University of Cape Town
Department of Mathematical Statistics

Outliers and Influence Under Arbitrary Variance

by
Robert Schall

A thesis prepared unter the supervision of
Dr. T. T. Dunne
in fulfilment of the requirements for the degree of
Doctor of Philosophy in Mathematical Statistics

University of Cape Town
1986

ACKNOWLEDGEMENTS

In German, the supervisor of a PhD student is called *Doktorvater*, and Dr. T. T. Dunne is in more than one respect the father of this thesis. My happy thanks to him for his encouragement, and participation, which was always as cheerful as it was skilful. His inexhaustible and contagious excitability made it a pleasure to work under his supervision.

Meinen Vorgesetzten in der Abteilung für Produktanalyse der Messerschmitt-Bölkow-Blohm GmbH danke ich für ihre Toleranz, besonders Herrn O. Pospischil, dem Boss, für seine Unterstützung.

Für die großartige Leistung von Herrn Czech beim Setzen des Manuskripts kann man nur staunende Bewunderung aufbringen. Ohne seine Geduld und sein Können hätte diese Arbeit nicht vollendet werden können.

Vir die proeflees, hoewel kommentaarloos, baie dankie aan ons geliefkoosde Honneurs-student.

ABSTRACT

Quadratic forms in a normal variate which are distributed as chi-square variables can be represented in terms of generalized inverses of the variance-covariance structure of the variate. The total sum of squares under the general linear model is decomposed into uncorrelated sums of squares which are distributed chi-squared, and a (possibly void) nonstochastic sum of squares.

Using a geometric approach to best linear unbiased estimation in the general linear model, the additional sum of squares principle, used to generate decompositions, can be generalized allowing for an efficient treatment of augmented linear models. The notion of the admissibility of a new variable is useful in augmenting models. Best linear unbiased estimation and tests of hypotheses can be performed through transformations and reparametrizations of the general linear model.

The theory of outliers and influential observations can be generalized so as to be applicable for the general univariate linear model, where three types of outlier and influence may be distinguished. The adjusted models, adjusted parameter estimates, and test statistics corresponding to each type of outlier are obtained, and data adjustments can be effected. Relationships to missing data problems are exhibited. A unified approach to outliers in the general linear model is developed. The concept of recursive residuals admits generalization.

The typification of outliers and influential observations in the general linear model can be extended to normal multivariate models. When the outliers in a multivariate regression model follow a nested pattern, maximum likelihood estimation of the parameters in the model adjusted for the different types of outlier can be performed in closed form, and the corresponding likelihood ratio test statistic is obtained in closed form. For an arbitrary outlier pattern, and for the problem of outliers in the generalized multivariate regression model, three versions of the EM-algorithm corresponding to three types of outlier are used to obtain maximum likelihood estimates iteratively.

A fundamental principle is the comparison of observations with a choice of distribution appropriate to the presumed type of outlier present. Applications are not necessarily restricted to multivariate normality.

CONTENTS

PREFACE

In accordance with the regulations for the Degree of PhD 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 comprises some ancillary results which are of use in later chapters.
Equivalence conditions for the chi-squaredness and independence of quadratic forms in a normal variate are given, in terms of generalized inverses of the variance-covariance structure of the variate.
The chi-squaredness of the quadratic from $\hat{\mathbf{y}}'\mathbf{V}^{-}\hat{\mathbf{y}}$ is examined, and an analysis of variance result for the general linear model is obtained in such a way that the total sum of squares in a linear model is decomposed into uncorrelated sums of squares, which are distributed chi-squared, and a nonstochastic sum of squares.
Two theorems of Dunne (1982) and Chipman (1964) are generalized.
In the variance components model, a necessary and sufficient condition is derived for the best linear unbiased estimator of an estimatable parametric function to be independent of the variance components.

Chapter 2 summarizes the theory of best linear unbiased estimation and testing of hypotheses in the general linear model, providing a theoretical framework for the subsequent chapters.
Generally, a geometric approach to estimation and tests of hypotheses is used.
In Section 2.1 the row and column spaces of matrices in a linear model which is possibly not of full rank are decomposed, and specifically the sure equations in a singular linear model are presented in a canonical form.
The term *admissibility* of a new variable is defined for a singular linear model to be augmented by that new variable. In Lemma 2.31 an equivalence condition for the admissibility of a variable is derived.
The additional sum of squares principle is generalized for the general linear model in Theorem 2.33, allowing for an efficient treatment of augmented linear models. Augmenting the linear model by dummy variables leads to downdating formulae for the general linear model, and in a later chapter three different types of downdating are distinguished.
In Section 2.6 two approaches to best linear unbiased estimation and tests of hypotheses in the general linear model are given, leading to several algorithms.

Chapter 3 treats the problem of outliers and influential observations in the general linear model.
Three types of outlier are distinguished, and the corresponding adjusted models, adjusted parameter estimates, and test-statistics are derived. Thus the work of Dunne (1982) is extended, and the work of Gentleman and Wilk (1975), John and Draper (1978) and Cook and Weisberg (1979) is generalized for the general linear model. The idea of recursive residuals is similarly generalized. A unified approach to outliers in the general linear model is presented.

Corresponding to three types of outlier, the influence measures of Cook (1977) and Andrews and Pregibon (1978) are generalized for the general linear model, yielding three versions of each influence measure. The statistic of Andrews and Pregibon is generalized for the case where influence on only a specified subset of linear functions of the parameter vector in a linear model is considered.
Outliers and influential observations in the variance components model are briefly discussed.

Chapter 4 extends the ideas and methods of the third chapter to normal multivariate models. Emphasis is given to maximum likelihood estimation in models adjusted for outliers. When the outliers in a multivariate regression model follow a nested pattern, maximum likelihood estimation of the parameters in the model and the resulting likelihood ratio test statistic are obtained in closed form. For an arbitrary outlier pattern, and for the problem of outliers in the generalized multivariate regression model, the corresponding versions of the EM-algorithm are presented with respect to three types of outlier, to perform maximum likelihood estimation iteratively.
The iris data of Fisher (1936) is examined for the presence of possible outliers or anomalous observations.

Chapter 5 presents a general approach to outliers without the assumption of multivariate normality. Two types of outlier, additive and distributional, are distinguished in this general case. The regression formulation for those types of outlier, which was applied in normal linear models, as presented in earlier chapters, turns out to be a special manifestation of a fundamental principle.

At the beginning of chapters and subchapters, summaries of their respective content are presented, and new results and extensions of known results are indicated in greater detail, where appropriate. Theorems and corollaries are, where possible, attributed to original sources. Where the candidate's name appears behind the heading of a result, prior research is acknowledged in context, and it is not intended to suggest that a specified result is new in its entirety.

München, August 1986 R. Schall

CHAPTER 1

Ancillary Results

1.1 ON QUADRATIC FORMS

We give a representation of all quadratic forms $\mathbf{y}'\mathbf{Q}\mathbf{y}$ in a normal variable $\mathbf{y}$ which are (possibly non-central) chi-squared in distribution, in terms of generalized inverses of the variance-covariance structure $\mathbf{V}$ of $\mathbf{y}$.

