21) with their corresponding generalized inverses partitioned as in (2.211). Taking C_i as in (2.212), we have

$$(2.238)\quad \begin{bmatrix} U_iC_i'B_{ii}C_iU_i' & U_iC_i'B_{ii}C_iW_i' \\ W_iC_i'B_{ii}C_iW_i' & W_iC_i'B_{ii}C_iW_i' \end{bmatrix}$$

$$= \begin{bmatrix} B_{jj}B_{jk} \\ B_{kj}B_{kk} \end{bmatrix} - \begin{bmatrix} V_i^{jj}V_i^{jk} \\ V_i^{kj}V_i^{kk} \end{bmatrix}$$

for i,j,k as before.

Proof: By Lemma 2.21, applied twice, we have

$$(2.239)\quad CC^g = \begin{bmatrix} S_{11}S_{11}^g & 0 & 0 \\ 0 & S_{22}S_{22}^g & 0 \\ 0 & 0 & S_{33}S_{33}^g \end{bmatrix} = C^gC$$

Thus, for example, (2.215) and (2.217) hold as before, with the notation changes of (2.237). However (2.218) is replaced by

$$(2.240)\quad S_{33}^gS_{33} - B_{33}C_{33} = B_{31}C_{13} + B_{32}C_{23} \quad .$$

The generalized Schur identity yields

$$(2.241)\quad B_{11}^g = A_{11} - A_{12}A_{22}^gA_{21} \quad , \qquad \text{and}$$

$$(2.242)\quad V_3^g = \begin{bmatrix} B_{11}B_{12} \\ B_{21}B_{22} \end{bmatrix} - \begin{bmatrix} B_{13} \\ B_{23} \end{bmatrix}(B_{33}^g)[B_{31}B_{32}] \; ,$$

and following the previous sequence of proof

$$(2.243)\quad V_3^g\begin{bmatrix} C_{13} \\ C_{23} \end{bmatrix} = \begin{bmatrix} -B_{13}C_{33} \\ B_{23}C_{33} \end{bmatrix} - \begin{bmatrix} B_{13} \\ B_{23} \end{bmatrix}(B_{33})^g(S_{33}^gS_{33}-B_{33}C_{33})$$

$$= -\begin{bmatrix} B_{13} \\ B_{23} \end{bmatrix}B_{33}^g$$

since, the Schur identity implies

$$(2.244) \quad R(B^g_{33}) = R(S_{33})$$

under the given rank conditions. Thus

$$(2.245) \quad V^g_3 \begin{bmatrix} C_{13} \\ C_{23} \end{bmatrix} B_{33} [C_{31} C_{32}] V^g_3$$

$$= \begin{bmatrix} B_{13} \\ B_{23} \end{bmatrix} B^g_{33} [B_{31} B_{32}]$$

$$= \begin{bmatrix} B_{11} B_{12} \\ B_{21} B_{22} \end{bmatrix} - V^g_3$$

from (2.242). The result follows after partitioning, and the proofs of the other two cases are analogous. □

A similar result to Corollary 2.20.1 is easily shown, and thus extends the Linhart-Zucchini conjecture to models where the X matrix can be partitioned into sets of elements from virtually disjoint subspaces.

CHAPTER 3

ESTIMATION IN THE LINEAR MODEL

As indicated in Chapter 1, the known quantities in the model (1.1) are $\underline{y}$ and X. In general, the relations

$$\underline{y} = X\underline{\beta} \quad , \tag{3.1}$$

called the observational equations (OE's), are inconsistent because of the intrusion of the error terms if the model is valid, or because of the gratuitous imposition of the column-space $C(X)$ as a restriction on the variate $\underline{y}$ by (3.1) when the model is invalid.

Generalizing the least squares ideas of Gauss (1816, 1821, 1823, 1826), Aitken (1934, 1935) was led to suggest, *inter alia,* that for the case of variance-covariance structure $V = \sigma^2 I$, $\underline{\beta}$ in (3.1) be assigned as an ordinary least squares (OLS) solution $\underline{b}$, i.e. $\underline{b}$ satisfying

$$\frac{\partial}{\partial \underline{\beta}} (\underline{y}-X\underline{\beta})'(\underline{y}-X\underline{\beta})\Big]_{\underline{\beta}=\underline{b}} = 0 \tag{3.2}$$

so that a minimum is achieved for the sum of squares

$$(\underline{y}-X\underline{\beta})'(\underline{y}-X\underline{\beta}) \quad . \tag{3.3}$$

The set of such $\underline{b}$ is precisely that with

$$(X'X)\underline{b} = X'\underline{y} \quad . \tag{3.4}$$

The equations (3.4) are called the normal equations (NE's) and are conventionally written in the alternative form

(3.5) $\quad S\underline{b} = \underline{g}$

when $S = X'X$ and $\underline{g}$ is the right-hand side of (3.4).

The NE's are consistent in every case, because $C(S) = C(X')$, so that solutions $\underline{b}$ always exist. If X has full column-rank k, then $\underline{b}$ is unique, and may be described as an estimate of $\underline{\beta}$. However if $r(X) = r < k$, then an entire class of solutions may be described by

(3.6) $\quad \underline{b} = S^{g_1}X'\underline{y} + (I-S^{g_1}S)\underline{c}$

where $\underline{c}$ is an arbitrary vector in the k-dimensional Euclidean space R^k, and S^{g_1} is any g_1-inverse of S. There is therefore no linear function of the observed variate-values $\underline{y}$ available in this approach which would assign a unique value to $\underline{b}$, and thus $\underline{b}$ cannot be regarded as an estimate, in the modern sense.

The additional assumption in (1.1) of multivariate normality for $\underline{\varepsilon}$ allows an attempt at a maximum likelihood estimate (MLE) for $\underline{\beta}$. In the $V = \sigma^2 I$ case, we obtain the same indeterminate NE's, and again an estimate $\underline{b}$ is obtainable if and only if $r(X) = r = k$. A wider class of maximum likelihood estimation procedures, through generalized least squares (GLS) methods, and the assumption of a wider class of variance-covariance structures will be discussed in a later chapter. As far as the $V = \sigma^2 I$ case is concerned, Cochran, in private correspondence with Rayner, states that he is quite sure R.A. Fisher knew of the least squares (OLS) method of solving for $\underline{\beta}$, and its relationship to MLE, as early as 1934, and that no one else (knew the relationship). It is however reasonable to infer that the problem of the non-uniqueness of $\underline{b}$ did not bother early practitioners unduly, since the problem was only theoretically explained by Bose (1944). Since virtually all linear models arising from experimental designs are not of full rank, non-uniqueness of $\underline{b}$ in (3.4) might have been a serious matter. On this point Rayner notes, in private communication:

> *"It is a tribute to the genius of early statisticians that practical problems of analysis of experimental data were overcome long before the theory of the non-full rank model was properly developed, even though many often had little idea of how their methods worked."*

Before proceeding to estimation *per se*, we may consider the complete system of k-tuples of linear functions of the observations given by $G\underline{y}$, where G is $(k\times n)$, of full row-rank, and such that premultiplying on (3.1),

$$G\underline{y} = GX\underline{\beta} \tag{3.7}$$

is now consistent. If $\underline{\beta}$ in (3.7) were to have a unique solution $\underline{b}$, a necessary and sufficient condition is the existence of G satisfying the (left-inverse) property

$$GX = I_k \tag{3.8}$$

However necessary conditions for (3.8) are that X have full column-rank k and that

(3.9) $G = (X'X)^{-1}X'$.

This yields nothing other than a special case of (3.6) for $\underline{b}$. But if $r(X) = r < k$, the idea of obtaining any linear function of the observed variate-values $\underline{y}$, which assigns $\underline{\beta}$ a unique value over all admissible G in (3.7), leads nowhere, and thus a more general approach than the NE's also fails. For the uninitiated one would nevertheless at this point be able, as it were, to offer an *ex post facto* glimmer of hope. On the one hand if G in (3.7) can be taken $(r \times n)$ of full row-rank, such that

(3.10) $GX = [I_r : A]$

then a conformable sub-vector of $\underline{\beta}$ may be artificially taken as zero, and the remainder of $\underline{\beta}$ assigned the unique value $G\underline{y}$, based on the observations. If GX does not satisfy (3.10) but requires only column permutations to achieve the form, the same result can be applied, and it would not be difficult to generalize the approach by admitting non-singular matrices in place of I_r. Secondly, from the nature of (3.6) it is clear that $X\underline{b}$ is unique, over all g_1-inverses S, and all vectors $\underline{c}$, by (2.12) and (2.13). Thus $\underline{\theta}'\underline{b}$ is unique if $\underline{\theta}$ is in the column-space $C(X')$, and there exists a whole space of linear functions of $\underline{\beta}$ which may be assigned a unique value from the NE's *per se*. Essentially we have a "black-box" mechanism, whose properties are extensively specified in the literature, and discussed in what follows.

3.1 ESTIMABILITY AND UNBIASEDNESS

Under the special case of (1.1) assuming (1.4), we may define the "estimable functions" $\underline{\theta}'\underline{\beta}$ to be precisely those for which $\underline{\theta}'\underline{b}$ is unique over all solutions $\underline{b}$ in the NE's (3.5). The definition is due to Bose (1944).

In view of the introductory remarks it is apparent that the following properties are equivalent characterizations of estimable functions

(3.11) $\underline{\theta}' = \underline{a}'X$, for some $\underline{a}$;

(3.12) $\underline{\theta}' = \underline{\theta}'X^{g_1}X$; or

(3.13) $\underline{\theta}' = \underline{\theta}'S^{g_1}S$.

These relations of Pringle and Rayner (1971) imply that the maximum number of linearly independent estimable functions is precisely r, the dimension of the space of all such $\underline{\theta}$, so that it is only in full-rank model with $r(X) = k$ that estimability is a property of all linear functions of $\underline{\beta}$, and hence of $\underline{\beta}$ itself.

Nonetheless $X\underline{\beta}$ is always estimable, and the "fitted values" $\hat{y}$, obtained as

(3.14) $\hat{\underline{y}} = X\underline{b}$

are unique. We note that if $\underline{\gamma}'$ is not in the row-space $R(X)$ of X, by (3.11) $\underline{\gamma}'\underline{\beta}$ is not estimable, i.e. $\underline{\gamma}'\underline{b}$ is not unique over all $\underline{b}$ from (3.6).

The criteria (3.11) to (3.13) may be applied to a set of estimable functions $T\underline{\beta}$, yielding

(3.15) $T = AX$, for some A ;

(3.16) $T = TX^{g_1}X$; or

(3.17) $T = TS^{g_1}S$.

None of these criteria, nor the one-dimensional forms (3.12) to (3.14) are particularly suitable for computer applications. When estimability needs to be verified, these criteria involve checking elements of T against, for example, elements of $TS^{g_1}S$ while attempting at the same time to allow for computer rounding error between corresponding terms. Such error arises even in experimental design models when rationals require infinite decimal expansion, and the available precision influences the entries of S^{g_1} and/or $TS^{g_1}S$.

Milliken (1971) was therefore led to suggest a criterion for estimability which involved only an integer, at least in so far as theoretical examination apparently indicates. He shows that a necessary and sufficient condition for estimability of $T\underline{\beta}$, when T is $q \times k$ of rank $s \leq r \leq k$ is given by

(3.18) $r[(X(I-T^gT)] = r-s$

or equivalently

(3.19) $tr\{X(I-T^gT)[X(I-T^gT)]^g\} = r-s$.

Since T and $I-T^{g_1}T$ have virtually disjoint row-spaces, and span R^k, the g-inverses in (3.18) and (3.19) can be replaced by arbitrary g_1-inverses.

Even that modification does not save Milliken's criteria, which has been critized by Rayner (1977). In the first instance it is not clear that the rank of R will be known (except for T with few rows), though known X-rank is plausible in a wide class of experimental designs. Further there is no guarantee that T^{g_1} will not also be subject to rounding error as in the calculation of g_1-inverses for X and S. Even if the effect of the trace operator in (3.19) implies some type of cumulative cancellation of rounding errors, it involves the "inversion" of two matrices, one $q \times k$ and the other $n \times k$, where the latter at least is larger than S. Presumably there is more latitude for rounding error to influence (3.19) than in

(3.20) $$tr(S(I-T^{g_1}T)[S(I-T^{g_1}T)]^{g_1}) = r-s \quad .$$

In any event, regardless of ranks,

(3.21) $$\begin{aligned} tr[(I-S^{g_1}S)'T'T(I-S^{g_1}S)] &= tr[T(I-S^{g_1}S)(I-S^{g_1}S)'T'] \\ &= tr[T(I-S^{g}S)T'] \\ &= 0 \end{aligned}$$

involves simply the sum of the squares of the elements in the matrix $T(I-S^{g_1}S)$, and is easily programmed as an estimability criterion. Finally, it is likely that a row by row examination of $T(I-S^{g_1}S)$ is the best approach in that non-estimable functions will be identified individually.

Golub and Styan (1973) and Rayner (1977) have suggested the criterion

(3.22) $$r[T' : X'] = r(X') \quad .$$

By the methods of Golub and Styan (3.22) would involve duplication of calculations, and is less efficient than (3.21) computationally.

We turn to unbiasedness and its relation to estimability. An estimation $\hat{\underline{\beta}} = G\underline{y} + \underline{d}$ is said to be a (linear) unbiased estimator of $\underline{\beta}$ if and only if

$$(3.23) \quad E(\hat{\underline{\beta}}) = \underline{\beta} \quad .$$

From the preceding remarks it is evident that if

$$(3.23) \quad E(\hat{\underline{\beta}}) = GX\underline{\beta} + \underline{d} = \underline{\beta}$$

is to hold over the entire parameter space R^k, of $\underline{\beta}$, then

$$(3.24) \quad \underline{d} = \underline{0} \quad , \qquad \text{and}$$

$$(3.25) \quad GX = I_k \quad .$$

Thus an unbiased estimator of $\underline{\beta}$ exists if and only if X has full column-rank k. We may however define an estimator $\underline{\theta}'\hat{\underline{\beta}}$ of $\underline{\theta}'\underline{\beta}$ to be unbiased if and only if, regardless of $\underline{\beta}$,

$$(3.26) \quad E(\underline{\theta}'\hat{\underline{\beta}}) = \underline{\theta}'\underline{\beta} \quad .$$

Equivalently,

$$(3.27) \quad E(\underline{\theta}'\hat{\underline{\beta}}) = \underline{\theta}'GX\underline{\beta} + \underline{\theta}'\underline{d} = \underline{\theta}'\underline{\beta}$$

over the parameter space, and the condition is additive over $\underline{\theta}'$. Setting $\underline{\beta}$ equal to zero, implies

$$(3.28) \quad \underline{\theta}'\underline{d} = 0 \quad , \quad \text{for all admissible } \underline{\theta} \ ,$$

and additivity over $\underline{\beta}$. Now take $\underline{\beta}$ over the parameter space so as to describe the admissible $\underline{\theta}$ by

$$(3.29) \quad \underline{\theta}'GX = \underline{\theta}' \quad .$$

This means that an unbiased estimator of $\underline{\theta}'\underline{\beta}$ may only be defined if (3.11) holds. The estimability of $\underline{\theta}'\underline{\beta}$ and the existence of an unbiased estimat of $\underline{\theta}'\underline{\beta}$ are thus equivalent.

The definition of an unbiased estimator $T\underline{\beta}$ is a simple extension, and the equivalence of its existence to estimability of $T\underline{\beta}$ is well-known.

Pringle and Rayner (1971) point out that the entire class of linear estimators $\hat{\underline{\beta}} = G\underline{y} + \underline{d}$, for which $\underline{\theta}'\hat{\underline{\beta}}$ is an unbiased estimator of $\underline{\theta}'\underline{\beta}$ over all $\underline{\theta}'$ in the row-space $R(X)$, is obtained by taking $X\underline{d} = \underline{0}$ and G as any X^{g_1}. A sub-class of such estimators is given by taking the $\underline{b}$ of (3.6), in which G is any $X^{g_{13}}$. The question arises as to whether or not this sub-class acquires any further specific properties consequent on the restricted choices of G, and if so, what those properties imply about the non-empty set of linear unbiased estimators $\underline{\theta}'\underline{b}$ of the estimable function $\underline{\theta}'\underline{\beta}$.

However even the full-rank case of the normal equations (3.4), solved theoretically as

(3.30) $\quad \underline{b} = (X'X)^{-1}X'\underline{y}$

may not in practice prove computationally stable. Golub and Styan (1973) used the term *ill-conditioned* to describe a full column-rank matrix X, such that a "small" change in X can induce a correspondingly "large" change in $(X'X)^{-1}$, and thus in the solution (3.30). When these matrices arise in practice, the accumulation of round-off errors in the algorithm procedure used, constitute "small" changes in X or X'X. It is well-known that apparent solutions $\underline{b}$ to (3.11) can yield high relative error RE, obtained as

$$(3.31) \qquad RE = \|\underline{b}^* - \underline{b}\| / \|\underline{b}^*\|$$

where $\underline{b}^*$ is the exact solution, and $\|\cdot\|$ is the Euclidean norm.

The basic structure or singular value decomposition of X yields basic or singular values $sg_i(X)$ for X, as the positive square-roots of the eigenvalues of X'X and XX'. We may then write

$$(3.32) \qquad RE = \frac{\|S^{-1}.S(\underline{b}^*-\underline{b})\|}{\|S^{-1}.S\underline{b}^*\|} \leq \frac{\lambda_{max}(S^{-1}).\|S(\underline{b}^*-\underline{b})\|}{\lambda_{min}(S^{-1}).\|S\underline{b}^*\|}$$

$$\leq k^2(X).\frac{\|X'\underline{y}-X'X\underline{b}\|}{\|X'\underline{y}\|}$$

where the *condition number*

$$(3.33) \qquad k(X) = (\lambda_{max}(S)/\lambda_{min}(S))^{\frac{1}{2}}$$

is a measure of the ill-conditioning of X. Now (3.32) provides a bound for the relative error, and it is clear that if $k(X)$ is large then RE may be large. Golub and Styan aver that RE is likely to be so when $k(X)$ in (3.32) is replaced by

$$(3.34) \qquad (r_{11}/r_{qq})^2 \leq k^2(X)$$

and the new right-hand term is large. The quantities r_{11} and r_{qq} are obtained by use of Householder transformations

H on X, where H is of the form

$$(3.35) \quad H = I - 2(\underline{u}'\underline{u})^{-1}\underline{u}\underline{u}' \quad .$$

That claim appears to be incorrect, because the bound (3.32) is the product of two terms, the square of the condition number, and a measure of the accuracy (on the given matrix X) of the algorithm used. Householder transformations of X are just as ill-conditioned as X itself. If the algorithm is computationally stable, the relative error RE can be zero even when the condition number is large. In the opinion of this author, the role of the condition number is (3.32) is to reflect, in a simple way, the degree of extrapolation from

$$(3.5) \quad S\underline{b} = H\Delta H'\underline{b} = X'\underline{y} = \underline{g}$$

where Δ is $\mathrm{Diag}(\lambda_i(S))$, to the modified form

$$(3.30) \quad \frac{tr(S)}{q} \cdot I\underline{b} = \frac{tr(S)}{q} \cdot H\Delta^{-1}H'X'\underline{y} \quad .$$

Such extrapolation will exacerbate the effect of any procedure which is not stable. This interpretation will be seen to be consistent with results on variance of estimators discussed in Section 3.2.

The notion of estimability in (3.17) and (3.21) throws some light on the adequacy of a particular algorithm for solving a given set of NE's (3.4). In the remarks of Section 3.3 on residuals after fitting $\hat{\underline{\beta}}$ or $X\hat{\underline{\beta}}$ in the model (1.1), we discuss their relationship with computational accuracy and estimability.

3.2 BEST LINEAR UNBIASED ESTIMATION

Consider the entire class of unbiased estimators $\underline{\theta}'\hat{\underline{\beta}}$. We now examine well-known theory and restate the conditions for minimum variance within the class. Let

(3.36) $\underline{w}' = \underline{\theta}'G$,

then

Theorem 3.1 (Zyskind, 1967)

An unbiased estimate $\underline{w}'\underline{y} = \underline{\theta}'G\underline{y}$ of $\underline{\theta}'\underline{\beta}$ has minimum variance if and only if $\underline{w}$ is in the column-space $C(X)$.

Proof. The assumption $V = \sigma^2.I$ yields

(3.37) $\text{var}(\underline{w}'\underline{y}) = \underline{\theta}'GG'\underline{\theta}.\sigma^2 = \underline{w}'\underline{w}.\sigma^2$

which is positive semi-definite in $\underline{\theta}$. We require $\underline{w}'\underline{w}$ minimal subject to $\underline{w}'X = \underline{\theta}'GX = \underline{\theta}'$. Solving with Lagrange multiplier $\underline{\lambda}$ we obtain the requirement

(3.38) $\underline{w} = X\underline{\lambda}$,

and under this condition the minimum is

(3.39) $\text{var}(\underline{w}'\underline{y}) = \underline{\theta}'GX\underline{\lambda}.\sigma^2 = \underline{\theta}'\underline{\lambda}.\sigma^2$. □

An unbiased estimator $\underline{w}'\underline{y}$ of $\underline{\theta}'\underline{\beta}$ with minimum variance is said to be a best linear unbiased estimator (BLUE) of $\underline{\theta}'\underline{\beta}$. Since $\underline{\theta}'$ is any vector in the row-space $R(X)$, the condition may be equivalently described by

Theorem 3.2 (Pringle and Rayner, 1971)

A BLUE of $\underline{\theta}'\underline{\beta}$ is given by $\underline{\theta}'G\underline{y}$ if and only if

(3.40) $XGX = X$, and

(3.41) $XG = (XG)'$

Proof. From (3.29) we obtain (3.40). Then Theorem 3.1 implies $\underline{w} = G'\underline{\theta}$ is in the column-space of X, for all $\underline{\theta}'$ in the row-space of X, which is equivalent to

(3.42) $XX^{g_1}G'X' = G'X'$.

Taking $X^{g_1} = G$ from (3.40) we have from the symmetry of the left-hand side, and (3.41) follows from

(3.43) $XGG'X' = G'X'$. □

The conditions (3.40) and (3.41) are precisely the conditions for G to be a g_{13}-inverse of X, so that an estimable function has a BLUE $\underline{w}'\underline{y} = \underline{\theta}'G\underline{y}$, and conversely. Moreover, Pringle and Rayner show that (3.40) and (3.41) imply

(3.44) $X'XG = X'G'X' = X'$, where

(3.45) $G = (X'X)^{g_1}X' + [I-(X'X)^{g_1}X'X]Z$, for arbitrary Z,

which in turn yields

(3.46) $XG = X(X'X)^{g_1}X' = XS^gX' = XX^g$.

Thus XG is unique symmetric and idempotent regardless of the choice of S^{g_1} in (3.45) and (3.46), and the nature of (3.46) guarantees the estimability of $\underline{\theta}'\underline{\beta}$ when a BLUE $\underline{w}'\underline{y}$ of $\underline{\theta}'\underline{\beta}$ exists by

<u>Theorem 3.3</u> (Pringle and Rayner, 1971)

A BLU estimate of $\underline{\theta}'\underline{\beta}$ is uniquely given by $\underline{\theta}'\underline{b}$ where $\underline{b}$ is any solution to the normal equations, and

(3.47) $\underline{\theta}'\underline{b} = \underline{\theta}'X^g\underline{y} = \underline{\theta}'S^gX'\underline{y} = \underline{\theta}'S^g\underline{g}$.

Proof: Directly from (3.6), (3.45) and (3.46), we have the result. □

<u>Theorem 3.4</u> (Dunne)

There exists one and only one $\underline{w}$ such that $\underline{w}'\underline{y}$ is a BLU estimate of $\underline{\theta}'\underline{\beta}$.

Proof: From (3.36), and the existence of $X^{g_{13}}$, the set of such $\underline{w}$ is non-empty. By (3.45) and (3.46), $\underline{w}$ is unique and may be defined by

(3.48) $\underline{w}' = \underline{\theta}'X^g = \underline{\theta}'S^gX'$. □

The following consequence is well-known.

<u>Corollary 3.4.1</u>. The fitted values $\hat{\underline{y}} = X\hat{\underline{\beta}}$ are unique and

(3.49) $\hat{\underline{y}} = X\hat{\underline{\beta}} = X\underline{b} = XS^gX'\underline{y}$.

Proof: Vary $\underline{\theta}$ in (3.48) over $C(X')$, by taking X itself. Then (3.48) and Theorem 3.3 give the equality of $X\hat{\underline{\beta}}$ and (3.14). □

<u>Theorem 3.5</u> (Rao, 1967)

Let $\underline{\beta} = G\underline{y} + \underline{d}$; then $\hat{\underline{\beta}}$ is an OLS solution to the OE's (3.1) if and only if $X\underline{d} = \underline{0}$ and G satisfies Theorem 3.2.

Proof: The condition holds if and only if

(3.50) $S\hat{\underline{\beta}} = SG\underline{y} + S\underline{d} = X'\underline{y}$, for all $\underline{y}$;

thus $S\underline{d} = \underline{0}$ and $X\underline{d} = \underline{0}$, and

(3.51) $SG = X'$

which is equivalent to (3.45) and (3.46). □

We may therefore speak of *the* BLU estimator of an estimable function, and use $\underline{b}$, a solution to the NE's, and $\hat{\underline{\beta}}$ interchangeably in that context. The above results specify that the sub-class of estimators referred to at the end of Section 3.1, is the sub-class leading to the unique minimum variance unbiased estimator of each estimable function. Thus we may always obtain the minimum variance from

Theorem 3.6 (Pringle and Rayner, 1971)

The BLU estimate $\underline{w}'\underline{y} = \underline{\theta}'\underline{b}$ of $\underline{\theta}'\underline{\beta}$ has variance

(3.52) $\sigma^2.\underline{\theta}'S^g\underline{\theta} = \sigma^2.\underline{\theta}'S^{g_1}\underline{\theta}$,

which is unique over all g_1-inverses.

Proof: The result follows from (3.37), (3.45) and (3.46). □

Consequently, in so far as estimable functions are concerned we may proceed as though

(3.53) $\text{var}(\hat{\underline{\beta}}) = \sigma^2.S^{g_1}$,

generalizing the known relation with full-rank X, viz.

(3.54) $\text{var}\ \hat{\underline{\beta}} = \sigma^2.S^{-1}$.

From (3.52) and letting $\underline{\theta}'$ vary over the row-space $R(X)$, it will be easily seen that not only are the minimum variances of estimable functions exactly specified, but also the unique covariances of any two estimable functions, as in

(3.55) $\text{cov}(\underline{w}_1'\underline{y}, \underline{w}_2'\underline{y}) = \sigma^2\underline{w}_1'\underline{w}_2 = \sigma^2.\underline{\theta}_1'S^{g_1}\underline{\theta}_2$

over all g_1-inverses. The interchangeability of S^{g_1} for S^g and $X^{g_{13}}$ for X^g throughout the theory of BLUE's allows for a large number of possible algorithms for solving the normal equations, some of which may be derived from Section 3.4.

We note that when the matrix X is of full column-rank, but is ill-conditioned, the singular values of X or equivalently, the eigenvalues of S are not all equal and the ratio $k^2(X)$ from (3.33) is large. Even for a computationally stable algorithm, we obtain (3.54) as the variance matrix for

(3.56) $\quad \hat{\underline{\beta}} = (X'X)^{-1}X'\underline{y}$,

and so for $X'X = H\Delta H'$ as in (3.36),

(3.57) $\quad H'\hat{\underline{\beta}} = \Delta^{-1}H'X'\underline{y}$, and

(3.58) $\quad \mathrm{var}(H'\hat{\underline{\beta}}) = \sigma^2.\Delta^{-1}$

This is equivalent to having uncorrelated $\underline{h}_i'\hat{\underline{\beta}}$, with

(3.59) $\quad \mathrm{var}(\underline{h}_i'\tilde{\underline{\beta}}) = \sigma^2.\lambda_i^{-1}(S)$.

At least one of these unit-length linear functions of $\hat{\underline{\beta}}$ is subject to a large variance. This is similar to having one or more $\underline{h}_i'\hat{\underline{\beta}}$ in a region of extrapolation relative to the remaining unit-length linear functions $\underline{h}_j'\underline{\beta}$. To render estimation of the function $\underline{h}_i'\underline{\beta}$ less sensitive to extrapolation we would require one or more observations from that direction. If such were available then we augment X and $\underline{y}$ as

(3.60) $\quad X^* = \begin{bmatrix} X \\ t\underline{h}_i' \end{bmatrix}$, and

(3.61) $\quad \underline{y}^* = \begin{bmatrix} \underline{y} \\ y_{n+1} \end{bmatrix} = X^*\underline{\beta} + \underline{\varepsilon}^*$

respectively. Now the normal equations become

(3.62) $\quad (X^{*\prime}X^*)\hat{\underline{\beta}} = X^{*\prime}\underline{y}^*$

with solution, using Kronecker ∂_{ij},

(3.63) $\quad \hat{\underline{\beta}} = H.\mathrm{Diag}(\lambda_j+\partial_{ij}t^2)^{-1}.H'X^{*\prime}\underline{y}$.

Now the revised estimate of $H'\underline{\beta}$ is

(3.64) $\quad H'\hat{\underline{\beta}} = \mathrm{Diag}(\lambda_j+\partial_{ij}t^2)^{-1}(X'\underline{y}+ty_{n+1}\underline{h}_i)$,

so that $\underline{h}_i'\hat{\underline{\beta}}$ has variance

(3.65) $\quad \mathrm{var}(\underline{h}_i'\hat{\underline{\beta}}) = (\lambda_i+t^2)^{-1}.\sigma^2 << \lambda_i^{-1}.\sigma^2$

for even moderate t in the worst cases.

We are thus led to a notion of ill-conditioning as disparities between the eigenroots $\lambda_i(S)$, and may define such rough and ready measures as

(3.66) $\quad \|\Delta - \frac{tr\Delta}{k}.I_k\| = \Sigma(\lambda_i-\bar{\lambda})^2$

$$= tr(\Delta^2) - tr^2(\Delta)/k$$

$$= tr(S^2) - tr^2(S)/k \quad ,$$

a Froebenius norm, or similarly

(3.67) $\quad \|\Delta^{-1} - \frac{tr\Delta^{-1}}{k}.I_k\| = tr(S^{-2}) - tr^2(S^{-1})/k$.

Minimizing the above measures is much the same problem as minimizing the generalized variance of $\hat{\underline{\beta}}$, given in

Anderson (1958, p.166) as

$$(3.68)\qquad |\mathrm{var}(\hat{\underline{\beta}})| = \sigma^{2k}.|X'X|^{-1} \quad ,$$

over available designs X.

Both (3.67) and (3.68) are easily generalized to the case X not of full column-rank, after reparametrization of the model given in a later discussion in Section 3.4. Essentially these remarks for the full column-rank case amount to noting that even when an estimate is technically available, it may not be advisable to view it in isolation from its estimated variance, and that is precisely what the RE of (3.32) does in practice.

3.3 ESTIMATION OF THE SCALE PARAMETER σ^2

Since $\hat{\underline{y}}$ is the BLUE of $X\underline{\beta}$, it is clear that the expression

$$(3.69)\qquad \hat{\underline{\varepsilon}} = \underline{y}-\hat{\underline{y}} = \underline{y}-X\underline{b} = (I-XS^{g}X')\underline{y}$$

approximates $\underline{\varepsilon}$ is some sense. It is also easily verified that

$$(3.70)\qquad X'\hat{\underline{\varepsilon}} = X'(I-XS^{g}X')\underline{y} = \underline{0} \quad .$$

This means that the linear combinations of the $\hat{\underline{\varepsilon}}$ corresponding to the observation points of experimental designs, *inter alia*, will always be zero. Since $\hat{\underline{\varepsilon}}$ is unique (for a given $\underline{y}$) we may say it estimates $\underline{\varepsilon}$, in the model (1.1) subject to (1.2) and (1.4). Both are however random variables, and $\hat{\underline{\varepsilon}}$ satisfies

$$(3.71)\quad E(\hat{\underline{\varepsilon}}) = E((I-XS^gX')\underline{y}) = E((I-XS^gX')\underline{\varepsilon}) = \underline{0} \quad ,$$

and

$$(3.72)\quad \mathrm{var}(\hat{\underline{\varepsilon}}) = \sigma^2.(I-XS^gX')$$

because of the symmetry and idempotency of (3.41).

From (3.70) it is permissible to partition the sum of squares of the observations $\underline{y}'\underline{y}$ as

$$(3.73)\quad \underline{y}'\underline{y} = (\hat{\underline{y}}+\hat{\underline{\varepsilon}})'(\hat{\underline{y}}+\hat{\underline{\varepsilon}}) = \hat{\underline{y}}'\hat{\underline{y}} + \hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}} \quad ,$$

and equivalently as the orthogonal separation

$$(3.74)\quad \underline{y}'\underline{y} = \underline{y}'XS^gX'\underline{y} + \underline{y}'(I-XS^gX')\underline{y} = \underline{b}'S\underline{b} + (\underline{y}'\underline{y}-\underline{b}'S\underline{b}) \quad ,$$

where $\underline{b}$ is a solution to the NE's, i.e. $\underline{b}$ from (3.6). Each term in the separations is unique. Defining the sum of squares for error SS(E) by

$$(3.75)\quad SS(E) = \underline{y}'(I-XS^gX')\underline{y} = \hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}}$$

we have the well-known

Theorem 3.7 (Aitken, 1940)

The assumption of normality, i.e. $\underline{y} \sim N(X\underline{\beta},\sigma^2 I)$ implies that SS(E) has a central $\sigma^2.\chi^2_f$ distribution with degrees of freedom $f = n-r$.

Proof: Taking Q as $(I-XS^gX')/\sigma^2$ and $\underline{\mu} = E(\underline{y}) = X\underline{\beta}$, the quadratic form (3.75) in $\underline{y}$ satisfies the conditions of Theorem 2.18.

However an alternative proof may be given, and we give it in order to introduce a device which will lead to considerable simplification of many proofs in the ensuing sections and chapters. From (3.75) and the idempotency of (3.72) up to a scale parameter, the result is immediate for the quadratic form in $\hat{\underline{\varepsilon}}$. Under the assumptions of Theorem 3.7, this is equivalent to taking $Q = I/\sigma^2$ and $\underline{\mu} = E(\hat{\underline{\varepsilon}}) = \underline{0}$ in Theorem 2.18.

In either case the

$$(3.76) \quad f = r(I-XS^gX') = tr(I-XS^gX')$$
$$= tr(I) - tr(SS^g)$$
$$= n - r \quad .$$
□

Corollary 3.7.1. An unbiased estimator $\hat{\sigma}^2$ of σ^2 is

$$(3.77) \quad \hat{\sigma}^2 = (\hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}})/(n-r) \quad .$$

Proof: The result is well-known, and is independent of any distributional assumptions. Rao (1973, p.228) points out that under the normality assumption, this quadratic estimation $\hat{\sigma}^2$ is a minimum variance quadratic unbiased estimator (MINQUE) of σ^2. □

Equation (3.70) suggests a check for algorithmic accuracy in the solution of the normal equations by evaluating

$$(3.78) \quad X'(I-XS^*X') = (I-SS^*)X' \quad ,$$

or the transpose of (3.78), where S^* is the computed g_1-inverse of S. By (3.17) and (3.21), we require that

$$(3.79) \quad X(I-S^*S) = 0 \;, \quad \text{or} \quad S(I-S^*S) = 0 \quad ,$$

which amounts to a check of the estimability of $X\underline{\beta}$ under the given algorithm, with

$$(3.80) \quad \| X(I-S^*S) \| = tr(SS^*SS^*S+S) - 2tr(SS^*S)$$

as a scalar-valued index of accuracy. Since (3.80) may be computationally inefficient, the form of (3.21) suggests that it may be worthwhile taking the index $tr(S) - tr(SS^*S)$

as a first approximation. For X $n \times k$ of rank r this will involve a further set of calculations, but very much fewer than the number of calculations used in the formation and solution of the normal equations. If a program to solve normal equations and perform hypothesis tests is to apply to non-full rank X, e.g. to experimental designs, it may usefully include a check for estimability of the form (3.17). In that case

$$(3.79) \quad S(I-S^*S) = 0$$

may be checked as a matter of routine, and a measure established of the effective round-off error of the given algorithm for the particular matrix X, as its trace.

Note that (3.80) as criterion will apply for X, and hence S, of arbitrary rank. The validity of the equation is perhaps easiest to verify when S^* is of minimal rank r, e.g. when S is inverted by a Cholesky method and rows corresponding to singularities are dropped from S, together with the corresponding column. In such cases we have S^* of the form

$$(3.81) \qquad S^* = \begin{bmatrix} S^*_{11} & 0 \\ 0 & 0 \end{bmatrix}$$

with S^*_{11} the computed inverse of S_{11}.