Khatri (1962, 1963) showed that for $\mathbf{y} \sim N(\mu, \mathbf{V})$ the quadratic form $\mathbf{y}'\mathbf{Q}\mathbf{y}$, $\mathbf{Q}$ symmetric, follows a $\chi^2_r(\lambda)$ distribution if and only if

$$\mathbf{VQVQV} = \mathbf{VQV} \tag{1.1}$$

$$\mathbf{VQVQ}\mu = \mathbf{VQ}\mu \tag{1.2}$$

$$\mu'\mathbf{QVQ}\mu = \mu'\mathbf{Q}\mu \tag{1.3}$$

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

$$r = rank(\mathbf{VQV}) = trace(\mathbf{QV}) \ , \quad \text{and} \tag{1.4}$$

$$\lambda = \mu'\mathbf{Q}\mu \ . \tag{1.5}$$

In the following we give a representation of all $\mathbf{Q}$ to satisfy (1.1), in terms of generalized inverses of $\mathbf{V}$, and using this representation we give equivalent conditions for (1.2) to hold in addition to (1.1) and for (1.3) to hold in addition to (1.1) and (1.2). Some matrix equations related to (1.1) are also considered in the development, as well as conditions for the independence of two quadratic forms.

For earlier work on the problem considered here and related problems see Mitra (1968), Bhimasankaram and Majumdar (1980) and Baksalary, Hanke and Kala (1980).

Chi-squaredness of quadratic forms

Let $C(\mathbf{X})$ denote the column space of a matrix $\mathbf{X}$. A g_1-inverse $\mathbf{X}^-$ of a matrix $\mathbf{X}$ satisfies $\mathbf{X}\mathbf{X}^-\mathbf{X} = \mathbf{X}$, and a g_2-inverse $\mathbf{X}^=$ of a matrix $\mathbf{X}$ satisfies $\mathbf{X}^=\mathbf{X}\mathbf{X}^= = \mathbf{X}^=$.

Theorem 1.1 (Schall and Dunne, 1986b)

Let $\mathbf{V}$ be a symmetric and nonnegative definite matrix. Then a necessary and sufficient condition for a symmetric matrix $\mathbf{Q}$ to satisfy $\mathbf{VQVQV} = \mathbf{VQV}$ is that $\mathbf{Q}$ can be written as

$$\mathbf{Q} = \mathbf{V}^{-} - \mathbf{V}^{=} \tag{1.6}$$

where $\mathbf{V}^{-}$ is a symmetric g_1-inverse of $\mathbf{V}$ and $\mathbf{V}^{=}$ is a symmetric g_2-inverse of $\mathbf{V}$ such that $C(\mathbf{V}^{=}) \subset C(\mathbf{V})$.

Proof: Let

$$\mathbf{V} = \mathbf{P} \begin{bmatrix} \Delta & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{P}' \tag{1.7}$$

be the (complete) singular value decomposition of $\mathbf{V}$, i.e. $\mathbf{P}$ is an orthogonal matrix of the same order as $\mathbf{V}$ and Δ is a diagonal matrix containing the non-zero eigenvalues of $\mathbf{V}$.

Further let

$$\mathbf{Q}^* = \mathbf{P}'\mathbf{QP} = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \\ \mathbf{Q}_2' & \mathbf{Q}_3 \end{bmatrix} \tag{1.8}$$

be conformably partitioned.

Then we have with (1.7), (1.8):

(1.9) $\mathbf{VQVQV} = \mathbf{VQV}$

$$\Leftrightarrow \begin{bmatrix} \Delta & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \\ \mathbf{Q}_2' & \mathbf{Q}_3 \end{bmatrix} \begin{bmatrix} \Delta & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \\ \mathbf{Q}_2' & \mathbf{Q}_3 \end{bmatrix} \begin{bmatrix} \Delta & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} =$$

$$= \begin{bmatrix} \Delta & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \\ \mathbf{Q}_2' & \mathbf{Q}_3 \end{bmatrix} \begin{bmatrix} \Delta & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

$$\Leftrightarrow \mathbf{Q}_1\Delta\mathbf{Q}_1 = \mathbf{Q}_1$$

$$\Leftrightarrow \Delta^{1/2}\mathbf{Q}_1\Delta\mathbf{Q}_1\Delta^{1/2} = \Delta^{1/2}\mathbf{Q}_1\Delta^{1/2}$$

$$\Leftrightarrow \Delta^{1/2}\mathbf{Q}_1\Delta^{1/2} \quad \text{idempotent}$$

$$\Leftrightarrow \mathbf{Q}^* \text{ can be written as } \mathbf{Q}^* = \begin{bmatrix} \Delta^{-1/2}\overline{\mathbf{Q}}_1\Delta^{-1/2} & : \mathbf{Q}_2 \\ \mathbf{Q}_2' & : \mathbf{Q}_3 \end{bmatrix}$$

where $\mathbf{Q}_2$ is arbitrary, $\mathbf{Q}_3$ is an arbitrary symmetric matrix and $\overline{\mathbf{Q}}_1$ is an arbitrary symmetric idempotent matrix

$$\Leftrightarrow \mathbf{Q} \text{ can be written as}$$

$$\mathbf{Q} = \mathbf{P} \begin{bmatrix} \Delta^{-1} & \mathbf{Q}_2 \\ \mathbf{Q}_2' & \mathbf{Q}_3 \end{bmatrix} \mathbf{P}' - \mathbf{P} \begin{bmatrix} \Delta^{-1/2}(\mathbf{I} - \overline{\mathbf{Q}}_1)\Delta^{-1/2} & : \mathbf{0} \\ \mathbf{0} & : \mathbf{0} \end{bmatrix} \mathbf{P}'$$

$$= \mathbf{V}^{-} - \mathbf{V}^{=} .$$

Clearly $\mathbf{V}^-$ is a g_1-inverse of $\mathbf{V}$ and $\mathbf{V}^=$ is a symmetric g_2-inverse of $\mathbf{V}$ such that $C(\mathbf{V}^=) \subset C(\mathbf{V})$. Thus any $\mathbf{Q}$ satisfying (1.1) can be written in the form (1.6). On the other hand, it is easy to show that any matrix $\mathbf{Q}$ of the form (1.6) satisfies (1.1). This completes the proof of the lemma.

□

If $\mathbf{y} \sim N(\mathbf{0}, \mathbf{V})$ the lemma typifies all quadratic forms $\mathbf{y}'\mathbf{Q}\mathbf{y}$ which are distributed chi-squared, since (1.2) and (1.3) are trivially satisfied in this case.

It is interesting to note the reduction of (1.6) in the cases of nonsingular $\mathbf{V}$ and $\mathbf{V} = \mathbf{I}$ to well-known matrix types.

We note further that for (1.1) to hold it is sufficient that $\mathbf{V}^=$ in (1.6) is any symmetric g_2-inverse of $\mathbf{V}$. The column space condition $C(\mathbf{V}^=) \subset C(\mathbf{V})$ can therefore be dropped. In any event, being a symmetric g_2-inverse of a symmetric and nonnegative definite matrix, $\mathbf{V}^=$ is necessarily nonnegative definite.

Along the lines of the proof of Theorem 1.1 it can be shown that a general (not necessarily symmetric) solution $\mathbf{Q}$ to (1.1) can be written as $\mathbf{Q} = \mathbf{V}^- - \mathbf{V}^=$, where $\mathbf{V}^-$ and $\mathbf{V}^=$ are arbitrary g_1- and g_2-inverses respectively of $\mathbf{V}$.

This is still true when $\mathbf{V} = \mathbf{B}$ is any matrix. Thus a general solution $\mathbf{X}$ to the matrix equation $\mathbf{BXBXB} = \mathbf{BXB}$ can be written as $\mathbf{X} = \mathbf{B}^- - \mathbf{B}^=$, i.e. as the difference of general solutions $\mathbf{W}$ and $\mathbf{Y}$ to the equations $\mathbf{BWB} = \mathbf{B}$ and $\mathbf{YBY} = \mathbf{Y}$ respectively.

This contrasts with the result of Mitra (1968) who showed that $\mathbf{X}$ can be written as the sum of general solutions $\mathbf{Y}$ and $\mathbf{W}$ to $\mathbf{YBY} = \mathbf{Y}$ and $\mathbf{BWB} = \mathbf{0}$ respectively.

Finally, since the nonnegative definite solutions $\mathbf{Q}$ to $\mathbf{VQVQV} = \mathbf{VQV}$ and $\mathbf{QVQVQ} = \mathbf{QVQ}$ coincide for nonnegative definite $\mathbf{V}$ (Bhimasankaram and Majumdar, 1980), we can conclude that any nonnegative definite solution $\mathbf{Q}$ to $\mathbf{QVQVQ} = \mathbf{QVQ}$ can be written as $\mathbf{Q} = \mathbf{V}^- - \mathbf{V}^=$ where $\mathbf{V}^-$ is a nonnegative definite g_1-inverse of $\mathbf{V}$ and $\mathbf{V}^=$ is a nonnegative definite g_2-inverse of $\mathbf{V}$ with $C(\mathbf{V}^=) \subset C(\mathbf{V})$.

Using the representation (1.6) we can now give equivalent conditions for (1.2) to hold in addition to (1.1) and for (1.3) to hold in addition to (1.1) and (1.2).

Corollary 1.1.1

If $\mathbf{Q}$ is of the form (1.6) then

(1.10) $$\mathbf{VQVQ}\mu = \mathbf{VQ}\mu \iff \mathbf{V}^=\mu = \mathbf{V}^=\mathbf{V}\mathbf{V}^-\mu$$

If (1.10) holds and $\mathbf{Q}$ is of the form (1.6) then

(1.11) $$\mu'\mathbf{QVQ}\mu = \mu'\mathbf{Q}\mu \iff \mu'\mathbf{V}^-\mathbf{V}\mathbf{V}^-\mu = \mu'\mathbf{V}^-\mu \, .$$

Proof: With (1.6) we have

(1.12) $$\mathbf{VQ} = \mathbf{VV}^{-} - \mathbf{VV}^{=}, \quad \text{and}$$

(1.13) $$\mathbf{VQVQ} = \mathbf{VV}^{-} - \mathbf{VV}^{=}\mathbf{VV}^{-}$$

which proves (1.10), and since

(1.14) $$\mathbf{QVQ} = \mathbf{V}^{-}\mathbf{VV}^{-} - \mathbf{V}^{-}\mathbf{VV}^{=} - \mathbf{V}^{=}\mathbf{VV}^{-} + \mathbf{V}^{=} \quad .$$

Now (1.11) is proved using (1.10).

□

In summary, with Theorem 1.1 and its corollary we have proved:
If $\mathbf{y} \sim N(\mu, \mathbf{V})$ the quadratic form $\mathbf{y}'\mathbf{Qy}$ is distributed chi-squared if and only if

(1.15) (i) $\mathbf{Q}$ can be written as $\mathbf{Q} = \mathbf{V}^{-} - \mathbf{V}^{=}$ and

(ii) $\mathbf{V}^{=}\mu = \mathbf{V}^{=}\mathbf{V}\mathbf{V}^{-}\mu$ and

(iii) $\mu'\mathbf{V}^{-}\mathbf{V}\mathbf{V}^{-}\mu = \mu'\mathbf{V}^{-}\mu$.

Independence of quadratic forms

Khatri (1962, 1963) showed that two quadratic forms $\mathbf{y}'\mathbf{Q}_1\mathbf{y}$ and $\mathbf{y}'\mathbf{Q}_2\mathbf{y}$ in a normal variable $\mathbf{y} \sim N(\mu, \mathbf{V})$ are independently distributed if and only if

(1.16) $$\mathbf{VQ}_1\mathbf{VQ}_2\mathbf{V} = \mathbf{0}$$

(1.17) $$\mathbf{VQ}_1\mathbf{VQ}_2\mu = \mathbf{0}$$

(1.18) $$\mathbf{VQ}_2\mathbf{VQ}_1\mu = \mathbf{0}$$

(1.19) $$\mu'\mathbf{Q}_1\mathbf{VQ}_2\mu = \mathbf{0}$$

We note that conditions (1.17) through (1.19) are trivially satisfied when (i) $\mathbf{y} \sim N(\mathbf{0}, \mathbf{V})$ or (ii) (1.16) holds and $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are nonnegative definite. Thus we obtain the following corollary.

Corollary 1.1.2

Let $\mathbf{Q}_1 = \mathbf{V}_1^{-} - \mathbf{V}_1^{=}$ and $\mathbf{Q}_2 = \mathbf{V}_2^{-} - \mathbf{V}_2^{=}$ where $\mathbf{V}_j^{-}$, $\mathbf{V}_j^{=}$ as in (1.6), $j = 1,2$. Let further $\mathbf{y} \sim N(\mathbf{0}, \mathbf{V})$ or $\mathbf{Q}_1$, $\mathbf{Q}_2$ nonnegative definite.