The variance of $\hat{\underline{\varepsilon}}$ given in (3.72) led Theil (1971) to consider the problem of obtaining a set of residuals $\tilde{\underline{\varepsilon}}$ which were linear in $\underline{y}$, unbiased, homoscedastic and uncorrelated (with scalar variance-covariance matrix), therefore satisfying

$$(3.82) \qquad \tilde{\underline{\varepsilon}} = P\underline{y}$$

$$(3.83) \qquad E(\tilde{\underline{\varepsilon}}) = \underline{0}$$

$$(3.84) \qquad var(\tilde{\underline{\varepsilon}}) = \sigma^2.PP' = \sigma^2.I_\ell \quad .$$

Since the import of (3.83) is that

$$(3.85) \qquad PX\underline{\beta} = \underline{0}$$

over the parameter space, a vector $\tilde{\underline{\varepsilon}}$ of maximal order is obtained for

$$(3.86) \qquad \ell = n-r \quad ,$$

and P an $(\ell \times n)$ matrix of orthogonal row eigen-vectors corresponding to unity eigen-value of $(I-XS^{g_1}X')$. Such LUS residuals $\tilde{\underline{\varepsilon}}$ are not uniquely defined. Golub and Styan (1973) show that their method of Householder transformations leads in each case to a set of LUS residuals, from

$$(3.87) \qquad H\underline{y} = H(X\underline{\beta}+\underline{\varepsilon})$$

$$= \begin{bmatrix} R \\ 0 \end{bmatrix}\underline{\beta} + \begin{bmatrix} H_1 \\ H_2 \end{bmatrix}\begin{bmatrix} \underline{\varepsilon}_1 \\ \underline{\varepsilon}_2 \end{bmatrix}$$

$$= \begin{bmatrix} R\underline{\beta} + H_1\underline{\varepsilon} \\ \underline{0} \quad \underline{\tilde{\varepsilon}} \end{bmatrix} \quad , \qquad \text{and}$$

(3.88) $H^2 = HH' = I$

for H as in (3.35). For a given ordering of the observations $\underline{y}$, it can be shown that $\underline{\tilde{\varepsilon}}$ is such that

(3.89) $E(\underline{\tilde{\varepsilon}}-\underline{\hat{\varepsilon}}_2)'(\underline{\tilde{\varepsilon}}-\underline{\hat{\varepsilon}}_2))$

is minimal, so that we may use the term "best" in that sense, and say that $\underline{\tilde{\varepsilon}}$ is a BLUS residual vector (Judge, Griffiths, Carter Hill and Tsoung-Chao Lee, 1980). Golub and Styan also show that

(3.90) $\underline{\tilde{\varepsilon}}'\underline{\tilde{\varepsilon}} = \underline{\hat{\varepsilon}}'\underline{\hat{\varepsilon}}$

and note that (3.84) allows testing for serial correlation in $\underline{y}$ through examining autocorrelations in $\underline{\tilde{\varepsilon}}$.

3.4 REPARAMETRIZATION AND IMPOSED LINEAR RESTRICTIONS

Graybill (1961, pp.235-241) and Pringle and Rayner (1971, pp. 88-98) provide extensive summaries of the definitions and results, and the relationship with estimability.

Let the vector of parameters be transformed to $\underline{\beta}^0 = L\underline{\beta}$, with the contragredient transformation of X to $X^0 = XU$, where U is $k\times q$, L is $q\times k$ and

(3.91) $X^0\underline{\beta}^0 = X\underline{\beta}$.

A reparametrixation is defined by the transformations U and L which satisfy (3.91). It is clear that a necessary and sufficient condition for a reparametrization is that

(3.92) $XUL = X$.

Then $R(X) \subset R(L)$, and solving for U, we have from (2.13) and (2.14)

(3.93) $U = X^g X L^g + Z - X^g X Z L L^g$, for arbitrary Z.

U may always to taken as L^{g_1}. Now

(3.94) $XU = X[L^g + Z(I - LL^g)]$,

and from (3.92),

(3.95) $r(X) \geq r(XU) = r(X^0) \geq r(XUL) = r(X)$.

On this definition a reparametrized model has the same rank as the original model:

(3.96) $r(X^0) = r(X)$.

If a model is replaced by a model of lower rank, it is a reduced model not a reparametrization, and such cases are discussed under Section 3.5 and elsewhere, e.g. in cases where a partitioned model

(3.97) $$X\underline{\beta} = [X_1 : X_2]\begin{bmatrix}\underline{\beta}_1 \\ \underline{\beta}_2\end{bmatrix}$$

has $\underline{\beta}_2$ dropped from the model, and the new model is

(3.98) $$X\begin{bmatrix}\underline{\beta}_1 \\ \underline{0}\end{bmatrix} = X_1\underline{\beta}_1 = [X_1 : 0]\begin{bmatrix}\underline{\beta}_1 \\ \underline{\beta}_2\end{bmatrix} .$$

<u>Theorem 3.8</u> (Pringle and Rayner, 1971)

A reparametrized model is equivalent in every way to the original, in respect of estimation.

Proof: Estimability coincides, since if for some $\underline{a}$,

$$(3.99) \qquad \underline{\theta}'\underline{\beta} = \underline{a}'X\underline{\beta} = \underline{a}'X^0\underline{\beta}^0 = \underline{\theta}^{0\prime}\underline{\beta}^0 \quad ,$$

condition (3.11) is satisfied in each case. The BLU estimates coincide, because the normal equations are reparametrized to

$$(3.100) \qquad S^0\underline{b}^0 = X^{0\prime}\underline{y} = \underline{g}^0 \quad , \qquad \text{or}$$

$$(3.101) \qquad U'SU\underline{b}^0 = U'X'\underline{y} \quad .$$

Any solution $\underline{b}$ to the original normal equations yields $\underline{b}^0 = L\underline{b}$ as a solution to (3.101), and conversely, in (3.101) yields $\underline{b} = U\underline{b}^0$. Thus the BLUE OF $\underline{\theta}^{0\prime}\underline{\beta}^0$ and $\underline{\theta}'\underline{\beta}$ is given by

$$(3.102) \qquad \underline{\theta}^{0\prime}\underline{b}^0 = \underline{a}'X^0\underline{b}^0 = \underline{a}'X^0L\underline{b} = \underline{a}'X\underline{b} = \underline{\theta}'\underline{b} \quad ,$$

with equal variances from Theorem 2.8 in

$$\begin{aligned}(3.103) \qquad \mathrm{var}(\underline{a}'X^0\underline{b}^0) &= \sigma^2.\underline{a}'X^0(S^0)^{g_1}X^{0\prime}\underline{a} \\ &= \sigma^2.\underline{a}'X^0(LS^{g_1}L')X^{0\prime}\underline{a} \\ &= \sigma^2.\underline{a}'XS^{g_1}X'\underline{a} \\ &= \mathrm{var}\,(\underline{a}'X\underline{b}) \quad .\end{aligned}$$

Similarly the sum of squares for the fitted values coincide, since

$$\begin{aligned}(3.104) \qquad \hat{\underline{y}}_0'\hat{\underline{y}}_0 &= \underline{y}'X^0(S^0)^{g_1}X^{0\prime}\underline{y} \\ &= \underline{y}'XS^{g_1}X'\underline{y} \\ &= \hat{\underline{y}}'\hat{\underline{y}} \quad . \qquad \square\end{aligned}$$

Rayner, in unpublished lecture notes, defines a reparametrization to be estimable if $\underline{\beta}^0 = L\underline{\beta}$ is itself estimable under the original model. In that case $R(X) \subset R(L) \subset R(X)$, and we will write T in place of L. All linear functions of $\underline{\beta}^0$

will be estimable, and the BLUE of $\underline{a}'\underline{\beta}^0$ is

(3.105) $\underline{a}'\underline{b}^0 = \underline{a}'T\underline{b}$

with variance $\sigma^2\underline{a}'TS^{g_1}T'\underline{a}$. Note that T need not have full row-rank r. However T is $q\times k$ so that $q\geq r$ and $k\geq r$. The full row-rank (estimable) reparametrization of Graybill (1961) simply takes $q = r \leq k$. In that case U and XU have full column-rank, by (3.63). Also

(3.106) $XUTU = XU$

implies TU is a right-inverse of XU, and is therefore of full rank with

(3.107) $TU = I_r$, and

(3.108) $U = T^{g_{123}}$.

This means that in (3.93), $X^gX = T^gT$, the unique projection fixing the row-space of X, and Z admits only the g_{123}-inverses of T. Similarly (3.94) reduces to

(3.109) $XU = XT^g$.

For a given T, Graybill finds U by what is effectively a special case of Theorem 2.10, as

(3.110) $U = (T'T+H'H)^{-1}T'$.

Clearly in the full-rank reparametrization S^0 is non-singular, but it is not necessarily diagonal. Suppose, as in fitting orthogonal polynomials to factorial designs with a factor at equally spaced intervals, diagonal S^0 and hence diagonal $(S^0)^{-1}$ are desired. Whereas previously L or T

was selected, this situation amounts to the selection of U, after which an appropriate L is established. From

(3.111) $XUL = X$, and

(3.112) $U'SUL = U'S$

we have, for arbitrary W as in (2.13) and (2.14)

(3.113) $L = (S^0)^{g_1}U'S + (I-(S^0)^{g_1}S^0)W$.

Though U may have orthogonal columns, T need not have orthogonal rows. It may however occur, depending on the nature of S, or equivalently X, and of U. For instance U may comprise orthogonal eigen-vectors of S in which case we may take

(3.114) $L = (S^0)^{g_1}.S^0.U'$.

An orthogonal full-rank estimable reparametrization takes T satisfying

(3.115) $T = (S^0)^{g}U'S$.

If U is again such (3.114) holds we have T of the form

(3.116) $T = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} U'$.

Historically, special applications of reparametrization considerably simplified the algebra of estimation through manipulation of the normal equations, including cases for non-full rank. Whatever (full-rank estimable) reparametrizations were applied seem to have their origin in the imposition of linear restrictions (Yates, 1933, 1934; Yates and Hale, 1939; Plackett, 1950; Quenouille, 1950;

Kempthorne, 1952; Pringle and Rayner, 1971) which are examined below. Problems relating to the fact that certain reparametrizations involve T that extrapolates from the data may be an area for further research, since simply performing tests of hypotheses on $\underline{\beta}^0$ need not necessarily be an appropriate criterion. However such an analysis of T is essentially a problem in experimental design where theory and methods beyond the scope of this thesis, will usefully apply.

Consider X not of full rank, and the problem of finding a minimum variance, linear conditionally unbiased estimator (BLICUE) $\hat{\underline{\beta}}^*$, for $\underline{\beta}$ subject to the condition

$$(3.117) \quad L\underline{\beta} = \underline{c} \quad ,$$

where L is $q \times k$ with rows fully complementary to $R(X)$. If $q = n-r$, the consistency of (3.117) is guaranteed, and for $q > n-r$, take $\underline{c}$ in $C(L)$. For instance the "usual" restraints in analysis of variance take $\underline{c} = \underline{0}$.

The OLS principle is used to minimize

$$(3.118) \quad (\underline{y}-X\underline{\beta})'(\underline{y}-X\underline{\beta}) \quad ,$$

at a point $\hat{\underline{\beta}}^*$, in the hyperplane defined by (3.117). With a Lagrange multiplier $\underline{\lambda}$ we obtain an increased set of normal equations

$$(3.119) \quad \begin{bmatrix} S & L' \\ L & 0 \end{bmatrix} \begin{bmatrix} \hat{\underline{\beta}}^* \\ \underline{\lambda} \end{bmatrix} = \begin{bmatrix} \underline{g} \\ \underline{c} \end{bmatrix} \quad .$$

Since L was fully complementary to S, the equations have a unique solution, regardless of the rank of S. Pringle and

Rayner (1971) use special cases of Theorems 2.10 and 2.15 to show the following three results.

Theorem 3.9

The BLICUE $\hat{\underline{\beta}}^*$ of $\underline{\beta}$, subject to $L\underline{\beta} = \underline{c}$ is given by

(3.120) $\hat{\underline{\beta}}^* = K^{-1}X'\underline{y} + K^{-1}L'\underline{c}$

where $K^{-1}X'$ and $K^{-1}L'$ are g_{123}-inverses of X and L respectively, and K is obtained as in (2.66).

Theorem 3.10

The necessary and sufficient conditions for $\tilde{\underline{\beta}} = G\underline{y} + \underline{d}$ to be a LICUE of $\underline{\beta}$, with $L\underline{\beta} = \underline{c}$, are

(3.121) $GX = I - WL$, and

(3.122) $W\underline{c} = \underline{d}$

for some W.

Theorem 3.11 (Chipman, 1964)

If $\tilde{\underline{\beta}} = G\underline{y} + \underline{d}$ gives a LICUE $\underline{\theta}'\tilde{\underline{\beta}}$ for all estimable $\underline{\theta}'\underline{\beta}$, then $\tilde{\underline{\beta}}$ is a LICUE of $\underline{\beta}$ for some L. Further the BLUE $\underline{\theta}'\hat{\underline{\beta}}$ of $\underline{\theta}'\underline{\beta}$ is given by $\hat{\underline{\theta}}'\hat{\underline{\beta}}^*$, for $\hat{\underline{\beta}}^*$ from (3.120).

In consequence the following quantities coincide:

(i) the sums of squares for fitting $\hat{\underline{\beta}}$ and $\hat{\underline{\beta}}^*$,

(ii) the residual sums of squares $\underline{y}'(I-XK^{-1}X')\underline{y}$ and

$\underline{y}'(I-XS^{g}X')\underline{y}$,

(iii) the estimators of σ^2, and

(iv) the variance of the estimators $\underline{\theta}'\hat{\underline{\beta}}^*$ and $\underline{\theta}'\hat{\underline{\beta}}$ for estimable functions $\underline{\theta}'\underline{\beta}$.

Scheffe (1959, p.16) showed that for L as in (3.117), the method of imposed linear restrictions is equivalent to a replacement of parameters in a new model, i.e. $\underline{\beta}$ is replaced by $\underline{\beta}^*$ where

$$(3.123)\quad \begin{bmatrix} X \\ L \end{bmatrix} \underline{\beta}^* = \begin{bmatrix} X\underline{\beta} \\ \underline{c} \end{bmatrix} ,$$

or equivalently,

$$(3.124)\quad \underline{\beta}^* = K^{-1}S\underline{\beta} + K^{-1}L'\underline{c} .$$

For such a substitution to constitute a reparametrization it is sufficient that $\underline{c} = \underline{0}$, as Pringle and Rayner point out. However these relationships are special cases of

<u>Theorem 3.12</u> (Dunne)

An equivalent condition for any set of imposed linear restrictions to constitute an estimable reparametrization of the model (1.1) is that

$$(3.125)\quad L\underline{\beta} = \underline{c} = \underline{0} , \qquad \text{where}$$

$$(3.126)\quad \mathcal{R}(L) \cap \mathcal{R}(X) = \{\underline{0}\} .$$

Proof: We consider cases of (3.123) where $\begin{bmatrix} X \\ L \end{bmatrix}$ need not have full column rank, so that

$$(3.127)\quad \underline{\beta}^* = K^{g_1}S\underline{\beta} + K^{g_1}L'\underline{c} .$$

For reparametrization and (3.125) we require

$$(3.128)\quad LK^{g_1}L'\underline{c} = LL^{g}\underline{c} = \underline{0}$$

by Theorem 2.10. Then $\underline{c}$ is in the space $\mathcal{C}(L)$ implies

(3.129) $LL^g\underline{c} = \underline{c} = 0$.

Sufficiency is obvious from (3.127). □

<u>Theorem 3.13</u> (Dunne)

The equivalent condition to

(3.130) $\underline{\beta}^* = K^{g_1}S\underline{\beta}$

giving a full rank estimable reparametrization is that the rows of K^{g_1} contain a basis for the zero row eigen-space of S.

Proof: Partition $\underline{\beta}^*$ as $\begin{bmatrix}\underline{\beta}_1^* \\ \underline{\beta}_2^*\end{bmatrix}$. Then we require

(3.131) $\underline{\beta}_2^* = \underline{0}$

over the parameter space. □

The import of Theorem 3.13 is that L must be chosen in the orthogonal complement of $R(X)$. This condition is not sufficient, unless the rows of L span that complement.

Orthogonal reparametrization is essentially a device applied to full-rank models, and not equivalent to any set of imposed linear constraints. When a full-rank estimable reparametrization has been effected, orthogonality is achieved by taking U and T as in (3.114) and (3.115).

Early practitioners used L of maximum rank, to obtain the unique solution $\hat{\underline{\beta}}^*$ by a number of different procedures. Examples include

(i) full rank reparametrization as in 2^n factorial designs,

(ii) bordering S with L to form (3.119) as in randomized blocks with the "usual" restrictions,

(iii) dropping surplus parameters, as when setting $\mu = 0$ in a simple random design,

(iv) augmenting S in the normal equations by $L\hat{\underline{\beta}} = \underline{c}$, with $\underline{c}$ usually $\underline{0}$.

The last is effectively a dropping of $\underline{\lambda}$ in (3.119).

A further method is possible, by extending the OE's (3.1) to

(3.132) $$\begin{bmatrix} \underline{y} \\ \underline{0} \end{bmatrix} = \begin{bmatrix} X \\ L \end{bmatrix} \underline{\beta}$$

with singular variance-covariance matrix given by

(3.133) $$V = \sigma^2 \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} ,$$

and solving for the generalized least squares (GLS) estimates, by methods described in Chapter 5. It will transpire that the solution for the BLUE of $\underline{\beta}$ or $X\underline{\beta}$ in (3.132) is the same as the BLICUE of $\underline{\beta}$ or $X\underline{\beta}$ subject to $L\underline{\beta} = \underline{0}$. Similarly for $L\underline{\beta} = \underline{c}$.

If the condition on rank invariance is relaxed, so that we obtain a reduced model, and not a reparametrized model, the imposed linear restrictions intrude into the space of estimable functions. The foregoing theory no longer applies. We examine such cases in Section 3.5 and Section 3.6. Theorems 3.9 to 3.11 are special cases of more general results.

It should be noted that the preservation of the rank of the model does not constitute a sufficient condition for reparametrization. Consider the partitioned model

$$(3.134)\quad \underline{y} = [X_1 : X_2]\begin{bmatrix}\underline{\beta}_1\\ \underline{\beta}_2\end{bmatrix} + \underline{\varepsilon} \quad .$$

In (3.134) we may choose to replace $X_2\underline{\beta}_2$ by the estimable function

$$(3.135)\quad (I - X_1(X_1'X_1)^{g_1}X')X_2\underline{\beta}_2 = F\underline{\beta}_2 \quad .$$

However if $X_2\underline{\beta}_2$ is itself not estimable in (3.134), then the new model

$$(3.136)\quad \underline{y} = [X_1 : F]\begin{bmatrix}\underline{\beta}_1\\ \underline{\beta}_2\end{bmatrix} + \underline{\varepsilon}$$

is not equivalent to (3.134), despite the fact that

$$(3.137)\quad r[X_1 : X_2] = r(X_1) + r(F) = r[X_1 : F] \quad .$$

On the other hand, if $X_2\underline{\beta}_2$ is estimable, then

$$(3.138)\quad r[0 : X_2] = r[0 : F] \quad , \qquad \text{and}$$

$$(3.139)\quad R(F) = R(X_2)$$

so that

$$(3.140)\quad R([X_1 : X_2]) = R([X_1 : F]) \quad ,$$

and a reparametrization is assured.

3.5 PRIOR LINEAR CONSTRAINTS

In the preceding section, the constraint matrix L was limited to rows corresponding to non-estimable functions.

has a LUE $\underline{a}_1'\underline{y} + d$ if and only if $\underline{\theta}$ is in $R(K)$, where

The constraints were applied simply to facilitate the calculation of a particular solution to the normal equations. We examine and extend the results of Rao (1971, pp.231-233) and Pringle and Rayner (1971, pp.98-101) for the situation where $\underline{\beta}$ is presumed *a priori* to satisfy $L\underline{\beta} = \underline{c}$, and these restrictions are not just the device of the preceding section. Accordingly it is possible that L intrudes into the space of estimable functions, i.e. into the space $R(X)$. For instance, in analysis of variance, testing for the additional effect of fitting a subvector of $\underline{\beta}$ amounts to comparing fitted values from the usual OLS solution of the normal equations with those subject to additional restrictions of the form

$$(3.141) \quad L\underline{\beta} = [0 : I]\underline{\beta} = \underline{0} \quad ,$$

where L will usually have

$$(3.142) \quad R(L) \cap R(X) \neq \{\underline{0}\} \ .$$

In this section, consistency of the constraints is assumed, but L may or may not have full row-rank.

Minimizing $(\underline{y}-X\underline{\beta})'(\underline{y}-X\underline{\beta})$ subject to

$$(3.143) \quad L\underline{\beta} = \underline{c}$$

we obtain with a Lagrange multiplier $\underline{\lambda}$,

$$(3.144) \quad \begin{bmatrix} S & L' \\ L & 0 \end{bmatrix} \begin{bmatrix} \hat{\underline{\beta}}_0 \\ \underline{\lambda} \end{bmatrix} = \begin{bmatrix} \underline{g} \\ \underline{c} \end{bmatrix}$$

and solve by means of (2.123) in Theorem 2.14. Then $\underline{\theta}'\underline{\beta}$ has a LUE $\underline{a}_1'\underline{y} + d$ if and only if $\underline{\theta}'$ is in $R(K)$, where

(3.145) $K = S + L'L$,

whence

(3.146) $\underline{\theta}' = \underline{a}_1'X + \underline{a}_2'L$, and

(3.147) $\underline{a}_2'\underline{c} = d$.

Theorem 3.14 (Pringle and Rayner, 1971).

The BLUE of $\underline{\theta}'\underline{\beta}$ in the model constrained by (3.143) is given by $\underline{\theta}'\hat{\underline{\beta}}_0$, where $\hat{\underline{\beta}}_0$ is any solution to (3.144), and is unique over all $\hat{\beta}_0$. For G_{11} as the leading submatrix in (2.123),

(3.148) $\text{var}(\underline{\theta}'\hat{\underline{\beta}}_0) = \sigma^2.\underline{\theta}'G_{11}\underline{\theta}$.

The uniqueness of $\underline{\theta}'\hat{\underline{\beta}}_0$ is said to have an algebraic proof that is complicated. However, from (3.146) we may write for some suitable $\underline{a}$,

(3.149) $\underline{\theta}' = \underline{a}'K$.

Then it follows from (2.123) that, for all such $\underline{a}$,

(3.150) $$\begin{aligned}\text{var}(\underline{\theta}'\hat{\underline{\beta}}_0) &= \text{var}(\underline{a}'K\hat{\underline{\beta}}_0) = \text{var}(\underline{a}'KG_{11}X'\underline{y}) \\ &= \sigma^2.\underline{a}'KG_{11}SG_{11}K\underline{a} \\ &= \sigma^2.\underline{a}'KG_{11}K\underline{a} \\ &= \sigma^2.\underline{a}'(K-L'R^gL)\underline{a} \\ &= \sigma^2.\underline{\theta}'(K^g-K^gL'R^gLK^g)\underline{\theta}\end{aligned}$$

since $\mathcal{R}(R) = \mathcal{R}(LK^gL') = \mathcal{R}(L)$, by applying Corollary 2.8.3. The fitted model induces a residual variate $\hat{\underline{\varepsilon}}_0$ defined by

(3.151) $\hat{\underline{\varepsilon}}_0 = (\underline{y}-\hat{\underline{y}}_0) = \underline{y} - X\hat{\underline{\beta}}_0$

$$= (I-XG_{11}X')\underline{y} - XK^gL'R^g\underline{c} \quad .$$

Theorem 3.15 (Rao, 1973)

Under the constraints (3.143) and assuming normality, the quadratic form $\hat{\underline{\varepsilon}}_0'\hat{\underline{\varepsilon}}_0$ has central $\sigma^2.\chi_f^2$ distribution with degrees of freedom

(3.152) $f = n - r(K) + r(L) \quad .$

Proof: (Dunne). In view of the relation $\underline{\theta}'K^g K = \underline{\theta}'$, we have from Theorem 2.14, with G_{11} and G_{12} the corresponding submatrices of the partitioned inverse (2.123)

(3.153) $X = XK^gK$

$$= X(K^g-K^gL'R^gLK^g)K + XK^gL'R^gL$$

$$= XG_{11}K + XG_{12}L \quad .$$

Solving (3.144), for arbitrary $\underline{z}$,

(3.154) $\hat{\underline{\beta}}_0 = G_{11}X'\underline{y} + G_{12}\underline{c} + (I-K^gK)\underline{z}$

implies that $\hat{\underline{y}}_0$ is uniquely given by

(3.155) $\hat{\underline{y}}_0 = X\hat{\underline{\beta}}_0 = XG_{11}X'\underline{y} + XG_{12}\underline{c}$

$$= XG_{11}X'\underline{y} + XK^gL'R^g\underline{c} \quad .$$

This verifies (3.151), and we note that $XG_{11}X'$ is idempotent from

(3.156) $XG_{11}X'XG_{11}X' = XG_{11}(K-L'L)G_{11}X'$

$$= XG_{11}X'$$

by successive applications of

$$(3.157)\quad LG_{11}X' = L(K^g - K^gL'R^gLK^g)X'$$
$$= 0 \quad ,$$

since $R(L') = R(R)$. Thus

$$(3.158)\quad E(\hat{\underline{\varepsilon}}_0) = E(\underline{y}) - E(\hat{\underline{y}}_0)$$
$$= X\underline{\beta} - XG_{11}S\underline{\beta} - XK^gL'R^g\underline{c}$$
$$= X\underline{\beta} - (X\underline{\beta} - XK^gL'R^gL\underline{\beta}) - XK^gL'R^g\underline{c}$$
$$= \underline{0} \quad , \qquad \text{and}$$

$$(3.159)\quad \text{var}(\hat{\underline{\varepsilon}}_0) = \sigma^2.(I - XG_{11}X') \quad ,$$

which is idempotent up to a scale parameter, from (3.156). Then for $Q = I$ in Theorem 2.18, the quadratic form $\hat{\underline{\varepsilon}}_0' I \hat{\underline{\varepsilon}}_0$ has $\sigma^2.\chi^2_f$ distribution with

$$(3.160)\quad f = r(I - XG_{11}X') = tr(I - XG_{11}X')$$
$$= n - tr[G_{11}(K - L'L)]$$
$$= n - tr(K^gK) + tr(K^gL'R^gLK^gK)$$
$$= n - r(K) + r(R)$$
$$= n - r(K) + r(L)$$

<u>Corollary 3.15.1</u> (Rao, 1973)

An unbiased estimator of σ^2 is given by

$$(3.161)\quad \hat{\sigma}_0^2 = (\underline{y}'(I - XG_{11}X')\underline{y} + \underline{c}'R^g\underline{c} - \underline{c}'\underline{c})/f \; .$$

Proof: (Dunne). Noting $\underline{c}$ is in $C(L) = C(R)$, we simplify

$$(3.162)\quad \hat{\underline{\varepsilon}}_0'\hat{\underline{\varepsilon}}_0 = \underline{y}'(I - XG_{11}X')\underline{y} + \underline{c}'R^gLK^g(K - L'L)K^gL'R^g\underline{c}$$
$$= \underline{y}'(I - XG_{11}X')\underline{y} + \underline{c}'R^g\underline{c} - \underline{c}'\underline{c} \; . \qquad \square$$

Rao gives (3.161) in another form and writes $r\begin{bmatrix} X \\ L \end{bmatrix}$ in place of $r(K)$ in (3.160). We note that (3.160) reduces to $n-r$ if and only if L is partly or fully complementary to X, in which case $\hat{\underline{\varepsilon}}_0$ reduces to

(3.163) $\hat{\underline{\varepsilon}}_0 = (I-XG_{11}X')\underline{y} - \underline{0}$

$$= (I-XK^gX')\underline{y}$$

$$= (I-XS^{g_1}X')\underline{y}$$

$$= \hat{\underline{\varepsilon}} \quad ,$$

and the theory of imposed linear restrictions in Section 3.4 is corroborated. However (3.163) holds providing

(3.164) $XK^gL'R^g\underline{c} = \underline{0}$, thus

(3.165) $LK^g(K-L'L)K^gL'R^g\underline{c} = \underline{0}$, and

(3.166) $\underline{c} = RR^g\underline{c} = R\underline{c}$.

For (3.166) to hold for arbitrary $\underline{c}$ in consistent restrictions (3.143), the equivalent condition is that R is idempotent, or that

(3.167) $L = LK^{g_1}L'L$,

for all g_1-inverses of K. In turn this holds if and only if (3.126) is satisfied.

<u>Theorem 3.16</u> (Dunne)

The variates $X\hat{\underline{\beta}}_0$ and $\hat{\underline{\varepsilon}}_0$ are uncorrelated.

Proof: From (3.151) and (3.156),

$$(3.168) \quad \mathrm{cov}(\hat{\underline{\varepsilon}}_0, \hat{\underline{y}}_0) = E(\hat{\underline{\varepsilon}}_0 . \hat{\underline{y}}_0') - E(\hat{\underline{\varepsilon}}_0).E(\hat{\underline{y}}_0')$$

$$= (I - XG_{11}X')XG_{11}X' - 0$$

$$= 0 \qquad \square$$

Theorem 3.17 (Dunne)

Let $\hat{\underline{\varepsilon}}$ be the OLS residuals obtained as in (3.69) ignoring the constraints. The variables $(\hat{\underline{\varepsilon}}_0 - \hat{\underline{\varepsilon}})$ and $\hat{\underline{\varepsilon}}$ are uncorrelated. Under the assumption of normality, and subject to the constraints,

$$(3.169) \quad F = \frac{\hat{\underline{\varepsilon}}_0'\hat{\underline{\varepsilon}}_0 - \hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}}}{\hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}}} \cdot \frac{f_2}{f_1}$$

is distributed as central $F(f_1, f_2)$ with degrees of freedom

$$(3.170) \quad f_1 = r(X) + r(L) - r(K) \quad , \quad \text{and}$$

$$(3.171) \quad f_2 = n - r(X) = n - r \quad .$$

Proof: From (3.69) and (3.151) we examine

$$(3.172) \quad (XS^gX' - XG_{11}X')E(\underline{y}\underline{y}')(I - XS^gX')$$

$$-[(XS^gX' - XG_{11}X')E(\underline{y}) - XK^gL'R^g\underline{c}]E(\underline{y}')(I - XS^gX') \quad ,$$

the covariance of $(\hat{\underline{\varepsilon}}_0 - \hat{\underline{\varepsilon}})$ and $\hat{\underline{\varepsilon}}$. We note that the restrictions do not affect the variance of $\underline{y}$, and that

$$(3.173) \quad \mathrm{cov}(\hat{\underline{\varepsilon}}_0 - \hat{\underline{\varepsilon}}, \hat{\underline{\varepsilon}}) = 0$$

if $E(\underline{y})$ is in the column space $C(X)$. Given the restrictions (3.143), we may equivalently write

$$(3.174) \quad \begin{bmatrix} L_1 \\ L_2 \end{bmatrix} \underline{\beta} = AL\underline{\beta} = A\underline{c} = \begin{bmatrix} \underline{a}_1 \\ \underline{a}_2 \end{bmatrix}$$

where A has full row-rank, the rows of L_1 are a basis for the space $\mathcal{R}(L) \cap \mathcal{R}(X)$, and the rows of L_2 form a space complementary to $\mathcal{R}(L_1)$, and hence to $\mathcal{R}(X)$. Further we may expand L_1 to obtain L_0 such that

$$(3.175) \quad T = \begin{bmatrix} L_0 \\ L_1 \end{bmatrix}$$

is an $(r \times k)$ matrix of full row-rank r, and has $\mathcal{R}(T) = \mathcal{R}(X)$. We may reparametrize $X\underline{\beta}$ as

$$(3.176) \quad X\underline{\beta} = XUT\underline{\beta} = XU\begin{bmatrix} L_0 \\ L_1 \end{bmatrix}\underline{\beta} = XU\begin{bmatrix} \underline{\beta}_1^* \\ \underline{\beta}_2^* \end{bmatrix}$$

where one choice for U is given by (3.109). Given the consistency of the restrictions (3.143), it is clear that their effect is simply to reduce (3.176) to

$$(3.177) \quad X\underline{\beta}\Big]_{L\underline{\beta}=c} = XUL_0\underline{\beta} + XU\underline{a}_1 \quad ,$$

proving (3.173). In view of Theorems 3.7 and 3.15 it suffices to show

$$(3.178) \quad (\hat{\underline{\varepsilon}}_0 - \hat{\underline{\varepsilon}})'(\hat{\underline{\varepsilon}}_0 - \hat{\underline{\varepsilon}}) = \hat{\underline{\varepsilon}}_0'\hat{\underline{\varepsilon}}_0 - \hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}} \quad ,$$

which follows from

$$(3.179) \quad (I - XS^gX')\hat{\underline{\varepsilon}}_0 = (I - XS^gX')[(I - XG_{11}X')\underline{y} + XK^gL'R^g\underline{c}] = (I - XS^gX')\underline{y} + \underline{0} = \hat{\underline{\varepsilon}}$$

□

It remains a question whether (3.169) represents a test-statistic for any hypothesis associated in some way with the *a priori* restrictions (3.143). This problem will be discussed in Chapter 4. We draw attention to (3.174) and (3.176) and note that Section 3.4 showed that taking $\underline{a}_1 = \underline{0}$ in (3.174) amounts to a further reparametrization of (3.176), from

(3.180) $$XU\begin{bmatrix}\underline{\beta}_1^* \\ \underline{\beta}_2^*\end{bmatrix} = [X_1^* : X_2^*]\begin{bmatrix}\underline{\beta}_1^* \\ \underline{\beta}_2^*\end{bmatrix} \quad , \qquad \text{to}$$

(3.181) $$X_1^*\underline{\beta}_1^* + X_2^*\underline{a}_1 = X_1^*\underline{\beta}_1 \quad , \qquad \text{whereas}$$

$\underline{a}_2 = 0$ in (3.174) amounts to a reduction of the model. In general, the relation $L_2\underline{\beta} = \underline{a}_2$ in (3.174) affects only the algebra and specifically the choice of $\hat{\underline{\beta}}^*$ from OLS estimation in (3.180), but the fitted values $\hat{\underline{y}}$ and $\hat{\underline{y}}_0$ are unaffected by $\underline{a}_2$ and determined only by the value $\underline{a}_1$ and the space $R(L_1)$. We will prove this result in Chapter 4.

It is clear that all the g-inverses in (3.148) through to (3.181) may be replaced by arbitrary g_1-inverses of the matrices in question, but that all the expressions have the unique value stated.

Goldman and Zelen (1964) introduced the term *pre-estimable* to specify those functions which we have described as estimable, viz. those derived from the space $R(X)$, and use the term estimable to include the wider set of linear functions derived from $R(K)$. They considered only the case of K non-singular, so that their use of estimability includes the conditional estimability described here in terms of (3.146) or (3.149). They reduce the restraints to full-rank pre-estimable and conditionally estimable sets, which are linearly independent, and using non-singular transformations, they obtain equivalent but less explicit special cases of Theorems 3.14 and 3.15. An alternative approach to Theorem 3.14 involves the separation of the row-space $R(S)$ into the subspace $R(L)$ and the subspace orthogonal to $R(L)$, but also leads to a less explicit special case.

Chipman (1964) considers X not of full column-rank with L fully (row) complementary to X, and quotes an earlier version of the Goldman and Zelen special cases. In both approaches $\hat{\underline{\beta}}_0$ is uniquely defined and is uncorrelated with $\hat{\underline{\varepsilon}}_0$, as a special case of Theorem 3.16, since $X\hat{\underline{\beta}}_0$ and $L\hat{\underline{\beta}}_0$ are fixed and span the space of all linear combinations $\hat{\underline{\beta}}_0$ including $I\hat{\underline{\beta}}_0$ itself.

Chipman relates the methods of this section to estimation under criteria other than those of BLU estimation. In Section 3.7 we consider some of those criteria.

3.6 REDUCED MODELS

In a preceding section on reparametrization and in (2.176) and (3.177), the effect of linear restrictions which intrude into the space of estimable functions was associated with a reduction of the rank of the model. Rao (1973, p.231) considers substitution into the model (1.1) of a set of restrictions, say

(3.143) $L\underline{\beta} = \underline{c}$.

Following the method of Rayner, in unpublished lecture notes, we solve for $\underline{\beta}$ and set the solution $\underline{\beta}^*$ as

(3.182) $\underline{\beta}^* = L^{g_1}\underline{c} + U\underline{\tau}$

where $\underline{\tau}$ is an arbitrary $(t\times 1)$ vector, and U is $(k\times t)$ such that $LU = 0$. For instance, we may take U to $(I_k - L^{g_1}L)$ for $t = k$, or as the orthogonal complement of L' for $t = k - r(L)$. The model (1.1) becomes

(3.183) $\underline{y} = XL^{g_1}\underline{c} + XU\underline{\tau} + \underline{\varepsilon}$, or

(3.184) $\underline{y}^* = (\underline{y} - XL^{g_1}\underline{c}) = X^*\underline{\tau} + \underline{\varepsilon}$

for $X^* = XU$. The fitted values for this model, after OLS solution, are

(3.185) $\hat{\underline{y}}^* = X^*\hat{\underline{\tau}}$, or

(3.186) $\hat{\underline{y}} = X^*\hat{\underline{\tau}} + XL^{g_1}\underline{c}$.

These values are unique over all choices of L^{g_1} in (3.182) by application of Corollary 2.4.1. Without loss of generality we may write (3.182) as

(3.187) $\underline{\beta}^* = L^{g}\underline{c} + (I - L^{g}L)\underline{\tau} = (K^{g}L'R^{g}\underline{c}) + (I - K^{g}L'R^{g}L)\underline{\tau}$

so that (3.184) is a reduced rank model if and only if

(3.142) $\mathcal{R}(L) \cap \mathcal{R}(X) \neq \{\underline{0}\}$.

If the spaces are virtually disjoint we may take

(3.188) $L^{g_1} = (X'X + L'L)^{g_1}L' = K^{g_1}L'$

from Theorem 2.10 and obtain

$$(3.189)\quad X\underline{\beta}^* = XK^{g_1}L'\underline{c} + X(I-K^{g_1}L'L)\underline{\tau}$$

$$= XK^{g_1}L'\underline{c} + X(K^{g_1}K-K^{g_1}L'L)\underline{\tau}$$

$$= \underline{0} + XK^{g_1}X'X\underline{\tau}$$

$$= X\underline{\tau} \quad .$$

It is clear that the choice of $\underline{\beta}^*$ as

$$(3.190)\quad \underline{\beta}^* = K^{g_1}L'\underline{c} + K^{g_1}S\underline{\tau}$$

is a reparametrization if and only if $\underline{c} = \underline{0}$, as required by Theorem 3.12.