Then the two quadratic forms $\mathbf{y}'\mathbf{Q}_1\mathbf{y}$ and $\mathbf{y}'\mathbf{Q}_2\mathbf{y}$ are independently distributed if and only if

(1.20) $$(\mathbf{I} - \mathbf{VV}_1^{=})(\mathbf{I} - \mathbf{VV}_2^{=})\mathbf{V} = \mathbf{0} \, .$$

□

1.2 ON THE QUADRATIC FORM $\hat{\mathbf{y}}'\mathbf{V}^-\hat{\mathbf{y}}$

Rao (1971) showed that in the linear model $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ the best linear unbiased estimator (BLUE) $\mathbf{X}\hat{\beta} = \hat{\mathbf{y}}$ for $\mathbf{X}\beta$ is given by

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^-\mathbf{X}'\mathbf{V}^*\mathbf{y} \tag{1.21}$$
$$= \mathbf{X}\mathbf{A}\mathbf{y} \quad ,$$

where $\mathbf{V}^* = (\mathbf{V} + \mathbf{X}\mathbf{U}\mathbf{X}')^-$ is a g-inverse of $\mathbf{V}$ in the manner of Rao (1971). That is, $\mathbf{U}$ is an arbitrary matrix such that $C(\mathbf{V} + \mathbf{X}\mathbf{U}\mathbf{X}') = C([\mathbf{X} : \mathbf{V}])$ and $C(\mathbf{V}) \cap C(\mathbf{X}\mathbf{U}\mathbf{X}') = \{\mathbf{0}\}$. The matrix $\mathbf{V}^*$ need not be symmetric for the estimation process, but during this development we assume the symmetry of $\mathbf{V}^*$, which results in no loss of generality, and allows the application of conditions for chi-squaredness.

When $\mathbf{y}$ follows a normal distribution, the quadratic form $\hat{\mathbf{y}}'\mathbf{V}^-\hat{\mathbf{y}}$ is not in general, i.e. not for an arbitrary choice of a g-inverse $\mathbf{V}^-$ of $\mathbf{V}$ distributed (noncentral) chi-squared. In the following we determine the class of symmetric g-inverses $\mathbf{V}^-$ of $\mathbf{V}$ for which the quadratic form $\hat{\mathbf{y}}'\mathbf{V}^-\hat{\mathbf{y}}$ is distributed chi-squared, and we show that this class is not empty. An ANOVA for the linear model $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ in terms of chi-square statistics is given, such that the total sum of squares in the model is decomposed into independent sums of squares which are distributed chi-squared, and a nonstochastic sum of squares.

Theorem 1.2 (Schall)

If $\mathbf{V}^-$ is a symmetric g-inverse of $\mathbf{V}$, then the quadratic form $\hat{\mathbf{y}}'\mathbf{V}^-\hat{\mathbf{y}}$ is distributed chi-squared under the normal linear model $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ if and only if

$$\mathbf{X}'\mathbf{V}^-\mathbf{V}\mathbf{V}^-\mathbf{X} = \mathbf{X}'\mathbf{V}^-\mathbf{X} \quad , \quad \text{and} \tag{1.22}$$

$$\mathbf{V}\mathbf{V}^-\mathbf{X} = \mathbf{X}\mathbf{B} \quad , \quad \text{for some } \mathbf{B}. \tag{1.23}$$

The class of g-inverses $\mathbf{V}^-$ of $\mathbf{V}$ satisfying (1.22) and (1.23) is not empty.

Proof:

Sufficiency: We must check conditions (1.1) through (1.3), where it is obviously sufficient to show that $\mathbf{Q}\mathbf{V}\mathbf{Q} = \mathbf{Q}$. But we can write

$$\hat{\mathbf{y}}'\mathbf{V}^-\hat{\mathbf{y}} = \mathbf{y}'\mathbf{A}'\mathbf{X}'\mathbf{V}^-\mathbf{X}\mathbf{A}\mathbf{y} \quad , \quad \text{from (1.21)} \tag{1.24}$$
$$= \mathbf{y}'\mathbf{Q}\mathbf{y}$$

so that $\mathbf{Q} = \mathbf{A}'\mathbf{X}'\mathbf{V}^-\mathbf{X}\mathbf{A}$. It is well-known that

$$\mathbf{X}\mathbf{A}\mathbf{X} = \mathbf{X} \quad , \quad \text{and} \tag{1.25}$$

$$\mathbf{X}\mathbf{A}\mathbf{V}\mathbf{A}'\mathbf{X}' = \mathbf{X}\mathbf{A}\mathbf{V} \, . \tag{1.26}$$

Then

(1.27) $$\begin{aligned}\mathbf{QVQ} &= \mathbf{A'X'V^{-}XAVA'X'V^{-}XA} \\ &= \mathbf{A'X'V^{-}XAVV^{-}XA}\ , \quad \text{from (1.26)} \\ &= \mathbf{A'X'V^{-}XAXBA}\ , \quad \text{from (1.23)} \\ &= \mathbf{A'X'V^{-}XBA}\ , \quad \text{from (1.25)} \\ &= \mathbf{A'X'V^{-}VV^{-}XA}\ , \quad \text{from (1.23)} \\ &= \mathbf{A'X'V^{-}XA}\ , \quad \text{from (1.22)} \\ &= \mathbf{Q}\ .\end{aligned}$$

Let G denote the class of g-inverses $\mathbf{V}^{-}$ of $\mathbf{V}$ which satisfy the conditions (1.22) and (1.23). We proceed to show that this class is not empty. Let $\mathbf{V}^{*}$ be a g-inverse of $\mathbf{V}$ as given by (1.21), that is, $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X'V^{*}X})^{-}\mathbf{X'V^{*}y}$. Now take $\mathbf{V}^{-}$ as

(1.28) $$\mathbf{V}^{-} = \mathbf{V^{*}VV^{*}}\ .$$

Clearly, $\mathbf{V}^{*}$ is a g-inverse of $\mathbf{V}$ with

(1.29) $$\mathbf{V^{-}VV^{-}} = \mathbf{V^{*}VV^{*}VV^{*}VV^{*}} = \mathbf{V^{*}VV^{*}} = \mathbf{V^{-}}\ , \quad \text{and}$$

(1.30) $$\mathbf{VV^{-}X} = \mathbf{VV^{*}VV^{*}X} = \mathbf{VV^{*}X} = \mathbf{XB}\ , \quad \text{for some } \mathbf{B}.$$

Thus $\mathbf{V}^{-} = \mathbf{V^{*}VV^{*}}$ satisfies the conditions (1.22) and (1.23) of Theorem 1.2, and G is not empty.