Rayner (1976) examined the relation

$$(3.191)\quad \begin{bmatrix} X \\ L \end{bmatrix}\beta^* = \begin{bmatrix} X\underline{\beta} \\ \underline{c} \end{bmatrix}$$

solved as

$$(3.192)\quad \beta^* = \begin{bmatrix} X \\ L \end{bmatrix}^{g_1}\begin{bmatrix} X\underline{\beta} \\ \underline{c} \end{bmatrix}$$

$$= [F : (I-FX)L^{g_1}]\begin{bmatrix} X\underline{\beta} \\ \underline{c} \end{bmatrix} \quad ,$$ where

$$(3.193)\quad F = E^{g_{124}} = E'(EE')^{g_1} \quad ,$$ for

$$(3.194)\quad E = X(I-L^{g_{124}}L) \quad .$$

The validity of (3.192) in (3.191) is easily verified since

$$(3.195)\quad LE' = 0 \quad ,$$

$$(3.196)\quad XE' = EE' \quad ,$$ and

$$(3.197)\quad X(I-L^{g_1}L) = E(I-L^{g_1}L) \quad .$$

These results show that reduced models are equivalent in every way to prior restriction models, and that all the results of Section 3.4 and Section 3.5 apply *mutatis mutandis* to Section 3.6. The essential difference between the approaches is the stage at which the restrictions are introduced into the solution process.

3.7 ALTERNATIVE ESTIMATION PROCEDURES

It has been noted that the OLS solution for the OE's (3.1) is given by taking

(3.198) $\hat{\underline{\beta}} = G\underline{y}$, with

(3.199) $G = X^{g_{13}} = (X'X)^{g_1}X'$.

Bjerhammar (1958) showed that the minimum Euclidean norm of $\hat{\underline{\beta}}$ in the class generated by (3.198) is obtained for

(3.200) $G = X^{g} = (X'X)^{g}X'$.

In the full-rank linear model, the condition

(3.201) $GX = I_k$

is equivalent to having (3.198) unbiased for $\underline{\beta}$, and hence also a BLUE for the non-full rank case, the condition cannot be fulfilled. We may therefore use the BLICUE's of Section 3.4 or find that class of G which minimize the bias matrix.

(3.202) $B = (I-GX)(I-GX)'$

in the sense that its diagonal elements are minimal. Chipman (1964) discusses minimum bias estimators (LIMBE's) and shows that (3.202) is derived from minimizing the Frobenius norm associated with the relation

(3.203) $$E(\hat{\underline{\beta}}) - \underline{\beta} = (GX\underline{\beta}+\underline{b}) - \underline{\beta}$$

$$= (I-GX)\underline{\beta} + \underline{b} \quad .$$

The corresponding solution for G is then

(3.204) $$G = X^{g_{14}} \quad ,$$

and by Corollary 2.6.1,

(3.205) $$B = (I-X^gX)(I-X^gX)'$$

$$= (I-X^gX) \quad .$$

He also shows that $G\underline{y}$ is a BLICUE of $\underline{\beta}$ for $L\underline{\beta} = 0$ where L is any matrix with $R(L)$ orthogonal to $R(X)$. Rao (1971) states that minimizing (3.202) is equivalent to choosing the norm $\|\underline{e}_i(I-GX)\|$ to be minimal for each unit vector $\underline{e}_i$ of the form

(3.206) $$\underline{e}_i' = (e_{i1},\ldots,e_{in}) = (\ldots,\delta_{ij},\ldots) \quad ,$$

using the Kronecker delta. He generalizes the notion of LIMBE's to a wider class of norms, and thus to variance-covariance matrices other than $\sigma^2.I$ in (1.1), and defines best or minimum variance in this class by isolating G for which the diagonal elements of GVG' are minimal. Under the variance assumption here, this amounts to

(3.207) $$G = X^g$$

and $\hat{\underline{\beta}}$ is said to be a BLIMBE of $\underline{\beta}$. Rao's solution of the minimum restriction to be put on $\underline{\beta}$ so that $\underline{\beta}$ admits a LUE and hence a BLUE, are the conditions stated by Chipman for his special case of Theorem 2.10.

Hoerl and Kennard (1970a, 1970b) examined best linear estimation (BLE) by the method of ridge regression (biased) estimators. Chipman (1964) proposed a wider approach of assigning a prior density to $\underline{\beta}$, and proved special cases of Theorem 2.17. The minimum mean square error estimator (MMSEE or BLE) of $\underline{\beta}$ is obtained for $\hat{\underline{\beta}}$ taken as the posterior mean, and the matrix of mean square error (or risk) at the minimum is obtained as the posterior variance. Theorem 2.17 derives the general form of these quantities. Rao (1971) generalizes the remarks of Chipman and states without proof some special cases of Theorem 2.17. We will return to BLE's and to BLICUE's in a later chapter on residuals.

In Section 3.3 BLUS residuals were examined, and their relationship with autocorrelation noted. Other types, such as best augmented unbiased with scalar variance matrix (BAUS) residuals are discussed *inter alia* in Judge, et al. (1980), who give further references.

CHAPTER 4

HYPOTHESES AND PARTITIONED SUMS OF SQUARES

In this chapter the device of partitioning sums of squares into uncorrelated or orthogonal sums, is examined. The device is well-known in the literature, and serves to underpin hypothesis testing in the linear model. It is also clearly implicit in the estimation procedures, where from (3.73),

(4.1) $$\underline{y}'\underline{y} = (\hat{\underline{y}}+\hat{\underline{\varepsilon}})'(\hat{\underline{y}}+\hat{\underline{\varepsilon}}) = \hat{\underline{y}}'\hat{\underline{y}} + \hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}}$$
$$= \underline{y}'XS^gX'\underline{y} + \underline{y}'(I-XS^gX')\underline{y} \; .$$

Any such partitioning is a function of the X-matrix of the model (1.1), and the properties of partitioning are induced from the model assumptions. Trivial as these remarks may be, it is useful to see (4.1) as the first of a number of steps, and to consider the nature of further partitionings over one or both of its terms. For example, suppose

(4.2) $$\underline{y}'\underline{y} = (\hat{\underline{y}}_1'\hat{\underline{y}}_1+\hat{\underline{y}}_2'\hat{\underline{y}}_2) + (\hat{\underline{\varepsilon}}_1'\hat{\underline{\varepsilon}}_1+\hat{\underline{\varepsilon}}_2'\hat{\underline{\varepsilon}}_2)$$

for suitable choices of

(4.3) $$\hat{\underline{y}} = \hat{\underline{y}}_1 + \hat{\underline{y}}_2 \quad , \qquad \text{and}$$

(4.4) $$\hat{\underline{\varepsilon}} = \hat{\underline{\varepsilon}}_1 + \hat{\underline{\varepsilon}}_2 \quad .$$

Then it may be informative to consider the nature of the quantities $\hat{\underline{y}}_i$ singly, and similarly for $\hat{\underline{\varepsilon}}_i$. The $\hat{\underline{y}}_i$ may be associated with a reduced model, and the $\hat{\underline{\varepsilon}}_i$ with an increased model or the availability of additional observations.

In general, such partitionings can be usefully related to extensions, deletions or partitioning of the model (1.1) *per se* or a special case. The equivalent matrix partitions and restrictions will, as is well-known, highlight the implications of the partitionings in hypothesis testing.

4.1 TESTS OF HYPOTHESES

Rao (1973, p.167), notes that the ratio of two independently distributed $\sigma^2.\chi^2$ variates, over the ratio of their degrees of freedom, is a variate with a Fisher's F-distribution. Since $V = \sigma^2.I$ in (1.1), and $Q = XS^gX'$ is idempotent, we have from Theorem 2.18 and assuming normality, that

(4.5) $$\underline{\hat{y}}'\underline{\hat{y}} = \underline{y}'Q\underline{y} \sim \sigma^2.\chi^2_r(\lambda)$$

where r is $r(X)$ as before, and $\lambda = \underline{\beta}'S\underline{\beta}$, and

(4.6) $$\underline{\hat{\varepsilon}}'\underline{\hat{\varepsilon}} = \underline{y}'(I-Q)\underline{y} \sim \sigma^2.\chi^2_{n-r} \quad .$$

The latter term is central $(\lambda = 0)$ and independent of (4.5). Pringle and Rayner (1971, p.86), and other authors, suggest that a test of the hypothesis $\underline{\beta} = \underline{0}$ may be applied, using the statistic

(4.7) $$F = \frac{SS(\underline{\hat{y}})}{SS(\underline{\hat{\varepsilon}})} . \frac{(n-r)}{r} \quad .$$

However, Searle (1971, p.178) draws the distinction between the null hypothesis $H_0 : \underline{\beta} = \underline{0}$ involving non-estimable functions generally, and $H_0 : X\underline{\beta} = \underline{0}$, an hypothesis on estimable functions. Roy and Roy (1960), reported in Elston and Bush (1964), defined the notions of *weak* and *strong* testability. An hypothesis is *strongly* testable if it involves a relation on strictly estimable functions (e.g. $X\underline{\beta} = \underline{0}$), and otherwise is weakly testable. To the latter hypotheses there correspond a strongly testable sub-hypothesis, and a set of restrictions on $\underline{\beta}$ which involve non-estimable functions. Then we may orthogonally decompose $\underline{\beta}$ as $\underline{\beta} = (X^g X)\underline{\beta} + (I-X^g X)\underline{\beta}$ for any $\underline{\beta}$ in the parameter space.

Thus $\underline{\beta} = \underline{a}$, corresponds to $X^g X\underline{\beta} = X^g X\underline{a}$, and $(I-X^g X)\underline{\beta} = (I-X^g X)\underline{a}$. Equivalently $X\underline{\beta} = X\underline{a}$ is testable and $(I-X^g X)\underline{\beta} = (I-X^g X)\underline{a}$ are *a priori* restrictions which are strictly non-testable.

Searle (1971, pp.188-204) provides an extensive summary of hypothesis testing, under $V = \sigma^2 I$. Drawing from it and the Pringle and Rayner development, we will write $H_0 : L\underline{\beta} = \underline{c}$ to indicate a testable hypothesis, assuming consistency, and estimability.

Under H_0, the restrictions are applied *a priori* in the model. Minimizing

(3.118) $(\underline{y}-X\underline{\beta})'(\underline{y}-X\underline{\beta})$

subject to $L\underline{\beta} = \underline{c}$, is equivalent to solving (3.144) for L as described here. Corollary 2.14.1 in (3.155), and (2.43) imply that

(4.8) $\hat{\underline{\beta}}_0 = (S^{g_1}-S^{g_1}L'R^{g_1}LS^{g_1})X'\underline{y} + S^{g_1}L'R^{g_1}\underline{c}$, whence

(4.9) $\hat{\underline{y}}_0 = X\hat{\underline{\beta}}_0 = XS^g X'\underline{y} - XS^g L'R^g(LS^g X'\underline{y}-\underline{c})$

$= X\hat{\underline{\beta}} - XS^g L'R^g(L\hat{\underline{\beta}}-\underline{c})$

uniquely over all g_1-inverses in (4.8). Here R is taken as in (2.125) and $X\hat{\underline{\beta}}$ is the OLS solution (3.14). Defining $\hat{\underline{\varepsilon}}_0$ as in (3.151) we have

$$(4.10) \qquad \hat{\underline{\varepsilon}}_0 = \hat{\underline{\varepsilon}} + XS^gL'R^g(L\hat{\underline{\beta}}-\underline{c}) \qquad , \qquad \text{with}$$

$$(4.11) \qquad \hat{\underline{\varepsilon}}_0'\hat{\underline{\varepsilon}}_0 = \hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}} + (L\hat{\underline{\beta}}-\underline{c})'R^g(L\hat{\underline{\beta}}-\underline{c}) \qquad .$$

We may rewrite $\underline{y}'\underline{y}$ as

$$(4.12) \qquad \underline{y}'\underline{y} = \underline{y}_0'\underline{y}_0 + (L\hat{\underline{\beta}}-\underline{c})'R^g(L\hat{\underline{\beta}}-\underline{c}) + \hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}} \qquad ,$$

and using the conventional definitions, write

$$(4.13) \qquad SS(E_0) = \hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}} + (L\hat{\underline{\beta}}-\underline{c})'R^g(L\hat{\underline{\beta}}-\underline{c})$$
$$= SS(E) + SS(H)$$

respectively, in the manner of Rao (1965, pp.155-157). He, and Searle (1965), derive SS(H) for L with full row-rank. Their results are generalized to

<u>Theorem 4.1</u> (Pringle and Rayner, 1971, pp.86-88)

Under the assumption of normality, the ratio

$$(4.14) \qquad F = \frac{SS(H)}{SS(E)} \cdot \frac{n-r}{s}$$

has central $F(s,n-r)$ distribution, with $s = r(L)$, subject to the hypothesis $H_0 : L\underline{\beta} = \underline{c}$.

Proof (Dunne): From (4.10) and (3.69) we have

$$(4.15) \qquad \hat{\underline{\varepsilon}}'(\hat{\underline{\varepsilon}}_0-\hat{\underline{\varepsilon}}) = 0 \qquad , \qquad \text{and}$$

$$(4.16) \qquad \text{var}(\hat{\underline{\varepsilon}}).(\hat{\underline{\varepsilon}}_0-\hat{\underline{\varepsilon}}) = \sigma^2.(I-XS^gX').(\hat{\underline{\varepsilon}}_0-\hat{\underline{\varepsilon}})$$
$$= \underline{0} \qquad . \qquad \text{Thus}$$

$$(4.17) \qquad \text{cov}(\hat{\underline{\varepsilon}},\hat{\underline{\varepsilon}}_0-\hat{\underline{\varepsilon}}) = 0 \qquad , \qquad \text{and}$$

(4.18) $SS(H) = (\hat{\underline{\varepsilon}}_0 - \hat{\underline{\varepsilon}})'(\hat{\underline{\varepsilon}}_0 - \hat{\underline{\varepsilon}}) = \hat{\underline{\varepsilon}}_0'\hat{\underline{\varepsilon}}_0 - \hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}}$

follows a central $\sigma^2.\chi^2$-distribution, degrees of freedom $s = r(L)$, independent of $SS(E)$, if and only if $SS(E_0)$ is central $\sigma^2.\chi^2_{n-r+s}$, under H_0.

By Theorem 3.15, $SS(E_0)$ follows $\sigma^2.\chi^2_f(\lambda)$ where $\lambda = 0$ under H_0, and f as given by (3.152). However since $\mathcal{R}(L) \subset \mathcal{R}(X)$, we have

(4.19) $$f = n - r\begin{bmatrix} X \\ L \end{bmatrix} + r(L) = n - r(X) + s$$
$$= n-r+s \quad .$$

The additivity property of χ^2-variates gives the result for (4.14). □

The imposition of a null hypothesis is effectively a reduction of the model to the form

(4.20) $$X\underline{\beta} \equiv X(S^g - S^gL'R^gLS^g)S\underline{\beta}^* + XS^gL'R^g\underline{c}$$
$$= (X - XS^gL'R^gL)\underline{\beta}^* + XS^gL'R^g\underline{c}$$
$$= X\underline{\beta}^* - XS^gL'R^g(L\underline{\beta}^* - \underline{c}) \quad .$$

Thus the total SS for the reduced model is given by

(4.21) $$\underline{y}_*'\underline{y}_* = (\underline{y} - XS^gL'R^g\underline{c})'(\underline{y} - XS^gL'R^g\underline{c})$$
$$= \underline{y}'\underline{y} + \underline{c}'R^g\underline{c} - 2\underline{c}'R^gLS^gX'\underline{y} \quad .$$

The corresponding NE's are

(4.22) $$(S - L'R^gL)\hat{\underline{\beta}}^* = X'\underline{y} - L'R^gLS^gX'\underline{y} \quad ,$$

and solving for $\hat{\underline{\beta}}^*$ we obtain

(4.23) $$\hat{\underline{y}}_* = (X - XS^gL'R^gL)\hat{\underline{\beta}}^* \quad ,$$ and

(4.24) $\quad \hat{\underline{y}}_0 = \hat{\underline{y}}_* + XS^gL'R^g\underline{c}$.

Note that $L\underline{\beta} = \underline{c}$ is easily verified in (4.20). Also

(4.25) $\quad \underline{y}'\underline{y} = \hat{\underline{y}}_0'\hat{\underline{y}}_0 + (\hat{\underline{\varepsilon}}_0-\hat{\underline{\varepsilon}})'(\hat{\underline{\varepsilon}}_0-\hat{\underline{\varepsilon}}) + \hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}}$, and

(4.26) $\quad \underline{y}_*'\underline{y}_* = \hat{\underline{y}}_*'\hat{\underline{y}}_* + (\hat{\underline{\varepsilon}}_0-\hat{\underline{\varepsilon}})'(\hat{\underline{\varepsilon}}_0-\hat{\underline{\varepsilon}}) + \hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}}$.

Equation (4.20) reduces to

(4.27) $\quad X\underline{\beta} \equiv (X-XS^gL'R^gL)\underline{\beta}$

if and only if $\underline{c} = \underline{0}$, from (3.166).

4.2 ORTHOGONAL HYPOTHESES

Suppose that we may partition the hypothesis as

(4.28) $$\begin{bmatrix} L_1 \\ L_2 \end{bmatrix}\underline{\beta} = \begin{bmatrix} \underline{c}_1 \\ \underline{c}_2 \end{bmatrix}$$

with L_1 and L_2 such that

(4.29) $$R = \begin{bmatrix} L_1 \\ L_2 \end{bmatrix} S^{g_1}[L_1' \ : \ L_2'] = \begin{bmatrix} L_1S^gL_1' & 0 \\ 0 & L_2S^gL_2' \end{bmatrix} .$$

$$= \begin{bmatrix} R_1 & 0 \\ 0 & R_2 \end{bmatrix}$$

Clearly the estimable functions $L_1\underline{\beta}$ and $L_2\underline{\beta}$ have estimators $L_1\hat{\underline{\beta}}$ and $L_2\hat{\underline{\beta}}$ which are uncorrelated and hence orthogonal in the statistical sense. From (4.13) we have

$$(4.30) \quad SS(H) = (L\hat{\underline{\beta}}-\underline{c})'R^g(L\hat{\underline{\beta}}-\underline{c})$$

$$= (L_1\hat{\underline{\beta}}-\underline{c}_1)'R_1^g(L_1\hat{\underline{\beta}}-\underline{c}_1) + (L_2\hat{\underline{\beta}}-\underline{c}_2)'R_2^g(L_2\hat{\underline{\beta}}-\underline{c}_2)$$

$$= SS(H_1) + SS(H_2)$$

where H_1 and H_2 represent the two hypotheses of (4.28). We note from (4.29) that

$$(4.31) \quad R(L_1) \cap R(L_2) = \{\underline{0}\} \quad ,$$

for otherwise orthogonality is contradicted. This condition is not in general sufficient for (4.29). From (4.18), (4.25) and (4.26) we have

$$(4.32) \quad \underline{y}'\underline{y} = \hat{\underline{y}}_0'\hat{\underline{y}}_0 + SS(H_1) + SS(H_2) + \hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}}$$

$$= \hat{\underline{y}}_1'\hat{\underline{y}}_1 + SS(H_1) + \hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}}$$

$$= \hat{\underline{y}}_2'\hat{\underline{y}}_2 + SS(H_2) + \hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}}$$

where $\hat{\underline{y}}_i'\hat{\underline{y}}_i$ is the sum of squares of the fitted values under the hypothesis $H_i : L_i\underline{\beta} = \underline{c}_i$, for $i = 1,2$. By applying Theorem 4.1 to each case we have independent F-tests of the sub-hypotheses in question. Since the above decompositions of $SS(H)$ and $\underline{y}'\underline{y}$ extend to partitionings of L with up to $r = r(X)$ submatrices L_i in (4.28) through to (4.32), we may have the convenience of simultaneously testing many hypotheses which throw light upon underlying relations. For instance the foregoing allows the testing jointly and severally of linear, quadratic, cubic, ... effects of a factor at equally spaced levels. In that case the matrices L_i also correspond to the algebraically orthogonal row-eigenvectors of S, as well as being statistically orthogonal.

The question arises as to when (4.31) is sufficient for (4.29). Certainly if there exists A with

(4.33) $\mathcal{R}(A) \cap \mathcal{R}(L) = \{\underline{0}\}$, and

(4.34) $A'A + L'L = S$,

then Theorem 2.10 extends in a straightforward way to ensure that

(4.35) $$\begin{bmatrix} A \\ L_1 \\ L_2 \end{bmatrix} S^{g_1} [A' : L_1' : L_2'] = \begin{bmatrix} AS^gA' & 0 & 0 \\ 0 & L_1S^gL_1' & 0 \\ 0 & 0 & L_2S^gL_2' \end{bmatrix}$$

and (4.29) follows. We conjecture that the existence of such an A is necessary.

Seber (1980, pp.40-58) discusses orthogonal hypotheses in nested procedures and in experimental designs. A more general approach is given by Searle (1971, pp.199-204), but is restricted to taking $\underline{c} = \underline{0}$ in H_0.

4.3 NON-TESTABLE HYPOTHESES

It will be of interest in Section 4.4 to examine certain hypotheses which are strictly non-testable. Searle (1971, p.195) states that a test of hypotheses involving some non-estimable functions is only a test of the hypotheses made up of just the estimable functions in the original set. A proof is sketched for a single non-estimable function, with the estimable set strictly taken with full rank. The following theorems extend the result for consistent restrictions of any order and rank. They also establish algebraically the uniqueness of solutions over equivalent restrictions. The uniqueness is however widely known and usually handled by an appeal to geometric notions, such as the hyperspace determined by the restrictions.

Theorem 4.2 (Dunne)

Given a testable hypothesis

(4.36) $H_0 : L_1\underline{\beta} = \underline{c}_1$,

then under H_0, the fitted values $\hat{\underline{y}}_0$ from (4.24) are invariant over all additional restrictions

(4.37) $M\underline{\beta} = \underline{k}$, where

(4.38) $R(M) \cap R(X) = \{\underline{0}\}$.

Further, the sum of squares associated with the restrictions (4.36) and (4.37) is invariant over the restrictions and hence takes precisely the value associated with H_0 alone.

Proof: Since the joint restrictions

(4.39) $$L\underline{\beta} = \begin{bmatrix} L_1 \\ M \end{bmatrix} \underline{\beta} = \begin{bmatrix} \underline{c}_1 \\ \underline{k} \end{bmatrix} = \underline{c}$$

intrude into the space of estimable functions, we require the theory of Section 3.5. By applying Theorems 3.14 and 3.15 with (4.39) we have

(4.40) $$\hat{\underline{y}}_0 = XG_{11}X'\underline{y} + XK^g[L_1' : M']R^g \begin{bmatrix} \underline{c}_1 \\ \underline{k} \end{bmatrix}$$, where

(4.41) $K = [X' : L_1' : M']\begin{bmatrix} X \\ L_1 \\ M \end{bmatrix} = S + L_1'L_1 + M'M$. Now

(4.42) $R = \begin{bmatrix} L_1 \\ M \end{bmatrix} K^g [L_1' : M']$

$= \begin{bmatrix} L_1 K^g L_1' & 0 \\ 0 & MK^g M' \end{bmatrix}$

$= \begin{bmatrix} R_1 & 0 \\ 0 & R_2 \end{bmatrix}$, say

by virtue of Theorem 2.10 and

(4.43) $\begin{bmatrix} X \\ L_1 \end{bmatrix} K^g M' = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. Thus

(4.44) $R^g = \begin{bmatrix} R_1^g & 0 \\ 0 & R_2^g \end{bmatrix}$

by Theorem 2.13. Now, by (4.43),

(4.45) $XK^g[L_1' : M'] = XK^g[L_1' : 0]$, and

(4.46) $XG_{11}X' = XK^gX' - XK^gL_1'.R_1^g.L_1K^gX'$.

To complete the proof for $\hat{\underline{y}}_0$ we need only show that K^g may be replaced throughout (4.40) to (4.46) by $(S+L_1'L_1)^g$. This follows easily from Theorem 2.10 since K^g is a g_1-inverse of $(S+L'L)$, and thus

(4.47) $XK^gX' = X(S+L_1'L_1)^{g_1}X' = X(S+L_1'L_1)^gX'$,

from the invariance property (2.51) in Corollary 2.8.3. In turn, since L_1 is in $R(X)$

(4.48) $\quad XK^gL_1' = X(S+L_1'L_1)^gL_1' \quad .$

From (3.151) $\hat{\underline{\varepsilon}}_0$ is invariant over all choices of M in (4.37) and (4.38). By (4.12), (4.18) and (4.26), we have that

$$(4.49) \quad SS(H) = (L_1\hat{\underline{\beta}}-\underline{c}_1)'.R^g.(L_1\hat{\underline{\beta}}-\underline{c}_1)$$
$$= (L\hat{\underline{\beta}}_0-\underline{c})'.R^g.(L\hat{\underline{\beta}}_0-\underline{c}_1)$$

for $\hat{\underline{\beta}}$ any solution to the unrestricted normal equations, and $\hat{\underline{\beta}}_0$ is any $\hat{\underline{\beta}}$ which also satisfies (4.37). □

Corollary 4.2.1 (Dunne)

The F-statistic from Theorem 3.17, namely

$$(3.169) \quad F = \frac{\hat{\underline{\varepsilon}}_0'\hat{\underline{\varepsilon}}_0-\hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}}}{\hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}}} \cdot \frac{f_2}{f_1}$$

may be interpreted as the statistic associated with a null hypothesis $H_0 : L\underline{\beta} = \underline{c}$ as in (4.39), regardless of whether of not $L\underline{\beta}$ is strictly estimable.

We note that unless the separation is effected of the hypothesis functions $L\underline{\beta}$ into estimable and non-estimable parts, formulae such as (3.178) have to be used to calculate SS(H). Use of (4.11) will not in general lead to a unique quantity unless $\hat{\underline{\beta}}$ satisfies the non-estimability restrictions (4.37).

If $H_0 : H\underline{\beta} = \underline{h}$ is not of the form (4.39) it is clear that there exist matrices A, of many orders, such that

(4.50) $AH\underline{\beta} = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} H\underline{\beta} = \begin{bmatrix} L_1 \\ M \end{bmatrix} \underline{\beta} = \begin{bmatrix} A_1 & \underline{h} \\ A_2 & \underline{h} \end{bmatrix} = \begin{bmatrix} \underline{c}_1 \\ \underline{k} \end{bmatrix}$ has

(4.51) $R(A_1H) = R(H) \cap R(X)$, and

(4.52) $R(A_2H) \cap R(X) = \{\underline{0}\}$.

Theorem 4.2 applied to (4.39) gives an equivalent test to $H_0 : H\underline{\beta} = \underline{h}$, because the fitted values for H_0 and for (4.50) coincide. The solutions $\hat{\underline{\beta}}^*$ to the normal equations subject to H_0 form precisely the same set as solutions $\hat{\underline{\beta}}^*$ subject to (4.50), as we will shortly establish in Lemma 4.4. Moreover, in view of Theorem 4.2, we need only the matrix A, and in that case (4.49) becomes

(4.53) $SS(H) = (H\hat{\underline{\beta}}-\underline{h})'A_1'R_1^g.A_1(H\hat{\underline{\beta}}-\underline{h})$, for

(4.54) $R_1 = A_1H(A_1HS^gH'A_1')^gH'A_1'$

$= L_1(L_1S^gL_1')L_1'$.

It will also transpire that, from Lemmas 4.3 and 4.4,

(4.55) $R_1^{g_1} = B'(HS^gH')^gB$, for some B, and

(4.56) $SS(H) = (L_1\hat{\underline{\beta}}-\underline{c}_1)'R_1^g(L_1\hat{\underline{\beta}}-\underline{c}_1)$

$= (L_1\hat{\underline{\beta}}-\underline{c}_1)'(L_1S^gL_1')^g(L_1\hat{\underline{\beta}}-\underline{c}_1)$,

by substitution, and (2.52) in Corollary 2.8.3.

<u>Lemma 4.3</u> (Pringle and Rayner, 1971; p.50)

Consider S positive semidefinite with $L \subset R(S)$ and K and R as given by

(2.121) $K = S + L'L$,

(2.122) $R = LK^{g_1}L'$. Then

(4.57) $R^g = (LS^gL')^g(I+LS^gL') = (R^g)'$.

Proof: From Theorem 2.11

(4.58) $LK^gL' = LS^gL' - LS^gL'(I+LS^gL')^{-1}LS^gL'$

$= (LS^gL')(I+LS^gL')^{-1}(I+LS^gL'-LS^gL')$

$= (LS^gL')(I+LS^gL')^{-1}$.

By Theorem 2.2, (4.57) follows. □

Lemma 4.4 (Dunne)

Consider S, L, K and R as in Lemma 4.3. Let

(4.59) $K_0 = S + L'A'AL$

for any A such that

(4.60) $\mathcal{R}(AL) = \mathcal{R}(L)$. If

(4.61) $R_0 = ALK_0^g L'A'$, then

(4.62) $L'A'.R_0^gAL - L'A'AL = L'.R^gL - L'L$.

Proof: From (4.60) there exists B such that

(4.63) $BAL = L$.

By Lemma 4.3 we may simplify (4.61) to

(4.64) $R_0 = ALS^gL'A'(I+ALS^gL'A')^{-1}$, and

(4.65) $R_0^g = (ALS^gL'A')^g(I+ALS^gL'A')$.

A g_1-inverse of $(ALS^gL'A')$ is given by

(4.66) $(ALS^gL'A')^{g_1} = B'(LS^gL')^gB$. Thus

(4.67) $L'A'.R_0^gAL = L'A'.R_0^{g_1}AL$

$= L'A'B'(LS^gL')^gB(I+ALS^gL'A')AL$

$$= L'(LS^gL')^gL + L'A'AL \quad .$$

Similarly

(4.68) $$L'R^gL = L'(LK^gL')L$$

$$= L'(LS^gL')^g(I+LS^gL')L$$

$$= L'(LS^gL')^gL + L'L$$

The result follows. □

Corollary 4.4.1 (Dunne)

The form $XG_{11}X'$, for G_{11} the leading submatrix of (2.123), is unique over all equivalent restrictions.

Proof: We may equivalently examine

(4.69) $$SG_{11}S = SK^gS - SK^gL'R^gLK^gS$$

$$= (S-SK^gL'L) - SK^g(L'R^gL-L'L)$$

$$= S - (K-L'L)K^g(L'R^gL)$$

$$= S + L'L - L'R^gL \quad .$$

By (4.62), any equivalent set of restrictions involving say AL, must reduce to (4.69). □

This implies that all the fitted values are invariant. The $\hat{\underline{\beta}}^*$ are also equivalent sets because we may find a non-singular map of $G_{11}X'$ to the matrices

(4.70) $$K_0^gX' - K_0^gL'A'R_0^gALK_0^gX'$$

using (4.67) and the equality of the spaces $R(K)$ and $R(K_0)$. The effect of differing choices of $\underset{\sim}{M}$ needs no investigation. It is known from (4.45) and (4.46) that no changes are imposed on the fitted values.

Nonetheless the argument for the invariance of the set of solutions $\hat{\underline{\beta}}^*$ over all equivalent restrictions can be simply derived from a comparison of the form

(3.144) $$\begin{bmatrix} S & H' \\ H & 0 \end{bmatrix} \begin{bmatrix} \hat{\underline{\beta}} \\ \underline{\lambda} \end{bmatrix} = \begin{bmatrix} X'\underline{y} \\ \underline{c} \end{bmatrix} \quad , \qquad \text{and}$$

(4.71) $$\begin{bmatrix} S & H'A' \\ AH & 0 \end{bmatrix} \begin{bmatrix} \hat{\underline{\beta}} \\ B'\underline{\lambda} \end{bmatrix} = \begin{bmatrix} X'\underline{y} \\ A\underline{c} \end{bmatrix} \quad .$$

It is also apparent that g_1-inverses may replace g-inverses throughout (4.40) to (4.70).

If tests of hypotheses are constructed by means of Theorem 3.17 and the results of this section, the criterion must be understood as the sum of squares associated with the change in fitted values, and not as the sum of squares for deviations of $L\hat{\underline{\beta}}$ from $\underline{c}$. These two interpretations coincide only for the case of strongly testable hypotheses.

4.4 PARTITIONED LINEAR MODELS

We return to (4.7) and the remarks on testability, and examine the consequences on partitioned linear models of the form

(4.72) $$\underline{y} = X\underline{\beta} + \underline{\varepsilon} = [X_1 : X_2] \begin{bmatrix} \underline{\beta}_1 \\ \underline{\beta}_2 \end{bmatrix} + \underline{\varepsilon} \quad ,$$

a modification of (1.1). Such a conformable partition may reflect a natural order of complexity in the model, e.g. when X_2 corresponds to interactions in a factorial design, or an analysis of covariance problem. It may also reflect an intuitive or subjective partitioning of the regressors in a regression analysis. Further, it may simply correspond to a convenient partitioning which has algebraic advantages in the solution of normal equations and related operations. In any event, the conformably partitioned normal equations are

$$(4.73)\qquad \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \begin{bmatrix} \underline{\hat{\beta}}_1 \\ \underline{\hat{\beta}}_2 \end{bmatrix} = \begin{bmatrix} X_1'\underline{y} \\ X_2'\underline{y} \end{bmatrix} = \begin{bmatrix} \underline{g}_1 \\ \underline{g}_2 \end{bmatrix} \quad .$$

Pringle and Rayner (1971, pp.101-102) discuss premultiplication by

$$(4.74)\qquad \begin{bmatrix} I & 0 \\ -C & I \end{bmatrix}$$

for suitable C, such that a set of equations in $\underline{\hat{\beta}}_2$ eliminating $\underline{\hat{\beta}}_1$ can be formed. Such a choice of C is always possible as

$$(4.75)\qquad C = S_{21}S_{11}^{g_1} \quad .$$

This method generalizes the *sweep-out* method of Anderson and Bancroft (1952, p.280) or the *pivotal condensation* of Rao (1962) and Rohde (1964, pp.53-54).

A solution $\underline{\hat{\beta}}_2$ to the resulting consistent equations

$$(4.76)\qquad S_{22}^{*}\underline{\hat{\beta}}_2 = (S_{22}-S_{21}S_{11}^{g_1}S_{12})\underline{\hat{\beta}}_2 = \underline{g}_2^{*}$$

$$= \underline{g}_2 - X_2'(X_1S_{11}^{g_1}X_1')\underline{y}$$

$$= X_2'(I-X_1S_{11}^{g}X_1')\underline{y}$$

is a solution for the $\underline{\hat{\beta}}_2$ sub-vector in (4.73). Now

(4.77) $\quad \hat{\underline{\beta}}_2 = (S^*_{22})^{g_1}\underline{g}^*_2$ yields

(4.78) $\quad \hat{\underline{\beta}}_1 = (S^{g_1}_{11})\underline{g}_1 - (S^{g_1}_{11})S_{12}\hat{\underline{\beta}}_2$

$$= \underline{b}_1 - S^{g_1}_{11}S_{12}\hat{\underline{\beta}}_2 \quad ,$$

where $\underline{b}_1$ is a solution of the reduced model normal equations obtained from

(4.79) $\quad \underline{y} = X_1\underline{\beta}_1 + \underline{\varepsilon} \quad .$

Similarly, premultiplication by

(4.80) $\quad \begin{bmatrix} I & -S_{12}S^{g_1}_{22} \\ 0 & I \end{bmatrix}$

gives corresponding results for $\hat{\underline{\beta}}_2$ in terms of $\underline{b}_2$ obtained from the alternative reduced model. These results amount to two-stage techniques in solving normal equations. The nature of (4.77) and (4.78) imply economies of calculation in situations where augmenting a model with regressors, or deleting parameters, is under examination. We note that

$\underline{b}_i$ is a solution for $\underline{\beta}_i$ fitted alone (i = 1,2,)
$\hat{\underline{\beta}}_i$ is a solution for $\underline{\beta}_i$ simultaneously fitted last,

which explains the adjustments in (4.78). The partitioning process *per se* may be repeatedly applied to successive model partitions and their corresponding normal equations.

The usual problem associated with model (4.72) is whether or not $X_2\underline{\beta}_2$ may be dropped from the model. Again Pringle and Rayner (1971, p.104) suggest that this is a test of $\underline{\beta}_2 = \underline{0}$, whereas Searle (1971, p.189) follows the notion of strong testability and insists that the test is that of

$X_2\underline{\beta}_2 = \underline{0}$. Further, in the context of model (4.72), we may adapt his notation (p.246) and consider the problem as that of assessing the sums of squares for regression on $\underline{\beta}_2$ after fitting $\underline{\beta}_1$, i.e. $R(\underline{\beta}_2|\underline{\beta}_1)$. This notation reflects the nature of interpretation problem.

The model may now be considered subject to $\underline{\beta}_2 = \underline{0}$. Then, similarly to (4.7), we have

$$(4.81) \qquad \begin{bmatrix} X_2 \\ (I-X_2^g X_2) \end{bmatrix} \underline{\beta}_2 = \underline{0} \qquad ,$$

which may be represented by a reduced set of equations,

$$(4.82) \qquad A\underline{\beta}_2 = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \underline{\beta}_2 = \underline{0}$$

where A has full row-rank and $R(A_1) = R(X_2)$. However the function $X_2\underline{\beta}_2$ is not in general estimable. So that (4.39) may be replaced by

$$(4.83) \qquad \begin{bmatrix} (I-X_1S_{11}^g X_1')X_2 \\ I-S_{22}^{*g} S_{22}^* \end{bmatrix} \underline{\beta}_2 = \underline{0} \qquad , \qquad \text{where}$$

$$(4.84) \qquad S_{22}^* = X_2'(I-X_1S_{11}^{g_1}X_1')^2 X_2 = X_2'(I-X_1S_{11}^g X_1')X_2$$

and an appropriate reduction to a form of (4.71) can follow. Essentially we obtain the desired separation into strongly testable and other subhypotheses, for the general hypothesis $\underline{\beta}_2 = 0$. In (4.83), since

$$(4.85) \qquad (I-X_1S_{11}^g X_1')[X_1 : X_2] = [0 : (I-X_1S_{11}^g X_1')X_2]$$

we have a set of restrictions in estimable functions, whose rank is given by

(4.86) $r[X_2'(I-X_1S_{11}^gX_1')^2X_2] = r(S_{22}^*)$.

Noting that in Theorem 3.17 we obtain

(4.87) $f_1 = r(S_{22}^*)$, we have

(4.88) $f_2 = n-r[X_1 : X_2]$

$= n-r(X_1) - r(S_{22}^*)$.

Thus we obtain

Theorem 4.5 (Pringle and Rayner, 1971, pp.104-106)

The F-statistic for $H_0 : \underline{\beta}_2 = \underline{0}$ is given by

(4.89) $$\frac{\hat{\underline{\beta}}_2' S_{22}^* \hat{\underline{\beta}}_2}{\underline{y}'\underline{y}-\hat{\underline{\beta}}'S\hat{\underline{\beta}}} \cdot \frac{f_2}{f_1}$$

Proof: Essentially we need only substitute equivalent forms into (3.169). Solving for $\hat{\underline{\beta}}$ from (4.73),

(4.90) $\hat{\underline{\varepsilon}}_0'\hat{\underline{\varepsilon}}_0 - \hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}} = \hat{\underline{\beta}}_2' S_{22}^* \hat{\underline{\beta}}_2$

proves the result. □

In terms of underlying partitions of sums of squares as in (4.2) we note

(4.91) $\underline{y}'\underline{y} = \hat{\underline{y}}'\hat{\underline{y}} + \hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}} = \hat{\underline{\beta}}'S\hat{\underline{\beta}} + \hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}}$

$= (\underline{b}_1'S_{11}\underline{b}_1 + \hat{\underline{\beta}}_2'S_{22}^*\hat{\underline{\beta}}_2) + \hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}}$

$= \underline{b}_1'S_{11}\underline{b}_1 + (\hat{\underline{\beta}}_2'S_{22}^*\hat{\underline{\beta}}_2 + \hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}})$

for $\underline{b}_1$ as in (4.78). Thus $\hat{\underline{\beta}}_2'S_{22}^*\hat{\underline{\beta}}_2$ is the increase in the SS for residual values if $\underline{\beta}_2$ is dropped from the model (4.72), and is the increase in SS for fitted values when (1.1) is augmented by X_2 to form the extended model.

Zyskind (1964) noted that the equations (4.76) effectively give all the information to typify BLUE's of functions $\underline{\theta}'\underline{\beta}_2$. The hypothesis in Theorem 4.5 is therefore equivalent to $H_0 : S^*_{22}\underline{\beta}_2 = \underline{0}$, a strongly testable hypothesis in the full model (4.73).

A different approach of Tukey (1949, 1955) extends the nature of X_2 from a set of constants which may or may not be functionally dependent on the entries of X_1 (as for interaction effects or additional independent variates), to a function of $X_1\underline{\beta}_1$. The one degree-of-freedom test for non-additivity is based upon a model such as

(4.92) $$y_{ij} = \mu + \alpha_i + \beta_j + \lambda\alpha_i\beta_j + \varepsilon_{ij} \quad ,$$

or, equivalently, for suitable matrices,

(4.93) $$\underline{y} = [X_1 : \underline{x}_2]\begin{bmatrix}\underline{\beta}_1 \\ \lambda\end{bmatrix} + \underline{\varepsilon} \quad .$$

Milliken and Graybill (1970) have generalized the Tukey test to examine a model

(4.94) $$\underline{y} = [X_1 : X_2]\begin{bmatrix}\underline{\beta}_1 \\ \underline{\beta}_2\end{bmatrix} + \underline{\varepsilon}$$

where

(4.95) $$X_2 = X_2(X_1\underline{\beta}) = [f_{ij}(X_1\underline{\beta})] \quad .$$

To obtain X_2 they substitute $X_1\underline{b}_1$ obtained from the reduced model, into (4.95) to obtain

(4.96) $$\hat{X}_2 = [f_{ij}(X_1\underline{b}_1)] \quad ,$$ and

(4.97) $$\underline{y} = [X_1 : \hat{X}_2]\underline{\beta} + \underline{\varepsilon} \quad .$$

To test the hypothesis $\underline{\beta}_2 = \underline{0}$ in (4.94), by Theorem 3.17 is equivalent to testing

(4.98) $\quad H_0 : (I-X_1S^g_{11}X_1')X_2\underline{\beta} = \underline{0}$.

Then since with probability one we have

(4.99) $$r[X_1 : X_2] = r[X_1 : \hat{X}_2]$$
$$= r(X_1) + r[(I-X_1S^g_{11}X_1')\hat{X}_2]$$
$$= r(X_1) + r(\hat{S}^*_{22})$$

Theorem 4.5 holds with

(4.100) $$F = \frac{\hat{\underline{\beta}}_2'\hat{S}^*_{22}\hat{\underline{\beta}}_2}{\underline{y}'\underline{y}-\hat{\underline{\beta}}'S\hat{\underline{\beta}}} \cdot \frac{f_2}{f_1}$$, and

(4.101) $$f_1 = r(\hat{S}^*_{22}) = r(S^*_{22})$$.

Rao (1971, p.251) interprets this generalization in terms of an alternative model. Let

(4.102) $$\hat{\underline{e}} = (I-X_1S^g_{11}X_1')\underline{y}$$

be the residuals after fitting only the $X_1\underline{\beta}_1$ term in the model (4.94). Taking $\hat{X}_2$ as in (4.96), let

(4.103) $$M = (I-X_1S^g_{11}X_1')\hat{X}_2$$.

Consider the new model as

(4.104) $$\hat{\underline{e}} = M\underline{\beta}_2 + \underline{\varepsilon}_0$$

where $\underline{\varepsilon}_0$ has variance-covariance structure $\sigma^2.I$. Then (4.100) is a test of $\underline{\beta}_2 = \underline{0}$. He points out that $\hat{X}_2$ and $\hat{\underline{e}}$ are independent, since by (4.96) $\hat{X}_2$ is a function of $X_1\underline{b}_1$ and it is independent of $\hat{\underline{e}}$. However, since strictly speaking $\hat{\underline{e}}$ is in $C(I-X_1S^g_{11}X_1')$, $\underline{\varepsilon}_0$ has variance-covariance structure $\sigma^2(I-X_1S_{11}^{g}X_1')$, and for Rao's interpretation to be verified we require the theory for arbitrary variance matrices in the linear model. That theory is reviewed in Chapter 5. It transpires that the definition of $\hat{\underline{e}}$ in (4.102) and the idempotency of $(I-X_1S_{11}^{g}X_1')$ provide the justification.

Seber (1980, pp.59-60) also discusses modified hypotheses and notes that the assumptions of model (4.73) immediately imply that

$$\text{(4.105)} \qquad \text{cov}(X_1\underline{b}_1,\ (I-X_1S_{11}^{g}X_1')X_2\hat{\underline{\beta}}_2) = 0 \quad .$$

The model assumptions are expressed as space-conditions, so that Seber (*op. cit.*) in taking X_1 and X_2 of full column-rank, can write

$$\text{(4.106)} \qquad \text{cov}(\underline{b}_1,\ \hat{\underline{\beta}}_2) = 0 \quad .$$

From (4.76), (4.77), (4.91) and (4.102) we may write

$$\text{(4.107)} \qquad \hat{\underline{\beta}}_2'S_{22}^{*}\hat{\underline{\beta}}_2 = \underline{g}_2^{*\prime}(S_{22}^{*})^{g}\underline{g}_2^{*} = \hat{\underline{e}}'X_2(S_{22}^{*})^{g}X_2'\hat{\underline{e}} \quad , \qquad \text{and}$$

$$\text{(4.108)} \qquad \underline{y}'\underline{y} - \hat{\underline{\beta}}'S\hat{\underline{\beta}} = \underline{y}'\underline{y} - (\underline{b}_1'S_{11}\underline{b}_1 + \hat{\underline{\beta}}_2'S_{22}^{*}\hat{\underline{\beta}}_2) = \hat{\underline{e}}'\hat{\underline{e}} - \hat{\underline{\beta}}_2'S_{22}^{*}\hat{\underline{\beta}}_2 = \hat{\underline{e}}'\hat{\underline{e}} - \hat{\underline{e}}'X_2(S_{22}^{*})^{g}X_2'\hat{\underline{e}} \quad .$$

Now, following from John and Draper (1978) we may take X_2 of the form

$$\text{(4.109)} \qquad X_2 = \begin{bmatrix} 0 \\ I \end{bmatrix} \quad ,$$

and obtain a special case in which F serves as a test-statistic for additive outlier effects. In view of (4.89) the F-value is the same whether the full model is fitted first, or whether the reduced model is used with (4.107) and (4.108). We deal with the related theory in Chapter 6. Note that Theorem 3.17 implies that it is not necessary for S^*_{22} to have full column-rank, and thus, equivalently, in the model

$$(4.110) \qquad \underline{y} = \begin{bmatrix} \underline{y}_1 \\ \underline{y}_2 \end{bmatrix} = \begin{bmatrix} X_{11} & 0 \\ X_{12} & I \end{bmatrix} \begin{bmatrix} \underline{\beta}_1 \\ \underline{\beta}_2 \end{bmatrix} + \underline{\varepsilon} \qquad ,$$

it is not necessary that

(i) $I\underline{\beta}_2$ is estimable, or

(ii) $X_{12} \subset \mathcal{R}(X_{11})$.

Consequently we have a test of the effect of incorporating $\underline{y}_2$ against ignoring it as data appropriate to the model.

4.5 ANALYSIS OF COVARIANCE

In general the term 'analysis of variance' is applied to analyses, such as those arising from experimental designs, in which the entries of the X-matrix in a model represent qualitative distinctions between observations. When the entries correspond to quantitative differences between observations, the usual descriptive term for the same operations as above is 'regression analysis'. In either case the columns of the X-matrix describe qualitative or quantitative 'factors', or exogenous (independent) variates, respectively. The term 'analysis of covariance', describes a situation in which a model is applied that involves a mixture of both types of factors. As such we have a special case of the model (4.72), with X_1 corresponding to qualitative effects and X_2 to an additional quantitative variate. Seber (1977, pp.279-301; 1980, pp.61-65) gives a summary and examples.

Since it is reasonable to assume that

(4.111) $r[X_1 : X_2] = r(X_1) + r(X_2)$, with

(4.112) $r(X_2) = k_2$

where X_2 is $n \times k_2$, S^*_{22} will in general be non-singular, and $\underline{\beta}_2$ has a BLUE given by

(4.113) $\hat{\underline{\beta}}_2 = (S^*_{22})^{-1}\underline{g}^*_2$.

The test of hypothesis generated by (4.100) is therefore a test of the regression effects in the full model. In view of (4.76) and hence (4.107) and (4.108) it is easily seen that such a test

(i) examines the additional sums of squares due to fitting the quantitative variable, and

(ii) examines the regression of the residuals (after fitting qualitative factors) on the variables X^*_2 with

(4.114) $X^*_2 = (I - X_1 S^g_{11} X'_1) X_2$,

the (residual) orthogonal part of X_2.

If the regression effects are significant, then we may change stance as follows : allow X_1 to represent the quantitative factors, and X_2 the qualitative factors, then the F-test of (4.100) represents the additional effects of the qualitative factors after adjustment for regression. Equivalently it reflects the regression of the quantitative residuals on the (residual) orthogonal part of X_2. Moreover since X_2 is often an experimental design matrix, it may itself be subject to a partitioning. For instance, in a randomized blocks design, X_2 will typify block and treatment effects. Then equation (4.30) allows the orthogonal partitioning of the sum of squares due to the adjusted experimental effects, and corresponding separate F-tests of adjusted block and of adjusted treatment effects.

It is interesting to note that the foregoing remarks on analysis of covariance, and indeed, as is clear from (4.107), all examinations of additional effects in partitioned models, may be thought of as models

$$(4.115) \qquad \underline{y} = (I - X_1 S_{11}^{g} X_1') X_2 \underline{\beta} + \underline{\varepsilon}$$

$$= X_2^{*} \underline{\beta} + \underline{\varepsilon}$$

where $\text{var}(\underline{\varepsilon}) = \sigma^2 (I - X_1 S_{11}^{g_1} X_1')$. As with earlier remarks about statements of Rao (1971) recorded below (4.104), we may not take $\text{var}(\underline{\varepsilon}) = \sigma^2 . I$ without in general altering the sum of squares of the observations from $\underline{y}'\underline{y}$ to $\underline{y}'(I - X_1 S_{11}^{g_1} X_1')\underline{y}$.

Atkinson (1969) discussed the use of residuals as a concomitant variable. The yields of successive or adjacent experimental units may be affected by serial correlations. An analysis of covariance procedure is shown to give similar results to maximum likelihood estimators in a special case. Further research with two coefficients of correlation (e.g. one for rows and one for columns) may generalize the result. However problems of over-parametrization may arise if large numbers of residual concomitants are included. It is well-known in econometric theory that identifiability restraints must then be fitted, and it is not clear what restraints might be appropriate in the general experimental situation.

4.6 MISSING OBSERVATIONS AND ADDITIONAL DATA

Suppose that in the model (1.1), the partitioning

$$\begin{bmatrix} \underline{y}_1 \\ \underline{y}_2 \end{bmatrix} = \begin{bmatrix} X_{11} \\ X_{12} \end{bmatrix} \underline{\beta} + \begin{bmatrix} \underline{\varepsilon}_1 \\ \underline{\varepsilon}_2 \end{bmatrix} \tag{4.116}$$

has the observations $\underline{y}_2$ either inadequately observed, or lost. In an experimental design the fact that these observations are missing may have inconvenient consequences for estimation, hypothesis-testing and orthogonal partitioning of sums of squares. For instance, the matrix

$$S = X'X = [X'_{11} : X'_{12}] \begin{bmatrix} X_{11} \\ X_{12} \end{bmatrix} = (S_{11} + S_{22}) \tag{4.117}$$

usually has important design properties, and in general the matrix S_{11} will not preserve those properties. It is nonetheless possible to examine the observations directly under the model

$$\underline{y}_1 = X_{11}\underline{\beta} + \underline{\varepsilon}_1 \quad , \tag{4.118}$$

and test appropriate hypotheses in the strong or weak sense.

Before the advent of the electronic calculator or computer this course of action would have given rise to considerable arithmetic problems. Yates (1933) suggested the device of finding the minimum value of

$$(4.119) \quad \underline{y}'(I-XS^{g}X')\underline{y} = (\underline{y}-X\hat{\underline{\beta}})'(\underline{y}-X\hat{\underline{\beta}})$$

at $\underline{y}_2 = \hat{\underline{y}}_2$, and a suitable $\hat{\underline{\beta}}$ value. Then model (4.116) is applied with $\hat{\underline{y}}_2$ in place of the missing $\underline{y}_2$ and the conventional analysis is facilitated. The sub-vector $\hat{\underline{y}}_2$ is however not unique unless $X_{12}\underline{\beta}$ is estimable in the model (4.118), i.e. $X_{12} \subset R(X_{11})$. Seber (1977, p.291) expresses this property as

$$(4.120) \quad r(X_{11}) = r(X) = r\begin{bmatrix} X_{11} \\ X_{12} \end{bmatrix} \quad .$$

In that case, Cramer (1972) predicts the missing values as

$$(4.121) \quad \underline{y}_2 = X_{12}\underline{\beta} = X_{12}(X_{11}'X_{11})^{g_1}X_{11}'\underline{y} \quad .$$

By (4.117) adding $X_{11}'\underline{y}_1$ to $X_{12}'\hat{\underline{y}}_2$ yields

$$(4.122) \quad S\hat{\underline{\beta}} = X'\underline{y} = X'\begin{bmatrix} \underline{y}_1 \\ \hat{\underline{y}}_2 \end{bmatrix} \quad ,$$

we obtain $\hat{\underline{\beta}}$ as a solution to both the substituted form of (4.116) and (4.118), though $\hat{\underline{\beta}}$ is not necessarily unique. However given a choice of $\underline{y}_2$, $X\hat{\underline{\beta}}$ is unique and we may note that the residuals for the substituted vector in (4.122) are all zero, and that the residual sum of squares is precisely that of the model (4.118). Special cases of these results have been examined by Wilkinson (1958) and Cochran and Cox (1957) who give formulae for estimating a single missing observation in any one of a wide set of experimental designs.

Bartlett (1937), as reported in Seber (1977, p.297), suggested that $\underline{y}_2$ should be set equal to arbitrary values and that model (4.110) be then applied to the augmented data. The influence of the arbitrary values in removed by a co-variance analysis on the dummy concomitant variables. Specifically, if we take the arbitrary values as zero and thus examine

$$(4.123)\qquad \begin{bmatrix} \underline{y}_1 \\ \underline{0} \end{bmatrix} = \begin{bmatrix} X_{11} & 0 \\ X_{12} & I \end{bmatrix} \begin{bmatrix} \underline{\beta}_1 \\ \underline{\beta}_2 \end{bmatrix} + \begin{bmatrix} \underline{\varepsilon}_1 \\ \underline{\varepsilon}_2 \end{bmatrix}$$

we obtain

$$(4.124)\qquad \hat{\underline{\beta}}_2 = -X_{12}\hat{\underline{\beta}}_1 \qquad ,$$

which is unique if and only if (4.120) holds. It follows that we may take $-\hat{\underline{\beta}}_2$ as the estimated missing value in that case. However, if (4.120) fails as when an entire block of an incomplete block design is missing, no unique estimates are possible.

Once $\hat{\underline{y}}_2$ has been established it cannot be treated as an unchangeable estimate. For instance if an hypothesis test $H_0 : L\underline{\beta} = \underline{c}$ is to be applied to the observations, it will not be legitimate to apply H_0 on the vector $\begin{bmatrix} \underline{y}_1 \\ \hat{\underline{y}}_2 \end{bmatrix}$ unless $\hat{\underline{y}}_2$ is itself constructed from a value of $\hat{\underline{\beta}}$ which satisfies H_0. In such a case it will be found that the test statistic is biased upwards, as noted by Seber (1977, p.293). We may explain this fact as follows, in terms of the g-inverse approach. From (4.123) we solve the equations

$$(4.125)\qquad S^*\underline{\beta} = \begin{bmatrix} S & X_{12}' \\ X_{12} & I \end{bmatrix}\begin{bmatrix} \underline{\beta}_1 \\ \underline{\beta}_2 \end{bmatrix} = \begin{bmatrix} X_{11}'\underline{y} \\ 0 \end{bmatrix} = X'\underline{y}$$

subject to $L\underline{\beta}_1 = \underline{c}$. Equivalently we use Theorem 2.14 and its corollary to solve

$$(4.126)\qquad \begin{bmatrix} S^* & L' \\ L & 0 \end{bmatrix}\begin{bmatrix} \underline{\beta} \\ \underline{\lambda} \end{bmatrix} = \begin{bmatrix} X'\underline{y} \\ \underline{c} \end{bmatrix} .$$

Without loss of generality we may take $L \subset \mathcal{R}(S_{11}) \subset \mathcal{R}(S)$. Note that

$$(4.127)\qquad (S^*)^{g_1} = \begin{bmatrix} S_{11}^{g_1} & -S_{11}^{g_1}X_{12}' \\ -X_{12}S_{11}^{g_1} & I + X_{12}S_{11}^{g_1}X_{12}' \end{bmatrix}$$

from (2.94). Then R of (2.125) becomes

$$(4.128)\qquad R = [L : 0](S^*)^{g_1}\begin{bmatrix} L' \\ 0 \end{bmatrix} = LS_{11}^{g_1}L' = LS_{11}^{g}L' ,$$

and we need only consider

$$(4.129)\qquad \hat{\underline{\beta}} = G_{11}X'\underline{y} + G_{12}\underline{c}$$

where G_{11} and G_{12} are the corresponding entries of (2.126) for S^* in place of S. Substitution yields

$$(4.130)\qquad \hat{\underline{\beta}} = (S^*)^{g_1}X'\underline{y} - (S^*)^{g_1}\begin{bmatrix} L' \\ 0 \end{bmatrix}R^{g_1}(LS_{11}^{g_1}X_{11}'\underline{y}_1 - \underline{c}) , \quad \text{and}$$

$$(4.131)\qquad \begin{bmatrix} \hat{\underline{\beta}}_1 \\ \hat{\underline{\beta}}_2 \end{bmatrix} = \begin{bmatrix} S_{11}^{g_1}X_{11}'\underline{y}_1 \\ -X_{12}S_{11}^{g_1}X_{11}'\underline{y}_1 \end{bmatrix} - \begin{bmatrix} S_{11}^{g_1}L' \\ -X_{12}S_{11}^{g_1}L' \end{bmatrix}R^{g_1}(LS_{11}^{g_1}X_{11}'\underline{y}_1 - \underline{c}) .$$

Thus, under H_0, both $\hat{\underline{\beta}}_1$ and $\hat{\underline{\beta}}_2$ are modified from their forms in (4.124), though the equation is still satisfied.

It is clear from (4.131) that the hypothesis $L\underline{\beta}_1 = \underline{c}$ in (4.123) is equivalent to the hypothesis $L\underline{\beta} = \underline{c}$ in the model (4.118).

John and Prescott (1975) formulated the problem of outliers in terms of missing observations, and John (1978) has formulated a specific model. We defer discussion of that material to Chapter 6.

Plackett (1950) discusses the adjustments to estimates due to additional observations, under variance matrix $\sigma^2.I$, and with $r(X_{11}) = r(X)$. Beckman and Trussell (1974) give an alternative approach in deriving the distribution of an arbitrary studentized residual. Tietjen, Moore and Beckman (1973) suggest that if X does not have full column-rank then missing plot 'estimates' should not be hazarded. Fairfield Smith (1957) draws attention to the distinction between a missing plot estimate of an individual observation, and that of the mean of such observations. Though the estimates are equal their variances differ by σ^2.

Mitra and Bhimasankaram (1971) examined generalised inverses of partitioned matrices from a geometric viewpoint. They describe the recalculated least squares estimates for the addition and deletion of an observation or a parameter in the linear model.

CHAPTER 5

ARBITRARY VARIANCE MATRIX

Four distinct approaches were developed for problems of estimation and hypothesis testing in the general linear model. We examine these approaches and some of their inter-relationships, with a view to constructing a general test for outliers in Chapter 6 which will be applicable even when we relax the condition

(1.4) $V = \sigma^2.I$

in the model (1.1) to

(5.1) $\text{var}(\underline{\varepsilon}) = \sigma^2.V$.

Equivalently, the variance-covariance structure is assumed known up to a scale parameter, and we will require only that V is symmetric and non-negative definite. It is well known that $E(\underline{\varepsilon}) = \underline{0}$ and singular V in (5.1) imply that, with probability one (w.p.1)

(5.2) $\underline{\varepsilon} \in C(V)$. Thus

(5.3) $\underline{y} \in C[V : X]$, w.p.1 .

Unless otherwise stated, we will assume (5.3) is satisfied.

In Sections 2.2 and 2.3 singular multivariate normal distributions were discussed, and general conditions presented for chisquaredness and independence of quadratic forms in such variates. We will have recourse to these results in both this and the subsequent chapter.

Essentially, to parallel the development for Chapters 3 and 4, we must examine estimation of $\underline{\beta}$ or $X\underline{\beta}$, properties of generalized least squares (GLS) estimators and tests of hypotheses. We will present these issues within the context of the four approaches, while attempting to give an integrated overview of the results.

5.1 THE GOLDMAN-ZELEN METHOD

Pringle and Rayner (1971, pp.110-113) have given an explicit form to results, and a brief summary, of an extensive paper by Goldman and Zelen (1964). However while using their summary to introduce the material, we note that Goldman and Zelen proved a body of theory that is much more extensive than the summary suggests. Accordingly we present more of their results and establish some linking lemmas to the other sections of this chapter.

The first notion to be generalized is that of estimability. Goldman and Zelen (writing X' for the notation X here) transformed the model

(5.4) $\underline{y} = X\underline{\beta} + \underline{\varepsilon}$,

subject to (5.2), to

(5.5) $\underline{y}^* = X^*\underline{\beta} + \underline{\varepsilon}^* = \begin{bmatrix} X_1^* \\ X_2^* \end{bmatrix}\underline{\beta} + \underline{\varepsilon}^*$, where

(5.6) $$\underline{y}^* = \begin{bmatrix} \underline{y}_1^* \\ \underline{y}_2^* \end{bmatrix} = \begin{bmatrix} H_1 \\ H_2 \end{bmatrix} (X\underline{\beta} + \underline{\varepsilon}) \quad .$$ Thus

(5.7) $$\text{var}(\underline{y}^*) = \sigma^2 . \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}$$

for appropriate choices of H. Since BLUE's are known to be invariant over non-singular transformations (Mitra and Rao, 1968), such as H above, estimation may proceed along the lines of BLICU estimation developed in Section 3.4. Pringle and Rayner show that this effectively amounts to rewriting the problem of estimability and estimation in terms of the following theorem and three corollaries.

Theorem 5.1

The estimable functions $\underline{\theta}'\underline{\beta}$ in the model (5.4) are those for which unbiased estimators exist. The set of such functions is determined by taking $\underline{\theta} \in C(X')$. The BLU estimate of any estimable function $\underline{\theta}'\underline{\beta}$ is $\underline{\theta}'\hat{\underline{\beta}}$, where $\hat{\underline{\beta}}$ is any solution to

(5.8) $$\begin{bmatrix} X'V^gX & X'H_2' \\ H_2X & 0 \end{bmatrix} \begin{bmatrix} \underline{\beta} \\ \underline{\lambda} \end{bmatrix} = \begin{bmatrix} X'V^g \\ H_2 \end{bmatrix} \underline{y} \quad .$$

Corollary 5.1.1

The equations (5.8) represent minimizing the residual quadratic

(5.9) $$(\underline{y}-X\underline{\beta})'V^g(\underline{y}-X\underline{\beta}) \quad ,$$ subject to

(5.10) $$H_2X\underline{\beta} = H_2\underline{y} \quad .$$

Corollary 5.1.2

If X is in $C(V)$, then BLUE's $\underline{\theta}'\hat{\underline{\beta}}$ may be taken with $\hat{\underline{\beta}}$ satisfying the reduced form of (5.8) given by

$$(5.11) \quad X'V^{g}X\underline{\beta} = X'V^{g}\underline{y} \quad .$$

Corollary 5.1.3

Let

$$(5.12) \quad S = X'V^{g}X \quad , \quad \text{and}$$

$$(5.13) \quad L = H_2X \quad .$$

If $r(X_1^*) = r(X)$ in (5.5), then $\mathcal{R}(L) \subset \mathcal{R}(S)$, and (5.8) may be solved by means of (2.126).

It is always possible to solve (5.8) directly using Theorem 2.14 where S and L are taken as in (5.12) and (5.13), with

$$(5.14) \quad K = S + L'L \quad , \quad \text{and}$$

$$(2.122) \quad R = LK^{g_1}L' \quad .$$

Abbreviating the leading submatrices of the g_1-inverse of

$$(5.15) \quad \begin{bmatrix} X'V^{g}X & X'H_2' \\ H_2X & 0 \end{bmatrix}$$

obtained from (2.123), as G_{11} and G_{12}, Pringle and Rayner (1971, p.112) show that for

$$(5.16) \quad A = G_{11}X'V^{g} + G_{12}H_2$$

we obtain

$$(5.17) \quad XAX = X \quad , \quad \text{and}$$

(5.18) $VA'X' \subset C(X)$.

They argue that

(5.19) $LG_{11}' = 0$, but

(5.20) $LG_{11}'L' = 0$

is required because choices of K^{g_1} are arbitrary. We may however proceed as they do to show

(5.21) $VA'X' \subset C(XG_{11}X')$.

But, they miss the more explicit expression for the variance of the BLUE's given by

(5.22) $\text{var}(X\hat{\underline{\beta}}) = (XAVA'X')\sigma^2$, where

(5.23) $XAVA'X' = XAV = VA'X'$ (from (5.18)),

$$= XG_{11}X'V^gXG_{11}'X'$$

$$= XG_{11}X'(K-L'L)G_{11}'X'$$

$$= XG_{11}KG_{11}'X' = XG_{11}X'$$

$$= XK^gX' - XK^gL'R^gLK^gX'$$

where every g-inverse may be replaced by an arbitrary corresponding g_1-inverse. Observe that (5.23) implies V^g is a g_1-inverse of $XG_{11}X'$.

We may therefore show further results which would have allowed the Goldman-Zelen approach to be extended to a full examination of the general linear model (GLM).

<u>Theorem 5.2</u> (Zyskind and Martin, 1969)

The variates $\hat{\underline{y}} = X\hat{\underline{\beta}}$ and $\hat{\underline{\varepsilon}} = (\underline{y}-\hat{\underline{y}})$ are uncorrelated.

Proof: (Dunne)

From (5.22), we have

$$(5.24) \quad (I-XA)V(A'X') = 0 \quad . \qquad \square$$

Lemma 5.3 (Dunne)

The following rank identities hold:

$$(5.25) \quad r[V : X] - r(X) = r(V) - r(X) + r(H_2X).$$ However

$$(5.26) \quad r[V : X] - r(X) = r(V) - r(X'V^gX)$$

$$= r(V) - r(H_1X)$$

if and only if L in (5.13) is complementary to S in (5.12).

Proof: Premultiplying [V : X] by H yields

$$(5.27) \quad \begin{bmatrix} H_1 \\ H_2 \end{bmatrix} [V : X] = \begin{bmatrix} H_1V & : & H_1X \\ 0 & : & H_2X \end{bmatrix} \quad ,$$ with

$$(5.28) \quad r(V) = r(H_1V) \quad .$$

proving (5.25), since H is non-singular. Further

$$(5.29) \quad r(S) = r(X_1^*) = r(H_1X) \quad ,$$

and then

$$(5.30) \quad r(X) = r(H_1X) + r(H_2X)$$

if and only if

$$(5.31) \quad \mathcal{R}(H_1X) \cap \mathcal{R}(H_2X) = \{\underline{0}\} \quad . \qquad \square$$

Theorem 5.4 (Zyskind and Martin, 1969)

The residual quadratic $\hat{\underline{\varepsilon}}'V^{g_1}\hat{\underline{\varepsilon}}$ is invariant (w.p.1) over all g_1-inverses of V. Writing

$$(5.32) \quad s = r[V : X] - r(X) = r(V) - r(X) + r(H_2X) \quad ,$$ implies

$$(5.33) \quad E(\hat{\underline{\varepsilon}}'V^{g_1}\hat{\underline{\varepsilon}}) = s.\sigma^2 \quad .$$

Moreover with the normality assumption $\hat{\underline{\varepsilon}}'V^{g_1}\hat{\underline{\varepsilon}}$ has central $\sigma^2.\chi^2_s$ distribution.

Proof: (Dunne). Since $\hat{\underline{\varepsilon}}$ has

$$(5.34) \quad \text{var}(\hat{\underline{\varepsilon}}) = (I-XA)V(I-XA)' = V(I-XA)' \quad ,$$

we have $\hat{\underline{\varepsilon}}$ in $C(V)$ w.p.1. Invariance of the quadratic follows. Then, writing

$$(5.35) \quad \hat{\underline{\varepsilon}}'V^{g_1}\hat{\underline{\varepsilon}} = \underline{y}'Q\underline{y}$$ where

$$(5.36) \quad Q = (I-XA)'V^{g_1}(I-XA) \quad ,$$ we have

$$(5.37) \quad VQV = (I-XA)V = VQVQV \quad .$$

Now $QX = 0$ assures centrality and we need only

$$(5.38) \quad s = r(VQV) = r[(I-XA)V] \quad ,$$

so that by (5.17) we have

$$(5.39) \quad (I-XA)[V : X] = [VQV : 0] \quad ,$$ and

$$(5.40) \quad r(X) + r(VQV) = r[V : X]$$

to prove the result. □

Consequently we always have an unbiased estimator of σ^2 (Goldman and Zelen), but the distribution result is due to Zyskind and Martin (1969). The invariance appears to be due to Mitra and Rao (1968), though Khatri (1968) notes the invariance of

(5.41) $(\underline{y}-X\beta)'.V^{g_1}(\underline{y}-X\underline{\beta}) = \underline{\varepsilon}'V^{g_1}\underline{\varepsilon}$.

Khatri also observes that H_2 may be replaced by any orthogonal (full) complement F of V, since H_2 is only a choice of basis for such complements. On the other hand, F need not have full row-rank, in much the same way as (3.144) and (4.71) are equivalent. Further, Khatri indicates that replacing V^g in (5.8) by an arbitrary V^{g_1} does not change $X\hat{\underline{\beta}}$ (though it does in general change $\hat{\underline{\beta}}$). A formal proof may be obtained by substituting

(5.42) $V^{g_1} = H'\begin{bmatrix} I & D \\ E & C \end{bmatrix}H$

with arbitrary conformable C, D and E.

<u>Theorem 5.5</u> (Dunne)

The quadratic form $\hat{\underline{y}}'V^g\hat{\underline{y}}$ has non-central $\sigma^2.\chi^2_t(\lambda)$ distribution with

(5.43) $t = r(X'V^gX)$, and

(5.44) $\lambda = \underline{\beta}'X'V^gX\underline{\beta}$

if and only if for S and L as in (5.12) and (5.13),

(5.45) $SK^{g_1}L' = 0$.

Equivalently S and L have disjoint row-spaces.

Proof: Consider

(5.46) $\hat{\underline{y}}'V^g\hat{\underline{y}} = \underline{y}'Q\underline{y}$, where

(5.47) $Q = A'X'V^gXA$ for which

(5.48) $VQ = VQVQ$.

We require that for all $\underline{\beta}$ in the parameter space

(5.49) $\lambda = \underline{\beta}'X'V^gX\underline{\beta} = \underline{\beta}'X'V^gXG_{11}X'V^gX\underline{\beta}$

from Theorem 2.18. Thus either $\lambda = 0$, or

(5.50) $X'V^gX = X'V^gXG_{11}X'V^gX$.

In that case

(5.51) $S = (K-L'L)G_{11}(K-L'L)$

$= KG_{11}K$

$= K - L'R^gL$

$= S + L'L - L'R^gL$

implies that $R = LK^gL'$ is idempotent, and thus

(5.52) $LK^g(K-S)K^gL' = LK^gL'$, and

(5.45) $SK^gL' = 0$.

The results hold for all g_1-inverses of K. Under these conditions, either directly, or from Lemma 5.3, we have

(5.53) $t = tr(QV) = r(VQV)$

$= r(V) - r(X'V^gX)$. □

Condition (5.45) amounts to having the spaces $\mathcal{R}(S)$ and $\mathcal{R}(L)$ virtually disjoint. Since the rows of X are linearly dependent this condition will in general not be satisfied. Certainly if $\mathcal{C}(X) \subset \mathcal{C}(V)$ the theorem applies. It will later transpire that the inclusion relation is necessary.

We have previously noted V^g is a g_1-inverse of $XG_{11}X'$. For (5.52) to hold we have the equivalent condition that G_{11} is a g_1-inverse of $X'V^gX$, from (5.50).

The question arises as to whether a more convenient choice of V^{g_1} may be made in Theorem 5.1. This leads us directly to the Zyskind-Martin approach, which is examined in the next section. It will not be necessary to examine conditions for the independence of the quadratic forms $\hat{\underline{\varepsilon}}'V^{g_1}\hat{\underline{\varepsilon}}$ and $\hat{\underline{y}}'V^{g_1}\hat{\underline{y}}$ in view of Theorem 5.2.

Goldman and Zelen failed to note the extension of their methods to hypothesis testing under arbitrary variance matrix. For suppose we wish to examine

(5.54) $\quad H_0 : L\underline{\beta} = \underline{c}$

when the relations do not contradict the sure equations

(5.10) $\quad H_2X\underline{\beta} = H_2\underline{y}$.

Then whether or not the spaces $\mathcal{R}(L)$ and $\mathcal{R}(H_2X)$ intersect we may augment the model (5.4) to write

(5.55) $$\begin{bmatrix} \underline{y} \\ \underline{c} \end{bmatrix} = \begin{bmatrix} X \\ L \end{bmatrix}\underline{\beta} + \begin{bmatrix} \underline{e} \\ \underline{0} \end{bmatrix} ,$$

where $\begin{bmatrix} \underline{e} \\ \underline{0} \end{bmatrix}$ has mean zero and variance $\sigma^2.\begin{bmatrix} V & 0 \\ 0 & 0 \end{bmatrix}$. The equations (5.8) become

$$(5.56)\qquad \begin{bmatrix} X'V^gX & X'H_2' & L' \\ H_2X & 0 & 0 \\ L & 0 & 0 \end{bmatrix} \begin{bmatrix} \underline{\beta} \\ \underline{\lambda}_1 \\ \underline{\lambda}_2 \end{bmatrix} = \begin{bmatrix} X'V^g\underline{y} \\ H_2\underline{y} \\ \underline{c} \end{bmatrix}$$

for a new choice H_2^* satisfying

$$(5.57)\qquad H_2^* = \begin{bmatrix} H_2 & 0 \\ 0 & I \end{bmatrix} \quad .$$

Write

$$(5.58)\qquad p = r(X) - r\begin{bmatrix} X \\ L \end{bmatrix} + r\begin{bmatrix} H_2X \\ L \end{bmatrix} - r(H_2X) \quad .$$

Clearly, by applying Theorem 5.4 to model (5.55) and simplifying we may generate a quadratic form with degrees of freedom

$$(5.59)\qquad s + p = r(V) - r\begin{bmatrix} X \\ L \end{bmatrix} + r\begin{bmatrix} H_2X \\ L \end{bmatrix}$$

by application of (5.32), and an unbiased estimator of σ^2 as

$$(5.60)\qquad \hat{\sigma}^2 = \underline{\hat{e}}'V^{g_1}\underline{\hat{e}}/(s+p) \quad .$$

We may generalize Corollary 4.2.1 and a similar result of Rao (1975).