Necessity: We consider condition (1.3): it is necessary for $\hat{\mathbf{y}}'\mathbf{V}^{-}\hat{\mathbf{y}}$ to be distributed chi-squared that

(1.31) $$\begin{aligned}&\beta'\mathbf{X'A'X'V^{-}XAX}\beta = \beta'\mathbf{X'A'X'V^{-}XAVA'X'V^{-}XAX}\beta\ , \quad \text{for all } \beta \\ &\qquad \Leftrightarrow \quad \mathbf{X'V^{-}X} = \mathbf{X'V^{-}XAVV^{-}X}\ , \quad \text{from (1.25), (1.26)}\ ,\end{aligned}$$

But $\mathbf{VV^{-}X}$ can always be written as

(1.32) $$\mathbf{VV^{-}X} = \mathbf{XB} + \mathbf{VZC}\ , \quad \text{for some } \mathbf{B} \text{ and } \mathbf{C}$$

where $\mathbf{Z}$ is a matrix of maximum rank such that $\mathbf{X'Z} = \mathbf{0}$. It is well-known that

(1.33) $$\mathbf{XAVZ} = \mathbf{0}\ ,$$

so that we can write

(1.34) $$\begin{aligned}\mathbf{X'V^{-}XAVV^{-}X} &= \mathbf{X'V^{-}XAXB} + \mathbf{X'V^{-}XAVZC}\ , \quad \text{from (1.32)} \\ &= \mathbf{X'V^{-}XB}\ , \quad \text{from (1.25), (1.33)} \\ &= \mathbf{X'V^{-}VV^{-}X} - \mathbf{X'V^{-}VZC}\ , \quad \text{from (1.32)} \\ &= \mathbf{X'V^{-}VV^{-}X} - \mathbf{C'Z'VZC}\ , \quad \text{from (1.32)}\ .\end{aligned}$$

But any quadratic form which is distributed chi-squared is nonnegative definite w.p.1 as pointed out by Mitra (1968), so that $\mathbf{X}'\mathbf{V}^-\mathbf{X} \geq \mathbf{X}'\mathbf{V}^-\mathbf{V}\mathbf{V}^-\mathbf{X}$, and clearly $\mathbf{C}'\mathbf{Z}'\mathbf{V}\mathbf{Z}\mathbf{C} \geq \mathbf{0}$. Thus for (1.31) to hold it is necessary that $\mathbf{C}'\mathbf{Z}'\mathbf{V}\mathbf{Z}\mathbf{C} = \mathbf{0}$, or equivalently $\mathbf{V}\mathbf{Z}\mathbf{C} = \mathbf{0}$. Noting (1.32) this is in turn equivalent to $\mathbf{V}\mathbf{V}^-\mathbf{X} = \mathbf{X}\mathbf{B}$. Now using $\mathbf{V}\mathbf{V}^-\mathbf{X} = \mathbf{X}\mathbf{B}$ in (1.31) leads to $\mathbf{X}'\mathbf{V}^-\mathbf{V}\mathbf{V}^-\mathbf{X} = \mathbf{X}'\mathbf{V}^-\mathbf{X}$.

□

We note that (1.22) is equivalent with $\mathbf{V}^-\mathbf{V}\mathbf{V}^- = \mathbf{V}^-$, that is, $\mathbf{V}^-$ is a g_2- or reflexive type g-inverse of $\mathbf{V}$, if and only if $C([\mathbf{X} : \mathbf{V}]) = \mathbb{R}^n$, when $\mathbf{y}$ is a n-variate. Further, condition (1.22) is equivalent with the chi-squaredness of $\mathbf{y}'\mathbf{V}^-\mathbf{y}$, and (1.23) is equivalent with $(\mathbf{y}-\hat{\mathbf{y}})'\mathbf{V}^-\hat{\mathbf{y}} = \hat{\mathbf{e}}'\mathbf{V}^-\hat{\mathbf{y}} = \mathbf{0}$, since with $\hat{\mathbf{e}} \in C(\mathbf{V}\mathbf{Z})$ and $\hat{\mathbf{y}} \in C(\mathbf{X})$ we have that $\hat{\mathbf{e}} = \mathbf{V}\mathbf{Z}\gamma$ and $\hat{\mathbf{y}} = \mathbf{X}\lambda$ for some γ and λ. Then $\hat{\mathbf{e}}'\mathbf{V}^-\hat{\mathbf{y}} = \lambda'\mathbf{Z}'\mathbf{V}\mathbf{V}^-\mathbf{X}\lambda = \lambda'\mathbf{Z}'\mathbf{X}\mathbf{B}\lambda = \mathbf{0}$. Thus (1.22) is responsible for the chi-squaredness of $\mathbf{y}'\mathbf{V}^-\mathbf{y}$, and (1.23) allows for the decomposition of $\mathbf{y}'\mathbf{V}^-\mathbf{y}$ as $\mathbf{y}'\mathbf{V}^-\mathbf{y} = \hat{\mathbf{y}}'\mathbf{V}^-\hat{\mathbf{y}} + \hat{\mathbf{e}}'\mathbf{V}^-\hat{\mathbf{e}}$. But $\hat{\mathbf{e}}'\mathbf{V}^-\hat{\mathbf{e}}$ is distributed chi-squared for any g-inverse $\mathbf{V}^-$ of $\mathbf{V}$, and with the chi-squaredness of $\mathbf{y}'\mathbf{V}^-\mathbf{y}$ we have the chi-squaredness of $\hat{\mathbf{y}}'\mathbf{V}^-\hat{\mathbf{y}}$ as the difference of two chi-square variates and its independence of one of the variates.

We now consider the quadratic form $\hat{\mathbf{y}}'\mathbf{V}^*\hat{\mathbf{y}}$, where $\mathbf{V}^*$ is a g-inverse of $\mathbf{V}$ as given by (1.21). This quadratic form plays an important role in the analysis of variance (ANOVA) of the linear model $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$. It is well-known that the total sum of squares (SS) $\mathbf{y}'\mathbf{V}^*\mathbf{y}$ associated with the linear model $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ can be decomposed into uncorrelated sums of squares as

(1.35)
$$\begin{aligned} SS &= \mathbf{y}'\mathbf{V}^*\mathbf{y} \\ &= \hat{\mathbf{y}}'\mathbf{V}^*\hat{\mathbf{y}} + \hat{\mathbf{e}}'\mathbf{V}^*\hat{\mathbf{e}} \\ &= SSR + SSE\,, \end{aligned}$$

where $\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}}$. The sum of squares for error (SSE) $\hat{\mathbf{e}}'\mathbf{V}^*\hat{\mathbf{e}}$ is distributed chi-squared, but in general the sum of squares for regression (SSR) $\hat{\mathbf{y}}'\mathbf{V}^*\hat{\mathbf{y}}$ is not distributed chi-squared, since it does not in general satisfy condition (1.3), and consequently the total sum of squares $\mathbf{y}'\mathbf{V}^*\mathbf{y}$ also fails this condition. Dunne (1982) showed that $SSR = \hat{\mathbf{y}}'\mathbf{V}^*\hat{\mathbf{y}}$ and thus $SS = \mathbf{y}'\mathbf{V}^*\mathbf{y}$ is distributed chi-squared under the linear model $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ if and only if $C(\mathbf{X}) \subset C(\mathbf{V})$.