<u>Theorem 5.6</u> (Dunne)

The F-statistic

$$(5.61)\qquad F = \frac{\underline{\hat{e}}'V^{g_1}\underline{\hat{e}} - \underline{\hat{\varepsilon}}'V^{g_1}\underline{\hat{\varepsilon}}}{\underline{\hat{\varepsilon}}'V^{g_1}\underline{\hat{\varepsilon}}} \cdot \frac{s}{p} \quad ,$$

where $\underline{\hat{e}}$ is the residual vector $(\underline{y}-X\underline{\beta}_0)$ for $\underline{\hat{\beta}}_0$ from (5.56), has central $F(p,s)$ distribution under the null hypothesis (5.54), whether or not $L\underline{\beta}$ is strictly estimable. If $L\underline{\beta}$ is strictly estimable, (5.58) reduces to

$$(5.62) \qquad p = r\begin{bmatrix} H_2X \\ L \end{bmatrix} - r(H_2X) \qquad .$$

Proof: We proceed to show that

$$(5.63) \qquad \text{cov}(\hat{\underline{\varepsilon}},\hat{\underline{e}}-\hat{\underline{\varepsilon}}) = 0 \qquad , \qquad \text{and}$$

$$(5.64) \qquad \hat{\underline{e}}'V^{g_1}\hat{\underline{e}} = \hat{\underline{\varepsilon}}'V^{g_1}\hat{\underline{\varepsilon}} + (\hat{\underline{e}}-\hat{\underline{\varepsilon}})'V^{g_1}(\hat{\underline{e}}-\hat{\underline{\varepsilon}}) \qquad .$$

Then, assuming normality under H_0 we have the numerator as $(\hat{\underline{e}}-\hat{\underline{\varepsilon}})'V^{g_1}(\hat{\underline{e}}-\hat{\underline{\varepsilon}})$ as a difference between $\sigma^2.\chi^2$ variates. By its independence through (5.63) of one of the variates it must also have $\sigma^2.\chi^2$ distribution, and the result will follow from the separability properties of χ^2 variates.

The form of $\hat{\underline{e}}$ is given by the first row of

$$(5.65) \qquad \begin{bmatrix} \underline{y} \\ \underline{c} \end{bmatrix} - \begin{bmatrix} X \\ L \end{bmatrix} G^*_{11}[X' : L']\begin{bmatrix} V^g\underline{y} \\ \underline{c} \end{bmatrix}$$

where G^*_{11} is the leading sub-matrix for (5.56) corresponding to G_{11} for (5.15). Applying the relations (5.23) to the extended model (5.55) confirms that (5.65) is in the column-space of $\begin{bmatrix} V & 0 \\ 0 & 0 \end{bmatrix}$ and that $\hat{\underline{e}}$ is in $C(V)$. Also $(\hat{\underline{e}}-\hat{\underline{\varepsilon}})$ is in $C(V)$ and in $C(X)$. Thus

$$(5.66) \qquad \text{cov}(\hat{\underline{\varepsilon}},\hat{\underline{e}}-\hat{\underline{\varepsilon}}) = (I-XA)V(XA-XA^*_1)'$$

$$= (I-XA)X(A-A^*_1)V$$

$$= 0$$

from (5.22) and the parallel form for model (5.56), namely

$$(5.67)\qquad \begin{bmatrix} X \\ L \end{bmatrix} A^* \begin{bmatrix} V & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} X \\ L \end{bmatrix} [A_1^* : A_2^*] \begin{bmatrix} V & 0 \\ 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} V & 0 \\ 0 & 0 \end{bmatrix} [A_1^* : A_2^*]' [X' : L']$$

$$= \begin{bmatrix} XA_1^* V A_1^{*\prime} X' & 0 \\ 0 & 0 \end{bmatrix}$$

To show (5.64) we need only write

$$(5.68)\qquad \hat{\underline{e}} = \hat{\underline{\varepsilon}} + (\hat{\underline{e}} - \hat{\underline{\varepsilon}})$$

in the left-hand side and show that the mixed terms of the form are zero. Since every variate $\underline{\alpha}$ with $E(\underline{\alpha}) = \underline{0}$ has $\underline{\alpha}$ in $C[\mathrm{var}(\underline{\alpha})]$ w.p.1, we have

$$(5.69)\qquad (\hat{\underline{e}} - \hat{\underline{\varepsilon}}) = XA\underline{y} - (XA_1^*\underline{y} + XA_2^*\underline{c}) \in C(V) \cap C(X), \qquad \text{and}$$

$$(5.70)\qquad \hat{\underline{\varepsilon}}' V^{g_1} (\hat{\underline{e}} - \hat{\underline{\varepsilon}}) = \underline{a}'(I-XA)V.V^{g_1}V\underline{b} = \underline{a}'(I-XA).X\underline{d} = 0$$

for suitable choices of $\underline{a}$, $\underline{b}$ and $\underline{d}$, from (5.66). This completes the proof. We need only remark that p in (5.58) is found by subtraction of the terms derived from (5.32) for the quadratic forms under discussion. However p can be more simply expressed for any L in (5.54) as

$$(5.71)\qquad p = r\begin{bmatrix} H_2 X \\ L_1 \end{bmatrix} - r(H_2 X)$$

where L_1 gives a basis for the entire estimable part of L, i.e. of $R(L) \cap R(X)$. □

We may also extend the idea of the partitioned model to the arbitrary variance case. Suppose that

$$(5.72)\qquad \underline{y} = [X_1 : X_2]\begin{bmatrix}\underline{\beta}_1\\ \underline{\beta}_2\end{bmatrix} + \underline{\varepsilon}$$

is a partitioning of model (5.4). Ignoring $\underline{\beta}_2$, we may apply the methods of the foregoing theorems and find

$$(5.73)\qquad \tilde{\underline{y}} = X_1\tilde{A}\tilde{\underline{\beta}}_1 = X_1\tilde{G}_{11}X'V^g\underline{y} + X_1\tilde{G}_{12}H_2\underline{y}$$

$$(5.74)\qquad \tilde{\underline{\varepsilon}} = \underline{y} - \tilde{\underline{y}} = (I-X_1\tilde{A})\underline{y}$$

and the variance-covariance structure of $\tilde{\underline{\varepsilon}}$ is given by

$$(5.75)\qquad \text{var}(\tilde{\underline{\varepsilon}}) = \sigma^2.(I-X_1\tilde{A})V(I-X_1\tilde{A})' = \sigma^2.N, \text{ say.}$$

Note that by Theorem 5.4, the quadratic $\tilde{\underline{\varepsilon}}'V^g\tilde{\underline{\varepsilon}}$ follows a central $\sigma^2.\chi^2_q$ distribution, where

$$(5.76)\qquad q = r[V : X_1] - r[X_1] \quad .$$

Also q is the rank of the matrix N in (5.75).

Following the analysis of covariance approach of Section 4.5 and generalizing the relations (4.103) to (4.108), we model $\tilde{\underline{\varepsilon}}$ by

$$(5.77)\qquad \tilde{\underline{\varepsilon}} = (I-X_1\tilde{A})X_2\underline{\beta}_2 + \underline{e} = X_2^*\underline{\beta}_2 + \underline{e}$$

where $\underline{e}$ has variance-covariance structure $\sigma^2.N$. Now

$$(5.78)\qquad (I-X_1\tilde{A})[V : X_1 : X_2] = [N : 0 : X_2^*] \quad .$$

In general $C(X_2^*)$ is not in $C(N)$, so that Corollary 5.1.2 will not always apply. We may however use the structure of (5.8) for the model (5.77) and write

$$(5.79)\qquad \begin{bmatrix} X_2^{*\prime}N^gX_2^* & X_2^{*\prime}K_2' \\ K_2X_2^* & 0 \end{bmatrix}\begin{bmatrix}\underline{\beta}_2\\ \underline{\lambda}\end{bmatrix} = \begin{bmatrix} X_2^{*\prime}N^g \\ K_2\end{bmatrix}\tilde{\underline{\varepsilon}} \quad .$$

Then taking B in the manner of (5.16) as

(5.80) $$B = G^{*}_{11}X^{*\prime}_{2}N^{g} + G^{*}_{12}K_{2} \quad ,$$

we obtain

(5.81) $$X^{*}_{2}\underline{\tilde{\beta}}_{2} = X^{*}_{2}B\underline{\tilde{\varepsilon}} \quad , \qquad \text{and}$$

(5.82) $$\underline{\tilde{e}} = \underline{\tilde{\varepsilon}} - X^{*}_{2}B\underline{\tilde{\varepsilon}} \quad .$$

By Theorem 5.4, $\underline{\tilde{e}}'N^{g}\underline{\tilde{e}}$ has central $\sigma^{2}.\chi^{2}_{k}$ distribution where, by (5.32) and (5.78)

(5.83) $$\begin{aligned} k &= r(N : X^{*}_{2}) - r(X^{*}_{2}) \\ &= r[V : X] - r(X_{1}) - r(X^{*}_{2}) \\ &= r[V : X] - r(X) \\ &= s \quad . \end{aligned}$$

By Theorem 5.2, with (5.81) and (5.82), $\tilde{e}$ and $(\underline{\tilde{\varepsilon}}-\underline{\tilde{e}})$ are uncorrelated, and for some $\underline{a}$ and $\underline{b}$,

(5.84) $$\begin{aligned} (\underline{\tilde{\varepsilon}}-\underline{\tilde{e}})'N^{g}\underline{\tilde{e}} &= \underline{a}'NB'X^{*\prime}_{2}N^{g}(I-X^{*}_{2}B)N\underline{b} \\ &= \underline{a}'X^{*}_{2}\;BN(I-X^{*}_{2}B)'\underline{b} \\ &= 0 \end{aligned}$$

Finally, consider

(5.85) $$K = (I-X_{1}\tilde{A})'V^{g}(I-X_{1}\tilde{A}) \quad ,$$

which is a g_{1}-inverse of N. Theorem 5.4, with (5.74) and the idempotency of $X_{1}\tilde{A}$, allows us to write

(5.86) $$\begin{aligned} \underline{\tilde{\varepsilon}}'V^{g}\underline{\tilde{\varepsilon}} = \underline{\tilde{\varepsilon}}'N^{g_{1}}\underline{\tilde{\varepsilon}} &= \underline{\tilde{\varepsilon}}'N^{g}\underline{\tilde{\varepsilon}} \\ &= (\underline{\tilde{\varepsilon}}-\underline{\tilde{e}})'N^{g}(\underline{\tilde{\varepsilon}}-\underline{\tilde{e}})' + \underline{\tilde{e}}'N^{g}\underline{\tilde{e}} \end{aligned}$$

$$= \tilde{\underline{\varepsilon}}'Q\tilde{\underline{\varepsilon}} + \tilde{\underline{e}}'N^g\tilde{\underline{e}} \qquad , \qquad \text{where}$$

(5.87) $\quad Q = (X_2^*B)'N^gX_2^*B \quad .$ Because

(5.88) $\quad NQ = NQNQ$

the conditions for chisquaredness are satisfied (except for the non-centrality parameter) and we have proved a generalized "sweep-out" method in

<u>Theorem 5.7</u> (Dunne)

The F-statistic

(5.89) $$F = \frac{\tilde{\underline{\varepsilon}}'V^g\tilde{\underline{\varepsilon}} - \tilde{\underline{e}}'N^g\tilde{\underline{e}}}{\tilde{\underline{e}}'N^g\tilde{\underline{e}}} \cdot \frac{s}{q-s}$$

for residual vectors $\tilde{\underline{\varepsilon}}$ and $\tilde{\underline{e}}$ from (5.74) and (5.82), has central $F(q-s,s)$ distribution under the null hypothesis $H_0 : \underline{\beta}_2 = \underline{0}$, or more strictly when $X_2^*\underline{\beta}_2 = \underline{0}$ in (5.77).

Goldman and Zelen developed a special case of Theorem 5.7 for $V = I$, and examined an early but less general form of the reduced model method of Rayner (1976) discussed in Section 3.6. They also showed that Corollary 5.1.2 has a special case, namely

<u>Theorem 5.8</u>

If X and V satisfy either of the equivalent conditions

(5.90) $\quad X'V^g = BX' \quad ,$ or

(5.91) $\quad X' = BX'V$

for some non-singular B, then the OLS estimates of $X\underline{\beta}$

coincide with the GLS estimates.

Kempthorne (1976) and others have superceded the partial answer of Goldman and Zelen to the general conditions for $C(X)$ in $C(V)$ to allow the OLS and GLS solutions to coincide. These matters are examined in Section 5.5.

Finally they discussed restraints subject to uncertainty. This implies situations where restraints such as $T\underline{\beta} = \underline{m}$ in (5.54) are not precisely known but may be summarized as the value of a random vector $\hat{\underline{m}}$, which may be modelled by

(5.92) $E(\hat{\underline{m}}) = T\underline{\beta}$, and

(5.93) $\text{var}(\hat{\underline{m}}) = \sigma^2.U$, with

(5.94) $\text{cov}(\hat{\underline{m}},\underline{y}) = 0$.

The intention of the analysis is to fit the model (5.4) as though the given or observed restraints

(5.95) $T\hat{\underline{\beta}} = \hat{\underline{m}}$

were exact. The effect of such a process is clearly two-fold. On the one hand it extends the space of estimable functions (for which BLICUE's exist), and on the other, it restricts the values of certain pre-estimable functions while leaving others invariant, or "undisturbed" in the terminology of Goldman and Zelen. They comment that it may be important to identify the latter class of functions. If there is scepticism concerning the prior information $(\hat{\underline{m}})$ which may be available, they aver that it may be useful to include as far as possible in the space of undisturbed functions, all those functions for which the minimum variance estimator is of particular importance. Presumably this means that parts of the observed restraints (5.95) are ignored, or that some control may be possible *a priori* over the nature of T.

We may simplify and generalize their method using the reduced model forms of Rayner given in (3.183) through to (3.187). Thus

(5.96) $E(\underline{y}-XT^g\hat{\underline{m}}) = X(I-T^gT)\underline{\tau}$, with

(5.97) $\mathrm{var}(\underline{y}-XT^g\hat{\underline{m}}) = \sigma^2.(V+XT^gUT^{g'}X')$.

The estimated values of $X(I-T^gT)\underline{\tau}$ may be found by Theorem 5.1 applied to this new model, and the required estimates of $X\underline{\beta}$, from (3.186) will be given by

(5.98) $\hat{\underline{y}} = X(I-T^gT)\underline{\tau} + XT^g\hat{\underline{m}}$.

Again arbitrary g_1-inverses may replace the g-inverse without affecting the estimates of estimable functions. The estimable functions which are undisturbed are precisely those from the row-space of $X(I-T^gT)$, which has dimension given by

(5.99) $r\begin{bmatrix} X \\ T \end{bmatrix} - r(T)$.

We note that to use the above method when $V = I$ in the Goldman and Zelen case, or its generalized form consistency of (5.95) is assumed. Thus $C(U)$ is a subspace of $C(T)$, i.e. for some A

(5.100) $U = TAT'$.

This would be satisfied if $T\underline{\beta}$ was estimable in and estimated from some other model source, and failing that it would be necessary to augment (5.4) and write

$$(5.101) \qquad \begin{bmatrix} \underline{y} \\ \underline{m} \end{bmatrix} = \begin{bmatrix} X \\ T \end{bmatrix} \underline{\beta} + \underline{e}$$

where $\underline{e}$ has zero mean and variance structure given by

$$(5.102) \qquad \text{var}(\underline{e}) = \sigma^2 . \begin{bmatrix} V & 0 \\ 0 & U \end{bmatrix} \quad , \qquad \text{say.}$$

It is not clear how estimating $T\underline{\beta}$ ignoring $\underline{y}$ and then applying the estimated restraints (5.95) gives any advantage over the model (5.101). We conjecture that there is little, if any, nor is it equivalent to the generalized sweep-out of Theorem 5.7. In any event the nature of U and the relative sizes of the entries of V and U will presumably reflect the known or presumable degree of belief to be attached to the additional information.

5.2 THE ZYSKIND-MARTIN METHOD

It is clear from Corollary 5.1.2 that the equations

$$(5.11) \qquad X'V^g X\hat{\underline{\beta}} = X'V^g \underline{y}$$

do not lead to identical solutions $X\hat{\underline{\beta}}$ as those obtained from (5.8). Zyskind and Martin (1969) investigated the problem of finding a class of g_1-inverses V^* of V for which the "generalized normal equations", or GNE's

$$(5.103) \qquad X'V^* X\hat{\underline{\beta}} = X'V^* \underline{y}$$

yield solutions $\hat{\underline{\beta}}$ such that the BLUE of any estimable function $\underline{\theta}'\underline{\beta}$ is given by $\underline{\theta}'\hat{\underline{\beta}}$. We use the notation of Pringle and Rayner (1971, pp.114-117) in describing the method, but examine the theory more extensively.

Theorem 5.9

Let $\underline{b}$ be a solution to the equations

$$(5.104) \quad X'V^{g_1}X\hat{\underline{\beta}} = X\; V^{g_1}\underline{y} \quad .$$

Then $\underline{\theta}'\underline{b}$ is unique and an unbiased estimate (UE) of $\underline{\theta}'\underline{\beta}$ if and only if $\underline{\theta}'$ is in $R(X)$ and

$$(5.105) \quad r(X'V^{g_1}X) = r(X) \quad .$$

Proof: (Pringle and Rayner). For unbiasedness we require

$$(5.106) \quad E(\underline{\theta}'(X'V^{g_1}X)^{g_1}X'V^{g_1}\underline{y}) = \underline{\theta}'(X'V^{g_1}X)^{g_1}(X'V^{g_1}X)\underline{\beta}$$
$$= \underline{\theta}'\underline{\beta}$$

over the whole parameter space $\underline{\beta}$, and over the whole space of $\underline{\theta}'$. By the nature of (5.104) $\underline{\theta}'$ is in $R(X)$ and we may specify all such $\underline{\theta}'$. Thus

$$(5.107) \quad X(X'V^{g_1}X)^{g_1}(X'V^{g_1}X) = X$$

and (5.105) follows from

$$(5.108) \quad r(X) \leq r(X'V^{g_1}X) \leq r(X) \quad .$$

The converse follows easily. Uniqueness of $X\underline{b}$ and hence of all $\underline{\theta}'\underline{b}$ follows from (5.107). □

For an UE $\underline{w}'\underline{y}$ of $\underline{\theta}'\underline{\beta}$ to have minimum variance $\sigma^2.\underline{w}'V\underline{w}$, we minimize the variance subject to $\underline{w}'X = \underline{\theta}'$, or

$(\underline{w}'X-\underline{\theta}')\underline{\lambda} = 0$ for some Lagrange multiplier $\underline{\lambda}$. Zyskind (1967) showed that this implies

(5.109) $\quad V\underline{w} = X\underline{\lambda}$

i.e. $V\underline{w}$ is in $C(X)$ or $\underline{w}'V$ in $R(X')$. Thus Zyskind and Martin were led to solve for V^{g_1} in

(5.110) $\quad X(X'V^{g_1}X)^{g_1}X'V^{g_1}V = AX'$

for some A. We note (5.110) is equivalent to

(5.111) $\quad V(V^{g_1})'X = XB$

for some B. Denote a basis for $C(V) \cap C(X)$ by Q and extend this to a basis for $C(X)$ by adjoining R. Then

(5.112) $\quad X = QC_1 + RC_2$,

and we may always take

(5.113) $\quad C(R) \cap C(V) = \{\underline{0}\}$.

By Theorem 2.10 we may take

(5.114) $\quad V^* = (V+RC_2C_2'R_2')^{g_1}$. Thus

(5.115) $\quad V(V^*)'X = V(V^*)'QC_1 + V(V^*)'RC_2$

$$= QC_1 + 0$$

$$= XB$$

for some B, as required. Clearly a wide set of V^* is generated for each choice of Q and R. We may form the union of all such sets. Now, for any V^* we exhibit a corresponding choice of Q and R, given (5.111). Initially take any basis $[Q : R_0]$ for $C(X)$ as before. Then

$$(5.116)\quad XB = V(V^*)'X = V(V^*)'(QC_3+R_0C_4) \quad , \quad \text{say} = QC_3 + V(V^*)'R_0C_4 \quad .$$

implies that

$$(5.117)\quad V(V^*)'R_0C_4 = XB - QC_3 = QC_5 \quad , \quad \text{say} = V(V^*)'QC_5 \quad .$$

If QC_5 is zero, then any basis for R_0C_4, and specifically R_0 itself is a choice of R. If QC_5 is not zero then

$$(5.118)\quad X = Q(C_3-C_5) + (R_0C_4-QC_5) \quad , \quad \text{and}$$

$$(5.119)\quad V(V^*)'X = Q(C_3-C_5) = XB \quad .$$

Then we may take R as any basis for $(R_0C_4-QC_5)$.

Zyskind and Martin did not provide the typification of V^* in (5.114). Instead they exhibited a condition based on Corollary 2.2.1 under which a choice from V^{g_1} satisfying (5.111) was possible, and then showed that the condition could always be satisfied. Within the class of solutions to (5.111) they established further equivalent conditions to (5.107), which may again always be met. Essentially they provide a constructive but non-explicit characterization of all possible V^*, and noted that V^* need not be symmetric, even though V is. Pringle and Rayner (1971, p.116) have criticised the method for assuming that

$$(5.120)\quad V(V^{g_1})'R = 0$$

is necessary for (5.111). Though no explicit proof was given by Zyskind and Martin, the claim is in fact correct, as was shown in (5.118) and (5.119). We therefore have

Theorem 5.10 (Rao, 1971)

The necessary and sufficient conditions for a g_1-inverse V^* of V in the GNE's

(5.103) $X'V^*X\underline{\hat{\beta}} = X'V^*\underline{y}$

to yield the BLUE $\underline{\theta}'\underline{\hat{\beta}}$ of any estimable function $\underline{\theta}'\underline{\beta}$ are

(5.108) $r(X'V^*X) = r(X)$, and

(5.121) $V^* = (V+XUX')^{g_1}$.

Moreover the minimum variances may be obtained from

(5.122) $\text{var}(X\underline{\hat{\beta}}) = [X(X'V^*X)^{g_1}X' - XUX'].\sigma^2$.

Proof: In (5.114) suppose that $R = XP$, then

(5.123) $U = PC_2C_2'P'$ suffices.

The variance result follows by substituting $(V+XUX') - XUX'$ for V. □

We note that V^* need not be symmetric, though it may always be taken to symmetric. Further there is no loss of generality in specifying that $V + XUX'$ is symmetric.

Corollary 5.10.1 (Zyskind and Martin)

For all V^* in (5.103), subject to the conditions of Theorem 5.10 the solutions $\underline{\hat{\beta}}$ are identical. All the solutions $\underline{\hat{\beta}}$ also satisfy

(5.10) $H_2X\hat{\underline{\beta}} = H_2\underline{y}$,

so that the classes of solutions $\hat{\underline{\beta}}$ in (5.8) and (5.103) are identical.

Proof: From (5.108), there exists a nonsingular matrix A with

(5.124) $AX'V_1^*X = X'V_2^*X$

for any V^* matrices, V_1^* and V_2^*. By the uniqueness of estimable functions and the definition of V^*, (5.103) and (5.124) immediately imply that $AX'V_1^*\underline{y} = X'V_2^*\underline{y}$ for all $\underline{y}$.

Thus the sets of solutions $\hat{\underline{\beta}}$ to the equivalent systems of equations are identical. Further, from the model (5.4) $H_2X\underline{\beta}$ is estimable, and has an UE $H_2\underline{y}$ which has zero variance, which must therefore also be its BLUE. Equivalently (5.10) holds. Finally, from (5.41), the expression $(\underline{y}-X\underline{\beta})'W(\underline{y}-X\underline{\beta})$ is invariant over all choices of V^{g_1} for W. Thus its minimization with $W = V^g$ and subject to $H_2X\underline{\beta} = H_2\underline{y}$ is equivalent to minimizing with $W = V^*$ subject to the same conditions. In view of the fact that (5.10) has been shown for $\hat{\underline{\beta}}$ in (5.103), the conditions are void. Thus

(5.125) $\{\hat{\underline{\beta}}|(5.103)\} \subset \{\hat{\underline{\beta}}|(5.8)\}$

and it is not just the unique fitted values $X\hat{\underline{\beta}}$ from each equation that coincide. □

Moreover, as in Theorem 5.2 we may confirm the independence of $\hat{\underline{y}}$ and $\hat{\underline{\varepsilon}}$ by writing

$$(5.126) \quad D = X(X'V^*X)^{g_1}X'V^* \quad , \quad \text{whence}$$

$$(5.127) \quad DVD' = X(X'V^*X)^{g_1}X'V^*XB(X'V^*X)^{g_1'}X'$$

$$= XB(X'V^*X)^{g_1'}X'$$

$$= VD' = DV$$

from (5.115) and (5.107). Then, as in (5.24)

$$(5.128) \quad (I-D)VD' = 0 \quad .$$

Theorem 5.4 is proved for the quadratic form $\hat{\underline{\varepsilon}}'V^*\hat{\underline{\varepsilon}}$ by noting that from (5.128)

$$(5.129) \quad var(\hat{\underline{\varepsilon}}) = \sigma^2.(I-D)V(I-D)'$$

$$= \sigma^2.(V-DV) \qquad = \sigma^2.(V-VD')$$

$$= \sigma^2.N \quad , \quad \text{say}$$

and taking $Q = V^*$, the conditions of Theorem 2.18 are satisfied by

$$(5.130) \quad NV^*N = (I-D)VV^*V(I-D') = N \quad , \quad \text{with}$$

$$(5.131) \quad s = r(N) = tr(QN)$$

$$= tr(V^*V) - tr(V^*DV)$$

$$= r(V) - r(V^*DV) \quad .$$

To derive s as in (5.32) we note that from (5.126)

$$(5.132) \quad r(V^*DV) \geq r(X'V^*DV)$$

$$= r(X'V^*V)$$

$$\geq r(V^*DV)$$

with equality throughout, and then

$$(5.133)\quad r(X'V^*V) = \dim[C(X) \cap C(V)]$$
$$= r(X) - r(H_2X)$$
$$= r(V) + r(X) - r[V : X]$$

by the construction (5.114) of V^* and Theorem 2.10.

Paralleling the conditions of Theorem 5.5, the quadratic $\hat{\underline{y}}'V^*\hat{\underline{y}}$ does not in general follow a non-central $\sigma^2.\chi^2_t(\lambda)$ distribution. The non-centrality relation $\lambda = \underline{\mu}'Q\underline{\mu} = \underline{\mu}'QVQ\underline{\mu}$ of (2.171) is not satisfied. Substituting $(V+XUX') - XUX'$ for V in

$$(5.134)\quad \underline{\beta}'X'V^*(DVD')V^*X\underline{\beta} = \underline{\beta}'X'V^*X\underline{\beta} - \underline{\beta}'X'V^*XUX'V^*X\underline{\beta} \quad ,$$

we require that the subtracted term is zero over the entire parameter space, and thus

$$(5.135)\quad X'V^*XUX'V^*X = 0 \quad .$$

This with (5.107) is equivalent to

$$(5.136)\quad XUX' = 0$$

or, by virtue of the construction (5.112) that

$$(5.137)\quad C(X) \subset C(V) \quad .$$

This relation vindicates the claim below Theorem 5.5, that (5.137) is an equivalent condition for the theorem.

Zyskind *et al* (1964) show that the (relative) sum of squares associated with a strictly testable hypothesis $H_0 : L\underline{\beta} = \underline{c}$, assuming consistency with the sure equations

$$(5.10)\quad H_2X\underline{\beta} = H_2\underline{y} \quad ,$$

may be written as

(5.138) $SS(H_0) = (L_0\hat{\underline{\beta}}-\underline{c}_0)' \text{ var}(L_0\hat{\underline{\beta}})^{g_1}(L_0\hat{\underline{\beta}}-\underline{c}_0).\sigma^2$,

where $\hat{\underline{\beta}}$ is any solution to (5.103), and $L_0\underline{\beta} = \underline{c}_0$ represents that part of the hypothesis $L\underline{\beta} = \underline{c}$ which does not fall in the space $R(H_2X)$. $SS(H_0)$ in (5.138) has central $\sigma^2.\chi^2$ distribution under H_0, and is otherwise non-central $\sigma^2.\chi^2$, with degrees of freedom $r(L_0)$. Effectively, L_0 represents a set added to H_2X forming

(5.139) $$H = \begin{bmatrix} H_2X \\ L_0 \end{bmatrix}$$

with $R(L) \subset R(H) \subset R(X)$. By appealing to standardization methods to find canonical components of $\underline{\varepsilon}$ they prove that under the model (5.4),

<u>Theorem 5.11</u>

The F-statistic

(5.140) $$F = \frac{SS(H_0)}{\hat{\underline{\varepsilon}}' V^{g_1} \hat{\underline{\varepsilon}}} \cdot \frac{s}{r(L_0)}$$

is the statistic associated with the hypothesis $H_0 : L\underline{\beta} = \underline{c}$.

□

We may note the appropriate central and non-central distributions of (5.140). The question of how this relates to Theorem 5.6 is easily solved. From (5.139) and (5.62), by construction.

(5.141) $$p = r(L_0) = r\begin{bmatrix} H_2X \\ L \end{bmatrix} - r(H_2X)$$.

We need therefore to show that the numerators coincide. In the manner of (5.56) we write

$$\begin{bmatrix} X'V^gX & X'H_2' & L' \\ H_2X & 0 & 0 \\ L & 0 & 0 \end{bmatrix} \begin{bmatrix} \underline{\beta} \\ \underline{\lambda}_1 \\ \underline{\lambda}_2 \end{bmatrix} = \begin{bmatrix} X'V^g\underline{y} \\ H_2\underline{y} \\ \underline{c} \end{bmatrix} \quad . \tag{5.142}$$

In view of Corollary 5.10.1 we may replace (5.142) by

$$\begin{bmatrix} X'V^*X & L' \\ L & 0 \end{bmatrix} \begin{bmatrix} \underline{\beta} \\ \underline{\lambda} \end{bmatrix} = \begin{bmatrix} X'V^*\underline{y} \\ \underline{c} \end{bmatrix} \quad . \tag{5.143}$$

Then, without loss of generality, we assume V^* is symmetric and positive semi-definite (Pringle and Rayner, 1971, p.116), with (5.107) and the estimability of $L\underline{\beta} = AX\underline{\beta}$ implying $R(L) = R(AX) \subset R(X'V^*X)$. Solving (5.143) by applying the g_1-inverse of (2.126), where

(5.144) $\quad S = X'V^*X$ and

(5.145) $\quad R = LS^{g_1}L'$ correspondingly,

we obtain the exact parallels of (4.8) through to (4.13) in

$$\hat{\underline{\beta}}_0 = (S^{g_1}-S^{g_1}L'R^{g_1}LS^{g_1})X'V^*\underline{y} + S^{g_1}L'R^{g_1}\underline{c} \quad , \tag{5.146}$$

$$X\hat{\underline{\beta}}_0 = X\hat{\underline{\beta}} - XS^{g_1}L'R^{g_1}(L\hat{\underline{\beta}}-\underline{c}) \quad , \quad \text{and} \tag{5.147}$$

$$\hat{\underline{e}}'V^*\hat{\underline{e}} = \hat{\underline{\varepsilon}}'V^*\hat{\underline{\varepsilon}} + (L\hat{\underline{\beta}}-\underline{c})'R^{g_1}(L\hat{\underline{\beta}}-\underline{c}) \quad . \tag{5.148}$$

Invariance in Theorem 5.4 allows us to replace V^* in (5.146) by arbitrary V^{g_1}. The construction of R, which is in general not the variance-covariance structure of $L\hat{\underline{\beta}}$, nonetheless implies that R^{g_1} is also a g_1-inverse of that structure. Thus

$$(5.149) \quad \mathrm{var}(L\hat{\underline{\beta}})/\sigma^2 = \mathrm{var}(AX\hat{\underline{\beta}})/\sigma^2 \quad \text{for some } A$$

$$= ADVD'A' \quad , \text{ from } (5.128)$$

$$= AD(V+XUX')D'A' - ADXUX'D'A'$$

$$= AX(X'V^*X)^{g_1}X'A' - AXUX'A' \quad , \text{ from } (5.127)$$

$$= LS^{g_1}L' - LUL'$$

$$= LBVB'L' \quad , \text{ for some } B.$$

Now, by (5.145) we have

$$(5.150) \quad R = \mathrm{var}(L\hat{\underline{\beta}})/\sigma^2 + LUL' \quad .$$

By Theorem 2.10, for R^{g_1} to have the required property we need

<u>Lemma 5.12</u> (Dunne)

For any admissible U,

$$(5.151) \quad C(LBVB'L') \cap C(LUL') = \{0\}.$$

Proof: Observe that

$$(5.152) \quad DVD' + XUX' = XBVB'X' + XUX'$$

$$= X(X'V^*X)^{g_1}X' \quad .$$

By construction of U in (5.123), and from (5.128) we have

$$(5.153) \quad C(DVD') \subset C(V) \quad , \qquad \text{and}$$

$$(5.154) \quad C(V) \subset C(XUX') = \{0\} \ .$$

Similarly for the row-spaces. Thus

$$(5.155) \quad r(DVD') + r(XUX') = r[X(X'V^*X)^{g_1}X'] = r(X) \quad , \text{ and}$$

(5.156) $r(AXBVB'X') + r(AXUX') = r(AX)$

for any conformable A. Using the positive semi-definiteness of the matrices, we have from (5.156) that

(5.157) $\dim[C(LBVB'L') \cap C(LUL')]$

$= r(LBVL'L') + r(LUL') - r[LBVB'L' : LUL']$

$= r(AXBVB'X') + r(AXUX') - r(L)$

$= 0$

which proves the result. □

We have therefore an extension in

<u>Theorem 5.13</u> (Dunne)

The F-statistic

(5.158) $F = \frac{SS(H_0)}{\underline{\hat{\varepsilon}}'V^{g_1}\underline{\hat{\varepsilon}}} \cdot \frac{s}{p}$

where $SS(H_0)$ is obtained from

(5.159) $\underline{\hat{e}}'V^{g_1}\underline{\hat{e}} - \underline{\hat{\varepsilon}}'V^{g_1}\underline{\hat{\varepsilon}} = (L\underline{\hat{\beta}}-\underline{c})'R^{g_1}(L\underline{\hat{\beta}}-\underline{c})$.

and p as in (5.141) has central F distribution under H_0, when H_0 is false, it has non-central F-distribution if and only if

(5.160) $LUL' = 0$.

Proof: Taking Q as R^{g_1}, Lemma 5.12 implies that all the conditions of Theorem 2.18 are satisfied, except the non-centrality value λ when H_0 is false. In that case

(5.161) $\quad \underline{\mu}'Q\underline{\mu} = (L\underline{\beta}-\underline{c})'R^{g_1}(L\underline{\beta}-\underline{c})$

$$= (L\underline{\beta}-\underline{c})'R^{g_1}(R-LUL')R^{g_1}(L\underline{\beta}-\underline{c})$$

if and only if (5.160) holds, since H_0 is consistent and the ranks of R and L are equal. □

We note that (5.159) defines a unique value of $SS(H_0)$ whether or not H_0 is true. Replacing R^{g_1} by an arbitrary g_1-inverse of $K = \text{var}(L\hat{\underline{\beta}})/\sigma^2$ will give possibly differing values when H_0 is false, because the column-space $C(L)$ is in general only contained by $C(K)$ if (5.160) holds. On the other hand, suppose we specifically take K^{g_1} to be K^g, so that the conditions of Theorem 2.18 are immediately satisfied since

(5.162) $\quad K^g(R-LUL')K^g = K^gKK^g = K^g$

is sufficient. The notable fact is that K^g is not an R^{g_1} unless (5.160) holds. The apparent contradiction may be explained by the fact that the distribution of (5.159) was established in Theorem 5.6 for the case when H_0 is true. This theorem spells out the distributional properties when H_0 is false. It has been noted by Rao (1972a; 1973, p.302) that the terms

(5.163) $\quad (L\hat{\underline{\beta}}-\underline{c})'K^{g_1}(L\hat{\underline{\beta}}-\underline{c})$ and

(5.164) $\quad (L\hat{\underline{\beta}}-\underline{c})'R^{g_1}(L\hat{\underline{\beta}}-\underline{c})$

are equal for all L when $C(X) \subset C(V)$, i.e.

(5.165) $\quad XUX' = 0$.

We have strengthened the result showing equality (even when

H_0 is false) for any particular L if and only if

(5.166) $LUL' = 0$. Equivalently

(5.167) $LUX' = 0$.