The failure in general of the noncentrality parameter condition (1.3) is due to the almost sure contribution to the sum of squares made by $\mathbf{X}\mathbf{U}\mathbf{X}'$ being required in the model and in $\mathbf{V}^* = (\mathbf{V} + \mathbf{X}\mathbf{U}\mathbf{X}')^-$. The contribution is evident in writing $SS = \mathbf{y}'\mathbf{V}^*\mathbf{y}$ as

(1.36)
$$\begin{aligned} SS &= \mathbf{y}'\mathbf{V}^*\mathbf{y} \\ &= \mathbf{y}'\mathbf{V}^*(\mathbf{V} + \mathbf{X}\mathbf{U}\mathbf{X}')\mathbf{V}^*\mathbf{y} \\ &= \mathbf{y}'\mathbf{V}^*\mathbf{X}\mathbf{U}\mathbf{X}'\mathbf{V}^*\mathbf{y} + \mathbf{y}'\mathbf{V}^*\mathbf{V}\mathbf{V}^*\mathbf{y} \\ &= \mathbf{y}'\mathbf{V}^*\mathbf{X}\mathbf{U}\mathbf{X}'\mathbf{V}^*\mathbf{y} + \hat{\mathbf{y}}'\mathbf{V}^*\mathbf{V}\mathbf{V}^*\hat{\mathbf{y}} + \hat{\mathbf{e}}'\mathbf{V}^*\mathbf{V}\mathbf{V}^*\hat{\mathbf{e}} \\ &= \mathbf{y}'\mathbf{V}^*\mathbf{X}\mathbf{U}\mathbf{X}'\mathbf{V}^*\mathbf{y} + \hat{\mathbf{y}}'\mathbf{V}^*\mathbf{V}\mathbf{V}^*\hat{\mathbf{y}} + \hat{\mathbf{e}}'\mathbf{V}^*\hat{\mathbf{e}} \\ &= SSS + (SSR - SSS) + SSE \\ &= SSR + SSE\,. \end{aligned}$$

Note that the sure sum of squares $SSS = \mathbf{y}'\mathbf{V}^*\mathbf{X}\mathbf{U}\mathbf{X}'\mathbf{V}^*\mathbf{y}$ is invariant (w.p.1) over all observations from the linear model $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$, since $(\mathbf{y}_1 - \mathbf{y}_2) \in C(\mathbf{V})$, w.p.1 for a an arbitrary pair of observations $\mathbf{y}_1$ and $\mathbf{y}_2$, and thus

(1.38)
$$\begin{aligned}
&(\mathbf{y}_1 - \mathbf{y}_2)'\mathbf{V}^*\mathbf{X}\mathbf{U}\mathbf{X}' \\
&= (\mathbf{y}_1 - \mathbf{y}_2)'(\mathbf{V} + \mathbf{X}\mathbf{U}\mathbf{X}')^-(\mathbf{X}\mathbf{U}\mathbf{X}' + \mathbf{V} - \mathbf{V}) \\
&= (\mathbf{y}_1 - \mathbf{y}_2)'(\mathbf{V} + \mathbf{X}\mathbf{U}\mathbf{X}')^-(\mathbf{V} + \mathbf{X}\mathbf{U}\mathbf{X}') - (\mathbf{y}_1 - \mathbf{y}_2)'\mathbf{V}^*\mathbf{V} \\
&= (\mathbf{y}_1 - \mathbf{y}_2)' - (\mathbf{y}_1 - \mathbf{y}_2)' \\
&= \mathbf{0} .
\end{aligned}$$

Specifically we have

(1.39)
$$\begin{aligned}
SSS &= \mathbf{y}'\mathbf{V}^*\mathbf{X}\mathbf{U}\mathbf{X}'\mathbf{V}^*\mathbf{y} \\
&= \hat{\mathbf{y}}'\mathbf{V}^*\mathbf{X}\mathbf{U}\mathbf{X}'\mathbf{V}^*\hat{\mathbf{y}} \\
&= \beta'\mathbf{X}'\mathbf{V}^*\mathbf{X}\mathbf{U}\mathbf{X}'\mathbf{V}^*\mathbf{X}\beta , \quad \text{w.p.1} .
\end{aligned}$$

Finally, noting that the sum of squares $(SSR - SSS) = \hat{\mathbf{y}}'\mathbf{V}^*\mathbf{V}\mathbf{V}^*\hat{\mathbf{y}}$ is distributed chi-squared, from Theorem 1.2, we can give the ANOVA table corresponding to the decomposition (1.36).

Table 1.3 (Schall and Dunne)

ANOVA of the linear model $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$

SS	df	statistic	expectation
SSS	0	$\hat{\mathbf{y}}'\mathbf{V}^*\mathbf{X}\mathbf{U}\mathbf{X}'\mathbf{V}^*\hat{\mathbf{y}}$	$\beta'\mathbf{X}'\mathbf{V}^*\mathbf{X}\mathbf{U}\mathbf{X}'\mathbf{V}^*\mathbf{X}\beta$
$SSR - SSS$	$r(\mathbf{V}) - s$	$\hat{\mathbf{y}}'\mathbf{V}^*\mathbf{V}\mathbf{V}^*\hat{\mathbf{y}}$	$\beta'\mathbf{X}'\mathbf{V}^*\mathbf{V}\mathbf{V}^*\mathbf{X}\beta + (r(\mathbf{V}) - s)\cdot\sigma^2$
SSE	$s = r([\mathbf{X} : \mathbf{V}]) - r(\mathbf{X})$	$\hat{\mathbf{e}}'\mathbf{V}^*\hat{\mathbf{e}}$	$s \cdot \sigma^2$
SS	$r(\mathbf{V})$	$\mathbf{y}'\mathbf{V}^*\mathbf{y}$	$\beta'\mathbf{X}'\mathbf{V}^*\mathbf{X}\beta + r(\mathbf{V}) \cdot \sigma^2$

□

As noted above, SSR and SSE are uncorrelated, and with $(SSR - SSS)$ being distributed chi-squared we have that $(SS - SSS)$ is distributed chi-squared, as the sum of two independent chi-square variates. We may note that $(SS - SSS)$ and $(SSR - SSS)$ are invariant over the special choice of $\mathbf{V}^*$, since any observation $\mathbf{y}$ from the linear model $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ can be written as $\mathbf{y} = \mathbf{V}\lambda + \mathbf{X}\mathbf{U}\mathbf{X}'\xi$ for some λ and ξ, where $\mathbf{V}\lambda$ is invariant over the choice of $\mathbf{U}$ in $\mathbf{V}^* = (\mathbf{V} + \mathbf{X}\mathbf{U}\mathbf{X}')^-$, from $C(\mathbf{V}) \cap C(\mathbf{X}\mathbf{U}\mathbf{X}') = \{\mathbf{0}\}$. Thus

(1.40)
$$
\begin{aligned}
& SS - SSS \\
&= \mathbf{y}'\mathbf{V}^*\mathbf{V}\mathbf{V}^*\mathbf{y} \\
&= (\mathbf{V}\lambda + \mathbf{X}\mathbf{U}\mathbf{X}'\xi)'\mathbf{V}^*\mathbf{V}\mathbf{V}^*(\mathbf{V}\lambda + \mathbf{X}\mathbf{U}\mathbf{X}'\xi) \\
&= \lambda'\mathbf{V}\mathbf{V}^*\mathbf{V}\mathbf{V}^*\mathbf{V}\lambda \ , \quad \text{since } \mathbf{X}\mathbf{U}\mathbf{X}'\mathbf{V}^*\mathbf{V} = \mathbf{0} \quad \text{from (1.38)} \\
&= \lambda'\mathbf{V}\lambda
\end{aligned}
$$

is invariant over the special choice of $\mathbf{V}^* = (\mathbf{V} + \mathbf{X}\mathbf{U}\mathbf{X}')^-$, and so is similarly $SSR - SSS = \hat{\mathbf{y}}'\mathbf{V}^*\mathbf{V}\mathbf{V}^*\hat{\mathbf{y}}$. Thus the total sum of squares $SS = \mathbf{y}'\mathbf{V}^*\mathbf{y}$ under the general linear model $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ can be decomposed into a nonstochastic sum of squares SSS which in general depends on the special choice of $\mathbf{V}^*$, and a unique stochastic sum of squares $SS - SSS$. This stochastic sum of squares can in turn be uniquely decomposed into $(SSR - SSS)$ and SSE, the sum of squares for regression adjusted for the sure sum of squares, and the sum of squares for error $\hat{\mathbf{e}}'\mathbf{V}^*\hat{\mathbf{e}} = \hat{\mathbf{e}}'\mathbf{V}^-\hat{\mathbf{e}}$ which is known to be invariant over an arbitrary choice of a g-inverse $\mathbf{V}^-$ of $\mathbf{V}$.

1.3 A LEMMA ON A GENERALIZED INVERSE OF A MATRIX

After presenting a decomposition of the vector space $\mathbb{R}^P$, we show a lemma on a generalized inverse of a matrix. This lemma can be used to prove a theorem by Dunne (1982), which itself is a generalization of a result by Chipman (1964).

Lemma 1.4 (Schall, 1984)

Let the $n \times p$-matrix $\mathbf{X}$ be partitioned as

(1.41)
$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$$

where $\mathbf{X}_1$ is $(n-k) \times p$ and $\mathbf{X}_2$ is $k \times p$, and let $R(\mathbf{X}_1)$ and $R(\mathbf{X}_2)$ be the corresponding row spaces. Then

$$\text{(i)} \qquad \mathbb{R}^P = S_0 + S_1 + S_2 \ ,$$

where $S_0 = R(\mathbf{X}_1) \cap R(\mathbf{X}_2)$, $S_1 = R(\mathbf{X}_2)^\perp$ and $S_2 = R(\mathbf{X}_1)^\perp$. The spaces S_1 and S_2 are not necessarily disconnected.

$$\text{(ii)} \qquad S_1 \cap S_2 \cap R(\mathbf{X}) = \{\mathbf{0}\} \ .$$

Proof:

(i) It is sufficient to show that $S_0^\perp = S_1 + S_2$. The relationship $(S_1 + S_2) \subset S_0^\perp$ is trivial, and $S_0^\perp \subset (S_1 + S_2)$ is equivalent to $(S_1 + S_2)^\perp \subset S_0$. But $\ell \in (S_1 + S_2)^\perp$ implies that $\ell \perp R(\mathbf{X}_1)^\perp$ and $\ell \perp R(\mathbf{X}_2)^\perp$, then $\ell \in R(\mathbf{X}_1)$ and $\ell \in R(\mathbf{X}_2)$, and finally $\ell \in S_0 = R(\mathbf{X}_1) \cap R(\mathbf{X}_2)$.

(ii) Let $\ell \in S_1 \cap S_2 = R(\mathbf{X}_2)^\perp + R(\mathbf{X}_1)^\perp$. Then $\ell \in R(\mathbf{X})^\perp$, thus $\ell \notin R(\mathbf{X})$ when $\ell \neq \mathbf{0}$. □

Now a result on a generalized inverse of a matrix can be shown.

Lemma 1.5 (Schall, 1984)

Let $\mathbf{X}_1$ be a $(n-k)\times p$-matrix and $\mathbf{X}_2$ be a $k\times p$-matrix, $n > k$. Let further be $\ell \in (R(\mathbf{X}_1)^{\perp} + R(\mathbf{X}_2)^{\perp})$. Then

$$\text{(i)} \quad \ell'\mathbf{X}_1'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1+\mathbf{X}_2'\mathbf{X}_2)^{-}\mathbf{X}_2' = \mathbf{0}$$

$$\text{(ii)} \quad \ell'\mathbf{X}_1'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1+\mathbf{X}_2'\mathbf{X}_2)^{-}\mathbf{X}_1' = \ell'\mathbf{X}_1' .$$

Proof: Let be $\ell = \ell_1+\ell_2$ with $\ell_1 \in S_1 = R(\mathbf{X}_2)^{\perp}$ and $\ell_2 \in S_2 = R(\mathbf{X}_1)^{\perp}$. Then

(1.42)
$$\ell_1'\mathbf{X}_2' = \mathbf{0} , \quad \text{and}$$
$$\ell_2'\mathbf{X}_1' = \mathbf{0} ,$$

(i) $\ell'\mathbf{X}_1'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1 + \mathbf{X}_2'\mathbf{X}_2)^{-}\mathbf{X}_2'$

$= (\ell_1' + \ell_2')\mathbf{X}_1'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1 + \mathbf{X}_2'\mathbf{X}_2)^{-}\mathbf{X}_2'$

$= \ell_1'\mathbf{X}_1'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1 + \mathbf{X}_2'\mathbf{X}_2)^{-}\mathbf{X}_2' ,$ from (1.42)

$= \ell_1'\mathbf{X}_1'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1 + \mathbf{X}_2'\mathbf{X}_2)^{-}\mathbf{X}_2' + \ell_1'\mathbf{X}_2'\mathbf{X}_2(\mathbf{X}_1'\mathbf{X}_1 + \mathbf{X}_2'\mathbf{X}_2)^{-}\mathbf{X}_2'$

$= \ell_1'(\mathbf{X}_1'\mathbf{X}_1 + \mathbf{X}_2'\mathbf{X}_2)(\mathbf{X}_1'\mathbf{X}_1 + \mathbf{X}_2'\mathbf{X}_2')^{-}\mathbf{X}_2'$

$= \ell_1'\mathbf{X}_2'$

$= \mathbf{0} ,$ from (1.42).