It is precisely this result which with the assumption of consistency underpins the Zyskind *et al* result of Theorem 5.11. Since, from (5.111),

(5.168) $r(H_2X) \leq r(H_2[V+XUX']) = r(H_2XUX')$,

and then equality trivially, we may modify L to construct $L_0 = A_0X$ in (5.139), such that

(5.169) $L_0UX' = 0$

whenever (5.167) does not hold directly. Its effect is to guarantee that

(5.170) $\text{var}(L_0\hat{\underline{\beta}})/\sigma^2 = L_0(X'V^*X)^{g_1}L_0' - L_0UL_0'$

$= L_0(X'V^*X)^{g_1}L_0' = K_0$, say,

and hence that the mean of $(L_0\hat{\underline{\beta}}-\underline{c})$ is in the column-space $C(K_0)$. This last property yields the non-centrality condition

(5.171) $\lambda = \underline{\mu}'Q\underline{\mu} = \underline{\mu}'QK_0Q\underline{\mu}$

in every case. Theorem 5.11 is consequently a special case of Theorem 5.13. Zyskind and Martin note that it is equivalent to testing a second stage of a nested hypothesis, viz $L_0\underline{\beta} = \underline{c}_0$, given that the sure equations

(5.10) $H_2X\underline{\beta} = H_2\underline{y}$

are satisfied by the data, and by the subhypothesis. In fact the role of LUL' in Theorem 5.13 is to ensure that the sure equations are separated from the hypothesis itself, and thus specifically to define the column-spaces $C(L_0)$ within which $\underline{c}_0$ must fall to have

$$(5.172) \qquad (L\underline{\beta}=\underline{c}) \equiv \begin{bmatrix} L_0\underline{\beta} = \underline{c}_0 \\ H_2X\underline{\beta} = H_2\underline{y} \end{bmatrix}$$

and the required consistency. The relation

$$(5.166) \qquad LUL' = 0$$

also defines the classes of estimable functions in a given model for which $SS(H_0)$ may be derived equivalently from the method of subtraction (5.159) or the Rao term (5.163). The condition for consistency in the former case is simply that $\underline{c}$ is in $C(L)$. We deal with the latter case as Theorem 5.16 in the following section.

We may therefore extend the notion in Chapter 4 of a strongly testable hypothesis. Under the model (5.4), an estimable function $L\underline{\beta}$ and an internally consistent set of equations $H_0 : L\underline{\beta} = \underline{c}$, will be said to be strongly testable if and only if

$$(5.166) \qquad LUL' = 0$$

for any U constructed as in (5.123). This definition reduces to the strong testability concept of Chapter 4 when $var(\underline{\varepsilon}) = \sigma^2 I$. We note that $L\underline{\beta} = \underline{0}$ is not necessarily a strongly testable hypothesis when V is singular.

5.3 THE INVERSE PARTITIONED MATRIX (IPM) METHOD

Rao (1971) observed that the problem of minimizing $\underline{w}'V\underline{w}$ subject to $\underline{w}'X = \underline{\theta}'$ (and hence the variance of $\underline{w}'\underline{y}$ the UE of $\underline{\theta}'\underline{\beta}$) may be generally expressed as solving the equations

$$(5.173) \quad \begin{bmatrix} V & X \\ X' & 0 \end{bmatrix} \begin{bmatrix} \underline{w} \\ \underline{\lambda} \end{bmatrix} = \begin{bmatrix} \underline{0} \\ \underline{\theta} \end{bmatrix} ,$$

for some Lagrange multiplier $\underline{\lambda}$. By considering the non-explicit form of the g_1-inverse

$$(5.174) \quad \begin{bmatrix} V & X \\ X & 0 \end{bmatrix}^{g_1} = \begin{bmatrix} C_1 & : & C_2 \\ C_3 & & -C_4 \end{bmatrix} ,$$

important identities can be derived from

$$(5.175) \quad \begin{bmatrix} V & X \\ X' & 0 \end{bmatrix} \begin{bmatrix} C_1 & : & C_2 \\ C_3 & : & -C_4 \end{bmatrix} \begin{bmatrix} V & X \\ X' & 0 \end{bmatrix} = \begin{bmatrix} V & X \\ X' & 0 \end{bmatrix} , \qquad \text{or}$$

$$(5.176) \quad \begin{bmatrix} VC_1V+XC_3V+VC_2X'-XC_4X' & : & VC_1X+XC_3X \\ X'C_1V+X'C_2X' & : & X'C_1X \end{bmatrix} = \begin{bmatrix} V & X \\ X' & 0 \end{bmatrix}$$

In this expression we may replace C_2 by C_3', C_3 by C_2' and C_1 and C_4 by their own respective transposes, because the g_1-inverse relation (5.174) for a symmetric matrix V is also satisfied by the transpose of the g_1-inverse. Thus

$$(5.177) \quad \begin{bmatrix} VC_1'V+XC_2'V+VC_3'X'-XC_4'X' & : & VC_1'X+XC_2'X \\ X'C_1'V + X'C_3'X & : & X'C_1'X \end{bmatrix} = \begin{bmatrix} V & X' \\ X & 0 \end{bmatrix}$$

Rao (1973, pp.294-295) notes that the equations

(5.178) $\begin{bmatrix} V & X \\ X' & 0 \end{bmatrix}\begin{bmatrix} W \\ U \end{bmatrix} = \begin{bmatrix} 0 \\ X' \end{bmatrix}$

(5.179) $\begin{bmatrix} V & X \\ X' & 0 \end{bmatrix}\begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} X \\ 0 \end{bmatrix}$, and

(5.180) $\begin{bmatrix} V & X \\ X' & 0 \end{bmatrix}\begin{bmatrix} S \\ T \end{bmatrix} = \begin{bmatrix} V \\ 0 \end{bmatrix}$

are solvable, and we may take

(5.181) $\begin{bmatrix} W \\ U \end{bmatrix} = \begin{bmatrix} C_1 & : & C_2 \\ C_3 & : & -C_4 \end{bmatrix}\begin{bmatrix} 0 \\ X' \end{bmatrix} = \begin{bmatrix} C_2X' \\ -C_4X' \end{bmatrix}$ or $\begin{bmatrix} C_3'X' \\ -C_4'X' \end{bmatrix}$,

(5.182) $\begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} C_1X \\ C_3X \end{bmatrix}$ or $\begin{bmatrix} C_1'X \\ C_2'X \end{bmatrix}$, and

(5.183) $\begin{bmatrix} S \\ T \end{bmatrix} = \begin{bmatrix} C_1V \\ C_3V \end{bmatrix}$ or $\begin{bmatrix} C_1'V \\ C_2'V \end{bmatrix}$

<u>Lemma 5.14</u> (Rao, 1971)

The following results hold

(5.184) $VC_2X' = XC_4X' = XC_4'X' = VC_3'X'$;

$= XC_3V = XC_2'V$;

(5.185) $X'C_2X' = X' = X'C_3'X'$;

(5.186) $X'C_1X = 0$,

$VC_1'X = 0$ and $X'C_1V = 0$;

(5.187) $VC_1VC_1V = VC_1V = VC_1'V = VC_1'VC_1V$;

(5.188) $\mathrm{tr}(VC_1) = r[V : X] - r(X) = \mathrm{tr}(VC_1')$.

Finally over all choices of C_1 and C_4, we have invariance for VC_1V and for XC_4X' in (5.187) and (5.184).

The lemma is proved by continuous substitution of (5.181) through (5.183) into (5.178) through (5.180), and then examining simplifications of (5.176) and (5.177). The trace result is elegantly proved (*op. cit.* p.296) by noting

(5.189) $$r[V : X] + r(X) = r\begin{bmatrix} V & X \\ X' & 0 \end{bmatrix}$$

$$= tr\begin{bmatrix} V & X \\ X' & 0 \end{bmatrix}\begin{bmatrix} C_1 & : & C_2 \\ C_3 & : & -C_4 \end{bmatrix}$$

$$= tr(VC_1+XC_3) + tr(X'C_2)$$

$$= tr(VC_1) + r(X) + r(X) \quad .$$

Theorem 5.15 (Rao, 1971)

Let C_i be as defined in (5.174) for $i = 1,2,3,4$. Then

(i) The BLUE $X\hat{\underline{\beta}}$ of $X\underline{\beta}$ may be obtained from

(5.190) $$X\hat{\underline{\beta}} = XC_2'\underline{y} = XC_3\underline{y} \quad . \quad ,$$

and hence the BLUE of any estimable function $\underline{\theta}'\underline{\beta}$.

(ii) The variance-covariance structure of $X\hat{\underline{\beta}}$ is given by

(5.191) $$var(X\hat{\underline{\beta}}) = \sigma^2.XC_4X' \quad ,$$

and hence the variances and covariance of any pair of estimable functions $\underline{\theta}_1'\underline{\beta}$ and $\underline{\theta}_2'\underline{\beta}$.

(iii) An unbiased estimator of σ^2 is given by

(5.192) $$\hat{\sigma}^2 = \underline{y}'C_1\underline{y}/s \quad ,$$ where

(5.193) $$s = r[V : X] - r(X) \quad .$$

Under the assumption of normality, $\hat{\sigma}^2$ is distributed as as $\sigma^2.\chi_s^2$, and is independent of $X\hat{\underline{\beta}}$.

Proof: Unbiasedness follows from

(5.194) $E(X\hat{\underline{\beta}}) = XC_2^{\prime}X\underline{\beta} = X\underline{\beta}$, and

minimum variance is by construction. The uniqueness of UE's follows from the invariance of XC_4X' and

(5.195) $XC_2^{\prime}\underline{y} = XC_2^{\prime}X\underline{\beta} + XC_2^{\prime}V\underline{\alpha} = X\underline{\beta} + XC_2^{\prime}V\underline{\alpha}$, for some $\underline{\alpha}$

from the model (5.4). From (5.184) and (5.185)

(5.196) $\text{var}(XC_2^{\prime}\underline{y}) = \sigma^2.XC_2^{\prime}VC_2X' = \sigma^2.XC_2^{\prime}XC_4X'$

$= \sigma^2.XC_4X'$.

The well-known formula for the expected value of a quadratic form yields, with (5.186) and (5.188),

(5.197) $E(\underline{y}'C_1\underline{y}) = \underline{\beta}'X'C_1X\underline{\beta} + \sigma^2.\text{tr}(VC_1)$

$= \sigma^2.s$

so that, assuming normality (5.187) assures chi-squaredness and (5.186) assures centrality. The degrees of freedom are clearly s. By Corollary 2.18.2, we have the independence of $\hat{\sigma}^2$ and $X\hat{\underline{\beta}}$, since

(5.198) $XC_2^{\prime}VC_1[V : X] = [0 : 0]$

from (5.184) and (5.186).

The theorem effectively allows us to treat

(5.199) $\hat{\underline{\beta}} = C_2\underline{y}$

as though it was the BLUE of $\underline{\beta}$, subject to

(5.200) $\text{var}(\hat{\underline{\beta}}) = \sigma^2.C_4$.

Theorem 5.16 (Rao, 1972b)

Let $L\underline{\beta} = \underline{c}$ be the hypothesis to be tested. Define

(5.201) $\text{var}(L\hat{\underline{\beta}}) = \sigma^2.LC_4L' = \sigma^2.K$.

The hypothesis is consistent internally, and with the sure equations (5.10) if and only if, for

(5.202) $\underline{u} = L\hat{\underline{\beta}} - \underline{c}$, we have

(5.203) $KK^{g_1}\underline{u} = \underline{u}$.

Given the consistency as above then

(5.204) $$F = \frac{\underline{u}'K^{g_1}\underline{u}}{\hat{\underline{\varepsilon}}'V^{g_1}\hat{\underline{\varepsilon}}} \cdot \frac{s}{p}$$

has central $F(p,s)$ distribution under the hypothesis H_0 and non-central F distribution otherwise.

Proof: Certainly $\underline{u}$ in $C(K)$ is equivalent to (5.203) and implies that $\underline{c}$ is in $C(L)$ for internal consistency. Now consider all sure equations in $L\underline{\beta}$. If, for some $\underline{\alpha}$

(5.205) $\text{var}(\underline{\alpha}'L\hat{\underline{\beta}}) = 0$

then either $\underline{\alpha}'L = \underline{0}'$ or $\underline{\alpha}'LC_4L' = \underline{0}'$. In either case

(5.206) $\underline{\alpha}'(L\hat{\underline{\beta}}-L\underline{\beta}) = 0$.

Thus if $L\underline{\beta} = \underline{c}$ is to be consistent with the sure equations,

(5.207) $\underline{\alpha}'(L\hat{\underline{\beta}}-\underline{c}) = \underline{0}$, for all such $\underline{\alpha}$.

Thus $\underline{u}$ is in $C(K)$ is necessary. The quadratic form in $\underline{u}$ now satisfies directly all the conditions of Theorem 2.18 for central or non-central $\sigma^2.\chi^2$ distributions, because the defining property of K^{g_1} generates all the required equalities. The degrees of freedom are given, as before, by s as in (5.32) and

(5.208) $$p = r(KK^{g_1}K) = r(K) = r(LC_4L') = r\begin{bmatrix} L \\ H_2X \end{bmatrix} - r(H_2X) \quad ,$$

either directly or from (5.141).

Though Björck is credited in Rao (1975) with satisfactory algorithms for calculating C, it may be simpler to apply a check that LUL' satisfies

(5.160) $$LUL' = 0$$

and that $\underline{c}$ is in the column-space $C(L)$. Calculating U and then LUL' may be computationally simpler than solving (5.174) for C_i with $i = 1,2,3,4$.

Pringle and Chalton (1973) have noted that expressions for C_i may be derived by means of Theorem 2.14 by taking $V = XX'$ in place of $S + LL'$ in (2.121) and (2.122). These expressions are useful for the unbiased estimation of $X\underline{\beta}$ and σ^2, but fail to provide a test of hypothesis in the form of (5.204) for an arbitrary estimable function $L\underline{\beta}$. Rao (1978) has criticized Scobey (1975) for postulating that $V + c^2XX'$, for arbitrary non-zero c, will in general provide such a test. In fact it follows from Theorem 5.13 that Scobey's claim is valid if and only if $U = 0$ in (5.160), and thus $C(X)$ is in $C(V)$.

If H_2 is not available, then the method of Zyskind and

Martin, described here in terms of Rao's construction will provide a complete analysis of any hypothesis. This includes a check of consistency, and of strong testability besides the test-statistic itself. Calculation of U, V* and $(X'V^*X)^{g_1}$ may be computationally cumbersome. Similarly for the alternative method of Theorem 5.16. The question of simplifying the computations is partially the concern of Sections 5.5 and 5.6.

5.4 UNIFIED LEAST SQUARES (ULS)

We note that Section 5.1 addressed the estimation problem by a least squares approach minimizing

(5.9) $\quad (\underline{y}-X\underline{\beta})'V^g(\underline{y}-X\underline{\beta})$ subject to

(5.10) $\quad H_2X\underline{\beta} = H_2\underline{y}$.

In Section 5.2 the unconditional minimization of

(5.209) $\quad (\underline{y}-X\underline{\beta})'V^*(\underline{y}-X\underline{\beta})$

subject to V* a g_1-inverse of V and certain optimal properties of the minimum point $\hat{\underline{\beta}}$. The minimal point of

(5.210) $\quad \underline{w}'V\underline{w}$ subject to

(5.211) $\quad \underline{w}'X = \underline{\theta}'$

generates the IPM method of Section 5.3.

Rao (1971) achieved a unified theory of least squares by posing the question of the existence of a class of symmetric matrices M with the properties that

(a) the BLUE $X\hat{\underline{\beta}}$ of $X\underline{\beta}$ is given by $\hat{\underline{\beta}}$ minimizing

(5.212) $(\underline{y}-X\underline{\beta})'M(\underline{y}-X\underline{\beta})$

(b) an UE $\hat{\sigma}^2$ of σ^2 is given by R_0^2/s where

(5.213) $R_0^2 = (\underline{y}-X\hat{\underline{\beta}})'M(\underline{y}-X\hat{\underline{\beta}})$, with

(5.32) $s = r[V : X] - r(X)$

(c) assuming normality, an internally consistent hypothesis $H_0 : L\underline{\beta} = \underline{c}$, and R_1^2 the minimum value of (5.212) subject to H_0, the F-statistic

(5.214) $F = \dfrac{R_1^2-R_0^2}{R_0^2} \cdot \dfrac{s}{p}$, where

(5.208) $p = r\begin{bmatrix} H_2X \\ L \end{bmatrix} - r(H_2X)$

$= r[\mathrm{var}(L\hat{\underline{\beta}})]$

has $F(p,s)$ distribution.

The statements (a), (b) and (c) are the basic results of the least squares OLS method when $V = I$, or more generally when V is non-singular.

<u>Theorem 5.17</u> (Rao, 1971)

An equivalent condition to (a) is that

(5.215) $M = (V+XUX')^{g_1} + K$

where U and K are arbitrary symmetric matrices with

(5.216) $r[V : X] = r(V+XUX')$, and

(5.217) $X'K[V : X] = [0 : 0]$.

An equivalent condition to (a) and (b) is that K in (5.215) additionally satisfies

(5.218) $VKV = 0$.

In that case M simplifies to

(5.219) $M = (V+XUX')^{g_1}$

An equivalent condition to (a), (b) and (c) for an arbitrary testable hypothesis is that U in (5.215) additionally satisfies

(5.220) $XUX' = 0$.

In that case $C(X) \subset C(V)$ and M reduces as

(5.221) $M = V^{g_1}$.

Proof: We examine initially

(5.222) $X'MX\underline{\beta} = X'M\underline{y}$.

The existence of UE's of $X\underline{\beta}$ requires

(5.223) $r(X'MX) = r(X)$.

Then $M+K$ also satisfies (5.223) from (5.217). Minimum variance of (5.222) implies, from (5.109), that

(5.224) $VM'X = XQ = XPX$

for some Q, and because the row-spaces are the same, for some P, though $Q = PX$ is not necessary. Thus there exists U with

(5.225) $(V+XU'X')M'X = X$,

for instance, setting

(5.226) $U = (I-PX)(X'M'X)^{g_1}$ or

(5.227) $XU = (X-XQ)(X'M'X)^{g_1}$.

Now taking M as in (5.219) and noting $C(X)$ is in $C(V+XU'X')$, from (5.225) and (5.223) we obtain

(5.228) $C(XUX') \subset C(X) \subset C(V+XU'X')$,

(5.229) $V = (V+XU'X')(I-M'XU'X')$ and

(5.230) $C(V) \subset C(V+XU'X')$.

These expressions, with (5.225) allow us to write

(5.231) $C(V+XUX') = C(V+XU'X') = C[V : X]$,

using the invariance of rank under matrix transposition. Certainly we may take

(5.219) $M = (V+XUX')^{g_1}$

in (5.225). Then $M+K$ satisfies (5.225) if and only if

(5.232) $(V+XU'X')K'X = 0$

and premultiplying by $X'M'$ with (5.219) and (5.231) gives

(5.233) $X'K'X = 0 = X'KX$, and then

(5.234) $VK'X = 0$.

These conditions on K are clearly sufficient for $M+K$ to replace M in (5.224), and thus yeild BLUE's.

Now consider the expected value

(5.235) $E[(\underline{y}-X\hat{\underline{\beta}})'M(\underline{y}-X\hat{\underline{\beta}})] = 0 + \sigma^2.tr[MAVA']$ for

(5.236) $A = I - X(X'MX)^{g_1}X'M$:

Substituting we obtain, using the uniqueness relation (2.51)

(5.237) $AVA' = A(V+XUX')A' - AXUX'A'$

$$= (V+XUX') - X(X'MX)^{g_1}X' - 0$$

so that, from idempotency, and (5.231)

(5.238) $tr(MAVA') = r[V : X] - r(X)$.

The equivalent condition for $M+K$ to replace M in (5.238) simplifies by (5.233) and (5.234) to

(5.239) $tr(KAVA') = tr(KV) = 0$.

The property of chisquaredness for the expression

(5.213) $R_0^2 = (\underline{y}-X\hat{\underline{\beta}})'M(\underline{y}-X\hat{\underline{\beta}})$

follows from the fact that M is a g_1-inverse of AVA', and

(5.240) $X'M(AVA') = 0$.

The conditions of Theorem 2.18 are satisfied since Q need not be symmetric. For $M+K$ to replace M, from the previous relations on K we require only that

(5.241) $VKVKV = VKV$ and

(5.242) $VKV' = 0$

for the revised conditions of Theorem 2.18. The reduction of (5.241) to zero follows from the idempotency of $V^{\frac{1}{2}}KV^{\frac{1}{2}}$, and the fact that

(5.243) $tr(VK) = tr(V^{\frac{1}{2}}KV^{\frac{1}{2}}) = 0$.

Then $M+K$ is a g_1-inverse of $V + XUX'$, and (5.219) characterizes M.

Now if M is symmetric, then U may be taken to be symmetric. Also $M+K$ symmetric implies K is symmetric and that (5.234) and (5.242) are equivalent.

Finally the class of such M is non-empty since we may take U as in

(5.123) $U = PC_2C_2^{\frac{1}{2}}P'$ or

(5.244) $U = c^2.I$.

The required F-distribution and condition (5.220) have already been established as Theorem 5.13. □

The foregoing is essentially the work of Rao (1971). However, though the result for symmetric matrices is correctly stated, the case of non-symmetric M and the corresponding conditions as originally given in the quoted paper, depend on the validity of the second sentence of page 385, the last statement of the proof. We can neither prove nor disprove the claim, and have therefore typified the conditions differently from Rao, at (5.234), (5.239) and (5.242). We conjecture that the sentence in question is false, and note that in any event (5.242) is not necessary for the construction M+K to yield BLUE's.

5.5 OLSE - BLUE EQUIVALENCE

McElroy (1967) showed that a necessary and sufficient condition for the OLS estimate

(5.245) $\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{y}$

to be a BLUE of $\underline{\beta}$ when

(5.246) $\quad var(\underline{y}) = \sigma^2.V \neq \sigma^2.I$,

has full rank and the model includes a constant term, is that

(5.247) $\quad V = (1-\rho).I + \rho.\underline{1}\underline{1}'$,

where $0 \lesssim \rho < 1$ and $\underline{1}$ denotes a column-vector with all entries unity. Equivalently the observations have equal variances and equal non-negative correlation coefficients.

Balestra (1970) extended McElroy's result. Assume the matrix X is partitioned as

(5.248) $\quad X = [X_1 : X_2]$

where X_1 is known and X_2 is completely unspecified. Then an equivalent condition for the OLSE to coincide with the BLUE of $\underline{\beta}$ under (5.246) is that

(i) a subset of the eigenvectors of V span the column-space $C(X_1)$; and

(ii) the remaining eigenvectors of V correspond to a common eigenvalue, say λ. Then, Balestra claims

(5.249) $\quad V = \lambda I + C_1DC_1' - \lambda C_2C_2'$, where

(5.250) $\quad C = [C_1 : C_2]$

is an orthogonal matrix which gives the respective sets of eigenvectors and D is the appropriate diagonal matrix. We aver that (5.249) is false, since it implies

(5.251) $\quad VC_2 = \lambda C_2 = 0$

and contradicts the non-singularity of V. Furthermore it contradicts the McElroy special case (5.247).

Independently of the apparent contradiction, Balestra indicates that the variance and covariances estimated by the OLS yield the same results as the GLS method obtains, for the coefficients associated with X_2. However the OLS procedure may seriously over- or under-estimate the variances and covariances associated with the coefficients for X_1.

Rao (1968) showed that the OLSE of $X\underline{\beta}$ is BLUE, for arbitrary variance-covariance structure $\sigma^2.V$ if and only if one of the following equivalent conditions is satisfied:

(5.252) $X'VZ = 0$,

(5.253) $V = XA_1X' + ZA_2Z'$, or

(5.254) $V = I + XB_1X' + ZB_2Z'$

for arbitrary non-negative definite A_i and B_i, and Z any matrix with

(5.255) $\mathcal{R}(Z') = \mathcal{R}(I-XX^{g_1})$.

Rao's result generalizes a Zyskind (1967) formulation that exactly $r = r(X)$ eigenvectors of V span $C(X)$, and the Kruskal (1968) condition for equivalence of OLSE and BLUE of $X\underline{\beta}$, namely that for some Q,

(5.256) $VX = XQ$.

This is just Goldman and Zelen's (1964) result (5.91).

In the same paper Rao extends the theorem to obtain necessary and sufficient conditions for the coincidence of OLSE's and BLUE's on a specified subset of estimable functions. Finally he establishes the equivalent conditions for identical BLUE's with respect to different non-scalar variance-covariance structures V and V_0. It transpires that

(5.257) $V = XA_1X' + V_0ZA_2Z'V_0 + V_0$

is sufficient, with A_i and Z as in (5.253). The condition is necessary if

(5.258) V_0 is of full rank , or

(5.259) $C(X) \cup C(V_0Z) = C(I)$.

Kempthorne (1976) discusses the equivalence of the OLSE and BLUE of $X\underline{\beta}$ under arbitrary variance-covariance structure $\sigma^2.V$, using what is claimed to be an algebraically simpler construction than using conditional (i.e. generalized) inverses. His method amount, in effect, to solving explicitly (5.178). Then $W'\underline{y}$ is the BLUE of $X\underline{\beta}$, and W may be written as

(5.260) $W = (XX^g) + (I-XX^g)N$

where N is any solution of

(5.261) $(I-XX^g)V(I-XX^g)N + (I-XX^g)VXX^g = 0$.

This condition reduces to (5.252) if and only if the OLSE and BLUE of $X\underline{\beta}$ are identical. It is also noted that the BLUE of $X\underline{\beta}$ is given by a projection, i.e. by a symmetric idempotent transformation, if and only if N in (5.260) and (5.261) is restricted to choices of the form NXX^g, which exist whenever N exists. Equivalently the OLSE represents one method of calculating the BLUE.

Krämer (1980) formulates the problem of finding those $\underline{y}$, given full column-rank X and positive definite V, such that the OLSE of $\underline{\beta}$ and the BLUE are equal. The resulting condition appears to depend only on the positive definiteness of V, and therefore the result is easily generalized for X not of full-column rank, by noting that the OLSE and BLUE of $X\underline{\beta}$ will be equal. A proof follows directly from any full-rank reparametrization.

The import of the foregoing results is not only that certain estimates may coincide under a mistaken assumption of a scalar-variance-covariance structure. We also note that computational simplicities by means of the usual least-squares algorithms and methods may be available even when a particular structure is known. From a theoretical point of view there appears to be matter for further research on these issues.

5.6 COMPUTATIONAL ISSUES

Two further matters connected with possible computational convenience may be recorded here. Rao (1973a, 1975) and Harville (1981) have defined wider classes of BLUE's than the GLS estimates obtained in Sections 5.1 to 5.4. While the minimum variance estimates are unique, the assumption of a singular variance-covariance structure allows the class of estimators to be non-trivially extended.

From a completely different point of view, Kourouklis and Paige (1981) have suggested a radical departure from the generalized-inverse based methods of the proceeding sections. They extend to the singular V case the approach of Golub and Styan (1973), mentioned in Section 3.1 where $V = I$. Suppose that

(5.262) $V = FF'$

for some full column-rank F, where F is either known, or may be calculated from V by, say, a Cholesky factorization. Then writing the model as

(5.263) $\underline{y} = X\underline{\beta} + F\underline{u}$

where $\underline{u}$ has zero mean and scalar variance, the constrained least squares (CLS) method of Albert (1972) is used to minimize $\underline{u}'\underline{u}$, subject to (5.263). It is noted when the minimum point $(\hat{\underline{\beta}}, \hat{\underline{u}})$ exists, $\hat{\underline{u}}$ is unique. A solution exists if and only if (5.263) holds, and in that case it is shown to be such that

(5.264) $$\begin{bmatrix} V & X \\ X' & 0 \end{bmatrix} \begin{bmatrix} \hat{\underline{r}} \\ \hat{\underline{\beta}} \end{bmatrix} = \begin{bmatrix} \underline{y} \\ \underline{0} \end{bmatrix} ,$$

for $\hat{\underline{r}}$ with $V\hat{\underline{r}} = F\hat{\underline{u}}$. The method is shown to be numerically more stable over the examples studied, and general reliability and superiority of the method is postulated.

CHAPTER 6

OUTLIERS UNDER ARBITRARY VARIANCE

It has been noted that recent literature gives evidence of increased interest in outlier theory. In this chapter we are concerned with the generalized sweep-out method and its extension to a generalized hypothesis test for outliers. We first examine such a test for a specified subset, and then review the Pandora's box of searches for an unknown set of possible outliers of unknown size.

The test-statistic has interesting properties when applied to principal components and ridge-methods. These are examined along with the associated problem of the existence of influential observations, which may be outliers, or points of *leverage* in the regression relationship. A brief account is given of a possible Bayesian approach.

6.1 TESTS FOR OUTLIERS

Suppose that the model

(5.4) $\qquad y = X\underline{\beta} + \underline{\varepsilon}$, and

(5.1) $\qquad \text{var}(\underline{\varepsilon}) = \sigma^2 . V$

is adopted. It follows from Theorem 5.10 that the space of $\underline{y}$ is given by $C(V+XUX')$ with XUX' disjoint from $C(V)$. It has been noted that special case results are obtained when

(5.165) $XUX' = 0$,

inter alia Corollary 5.1.2 and Theorem 5.17.

If the above model is false because X and V do not define the appropriate space of possible observations $\underline{y}$, then it may happen that the sure equations

(5.10) $H_2X\underline{\beta} = H_2\underline{y}$

are contradicted. Equivalently, the equations are inconsistent for a given $\underline{y}$, or that

(6.1) $(V+XUX')V^*\underline{y} = \underline{y}$

is false for the given $\underline{y}$ and V^* a g_1-inverse from (5.121). Certainly if either (5.10) or (6.1) fails, then the model is inadequate. This may or may not be due to the presence of anomalous observations (in an otherwise satisfactory model).

The model may on the other hand correctly specify the space of observations $\underline{y}$, but the presence of one or more observations from tails of the underlying family of correlated distributions may seriously affect the resulting GLS estimates $X\hat{\underline{\beta}}$ of $X\underline{\beta}$, the UE $\hat{\sigma}^2$ of σ^2 or both.

Initially let us suppose that a particular subset of the observations is the subject of suspicion. Without loss of generality we may take the subset to be $\underline{y}_2$ in the partitioning of the model (5.4) to

(6.2) $$\begin{bmatrix} \underline{y}_1 \\ \underline{y}_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \underline{\beta} + \begin{bmatrix} \underline{\varepsilon}_1 \\ \underline{\varepsilon}_2 \end{bmatrix} ,$$ and

(6.3) $$\sigma^2.V = \sigma^2 \begin{bmatrix} V_{11} & : & V_{12} \\ V_{21} & : & V_{22} \end{bmatrix} .$$

Clearly the effect of an anomalous but legitimate sub-vector $\underline{\varepsilon}_2$ intrudes into the whole vector of observations $\underline{y}$, though it may not be as badly smeared across the observations as it will be over their GLS estimates $\hat{\underline{y}}$. Such a legitimate $\underline{\varepsilon}_2$ has an effect in the model which is described *a priori* by

(6.4) $$\begin{bmatrix} V_{12} \\ V_{22} \end{bmatrix} = V \begin{bmatrix} 0 \\ I \end{bmatrix} .$$

The removal of the second part of the partitioned data set would reduce the model to

(6.5) $$\underline{y}_1 = X_1\underline{\beta} + \underline{\varepsilon}_1 ,$$ with

(6.6) $$\text{var}(\underline{\varepsilon}_1) = \sigma^2.V_{11} .$$

If the estimates from the reduced data set were to differ markedly from those derived in a GLS analysis of the complete data, then the model, while correctly specifying the space of the observations as $C(V+XIX')$, may inadequately specify the mean $X\underline{\beta}$, or the variance $\sigma^2.V$. For instance

(6.7) $$C(H_1DH_1') = C(H_1EH_1')$$

for D and E full-rank diagonal matrices and H_1 a matrix of orthonormal columns. However we may take D and E such that the matrices clearly obey

(6.8) $$H_1DH_1' \neq H_1EH_1' .$$

We note the idempotency of the transformation of the observations $\underline{y}$ that leads to GLS estimates, and the fact that the GLS estimates of $X\hat{\underline{\beta}}$ and $\hat{\underline{\varepsilon}}$ are uncorrelated, from (5.17) and Theorem 5.2. We are thus led to examine the model

(6.9) $$\begin{bmatrix} \underline{y}_1 \\ \underline{y}_2 \end{bmatrix} = [X : W]\begin{bmatrix} \underline{\beta} \\ \underline{\lambda} \end{bmatrix} + \underline{\varepsilon}$$

$$= \begin{bmatrix} X_1 & : & V_{12} \\ X_2 & & V_{22} \end{bmatrix}\begin{bmatrix} \underline{\beta} \\ \underline{\lambda} \end{bmatrix} + \begin{bmatrix} \underline{\varepsilon}_1 \\ \underline{\varepsilon}_2 \end{bmatrix} .$$

Clearly this amounts to taking

(6.10) $$W = V\begin{bmatrix} 0 \\ I \end{bmatrix} .$$

The GLS estimates of $(X\underline{\beta}+W\underline{\lambda})$ in this model depend on the choice of U_0 and V^* such that

(6.11) $$V_0 = V + [X : W]U_0\begin{bmatrix} X' \\ W' \end{bmatrix}$$ has

(6.12) $$C(V_0) = C[V : X : W] .$$

It is apparent from (6.10) that we may take

(6.13) $$U_0 = \begin{bmatrix} U & 0 \\ 0 & 0 \end{bmatrix} ,$$ where

(6.14) $$C(V+XUX') = C[V : X] .$$ Then

(6.15) $$V^* = V_0^{g_1} = (V+XUX')^{g_1}$$ and

(6.16) $$[X : W]'V^*[X : W] = \begin{bmatrix} X'V^*X & W'V^*X \\ X'V^*W & W'V^*W \end{bmatrix}$$

$$= B \quad ,$$ say.

Recalling that V^* is a g_1-inverse of V from (5.121), and noting that

(6.17) $$W'V^*W = [0 : I]V[0 : I]' = V_{22} \quad ,$$

we may simplify (6.16), and find a g_1-inverse B^{g_1} by means of Theorem 2.12. Without loss of generality we take the terms Q and G_i in (2.91) and (2.94) as

(6.18) $$G_i = (X'V^*X)^g = G \quad , \quad \text{and}$$

(6.19) $$Q = V_{22} - WV^*X(X'V^*X)^{g_1}X'V^*W \quad .$$

From (5.126) through to (5.129), (6.19) reduces to

(6.20) $$Q = [0 : I]N[0 : I]' = N_{22}$$

where $\sigma^2 N$ is the variance structure of $\hat{\underline{\varepsilon}}$ obtained when $\underline{\lambda} = \underline{0}$ in (6.9). Accordingly

(6.21) $$B^{g_1} = \begin{bmatrix} G + GX'V^*WN_{22}^gW'V^*XG & : -GX'V^*WN_{22}^g \\ -N_{22}^gW'V^*XG & N_{22}^g \end{bmatrix} \quad .$$

We may write the GLS estimate of $\underline{y}$ under (6.9) as

(6.22) $$\tilde{\underline{y}} = [X : W]B^{g_1}[X : W]'V^*\underline{y} \quad .$$

As before write the GLS estimate of $\underline{y}$ under (5.4) as

(6.23) $$\begin{aligned} \hat{\underline{y}} &= X(X'V^*X)^gX'V^*\underline{y} = XGX'V^*\underline{y} \\ &= XC\underline{y} \quad , \quad \text{say.} \end{aligned}$$

Then substituting (6.21) and (6.23) into (6.22), we begin simplifying the expression. Note that

(6.24) $$\begin{aligned} N &= (I-XC)V(I-XC)' = (I-XC)V \\ &= (I-XC)(V+XUX')(I-XC)' \end{aligned}$$

$$= V + XUX' - X(X'V^*X)^{g_1}X'$$

$$= V + XUX' - XGX'$$

Then, after some tedious algebra collecting the terms in (6.22),

$$(6.25) \qquad \tilde{\underline{y}} = \hat{\underline{y}} + N\begin{bmatrix} 0 \\ I \end{bmatrix} N_{22}^{g}[0 : I]NV^*\underline{y}$$

$$= \hat{\underline{y}} + \underline{a} \qquad , \qquad \text{say.}$$

Thus for the usual definitions of

$$(6.26) \qquad \tilde{\underline{\varepsilon}} = \underline{y} = \tilde{\underline{y}} \qquad , \qquad \text{and}$$

$$(6.27) \qquad \hat{\underline{\varepsilon}} = \underline{y} - \hat{\underline{y}} \qquad ,$$

we have that

$$(6.28) \qquad \hat{\underline{\varepsilon}} - \tilde{\underline{\varepsilon}} = \tilde{\underline{y}} - \hat{\underline{y}} = \underline{a}$$

is independent of $\hat{\underline{y}}$ and $\tilde{\underline{\varepsilon}}$. Directly, or by applying Theorem 5.7, we have

<u>Theorem 6.1</u> (Dunne)

The F-statistic

$$(6.29) \qquad F = \frac{\hat{\underline{\varepsilon}}'N^{g_1}\hat{\underline{\varepsilon}} - \tilde{\underline{\varepsilon}}'N^{g_1}\tilde{\underline{\varepsilon}}}{\tilde{\underline{\varepsilon}}'N^{g_1}\tilde{\underline{\varepsilon}}} \cdot \frac{t}{s-t}$$

for the residual vectors $\hat{\underline{\varepsilon}}$ and $\tilde{\underline{\varepsilon}}$ in (6.26) and (6.27), and with the additional assumption of normality has central $F(s-t,t)$ distribution under the null hypothesis $H_0 : \underline{\lambda} = \underline{0}$, or more strictly, $N\begin{bmatrix} 0 \\ I \end{bmatrix}\underline{\lambda} = \underline{0}$. The degrees of freedom are given by

(6.30) $\quad s = r(N)$, and

(6.31) $\quad t = r(N) - r(N_{22})$.

Proof: The crux of the matter is that $\underline{a}$ is in $C(N)$, so that

(6.32)
$$\underline{a}'N^{g_1}\underline{a} = (\hat{\underline{\varepsilon}}-\tilde{\underline{\varepsilon}})'N^{g_1}(\hat{\underline{\varepsilon}}-\tilde{\underline{\varepsilon}})$$
$$= \hat{\underline{\varepsilon}}'N^{g_1}\hat{\underline{\varepsilon}} - \tilde{\underline{\varepsilon}}'N^{g_1}\tilde{\underline{\varepsilon}} \quad .$$

Further, since from (5.115) and (5.78)

(6.33) $\quad NV^*X = (I-XC)VV^*X = 0$,

we obtain, from (6.15),

(6.34)
$$NV^*VV^{*\prime}N = NV^*(V+XUX')V^{*\prime}N$$
$$= (I-XC)VV^*V(I-XC)'$$
$$= (I-XC)V(I-XC)' = N \quad .$$

It therefore follows that the variance of $\underline{a}$ is

(6.35) $\quad \text{var}(\underline{a}) = \sigma^2.N\begin{bmatrix}0\\I\end{bmatrix}N_{22}^{g}[0 : I]N$,

so that (6.32) is central $\sigma^2.\chi_f^2$ under H , with degrees of freedom

(6.36) $\quad f = r(N_{22}) = s-t$,

and is independent of $\tilde{\underline{\varepsilon}}$ and hence of the denominator. The degrees of freedom of the denominator are given by

(6.37) $$t = r[V : X] - r\left(X : V\begin{bmatrix}0\\I\end{bmatrix}\right)$$

$$= r[N : 0] - r\left(N\begin{bmatrix}0\\I\end{bmatrix}\right)$$

$$= r(N) - r(N_{22})$$ □

Corollary 6.1.1 (Dunne)

Under the model (6.9) the F-statistic (6.29) has non-central $F(s-t,t;\ \gamma)$ distribution where

(6.38) $$\gamma = \underline{\lambda}'N_{22}\underline{\lambda} \quad .$$

Proof: From (6.33) and (6.34) it is easy to verify that the quadratic form $\underline{a}'N^{g_1}\underline{a}$ of (6.32) satisfies all the properties of Theorem 2.18 and is therefore a $\sigma^2.\chi^2(\gamma)$ variate. Its independence of the denominator is not affected by the non-centrality, and the parameter γ simplifies directly to (6.38). □

We have therefore a proper hypothesis test of $H_0 : N\begin{bmatrix}0\\I\end{bmatrix}\underline{\lambda} = \underline{0}$, or, in an extension of the terminology of Searle discussed in Section 4.4, of the effect of fitting $V\begin{bmatrix}0\\I\end{bmatrix}$ last. Note that the condition for Corollary 5.1.2 is satisfied by the variate $\underline{a}$, which thus also obeys Theorem 5.13. However further simplifications arise. From (6.1) and (6.24) we have

$$(6.39)\qquad NV^*\underline{y} = (I-XC)VV^*\underline{y}$$

$$= (I-XC)(V+XUX')V^*\underline{y}$$

$$= (I-XC)\underline{y} = \hat{\underline{\varepsilon}} \quad .$$

Thus

$$(6.40)\qquad \underline{a} = N\begin{bmatrix}0\\ I\end{bmatrix}N^{g}_{22}[0 : I]\hat{\underline{\varepsilon}} \quad .$$

Partitioning $\underline{y}$ and $\tilde{\underline{y}}$ conformably with (6.40) we may write

$$(6.41)\qquad \begin{bmatrix}\tilde{\underline{y}}_1\\ \tilde{\underline{y}}_2\end{bmatrix} = \begin{bmatrix}\hat{\underline{y}}_1\\ \hat{\underline{y}}_2\end{bmatrix} + N\begin{bmatrix}0\\ I\end{bmatrix}N^{g}_{22}[0 : I]\begin{bmatrix}\hat{\underline{\varepsilon}}_1\\ \hat{\underline{\varepsilon}}_2\end{bmatrix}$$

and clearly, premultiplying by $[0 : I]$ yields

$$(6.42)\qquad \tilde{\underline{y}}_2 = \hat{\underline{y}}_2 + \hat{\underline{\varepsilon}}_2 = \underline{y}_2 \quad ,$$

whence

$$(6.43)\qquad \tilde{\underline{\varepsilon}}_2 = \underline{0} \quad .$$

Clearly the intuitive explanation, at the beginning of this section, which motivated the choice of W in (6.10) is vindicated. Its effect is to remove $\underline{y}_2$ from the estimates of $X\underline{\beta}$ and σ^2. It will now be important to determine whether or not $X\underline{\beta}$ remains estimable, but we defer this question to note

Corollary 6.1.2 (Dunne)

The F-statistic (6.29) may be written as

$$(6.44)\qquad F = \frac{\hat{\underline{\varepsilon}}_2' N^{g}_{22}\hat{\underline{\varepsilon}}_2}{\hat{\underline{\varepsilon}}' V^{g_1}\hat{\underline{\varepsilon}} - \hat{\underline{\varepsilon}}_2' N^{g}_{22}\hat{\underline{\varepsilon}}_2} \cdot \frac{s-m}{m}$$

where the degrees of freedom are given by

$$(6.45)\qquad m = r(N_{22}) = r[var(\hat{\underline{\varepsilon}}_2)] \quad ,$$

and

$$(6.46)\qquad s = r(N) = r[var(\hat{\underline{\varepsilon}})] \quad .$$

Proof: The numerator and denominator will reduce to (6.44) providing that (6.32) can be written as

$$(6.47) \qquad \underline{a}'N^{g_1}\underline{a} = \hat{\underline{\varepsilon}}_2' N^{g}_{22} \hat{\underline{\varepsilon}}_2 = \hat{\underline{\varepsilon}}_2' N^{g_{12}}_{22} \hat{\underline{\varepsilon}}_2$$

But the latter follows directly from substituting (6.40) into the left-hand side term. The rank result is the same as before, only noting the relationship to the variance matrices, which are deduced from (2.172) in Theorem 2.18. □

The form of F given by (6.44) is related to a number of well-known statistics associated with outlier searches in data. Taking $V = I$ throughout the development, and specifically in (6.44) reduces to the implicit F-ratio of a number of authors, as we will shortly exhibit. It will however be convenient to note some further properties.

In the model (6.2), suppose that the suspected observations represent functions which are estimable from the reduced-data model

$$(6.48) \qquad \underline{y}_1 = X_1\underline{\beta} + \underline{\varepsilon}_1 \quad .$$

Equivalently we take

$$(6.49) \qquad r(X) = r\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = r(X_1) \quad .$$

Now consider the effect of substituting for $\underline{y}_2$ a missing plot estimate obtained from the $\underline{y}_1$ observations alone. One such estimate in the general case may be to take

$$(6.50) \qquad \underline{y}_2 = \overset{\circ}{\underline{y}}_2 = A\overset{\circ}{\underline{y}}_1 = AX_1(X_1'V_{11}^{*}X_1)^{g_1}X'V_{11}^{*}\underline{y}_1$$

the GLS of $X_2\underline{\beta}$ in the model (6.48). Equation (6.49) guarantees the existence of some A such that

(6.51) $AX_1 = X_2$

and hence $\overset{o}{\underline{y}}_2$ is unique over all solutions $\overset{o}{\underline{\beta}}$. In the general situation this amounts to adding

(6.52) $\begin{bmatrix} \underline{y}_1 \\ \overset{o}{\underline{y}}_2 \end{bmatrix} - \begin{bmatrix} \underline{y}_1 \\ \underline{y}_2 \end{bmatrix} = \begin{bmatrix} 0 \\ \overset{o}{\underline{y}}_2 - \underline{y}_2 \end{bmatrix}$

to the term $\underline{y}$ in (6.25). The simplification of (6.52) to the unique expression (6.40) depends crucially on the observation term $\underline{y}$ being in the column-space C(V+XUX'). Equivalently, the uniqueness of $NV^*\underline{y}$ over all g -inverses V^* requires that the added vector in (6.52) also be in the space. Therefore, it is only under certain conditions, that a 'missing plots' type of approach leading to adjusted residuals can be successfully pursued. Certainly V = I leads to the Gentleman and Wilk (1975b) or John and Draper (1978) formulations. It appears that

(6.53) $V_{12} = 0$

will also ensure the space condition is satisfied. We conjecture from (6.53) that diagonal V is necessary for the development of successive independent adjusted residuals.

Taking V = I, we note that F in (6.29) and (6.44) reduces to

(6.54) $F = \dfrac{\hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}} - \tilde{\underline{\varepsilon}}'\tilde{\underline{\varepsilon}}}{\tilde{\underline{\varepsilon}}'\tilde{\underline{\varepsilon}}} \cdot \dfrac{t}{s-t}$

$$= \frac{\hat{\underline{\varepsilon}}_2' N_{22}^{g_1 2} \hat{\underline{\varepsilon}}_2}{\hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}} - \hat{\underline{\varepsilon}}_2' N_{22}^{g_1 2} \hat{\underline{\varepsilon}}_2} \cdot \frac{s-m}{m} \quad ,$$ where

(6.55) $N = I - X(X'X)^{g_1}X'$.

(6.56) $\tilde{\underline{\varepsilon}} = (I - X_1(X_1'X_1)^{g_1}X_1')\underline{y}_1$, and

(6.57) $\hat{\underline{\varepsilon}} = (I - X(X'X)^{g_1}X')\underline{y}$.

When X_2 has full-row rank and is of the form (6.51), it may happen that N_{22}, a principal submatrix of N in (6.54), is non-singular and

(6.58) $r(X_2) = r(N_{22}) = m$.

In that case, from (2.1) it follows that

(6.59) $N_{22}^{g_1} = N_{22}^{-1}$.

Certainly since N is always positive semidefinite whatever the choice of V, if we take suspicious observations one at a time, we have that the diagonal entries of N are positive and thus, for any appropriate i,

(6.60) $N_{22}^{g_1} = [n_{ii}]^{-1} = 1/n_{ii}$

the inverse i^{th} diagonal element of N. Then

(6.61) $$F = \frac{(\hat{\varepsilon}_i^2/n_{ii})(s-1)}{(\hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}} - \hat{\varepsilon}_i^2/n_{ii})} \quad ,$$ where

where s as in (6.46) reduces to $n - r(X)$. Writing

(6.62) $E = \hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}}$

we see that (6.60) represents the square of the statistic

(6.63) $$t_i = \frac{(\hat{\varepsilon}_i/\sqrt{n_{ii}E})(n-r-1)}{\sqrt{1 - (\hat{\varepsilon}_i^2/n_{ii}E)}} \quad ,$$

of Ellenberg (1973), which he names the i^{th} *standardized* residual.

Grubbs (1950) noted a monotonically related ratio for a simple random normal sample, i.e. for $r = 1$, and expresses it in terms of

(6.64) $T_i = \hat{\varepsilon}_i/\hat{\sigma}$

the i^{th} *studentized* residual of Pearson and Chandra Sekar (1936). They discuss a paper of Thompson (1935) who proved that t_i in (6.63) follow a Student's $t_{(n-2)}$ distribution in this special case. It appears that the names of the ratios (6.63) and (6.64) have become established by usage, even though the more recent results indicate that the names might more appropriately be conversely applied. There may have been computational simplicities which underpinned this usage, before the advent of the modern computer, but n_{ii} is constant in the random sample. Lund (1975) points out the usage and implies the interchange of the terms.

Dixon (1950, 1951) proposed and examined several types of ratio statistics, also noting the F-statistic (6.54) for double outliers, i.e. for X_2 of order $(2 \times k)$. He describes two forms of models for contamination in random samples which he labels *location* error and *scalar* error to distinguish aberrant observations from possible simple sources

(6.65) Normal $(\mu+\lambda\sigma, \sigma^2)$ and

(6.66) Normal $(\mu, \lambda^2\sigma^2)$.

These models serve to describe one- and two-directional outlier terms respectively. Tukey (1960) in discussing Anscombe (1960) and Daniel (1960) notes that mixtures of distributions may also serve as contamination models.

Daniel was concerned with the location of outliers in factorial experiments, giving specific attention to the possible existence of maverick constrasts in two-way layouts and 2^k factorial designs. Three possible sources of outliers listed are single large interactions in one cell, an extreme random error term or mistakes of a technical kind. Also noted is the advantage of full and even partial replication in factorial experiments, in assisting outlier detection. Gentleman and Wilk (1975b) examined two-way layouts under the presence of zero, one or two outliers.

Andrews (1971) established the joint distribution of the OLS ratios, the *normed* residuals

$$(6.67) \qquad \underline{u} = \hat{\underline{\varepsilon}}/\sqrt{\hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}}} = \hat{\underline{\varepsilon}}/\sqrt{E}$$

when $V = I$ in the general case, and under the assumption of normality. Significance levels are determined for some tests based on $\underline{u}$. Behnken and Draper (1972) determine the variances of individual residual terms and examine the variance patterns of various models, e.g. regression models with a constant term. They, like Andrews, suggest that the entries in (6.67) should be divided by the correcting constant, for example $\sqrt{n_{ii}}$. Tietjen, Moore and Beckman (1973) examined simple linear regression. Applying the correcting constant they examine a test statistic

$$(6.68) \qquad R_n = \max|\hat{\underline{\varepsilon}}_i/\hat{\sigma}n_{ii}^{\frac{1}{2}}|$$

for which critical values were obtained by a Monte Carlo study. They were particularly concerned with the possibility that the arrangement of the X's could have any effect on the critical values. There appeared to be such an effect but it was negligible for practical purposes and a table of Grubbs (1969) could be used with a minor modification.

It is the question of the maximum of the set of n correlated residual terms of whatever type that provides a stumbling block to the calculation of critical values. The Bonferroni inequality was applied by Ellenberg (1973) to obtain the conservative upper bound based on a significance level α/n for the $\max(t_i)$ in (6.63), following a suggestion of Stefansky (1972). Prescott (1975) examined critical values for a monotonic function of the F-statistic (6.54) for multiple outliers. His method falls into the same difficulty because the corresponding critical values of F are not available for the appropriate percentage points, e.g. $\alpha/\binom{n}{p}$ for p possible outliers. The excessively conservative nature of such approximations is noted by Hawkins (1980, p.61), *inter alia*.

Ellenberg (1976) suggested and examined a second order Bonferroni inequality to obtain upper and lower bounds for the percentage point associated with α/n in a test for a single outlier. He also established the equivalence in a general linear regression model of the statistics F of (6.61), t_i of (6.63) and T_i of (6.64). Doornbos (1981) has shown that the first Bonferroni inequality will yield a probability between $\alpha-\frac{1}{2}\alpha^2$ and α, for the critical value to be exceeded by $\max|t_i|$, on condition that all the correlations between residuals are smaller in absolute value than certain tabulated values.

Beckman and Trussell (1974) established the t-distribution of the standardized residual t_i of (6.63), for X of arbitrary rank $r < n$. They examined the effects on residuals and the sum of the squared residuals caused by adding an additional data point in a multiple regression model. Such a data point must be an observation whose mean is estimable in terms of the previous observations, otherwise the residual is zero. Their proof is a special case of the argument of Seber (1977, p.291) quoted in Section 4.6. However they assumed that X had full column rank. It is clear from Theorem 6.1 that X need not satisfy the condition. Essentially they define the i^{th} *recursive* residual, which is further examined by John and Draper (1978).

Gentleman and Wilk (1975b) have examined the issues of *masking* and *swamping* associated with testing for an unspecified set of possible outliers of unknown size. When more than one outlier is present, they can interact in such a way as to be impervious to direct methods of inspection. They suggest, that if aggregated properly, detection may be facilitated. However iterative approaches need not always be reliable. While the notion of many outliers may be poor in concept, they make the interesting observation that a subset of say k data points, with $k << n$, may be termed an outlier subset if k degrees of freedom are required for their joint explanation.

We pose the converse question: can a subset be deemed to be outlying if k degrees of freedom are not essential for their explanation? If so, for the $V = I$ case, are we back at model (6.9) and therefore a formulation of John and Draper (1978), or does this imply that we ought to restrict the search for suspicious subsets to those in which N has full-rank, and F in (6.54) has N_{22}^{-1} in place of N_{22}^{g1}? Hawkins, in personal discussions, has queried the appropriateness of the general form of (6.54) in a Gentleman and Wilks (1975b) type of search for the "k most likely outlier subset", which involves finding the largest value of F. On the other hand it is possible (although not highly likely) that outliers can occur at a set of design points which are linearly related, even in relatively balanced designs, with or without replications. It may be that in taking data subsets of size k, some non-comparable F-statistics are obtained, i.e. F statistics in which the degrees of freedom pairs are not all identical.

The existence of *clean* and *dirty* data subsets is noted by Gentleman and Wilks. Hawkins (1980, p.51-72) discusses multiple outliers in a random sample when the number of outliers is unknown (as is most likely to be the case). The *masking* effect (Murphy, 1951) or loss of power associated with presence of more outliers than the number suspected, originates in lower values for F caused by the contribution of unsuspected terms to the denominator. Thus an outlier (even an extreme observation) escapes detection because of the presence of other outliers. For this reason it is likely that nested and stepwise procedures of detection will differ in results. Hawkins (p.57) also notes the fact that a similar statistic E_k to F in (6.54), due to Tietjen and Moore (1972) and related through the well-known invertible transformations between the two kinds of β-distributions, is not robust in either direction against mis-specification of the number k of outliers. If F is also non-robust, then at least in a pre-screen of data, slight overestimation of k will be associated with high power for a number of outliers lower than k. *Swamping* (Fieller, 1976) describes the converse effect of several large errors contributing to the spurious declaration of a valid or typical observation as an outlier, i.e. as a member of an outlying set. Barnett and Lewis (1978, p.71) provide a simple example of each of the effects considering the sets

3, 4, 7, 8, 10, 13, 951

3, 4, 7, 8, 10, 949, 951 .

Hawkins (pp.63-67) notes that E_k is monotonically related to the squared partial multiple correlation coefficient used in regression to test the predictive power of a set of predictors. Thus E_k assesses the predictive power of $\begin{bmatrix} 0 \\ I \end{bmatrix}$ in the model (6.9) with $V = I$, and X representing only the constant term. The statistics $\underline{r}$ of (6.67) represent the simplest case of a statistic $T_{n:i}$ which may be used in a stepwise procedure to identify outliers. These $T_{n:i}$ are the equivalent of successive partial correlations in a multiple regression.

One such method is the recursive approach of John and Draper (1978), which amounts to a partitioning of F in (6.54) under the special case of X_2 and N_{22} of full row-rank, say k. Then for a given ordering of the last k rows of X, which presumably reflects an *a priori* ranking of aberrant random residuals, it is possible to generate a sequence of recursive residuals by defining at each i^{th} step the term r_{n+1-i} to be the residual of y_{n+1-i} from its estimate derived from the preceding observations for $i = 1,2,...,k$. These recursive residuals are independent, and may be adjusted or *normalized* so as to have common variance after adjustment. It is suggested that such partitionings be examined for a small subset of the $\binom{n}{k}$ possible F-statistics (6.54). Further it is claimed that the subset may be restricted to the $\binom{m}{k}$ partitionings of F-statistics obtained by considering (6.54) restricted to the m largest estimated residuals in the full model. A Monte Carlo study is presented with a view to approximating critical values of max F when $k = 0,1,2$. In the theory attention is restricted to X of full column-rank, and to two-way layouts, though neither of these assumptions is required under Theorem 6.1. Extension of the simulation to the $k = 3$ case is given in Draper and John (1980). Furnival and Wilson (1974) have suggested a "leaps and bounds" algorithm which locates max F and the subset of size k corresponding to it.

Rosner (1975) discusses the detection of many outliers and compares for a random sample, an equivalent to the F-statistic of (6.54) and other measures. The latter included kurtosis, studentized range and trimmed mean and standard deviation statistics. In his simulation the F-based statistic proved slightly superior to the rest. Hawkins (1980) gives attention to such measures, and to a more distributional formulation, namely slippage tests, which can yield as special cases the F-type test of some applications discussed in this chapter.

Thus far the discussion has been restricted to the $V = I$ variance-covariance structure. The problem of BLU estimation under non-singular V which is not scalar and allows correlated residuals was examined by Aitken (1933). This relaxation of the variance assumption has of course no effect on Theorem 6.1, nor the further relaxation of non-singularity. In any event, the estimated residuals, under whatever V, are correlated and have a singular structure. It is not apparent whether idempotency (as when $V = I$, or whenever the BLUE's and OLSE's coincide) yields any simplification of outlier detection beyond the result of the theorem.

The development of Theorem 6.1 arose in an interesting manner. Troskie and Dunne (1981) began considering the statistic t_i of (6.63) for the case of X not of full column-rank and $V \neq I$ but non-singular. The exact distributional results associated with (6.63) seemed appealing enough to warrant the examination. By adopting the equivalent F-statistic it was possible to generalize to the statistic (6.61) and (6.54) whether or not N_{22} had full rank, but without a linear model formulation. The key to this extension is that $\hat{\underline{\varepsilon}}_2$ has variance-covariance structure given by

(6.69) $\sigma^2.Q = \sigma^2.[0 : I]N[0 : I]' = \sigma^2.N_{22}$,

from (6.20). Thus directly the conditions of Theorem 2.18 are satisfied for the quadratic form $\hat{\underline{\varepsilon}}_2'N_{22}^{g_{12}}\hat{\underline{\varepsilon}}_2$ if the model (6.2) is correct, and

(6.70) $\hat{\underline{\varepsilon}}_2'N_{22}^{g_{12}}\hat{\underline{\varepsilon}}_2 \sim \sigma^2.\chi_f^2$

for f as in (6.36). The result uses the marginal distribution of $\hat{\underline{\varepsilon}}_2$ in $\hat{\underline{\varepsilon}}$, and writing

(6.71) $\hat{\underline{\varepsilon}}_2'N_{22}^{g_{12}}\hat{\underline{\varepsilon}}_2 = \hat{\underline{\varepsilon}}'\begin{bmatrix} 0 & 0 \\ 0 & N_{22}^{g_{12}} \end{bmatrix}\hat{\underline{\varepsilon}} = \hat{\underline{\varepsilon}}'N^*\hat{\underline{\varepsilon}}$, say ,

allows the quadratic form in $\hat{\underline{\varepsilon}}$ to be related to another quadratic in the same variable, namely

(6.72) $\hat{\underline{\varepsilon}}'V^{g_1}\hat{\underline{\varepsilon}} - \hat{\underline{\varepsilon}}_2'N_{22}^{g_{12}}\hat{\underline{\varepsilon}}_2 = \hat{\underline{\varepsilon}}'N^{g_1}\hat{\underline{\varepsilon}} - \hat{\underline{\varepsilon}}'N^*\hat{\underline{\varepsilon}}$.

They showed that

(6.73) $N(N^{g_1}-N^*)NN^*N = 0$

when $N_{22}^{g_{12}}$ is any g_{12}-inverse of N_{22}, and in consequence, Theorems 2.18 and 2.19 imply that the forms (6.71) and (6.72) are independent and that (6.72) is $\sigma^2.\chi_t^2$ for t as in (6.37). The last statement follows by noting that (6.73) yields

$$(6.74) \quad N(N^{g_1}-N^*)N(N^{g_1}-N^*)N = N(N^{g_1}-N^*)N \quad .$$

These statements amount only to the construction of an F-ratio given that the model (6.2) holds, and the statistic F was proposed as a diagnostic. However an unknown referee posed the problem of whether such a diagnostic could be justified in the absence of a hypothesis-test framework. This question led to the generalization of the John and Draper (1978) idea in (6.9), and the establishing of the non-central F-ratio of Corollary 6.1.1.

We note that attempting to fit a model

$$(6.75) \quad \begin{bmatrix} \underline{y}_1 \\ \underline{y}_2 \end{bmatrix} = \begin{bmatrix} X_1 & : & 0 \\ X_2 & : & I \end{bmatrix} \begin{bmatrix} \underline{\beta} \\ \underline{\lambda} \end{bmatrix} + \begin{bmatrix} \underline{\varepsilon}_1 \\ \underline{\varepsilon}_2 \end{bmatrix} = [X : J] \begin{bmatrix} \underline{\beta} \\ \underline{\lambda} \end{bmatrix} + \begin{bmatrix} \underline{\varepsilon}_1 \\ \underline{\varepsilon}_2 \end{bmatrix}$$

with singular V as in (6.3) will in general extend the space of observations from $C(V+XUX')$ to $C(V+XUX'+JJ')$ where

$$(6.76) \quad JJ' = \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix}$$

In this instance the model (6.75) may be interpreted as allowing for certain additive shifts, as opposed to rare observations in the underlying probability distribution. The generalized sweepout method of Theorem 5.7 applies with

(6.77) $s = r[V : X : J] - r[X : J]$, and

(6.78) $q = r[V : X] - r[X]$.

There does not appear to be a simplification of the F-ratio (5.89) associated with this partitioning to parallel the expression (6.54).

The question may be posed as to whether or not the independent recursive residual construction of John and Draper (1978) may be extended to the general case of F as in (6.54). While the sum of squares is easily shown to partition in a parallel way, it is not possible in general to interpret the recursive residuals as the deviation of an observation from an estimate based on the preceding data set. The discussion of the equations (6.48) through to (6.52), on pages 6.10 and 6.11 serve to explain this fact. This is not necessarily a great disadvantage, since the interpretation in the $V = I$ case is essentially a device which depends on the order in which observations are dropped.

6.2 DIAGNOSTIC INDICATORS

We now examine some measures of the ways in which characteristics of the X-matrix may influence parameter estimates in the possible presence of outliers. The preceding section dealt with X of arbitrary column-rank, and the development did not touch upon such well-known complications as near multicollinearities (and the consequent large variances associated with parametric functions). It therefore seems pertinent to examine how statistics similar to the

F-ratio of (6.54) may be developed for the multicollinearity techniques of principle component analysis and ridge regression. As such techniques give rise to biased estimates, it is to be expected that centrality conditions will either be violated or pose some difficulties as Troskie, Coutsourides and Jacobs (1980) point out.

Following Marquardt (1970), we assume that V is non-singular and that X is of full column rank, but subject to the condition that

(6.79) $\quad A = X'V^{-1}X$

is ill-conditioned. Then the OLS estimator of $\underline{\beta}$, is

(6.80) $\quad \hat{\underline{\beta}} = A^{-1}X'V^{-1}\underline{y}$

and is subject to large variances associated with some linear functions of $\hat{\underline{\beta}}$. Let $\lambda_1, \lambda_2, \ldots, \lambda_k$ be the eigenroots of A and $\underline{w}_1, \underline{w}_2, \ldots, \underline{w}_k$ a corresponding set of eigenvectors. Suppose that the first r of the roots are deemed significantly greater than zero, and that the remainder are small but non-zero. Let

(6.81) $\quad \Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k) = \begin{bmatrix} \Lambda_r & : & 0 \\ 0 & : & \Lambda_0 \end{bmatrix}$, and

(6.82) $\quad W = [\underline{w}_1, \underline{w}_2, \ldots, \underline{w}_k] = [W_r : W_0]$

conformably. By definition

(6.83) $\quad A = W\Lambda W'$. Let

(6.84) $\quad A^* = W_r\Lambda_r^{-1}W_r' = A^*AA^*$.

The principal component regression (PCR) estimate of $\underline{\beta}$ is defined to be

(6.85) $\hat{\underline{\beta}}_0 = A^*X'V^{-1}\underline{y}$. Thus

(6.86) $\hat{\underline{\varepsilon}}_0 = \underline{y} - X\hat{\underline{\beta}}_0 = (I-XA^*X'V^{-1})(X\underline{\beta} + \underline{\varepsilon})$

$= MX\underline{\beta} + M\underline{\varepsilon}$, where

(6.87) $M = I - R$, and

(6.88) $R = XA^*X'V^{-1}$. Then

(6.89) $\mathrm{var}(\hat{\underline{\varepsilon}}_0) = \sigma^2.N = \sigma^2.MVM'$.

Note that R and M are idempotent by virtue of (6.84), and have zero-matrix product from (6.87). Also

(6.90) $N = MV = VM' = N$.

<u>Theorem 6.2</u> (Troskie and Dunne, 1981)

The quadratic form $\hat{\underline{\varepsilon}}_0'V^{-1}\hat{\underline{\varepsilon}}_0$ is distributed as non-central $\sigma^2.\chi_f^2(\lambda)$ where

(6.91) $f = n - r$, and

(6.92) $\lambda = \underline{\beta}'(A-AA^*A)\underline{\beta} = \underline{\beta}'W_0\Lambda_0W_0'\underline{\beta}$.

Proof: From (6.88) and (6.89) we have,

(6.93) $NV^{-1}N = MVM' = N$. Also

(6.94) $NV^{-1}MX\underline{\beta} = M^2X\underline{\beta} = MX\underline{\beta}$

so that the conditions of Theorem 2.18 are satisfied with

(6.95) $f = \mathrm{tr}(NV^{-1}) = \mathrm{tr}(M) = n - \mathrm{tr}(A^*A) = n - r$ and

(6.96) $\lambda = \underline{\beta}'X'M'VMX\underline{\beta} = \underline{\beta}'(A-AA^*A)\underline{\beta}$

as required. □

We note that centrality holds for an arbitrary $\underset{\sim}{\beta}$ in the parameter space if and only if A^* is $A^{g_{12}}$, from (6.84) and (6.92).

Now let $\hat{\underline{\varepsilon}}_{02}$ be a subset of $\hat{\underline{\varepsilon}}_0$, say the last p, so that

(6.97) $\hat{\underline{\varepsilon}}_{02} = [0 : I]\hat{\underline{\varepsilon}}_0$, and

(6.98) $\text{var}(\hat{\underline{\varepsilon}}_0) = \sigma^2 . \begin{bmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{bmatrix}$, conformably.

Taking any g_1-inverse of N_{22} we may define

(6.99) $SS_1 = \hat{\underline{\varepsilon}}_0' V^{-1} \hat{\underline{\varepsilon}}_0 - \hat{\underline{\varepsilon}}_{02}' N_{22}^{g_{12}} \hat{\underline{\varepsilon}}_{02}$

$= \hat{\underline{\varepsilon}}_0' V^{-1} \hat{\underline{\varepsilon}}_0 - \hat{\underline{\varepsilon}}_0 N_0 \hat{\underline{\varepsilon}}_0$, where

(6.100) $N_0 = \begin{bmatrix} 0 & 0 \\ 0 & N_{22}^{g_{12}} \end{bmatrix}$.

Theorem 6.3 (Troskie and Dunne, 1981)

The variables $\hat{\underline{\varepsilon}}_{02}$ and SS_1 are independently distributed.

Proof: By Corollary 2.19.1 it is sufficient to observe that

(6.101) $[0 : I]N(V^{-1}-N_0)N = [0 : I]\begin{bmatrix} Q & 0 \\ 0 & 0 \end{bmatrix} = 0$, where

(6.102) $Q = N_{11} - N_{12}N_{22}^{g_{12}}N_{21}$, and thus

(6.103) $[0 : I]N(V^{-1}-N_0)MX\underline{\beta} = [0 : I]N(V^{-1}-N_0)NV^{-1}X\underline{\beta} = \underline{0}$

using (6.90).

Theorem 6.4 (Troskie and Dunne, 1981)

The quadratic form $\underline{\hat{\varepsilon}}_{02}' N_{22}^{g12} \underline{\hat{\varepsilon}}_{02}$ is distributed as non-central $\sigma^2 . \chi_h^2(\nu)$ where

(6.104) $h = r(N_{22})$, and

(6.105) $\nu = \underline{\beta}' X' M' N_0 M X \underline{\beta}$.

Proof: Writing $\underline{\hat{\varepsilon}}_{02}' N_{22}^{g12} \underline{\varepsilon}_{02}$ as $\underline{\hat{\varepsilon}}_0' N_0 \underline{\hat{\varepsilon}}_0$ from (6.90), we note

(6.106) $N_0 N N_0 = N_0$.

Thus all the conditions of the Theorem 2.18 follow and

(6.107) $h = r(N_0 N) = tr(N_0 N) = tr(N_{22}^{g12} N_{22})$

yields (6.103). The non-centrality parameter is directly

(6.105) $\nu = \underline{\beta}' X' M' N_0 M X \underline{\beta}$. □

The term (6.105) may be written in other forms if required, using (6.90) and (6.106). It follows that

(6.108) $$F = \frac{\underline{\hat{\varepsilon}}_0' N_0 \underline{\hat{\varepsilon}}_0}{\underline{\hat{\varepsilon}}_0' N^{g1} \underline{\hat{\varepsilon}}_0 - \underline{\hat{\varepsilon}}_0' N_0 \underline{\hat{\varepsilon}}_0} \cdot \frac{n-r-h}{h}$$

has doubly non-central F-distribution. Note that X has full rank k, in the strict sense, but that the roots $\lambda_{r+1}, \ldots, \lambda_k$ are close to zero. We may assume for practical purposes that the rank of X is r and that

(6.109) $E(\underline{\hat{\varepsilon}}_0) = MX\underline{\beta} = \underline{0}$, and thus

(6.110) $\lambda = \nu = \lambda - \nu = 0$,

and treat the F-ratio (6.108) as though it were central.

We are therefore lead to ask whether or not a similar development is possible for ridge regression (RR). Suppose that X is of full rank and $V = I$, but that the ill-conditioning of (6.79) applies to this special case. Now

(6.111) $A = X'X = S$

and the RR estimates of $\underline{\beta}$ are given by

(6.112) $\hat{\underline{\beta}}_R = (X'X+\ell I)^{-1}X'\underline{y}$

$= TX'\underline{y}$, say

where for the purposes of this development we take ℓ as given. Describing the source of the multicollinearity as before, we may modify (6.81) to (6.83) to note that

(6.113) $T = W.\mathrm{Diag}(\lambda_i+\ell)^{-1}.W'$. Then

(6.114) $X\hat{\underline{\beta}}_R = XTX'\underline{y}$, and

(6.115) $\hat{\underline{\varepsilon}}_R = (I-XTX')\underline{y}$.

are the RR estimates and residuals respectively. Here XTX' corresponds to R in (6.88) but is not idempotent. In fact

(6.116) $(XTX')^2 = XTX' - \ell XT^2X'$.

It follows, as may be verified by some tedious algebra, that no direct extension is possible for the RR case as a parallel of (6.108). An indirect extension for both the RR and PCR cases is possible when $V = I$, if the denominator mean square error (MSE) is taken to be the denominator MSE obtained by OLS as in (6.54). The following theorem justifies the use of t- or F-statistics obtained in this way, but tests based on them will have less power than the OLS counterparts.

Theorem 6.5 (Troskie and Dunne)

The quadratic form obtained by eliminating subvector $\underline{y}_2$,

(6.117) $\hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}} - \hat{\underline{\varepsilon}}_2' N_{22}^{g_{12}} \hat{\underline{\varepsilon}}_2$

is independent of the OLS, PCR and RR residuals $\hat{\underline{\varepsilon}}_2$, $\hat{\underline{\varepsilon}}_{02}$, and $\hat{\underline{\varepsilon}}_{R2}$ for the eliminated subvector when $V = I$.

Proof: From Theorem 2.19 it will suffice to note

(6.118) $[0 : I]I(N-NN^*N) = 0$

(6.119) $[0 : I](I-XA^*X')I(N-NN^*N) = 0$

(6.120) $[0 : I](I-XTX')I(N-NN^*N) = 0$, where

(6.121) $\text{var}(\hat{\underline{\varepsilon}}) = \sigma^2 N = \sigma^2(I-X(X'X)^{g_1}X')$ and

N^* is obtained from N as in (6.69). Clearly (6.119) and (6.120) reduce to (6.118), and this holds because

(6.122) $N - NN^*N = \begin{bmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{bmatrix} - \begin{bmatrix} N_{12}N_{22}^{g_1}N_{21} & N_{12} \\ N_{21} & N_{22} \end{bmatrix}$. □

Thus we may write a generalization of (6.108) as

(6.123) $$F = \frac{\hat{\underline{\varepsilon}}_0' N_0 \hat{\underline{\varepsilon}}_0}{\hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}} - \hat{\underline{\varepsilon}}_2' N_{22}^{g_{12}} \hat{\underline{\varepsilon}}_2} \cdot \frac{n-k-h}{h}$$

which has a simply non-central F-distribution. In the RR case

(6.124) $$F = \frac{\hat{\underline{\varepsilon}}_R' M^* \hat{\underline{\varepsilon}}_R}{\hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}} - \hat{\underline{\varepsilon}}_2' N_{22}^{g_{12}} \hat{\underline{\varepsilon}}_2} \cdot \frac{n-k-h}{p}$$, for

(6.125) $p = r(M^*)$, and

(6.126) $\text{var}(\hat{\underline{\varepsilon}}_R) = \sigma^2.M = \sigma^2(I-XTX'+\ell XT^2X')$

yields M^* in the manner of (6.69).

Cook (1977, 1979) devised a method of judging the contribution of an individual data point to the OLS estimate of $\underline{\beta}$ in regression models with $V = I$ and X full rank. In fact the method amounts to an examination of the difference of fitted values (and equivalently residuals), so the rank of X is not an issue. As a missing plot procedure it requires only that each row of X is in the row-space of the remaining rows, to be defined for each individual data point. The diagnostic statistic is Cook's distance

$$(6.127)\quad D_i = \frac{(\underline{\beta}_i^* - \hat{\underline{\beta}})' X' X (\underline{\beta}_i^* - \hat{\underline{\beta}})}{\hat{\underline{\varepsilon}}' \hat{\underline{\varepsilon}}} \cdot \frac{n-r}{r}$$

where $\hat{\underline{\beta}}$ is an OLSE of $\underline{\beta}$ obtained from all the observations and $\underline{\beta}_i^*$ an OLSE from the set excluding the i^{th} observation. It is derived from the idea that the normal theory $(1-\alpha)100\%$ confidence ellipsoid for $\underline{\beta}$ should contain $\underline{\beta}_i^*$, when it is uniquely defined. Cook suggests that D_i be treated like an $F(r,n-r)$ variate, and that we are concerned to have D_i relatively small. High values indicate that the data-point is unusual in some way. Cook has shown that

$$(6.128)\quad D_i = \frac{\hat{\underline{\varepsilon}}_i^2}{n_{ii} \cdot \hat{\underline{\varepsilon}}' \hat{\underline{\varepsilon}}} \cdot \frac{n-r}{r} \cdot \frac{1-n_{ii}}{n_{ii}}$$

$$= T_i^2 \cdot \frac{1}{r} \cdot \frac{\text{var}(\hat{\underline{y}}_i)}{\text{var}(\hat{\underline{\varepsilon}}_i)}$$

for T_i the i^{th} studentized residual as in (6.64), modified for changing variances $n_{ii} \cdot \sigma^2$. Note that

$$(6.129) \quad T_i^2 = \frac{t_i^2}{t_i^2+(n-r-1)} \cdot \frac{n-r}{r} \cdot \frac{1-n_{ii}}{n_{ii}}$$

where F is given by (6.61), so that

$$(6.130) \quad D_i = \frac{F}{F+(n-r-1)} \cdot \frac{n-r}{r} \cdot \frac{1-n_{ii}}{n_{ii}}$$

will be sensitive moderate F values when n_{ii} is small. Cook has generalized the measure D_i for statistics in which interest is focussed on only a specified subset of linear functions of $\underline{\beta}$, in the full rank model. It is again clear that if X does not have full column-rank, the same extension is easily established for a specified subset of estimable functions of $\underline{\beta}$, providing that a missing plot technique can be applied as before.

Andrews and Pregibon (1978) proposed another method of search for those outliers or observations that may potentially have a large influence on the parameter estimates. This procedure examines ratios $AP(\theta)$ of determinants, and therefore for the bivariate-case is a volume-based method. The paper points out that there are a large number of tests and procedures which operate sequentially on a most deviant (in some sense) observation, until some deviation falls below a particular threshhold, and without regard to the difference in influences on parameter estimates and predicted values which are often the prime focus of the analysis. If a point has almost no influence on the results there would appear to be little point in agonizing over the significance or nonsignificance of its deviation.

Given that variance estimation is of peripheral importance, their argument for directing attention to points with large *leverage* effects on the regression plane is compelling. Drawing from an analysis of their paper by Draper and John (1981), we may relate their statistic $AP(\theta)$ to the quantities examined in Section 6.1 and to the distance measure of Cook (1977). By definition,

(6.131) $AP(\theta) = |X_2^{*\prime}X_2^*| / |X_1^{*\prime}X_1^*|$, where

(6.132) $X_1^* = [X : \underline{y}]$,

(6.133) $X_2^* = [X : D : \underline{y}]$,

(6.134) $D = \begin{bmatrix} 0 \\ I \end{bmatrix}$, and

θ indexes the subset of observations assigned non-zero entries in D. In this example it is the last, say, p. Clearly $AP(\theta)$ is non-zero if and only if X_2^* has full column-rank. Draper and John show that

(6.135) $AP(\theta) = (1-\hat{\underline{\varepsilon}}_2^{\prime}N_{22}^{-1}\hat{\underline{\varepsilon}}_2/\hat{\underline{\varepsilon}}^{\prime}\hat{\underline{\varepsilon}}).|N_{22}|$

$= (1-p.F/[(n-r-p)+pF]).|N_{22}|$

and that an extended form of Cook's distance is

(6.136) $C(\theta) = \dfrac{\hat{\underline{\varepsilon}}_2^{\prime}N_{22}^{-1}\hat{\underline{\varepsilon}}_2}{\hat{\underline{\varepsilon}}^{\prime}\hat{\underline{\varepsilon}}} \, . \, \dfrac{n-r}{p} \, . \left(\dfrac{\hat{\underline{\varepsilon}}_2^{\prime}N_{22}^{-2}\hat{\underline{\varepsilon}}_2}{\hat{\underline{\varepsilon}}_2^{\prime}N_{22}^{-1}\hat{\underline{\varepsilon}}_2} - 1\right)$

where the latter has no obvious physical interpretation. Some applications are presented for θ indexing a single observation, and Draper and John conclude

(i) that $\hat{\underline{\varepsilon}}_2' N_{22}^{-1} \hat{\underline{\varepsilon}}_2$ will be large for deviant sets.

(ii) Cook's statistic is sensitive to observations which affect the fitted equation coefficients (and fitted values).

(iii) $|N_{22}|$ in $AP(\theta)$ is a spatial measure of the isolation of a set of design-points in the space $C(X)$. Low-values constitute sets which have leverage effects on the regression.

6.3 BAYESIAN APPROACHES

Well-known theory associated with a Bayesian view of the normal multiple regression model is presented in Zellner (1971, pp.70-81). We will restrict attention to informative priors for $\underline{\beta}$, and for σ^2.

If σ^2 is assumed known then Theorem 2.17 and its corollary define the distributions which allow all future observations from the same unchanging source to be modelled. In the same way that the LS methods compared estimates based on the data with the data, one possible approach is to examine the differences between the Bayes estimates for $X\underline{\beta}$ given $\underline{y}$, and $\underline{y}$ itself. Clearly there is a smearing effect if outliers are present among the $\underline{y}$-values. Equations (2.167) and (2.168) give the mean and variance of the appropriate conditional distribution, and it may be reasonable to compare the diagnostic F-ratio based on

$$(6.137)\quad \underset{\sim}{\tilde{\varepsilon}} = (\underline{y}-\tilde{\underline{\mu}}) \qquad , \qquad \text{namely}$$

$$(6.138)\quad F = \frac{\tilde{\underline{\varepsilon}}'\tilde{N}_0\tilde{\underline{\varepsilon}}}{\tilde{\underline{\varepsilon}}'\tilde{V}^{g_1}\tilde{\underline{\varepsilon}}-\tilde{\underline{\varepsilon}}'\tilde{N}_0\tilde{\underline{\varepsilon}}} \cdot \frac{r(\tilde{V})}{r(\tilde{N}_0)} \qquad , \qquad \text{where}$$

$$(6.139)\quad \tilde{N} = \begin{bmatrix} 0 & 0 \\ 0 & \tilde{V}_{22}^{g_{12}} \end{bmatrix} \quad .$$

If σ^2 is not known but a suitable inverted gamma distribution may be assigned as a prior, then we may proceed to establish a predictive distribution in the usual way, incorporating two natural conjugate priors into the model. Thus

$$(6.140)\quad P(\underline{\beta},\sigma) = P(\underline{\beta}|\sigma).P(\sigma) \qquad \text{where}$$

$$(6.141)\quad P(\underline{\beta}|\sigma) \to \frac{|A|^{\frac{1}{2}}}{\sigma^k} \cdot \exp - \frac{1}{2\sigma^2}(\underline{\beta}-\bar{\underline{\beta}})'A(\underline{\beta}-\bar{\underline{\beta}}) \qquad , \qquad \text{and}$$

$$(6.142)\quad P(\sigma) \to \frac{1}{\sigma^{w+1}} \cdot \exp - \frac{1}{2\sigma^2} w.c^2 \qquad , \qquad \text{for } w > 0 \ .$$

From the likelihood of $\underline{y}$ (or $\underline{\varepsilon}$) and the above priors we obtain the posterior pdf of $\underline{\beta}$ and σ given $\underline{y}$ as

$$(6.143)\quad P(\underline{\beta},\sigma|\underline{y}) \to \frac{1}{\sigma^{m+1}} \exp - \frac{1}{2\sigma^2}(Q) \qquad , \qquad \text{where}$$

$$(6.144)\quad m = n + k + w \qquad , \qquad \text{and}$$

$$(6.145)\quad Q = wc^2 + (\underline{\beta}-\bar{\underline{\beta}})'A(\underline{\beta}-\bar{\underline{\beta}}) + (\underline{y}-X\underline{\beta})'(\underline{y}-X\underline{\beta}) \quad .$$

Noting the substitution

$$(6.146)\quad \tilde{\underline{\beta}} = (A+X'X)^{-1}(A\bar{\underline{\beta}}+X'\underline{y}) \qquad ,$$

we may collect the terms and complete the square to obtain

$$(6.147)\quad Q = wc^2 + (\underline{\beta}-\tilde{\underline{\beta}})'(A+X'X)(\underline{\beta}-\tilde{\underline{\beta}}) + \underline{y}'\underline{y} + \bar{\underline{\beta}}'A\bar{\underline{\beta}} - \tilde{\underline{\beta}}'(A+X'X)\tilde{\underline{\beta}}$$

$$= R + S \qquad \text{where}$$

$$(6.148) \quad S = (\underline{\beta}-\tilde{\underline{\beta}})'(A+X'X)(\underline{\beta}-\tilde{\underline{\beta}}) \quad , \qquad \text{and}$$

$$(6.149) \quad R = Q - S \quad .$$

Integrating (6.143) with respect to σ yields the marginal density of $\underline{\beta}$ as

$$(6.150) \quad P(\underline{\beta}|\underline{y}) \to [R+S]^{-\frac{1}{2}m}$$

which is a multivariate Student t density with mean $\tilde{\underline{\beta}}$ from (6.146) and variance-covariance structure

$$(6.151) \quad \frac{(n+w)}{(n+w-2)} \cdot \left[\frac{R}{(n+w)} \cdot (A+X'X)^{-1}\right] \quad .$$

Now write

$$(6.152) \quad U = (A+X'X) \quad , \qquad \text{and}$$

$$(6.153) \quad M = U + \tilde{X}'\tilde{X} \quad .$$

<u>Lemma 6.6</u>

For M and U as above

$$(6.154) \quad (I-\tilde{X}M^{-1}\tilde{X}')^{-1} = I + \tilde{X}U^{-1}\tilde{X}' \quad .$$

Proof: By Theorem 2.11,

$$(6.155) \quad M^{-1} = U^{-1} - U^{-1}\tilde{X}'(I+\tilde{X}U^{-1}\tilde{X}')^{-1}\tilde{X}U^{-1} \quad .$$

Pre- and post-multiplying by $\tilde{X}$ and $\tilde{X}'$ respectively, and subtracting from I gives

$$(6.156) \quad I - \tilde{X}M^{-1}\tilde{X}' = (I+\tilde{X}U^{-1}\tilde{X}')^{-1}$$

after collecting the terms. Hence the result. □

For convenience in deriving the predictive pdf, let

(6.157) $K = I - \tilde{X}M^{-1}\tilde{X}'$,

(6.158) $L = I + \tilde{X}U^{-1}\tilde{X}'$, and

(6.159) $\underline{\beta}_0 = M^{-1}(U\tilde{\underline{\beta}}+\tilde{X}'\tilde{\underline{y}})$

where $\tilde{\underline{y}}$ is a set of future observations associated with $\tilde{X}$, and $\tilde{\underline{\beta}}$ is given by (6.146).

<u>Theorem 6.7</u> (Troskie and Dunne, 1980)

The predictive density of $\tilde{\underline{y}}$ is a multivariate Student t density with mean $\tilde{X}\tilde{\underline{\beta}}$ and variance-covariance structure

(6.160) $\frac{R}{n+w-2}(I-\tilde{X}M^{-1}\tilde{X}')^{-1} = \frac{R}{n+w-2}(I+\tilde{X}U\tilde{X}')$.

Proof: We may write directly

(6.161) $P(\tilde{\underline{y}}|\underline{\beta},\sigma,\underline{y},X,\tilde{X}) \to \frac{1}{\sigma^q} \exp - \frac{1}{2\sigma^2}(\tilde{\underline{y}}-\tilde{X}\underline{\beta})'(\tilde{\underline{y}}-\tilde{X}\underline{\beta})$

so that from (6.145) and (6.147) ,

(6.162) $P(\tilde{\underline{y}},\underline{\beta},\sigma|\underline{y},X,\tilde{X}) \to \frac{1}{\sigma^{m+q+1}} \exp - \frac{1}{2\sigma^2}(R+S+T)$, for

(6.163) $T = (\tilde{\underline{y}}-\tilde{X}\underline{\beta})'(\tilde{\underline{y}}-\tilde{X}\underline{\beta})$.

Integrating over σ yields

(6.164) $P(\tilde{\underline{y}},\underline{\beta}|\underline{y},X,\tilde{X}) \to [R+S+T]^{-\frac{1}{2}(m+q)}$

Now expand, complete the square and collect the terms to write

(6.165) $(R+S+T) = (R+\tilde{\underline{y}}'\tilde{\underline{y}}+\tilde{\underline{\beta}}'U\tilde{\underline{\beta}}+(\underline{\beta}-\underline{\beta}_0)'M(\underline{\beta}-\underline{\beta}_0)-\underline{\beta}_0M\underline{\beta}_0)$

paralleling the form (6.147). We require that (6.164) be integrated over $\underline{\beta}$. To achieve this we parallel (6.147) to (6.149) writing

(6.166) $Q_1 = (R+S+T) = R_1 + S_1$, and

(6.167) $S_1 = (\underline{\beta}-\underline{\beta}_0)'M(\underline{\beta}-\underline{\beta}_0)$.

Since it is well-known that a k-dimensional $\underline{x}$ distributed as multivariate Student t has density

(6.168) $$f(\underline{x}) = \frac{\nu^{\frac{1}{2}\nu}.\Gamma[\frac{1}{2}(\nu+k)].|V|^{\frac{1}{2}}}{\pi^{\frac{1}{2}q}.\Gamma(\frac{1}{2}\nu)} . [\nu+(\underline{x}-\underline{\theta})'V(\underline{x}-\underline{\theta})]^{-(\frac{k+\nu}{2})}$$

we obtain

(6.169) $$P(\tilde{\underline{y}}|\underline{y},X,\tilde{X}) \to [R+S+T]^{-\frac{1}{2}(n+w+q)}$$.

Finally we note from (6.157) and (6.158), together with collection of the terms, that

(6.170) $$[R+S+T] = [R + (\tilde{\underline{y}}-\tilde{X}\tilde{\underline{\beta}})'K(\tilde{\underline{y}}-\tilde{X}\tilde{\underline{\beta}})]$$

and clearly the required mean $\tilde{X}\tilde{\underline{\beta}}$ is recognised by way of (6.168), and the variance-covariance structure is

(6.171) $$\frac{R}{(n+w-2)} . K^{-1} = \frac{R}{(n+w-2)} . L$$

proving the theorem. □

It is clear that the theorem will generalize for singular variance-covariance structures if $|V|$ is interpreted in (6.168) as the product of non-zero eigenroots.

The principal of maximizing the predictive likelihood as a means of model selection is well-known. To apply the foregoing theorem, in principle, requires that it is possible to describe part of the data as clean, and the remaining subset as perhaps containing some unknown suspicious observations. Then a predictive density for future observations at the dirty data design points, based on the clean data, may allow judgements about the likelihood associated with the given observations, and about the appropriateness of the model.

B I B L I O G R A P H Y

AITKEN, A.C. (1933). On fitting polynomials to data with weighted and correlated errors. Proceedings of the Royal Society of Edinburgh, 54, 12-16.

AITKEN, A.C. (1934). On least squares and linear combination of observations. Proceedings of the Royal Society of Edinburgh, 55, 42-48.

AITKEN, A.C. (1935). See Aitken (1934). [Three dates are quoted e.g. Chipman and Rao (1964), Chipman (1964), and Goldman and Zelen (1964)].

AITKEN, A.C. (1940). On the independence of linear and quadratic forms in samples of normally distributed variates. Proceedings of the Royal Society of Edinburgh, 60, 40-46.

ALBERT, A. (1972). Regression and the Moore-Penrose Pseudo-inverse. Academic Press, New York.

ANDERSON, T.W. (1958). An Introduction to Multivariate Statistical Analysis. John Wiley and Sons, Inc., New York.

ANDERSON, R.L. and BANCROFT, T.A. (1952). Statistical Theory in Research. McGraw-Hill, New York.

ANDREWS, D.F. (1971). Significance tests based on residuals. Biometrika, 58, 139-148.

ANSCOMBE, F.J. (1960). Rejection of outliers. Technometrics, 2, 123-146.

ATKINSON, A.C. (1969). The use of residuals as a concomitant variable. Biometrika, 56, 33-41.

BALESTRA, P. (1970). On the efficiency of ordinary least-squares in regression models. Journal of the American Statistical Association, 65, 1330-1337.

BARNETT, V.D. and LEWIS, T. (1978). Outliers in Statistical Data. John Wiley and Sons, Chichester.

BARTLETT, M.S. (1937). Some examples of statistical methods of research in agriculture and applied biology. Journal of the Royal Statistical Society, Supplement, 4, 137-170 [quoted in Seber (1977)].

BECKMAN, R.J. and TRUSSELL, H.J. (1974). The distribution of an arbitrary studentized residual and the effects of updating in multiple regression. Journal of the American Statistical Association, 69, 109-201.

BEHNKEN, D.W. and DRAPER, N.R. (1972). Residuals and their variance patterns. Technometrics, 14, 101-111.

BEN-ISRAEL, A. (1981). History of the development of generalized inverses. Unpublished lecture notes.

BJERHAMMAR, A. (1958). A generalized matrix algebra. Kunglika Teknisa Högkolans Handlingar (Trans. Roy. Inst. Technology, Stockholm), 124, 1-32.

BOSE, R.C. (1944). The fundamental theorem of linear estimation (abstract). Proceedings of 31st Indian Science Congress, 4, 2-3 [quoted in Pringle and Rayner (1971), Rayner (1977)].

BOSE, R.C. (1959). Unpublished lecture notes on analysis of variance. University of North Carolina, Chapel Hill.

BOULLION, T.L. and ODELL, P.L. (1971). Generalized Inverse Matrices. Interscience Series. John Wiley and Sons, New York.

CHIPMAN, J.S. (1964). On least squares with insufficient observations. Journal of the American Statistical Association, 59, 1078-1111.

CHIPMAN, J.S. (1968). Specification problems in regression analysis. Proc. Symposium on Theory and Applications of Generalized Inverses of Matrices, 114-176. Mathematics Series No. 4, Texas Technological College, Lubbock, Texas.

CHIPMAN, J.S. and RAO, M.M. (1964). Projections, generalized inverses and quadratic forms. Journal of Mathematical Analysis and Applications, 9, 1-11.

COCHRAN, W.G. (1934). The distribution of quadratic forms in a normal system, with applications to analysis of covariance. Proceedings of the Cambridge Philosophical Society, 30, 178-191.

COCHRAN, W.G. and COX, G.M. (1957). Experimental Designs (2nd ed.). Wiley, New York.

COOK, R.D. (1977). Detection of influential observations in linear regression. Technometrics, 19, 15-18.

COOK, R.D. and WEISBERG, S. (1979). Finding influential cases in linear regression : a review. Unpublished mimeograph : Technical Report No. 338, University of Minnesota, School of Statistics, St. Paul, Minnesota.

CRAMER, E.M. (1972). Missing values in experimental design models. American Statistician, 26, (4), 58.

CRAMÉR, H. (1946). Mathematical methods of Statistics. Princeton University Press, Princeton.

DANIEL, C. (1960). Locating outliers in factorial experiments. Technometrics, 2, 149-156.

DIXON, W.J. (1950). Analysis of extreme values. Annals of Mathematical Statistics, 21, 488-506.

DIXON, W.J. (1951). Ratios involving extreme values. Annals of Mathematical Statistics, 22, 68-78.

DOORNBOS, R. (1981). Testing for a single outlier in a linear model. Biometrics, 37, 705-711.

DRAPER, N.R. and JOHN, J.A. (1980). Testing for three or fewer outliers in two-way tables. Technometrics, 22, 9-15.

DRAPER, N.R. and JOHN, J.A. (1981). Influential observations and outliers in regression. Technometrics, 23, 21-26.

DUNNE, T.T. and TROSKIE, C.G. (1980). Testing for outliers under arbitrary variance matrix in the general linear model. Unpublished mimeograph,Technical Report No. 8, University of Cape Town.

ELLENBERG, J.H. (1973). The joint distribution of the standardized least squares residuals from a general linear regression. Journal of the American Statistical Association, 68, 941-943.

ELLENBERG, J.H. (1976). Testing for a single outlier from a general linear regression. Biometrics, 32, 637-645.

ELSTON, R.C. and BUSH, N. (1964). The hypotheses that can be tested when there are interactions in an analysis of variance model. Biometrics, 20, 681-698.

FAIRFIELD SMITH, H. (1957). Missing plot estimates. Biometrics, 13, 115-118.

FIELLER, N.R.J. (1976). Some Problems related to the Rejection of Outlying Observations. Ph.D. Thesis, University of Sheffield [quoted in Barnett and Lewis (1978)].

FURNIVAL, G.M. and WILSON, Jr., R.W. (1974). Regression by leaps and bounds. Technometrics, 16, 499-511.

GALPIN, J.S. (1978). An investigation of methods of ridge regression. Technical report, TWISK9. NRIMS, CSIR, Pretoria.

GAUSS, C.F. (1816). Bestimmung der Genauigkeit der Beobachtungen. In: Werke, IV, (1973). Georg Olms Verlag, Hildesheim.

GAUSS, C.F. (1821). Theoria combinationis observationum erroribus minimis obnoxiae (pars prior). In: Werke, IV, (1973). Georg Olms Verlag, Hildesheim.

GAUSS, C.F. (1823). Theoria combinationis observationum erroribus minimis obnoxiae (pars posterior). In: Werke, IV, (1973). Georg Olms Verlag, Hildesheim.

GAUSS, C.F. (1826). Supplementum theoriae combinationis observationum erroribus minimis obnoxiae. In: Werke, IV, (1973). Georg Olms Verlag, Hildesheim.

GEISSER, S. and CORNFIELD, J. (1963). Posterior distributions for multivariate normal parameters. Journal of the Royal Statistical Society B, 25, 368-376.

GENTLEMAN, J.F. and WILK, M.B. (1975a). Detecting outliers in a two-way table. I. Statistical behaviour of outliers. Technometrics, 17, 1-14.

GENTLEMAN, J.F. and WILK, M.B. (1975b). Detecting outliers. II. Supplementing the direct analysis of residuals. Biometrics, 31, 387-410.

GOLDBERGER, A.A. (1964). Economic Theory. John Wiley and Sons, New York.

GOLDMAN, A.J. and ZELEN, M. (1964). Weak generalized inverses and minimum variance linear unbiased estimation. Journal of Research of the National Bureau of Standards, 68B, 151-172.

GOLUB, G.H. and KAHAN, W. (1964). Calculating the singular values and pseudo-inverse of a matrix. Technical report No. CS8, Computer Science Division, Stanford University.

GOLUB, G.H. and STYAN, G.P.H. (1973). Numerical computations for univariate linear models. Journal of Statistical Computation and Simulation, 2, 253-274.

GRAYBILL, F.A. (1961). An Introduction to Linear Statistical Models. (Volume 1). McGraw-Hill, New York.

GRUBBS, F.E. (1950). Sample criteria for testing outlying observations. Annals of Mathematical Statistics, 21, 27-58.

GRUBBS, F.E. (1969). Procedures for detecting outlying observations in samples. Technometrics, 11, 1-21.

GUTTMAN, I. and DUTTER, R. (1976). Procedures for investigating outliers when estimating in the general univariate linear situation - nonfull rank case. Communications in Statistics - Theory and Method, A5(9), 819-835.

HARRIS, Jr., W.A. and HELVIG, T.N. (1966). Marginal and conditional distributions for singular distributions. Publication of the Research Institute of Mathematical Scientists, Kyoto University, Ser. A, 1, 199-204.

HARVILLE, D.A. (1981). Unbiased and minimum-variance unbiased estimation of estimable functions for fixed linear models with arbitrary covariance structure. Annals of Statistics, 9, 633-637.

HAWKINS, D.M. (1980). Identification of Outliers. Monographs on Applied Probability and Statistics. Chapman and Hall, London.

HOERL, A.E. and KENNARD, R.W. (1970a). Ridge regression : Biased estimation for non-orthogonal problems. Technometrics, 12, 55-68.

HOERL, A.E. and KENNARD, R.W. (1970b). Ridge regression : Applications to nonorthogonal problems. Technometrics, 12, 69-82.

JOHN, J.A. (1978). Outliers in factorial experiments. Applied Statistics, 27, 111-119.

JOHN, J.A. and DRAPER, N.R. (1978). On testing for two outliers or one outlier in two-way tables. Technometrics, 20, 69-78.

JOHN, J.A. and PRESCOTT, P. (1975). Critical values to detect outliers in factorial experiments. Applied Statistics, 24, 56-59.

JOHNSON, W. and GEISSER, S. (1979). Assessing the predictive influence of observations. Unpublished mimeograph : Technical Report No. 355, University of Minnesota.

JOHNSON, W. and GEISSER, S. (1980). A predictive view of the detection and characterization of influential observations in regression analysis. Unpublished mimeograph : Technical Report No. 365, University of Minnesota.

JUDGE, G.G., GRIFFITHS, W.E., CARTER HILL, R. and TSOUNG-CHAO LEE (1980). The Theory and Practice of Econometrics. John Wiley and Sons, New York.

KEMPTHORNE, O. (1952). Design and analysis of experiments. Wiley and Sons, Inc., New York.

KEMPTHORNE, O. (1976). Best linear unbiased estimation with arbitrary variance matrix. In: Essays in Probability and Statistics (ed. Ikeda, S. et al). Shiako-Tsusho, Toyko.

KHATRI, C.G. (1962). Conditions for Wishartness and independence of second degree polynomials in a normal vector. Annals of Mathematical Statistics, 33, 1002-1007.

KHATRI, C.G. (1963). Further contributions to Wishartness and independence of second degree polynomials in normal vectors. Journal of the Indian Statistical Association, 1, 61-70.

KHATRI, C.G. (1968). Some results for the singular normal multivariate regression models. Sankhyā, Ser. A, 30, 267-280.

KOUROUKLIS, S. and PAIGE, C.C. (1981). A constrained least squares approach to the general Gauss-Markov linear model. Journal of the American Statistical Association, 76, 620-625.

KRÄMER, W. (1980). A note on the equality of ordinary least squares and Gauss-Markov estimates in the general linear model. Sankhyā, Ser.A, 42, 130-131.

KRUSKAL, W. (1968). When are Gauss-Markoff and least squares estimators identical? A co-ordinate free approach. Annals of Mathematical Statistics, 39, 70-75.

KRUSKAL, W., FERGUSON, T.S., TUKEY, J.W. and GUMBEL, E.J. (1960). Discussion of the papers of Messrs Anscombe and Daniel. Technometrics, 2, 157-166.

LINHART, H. and ZUCCHINI, W. (1981). On selecting the covariates in analysis of covariance. Submitted to South African Statistical Journal.

LONGLEY, J.W. (1967). An appraisal of least squares programs for the electronic computer from the point of view of the user. Journal of the American Statistical Association, 62, 819-841.

LUND, R.E. (1975). Tables for an approximate test for outliers in linear models. Technometrics, 17, 473-476.

MARQUARDT, D.W. (1970). Generalized inverses, ridge regression biased linear estimation and non-linear estimation. Technometrics, 12, 591-612.

McELROY, F.W. (1967). A necessary and sufficient condition that ordinary least-squares estimators be best linear unbiased. Journal of the American Statistical Association. 62, 1303-1304.

MILLIKEN, G.A. (1971). New criteria for estimability for linear models. Annals of Mathematical Statistics, 42, 1588-1594.

MILLIKEN, G.A. and GRAYBILL, F.A. (1970). Extensions of the general linear hypothesis model. Journal of the American Statistical Association, 65, 797-807.

MITRA, S.K. (1968). On a generalized inverse of a matrix and applications. Sankhyā, Ser.A, 30, 107-114.

MITRA, S.K. and RAO, C.R. (1968). Some results in estimation and tests of linear hypotheses under the Gauss-Markoff model. Sankhyā, Ser.A, 30, 281-290.

MITRA, S.K. and BHIMASANKARAM, P.(1971a). Generalized inverses of partitioned matrices and recalculation of least squares estimates for data and model changes. Sankhyā, 33, 395-410.

MITRA, S.K. and BHIMASANKARAM, P.(1971b). A characterization of Moore-Penrose inverse and related results. Sankhyā, Ser.A, 33, 411-416.

MOORE, E.H. (1920). On the reciprocal of the general algebraic matrix (abstract). Bulletin of the American Mathematical Society, 26, 394-5.

MURPHY, R.B. (1951). On testing for outlying observations. Ph.D. Thesis, Princeton University [quoted in Barnett and Lewis (1979), Hawkins (1980).]

PEARSON, E.S. and CHANDRA SEKAR, C. (1936). The efficiency of statistical tools and a criterion for the rejection of outlying observations. Biometrika, 28, 308-320.

PENROSE, R. (1955). A generalized inverse of matrices. Proceedings of the Cambridge Philosophical Society, 51, 406-13.

PLACKETT, R.L. (1950). Some theorems in least squares. Biometrika, 37, 149-157.

PLACKETT, R.L. (1960). Principles of Regression Analysis. Claredon Press, Oxford.

PRESCOTT, P. (1975). An approximate test for outliers in linear models. Technometrics, 17, 129-132.

PRINGLE, R.M. (1976). Unpublished lecture notes. Department of Biometry, University of Natal, Pietermaritzburg.

PRINGLE, R.M. and CHALTON, D.O. (1972). A look at Rao's unified theory of linear estimation. Unpublished mimeograph, Department of Biometry, University of Natal, Pietermaritzburg.

PRINGLE, R.M. and RAYNER, A.A. (1970). Expressions for generalized inverses of a bordered matrix with applications to problems in mathematical statistics. SIAM Review, 12, 107-115.

PRINGLE, R.M. and RAYNER, A.A. (1971). Generalized Inverse Matrices with applications to Statistics. Griffin, London.

QUENOUILLE, M.H. (1950). Computational devices in the application of least squares. Journal of the Royal Statistical Society, Ser.B, 12, 256-272.

RAO, C.R. (1962). A note on the generalized inverse of a matrix with applications to problems in mathematical statistics. Journal of the Royal Statistical Society, Ser. B, 24, 152-158.

RAO, C.R. (1965). The theory of least squares when the parameters are stochastic and its application to analysis of growth curves. Biometrika, 52, 447-458.

RAO, C.R. (1967). Calculus of generalized inverses of matrices. Part I : General Theory. Sankhyā, Ser.A, 29, 317-342.

RAO, C.R. (1968). A note on a previous lemma in the theory of least squares and some further results. Sankhyā, Ser.A, 30, 259-266.

RAO, C.R. (1971). Unified theory of linear estimation. Sankhyā, Ser. A, 33, 371-394.

RAO, C.R. (1972a). A note on the IPM method in the unified theory of linear estimation. Sankhyā, Ser.A, 34, 285-288.

RAO, C.R. (1972b). Some recent results in linear estimation. Sankhyā, Ser.B. 34, 369-378.

RAO, C.R. (1973). Linear Statistical Influence and its Applications (2nd ed.). John Wiley and Sons, Inc. New York.

RAO, C.R. (1973a). Representations of best linear unbiased estimators in the Gauss-Markoff model with a singular dispersion matrix. Journal of Multivariate Analysis, 3, 276-292.

RAO, C.R. (1975). On a unified theory of estimation in linear models - a review of recent results. In: Perspectives in Probability and Statistics, (J. Gani, ed.) Applied Probability Trust.

RAO, C.R. (1978). Least squares theory for possibly singular models. Canadian Journal of Statistics, 6, 1, 19-23.

RAO, C.R. and MITRA, S.K. (1971). Generalized Inverse of Matrices and its Applications. John Wiley and Sons, New York.

RAYNER, A.A. (1971). Algebra useful for Mathematical Biometry. Unpublished mineograph, Department of Statistics and Biometry, University of Natal, Pietermaritzburg.

RAYNER, A.A. (1974). Quadratic forms and degeneracy : a personal view. Unpublished mimeograph, Department of Statistics and Biometry, University of Natal, Pietermartizburg. Invited lecture, SASA Annual Conference, 1974.

RAYNER, A.A. (1976a). Changing linear constraints in a basic experimental design model. Unpublished mimeograph, Department of Statistics and Biometry, University of Natal, Pietermaritzburg.

RAYNER, A.A. (1976b). Unpublished lecture notes. Manuscript.

RAYNER, A.A. (1977). On Milliken's criterion for estimability in linear models. Unpublished mimeograph, Department of Statistics and Biometry, University of Natal, Pietermaritzburg.

RAYNER, A.A. and PRINGLE, R.M. (1967). A note on generalized inverses in the linear hypothesis not of full rank. Annals of Mathematical Statistics, 38, 271-273.

RAYNER, A.A. and PRINGLE, R.M. (1972). Let's drop the matrix D. Unpublished mimeograph, read at Joint Statistical Meetings, Montreal, August 1972.

RAYNER, A.A. and PRINGLE, R.M. (1976). Some aspects of the solution of singular normal equations with the use of linear restrictions. SIAM Journal of Applied Mathematics, 31, 449-460.

ROHDE, C.A. (1964). Contributions to the theory, computation and application of generalized inverses. Mimeograph No. 392, Institute of Statistics, University of North Carolina, Raleigh.

ROHDE, C.A. (1968). Special Applications of the theory of generalized matrix inversion to statistics. Proc. Symposium on Theory and Applications of generalized Inverses of Matrices, 239-266. Mathematics Series No. 4, Texas Technological College, Lubbock, Texas.

ROHDE, C.A. and HARVEY, J.R. (1965). Unified least squares analysis. Journal of the American Statistical Association, 60, 523-527.

ROY, S.N. and ROY, J. (1960). On testability in normal ANOVA and MANOVA with all 'Fixed Effects'. Jubilee Number of the Bulletin of the Calcutta Mathematical Society [quoted in Elston and Bush (1964)].

ROSNER, B. (1975). On the detection of many outliers. Technometrics, 17, 221-227.

SCAROWSKY, I. (1973). Quadratic forms in normal variates. M.Sc. Thesis, McGill University, Montreal.

SCHEFFÉ, H. (1959). The Analysis of Variance. John Wiley and Sons, New York.

SCOBEY, P. (1975). Singular Gauss-Markoff models. Canadian Journal of Statistics, 3, 1, 105-110.

SEARLE, S.R. (1965). Additional results concerning estimable functions and generalized inverse matrices. Journal of the Royal Statistical Society,Ser.B, 27, 480-490.

SEARLE, S.R. (1971). Linear Models. John Wiley and Sons, Inc., New York.

SEBER, G.A.F. (1977). Linear Regression Analysis. John Wiley and Sons, Inc., New York.

SEBER, G.A.F. (1980). The Linear Hypothesis : A General Theory (2nd ed.). Griffin, London.

STEFANSKY, W. (1971). Rejecting outliers by maximum normed residual. Annals of Mathematical Statistics, 42, 35-45.

STEFANSKY, W. (1972). Rejecting outliers in factorial designs. Technometrics, 14, 469-479.

THOMPSON, W.R. (1935). On a criterion for the rejection of observations and the distributions of the ratio of the deviation to the sample standard deviation. Annals of Mathematical Statistics, 6, 214-219.

TIETJEN, G.L. and MOORE, R.H. (1972). Some Grubbs-type statistics for the detection of several outliers. Technometrics, 55, 583-598.

TIETJEN, G.L. MOORE, R.H. and BECKMAN, R.J. (1973). Testing for a single outlier in simple linear regression. Technometrics, 15, 717-721.

TROSKIE, C.G. (1980). A note on the combined effects of multicollinearity, outliers, and variable selection procedures. Unpublished mimeograph, University of Cape Town.

TROSKIE, C.G., COUTSOURIDES, D. and JACOBS, M. (1980). Detection of outliers in the presence of multicollinearity. Unpublished mimeograph, Technical Report No. 3, University of Cape Town.

TROSKIE, C.G. and DUNNE, T.T. (1980). A Bayesian approach to the detection of outliers. Unpublished mimeograph, University of Cape Town.

TROSKIE, C.G. and DUNNE, T.T. (1981). Detection of outliers in the presence of multicollinearity. Unpublished mimeograph, Technical Report No. 13, University of Cape Town.

TUKEY, J.W. (1949). One degree of freedom for non-additivity. Biometrics, 5, 232-242.

TUKEY, J.W. (1955). Answer to Query 113. Biometrics, 11, 111-113.

TUKEY, J.W. (1960). Discussion of papers by Messrs Anscombe and Daniel. Technometrics, 2, 160-165.

WAMPLER, R.H. (1970). A report on the accuracy of some widely used least squares computer programs. Journal of the American Statistical Association, 65, 549-565.

WEISBERG, S. (1981). A statistic for allocating C_p to individual cases. Technometrics, 23, 27-31.

WILKINSON, G.N. (1958). Estimation of missing values for the analysis of incomplete data. Biometrics, 14, 257-286.

YATES, F. (1933). The analysis of replicated experiments when the field results are incomplete. Empire Journal of Experimental Agriculture, 1, 129-142.

YATES, F. (1934). The analysis of multiple classifications with unequal numbers in the different classes. Journal of the American Statistical Association, 29, 51-66.

YATES, F. and HALE, R.W. (1939). The analysis of Latin squares when two or more rows, columns, or treatments are missing. Journal of the Royal Statistical Society, Supplement, 6, 67-79.

ZELEN, M. and FEDERER, W.T. (1965). Application of the calculus for factorial arrangements. III : Analysis of factorials with unequal numbers of observations. Sankhyā, Ser.A, 27, 383-400.

ZELLNER, A. (1971). An Introduction to Bayesian Inference in Econometrics. John Wiley and Sons, New York.

ZUCCHINI, W. and LINHART, H. (1981). On selecting the covariates in analysis of covariance. [Submitted to South African Statistical Journal.]

ZYSKIND, G. (1967). On canonical forms, non-negative covariance matrices and best and simple least squares linear estimators in linear models. Annals of Mathematical Statistics, 38, 1092-1109.

ZYSKIND, G., KEMPTHORNE, O. WHITE, R.F., DAYHOFF, E.E. and DOERFLER, T.E. (1964). Research on analysis of variance and related topics. Aerospace Research Laboratories, Technical Report 64-193, Wright-Patterson Air Force Base, Ohio.

ZYSKIND, G. and MARTIN, F.B. (1969). A general Gauss-Markoff theorem for linear models with arbitrary non-negative covariance structure. SIAM: Journal of Applied Mathematics, 17, 1190-1202.