(ii) $\ell'\mathbf{X}_1'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1 + \mathbf{X}_2'\mathbf{X}_2)^{-}\mathbf{X}_1'$

$= (\ell_1' + \ell_2')\mathbf{X}_1'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1 + \mathbf{X}_2'\mathbf{X}_2)^{-}\mathbf{X}_1'$

$= \ell_1'\mathbf{X}_1'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1 + \mathbf{X}_2'\mathbf{X}_2)^{-}\mathbf{X}_1' ,$ from (1.42)

$= \ell_1'\mathbf{X}_1'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1 + \mathbf{X}_2'\mathbf{X}_2)^{-}\mathbf{X}_1' + \ell_1'\mathbf{X}_2'\mathbf{X}_2(\mathbf{X}_1'\mathbf{X}_1 + \mathbf{X}_2'\mathbf{X}_2)^{-}\mathbf{X}_1'$

$= \ell_1'(\mathbf{X}_1'\mathbf{X}_1 + \mathbf{X}_2'\mathbf{X}_2)(\mathbf{X}_1'\mathbf{X}_1 + \mathbf{X}_2'\mathbf{X}_2')^{-}\mathbf{X}_1'$

$= \ell_1'\mathbf{X}_1'$

$= \ell_1'\mathbf{X}_1' + \ell_2'\mathbf{X}_1'$

$= \ell'\mathbf{X}_1' .$

□

As a corollary to Lemma 1.5 we obtain

Theorem 1.6 (Dunne, 1982)

Let $\mathbf{X}_1$ be a $(n-k)\times p$-matrix and $\mathbf{X}_2$ be a $k\times p$-matrix, $n > k$ with $R(\mathbf{X}_1) \cap R(\mathbf{X}_2) = \{\mathbf{0}\}$. Then

(i) $\mathbf{X}_1'(\mathbf{X}_1'\mathbf{X}_1 + \mathbf{X}_2'\mathbf{X}_2)^-\mathbf{X}_2' = \mathbf{0}$

(ii) $(\mathbf{X}_1'\mathbf{X}_1 + \mathbf{X}_2'\mathbf{X}_2)^-\mathbf{X}_1'$ is a g-inverse of $\mathbf{X}_1$.

Proof: (Schall)

$S_0 = R(\mathbf{X}_1) \cap R(\mathbf{X}_2) = \{\mathbf{0}\}$, thus $(S_1 + S_2) = (R(\mathbf{X}_2)^\perp + R(\mathbf{X}_1)^\perp) = \mathbb{R}^P$ from Lemma 1.4. Thus the results (i) and (ii) of Lemma 1.5 hold for any $\ell \in \mathbb{R}^P$, which implies

$$\mathbf{X}_1'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1 + \mathbf{X}_2'\mathbf{X}_2)^-\mathbf{X}_2' = \mathbf{0}\,, \quad \text{and} \tag{1.43}$$

$$\mathbf{X}_1'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1 + \mathbf{X}_2'\mathbf{X}_2)^-\mathbf{X}_1' = \mathbf{X}_1'\,. \tag{1.44}$$

But (1.43) is equivalent with (i) and (1.44) is equivalent with (ii).

□

Theorem 1.6 is a generalization of a result by Chipman (1964) who required that $R(\mathbf{X}_1) \oplus R(\mathbf{X}_2) = \mathbb{R}^P$, the whole space.

1.4 BLU-ESTIMATION IN THE VARIANCE COMPONENTS MODEL

We consider the linear model (LM) $(\mathbf{y}, \mathbf{X}\beta, \mathbf{V} = \Sigma\sigma_i^2\mathbf{V}_i)$ with k variance components $(\sigma_1^2, \ldots, \sigma_k^2)$ and arbitrary nonnegative-definite and symmetric $\mathbf{V}_1, \ldots, \mathbf{V}_k$. In general the BLUE $\ell'\hat{\beta}$ of an estimable linear function $\ell'\beta$ of β is not independent of the variance components $(\sigma_1^2, \ldots, \sigma_k^2)$, as opposed to the general linear model $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ where the BLUE $\mathbf{X}\hat{\beta}$ of $\mathbf{X}\beta$ is known to be independent of σ^2 (which might be viewed as a single variance component). Thus in the $LM(\mathbf{y}, \mathbf{X}\beta, \Sigma\sigma_i^2\mathbf{V}_i)$ the BLUE $\ell'\hat{\beta}$ of $\ell'\beta$ in general is unknown if the variance components are unknown.

In the following we give a necessary and sufficient condition for the BLUE $\ell'\hat{\beta}$ of an estimable linear function $\ell'\beta$ of β to be independent of the variance components. In such a case the BLUE $\ell'\hat{\beta}$ of $\ell'\beta$ is known and can be computed assuming arbitrary values $(\alpha_1^2, \ldots, \alpha_k^2) \neq 0$ for the variance components $(\sigma_1^2, \ldots, \sigma_k^2)$.

Zyskind (1967) showed that in the $LM(\mathbf{y}, \mathbf{X}\beta, \mathbf{V})$, $\mathbf{V}$ arbitrary, a linear function $\mathbf{s}'\mathbf{y}$ of $\mathbf{y}$ is BLUE for its expectation $\mathbf{s}'\mathbf{X}\beta$ if and only if $\mathbf{V}\mathbf{s} \in C(\mathbf{X})$.

Further, Rao (1976) showed that the BLUE $\ell'\hat{\beta}$ of an estimable linear function $\ell'\beta$ of β in the $LM(\mathbf{y}, \mathbf{X}\beta, \mathbf{V})$, $\mathbf{V}$ arbitrary, can always be written as $\ell'\hat{\beta} = \ell'\mathbf{X}^-\mathbf{y}$ where $\mathbf{X}^-$ is the

g-inverse of $\mathbf{X}$ constrained by $\mathbf{VZ}$ ($\mathbf{Z}$ is a matrix of maximum rank such that $\mathbf{X}'\mathbf{Z} = \mathbf{0}$. Without loss of generality $\mathbf{Z}$ can be taken as $\mathbf{Z} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}')$, where $(\mathbf{X}'\mathbf{X})^-$ denotes an arbitrary g-inverse of $\mathbf{X}'\mathbf{X}$).

Using those results we can prove the following theorem:

Theorem 1.7 (Schall, 1984)

Consider the $LM(\mathbf{y},\mathbf{X}\beta,\Sigma\sigma_i^2\mathbf{V}_i)$ with variance components $(\sigma_1^2, \ldots, \sigma_k^2)$ and arbitrary nonnegative-definite and symmetric $\mathbf{V}_1, \ldots, \mathbf{V}_k$. The BLUE $\ell'\hat{\beta}$ of an estimable linear function $\ell'\beta$ of β is independent of the variance components $(\sigma_1^2, \ldots, \sigma_k^2)$ if and only if

$$\ell \in R(\mathbf{S}'\mathbf{X}) \,, \tag{1.45}$$

where $\mathbf{S}$ is a matrix of maximum rank such that

$$C(\mathbf{V}_i\mathbf{S}) \subset C(\mathbf{X}) \,, \quad \text{for all } i = 1, \ldots, k \,. \tag{1.46}$$

Proof: We prove the theorem for a $LM(\mathbf{y},\mathbf{X}\beta,\mathbf{V}=\sigma_1^2\mathbf{V}_1 + \sigma_2^2\mathbf{V}_2)$ with 2 variance components (σ_1^2, σ_2^2). The result for arbitrary $k > 2$ follows similarly.

Sufficiency:

$$\begin{aligned}
&(1.47) \qquad C(\mathbf{V}_1\mathbf{S}) \subset C(\mathbf{X}) \text{ and } C(\mathbf{V}_2\mathbf{S}) \subset C(\mathbf{X}) \\
&\Leftrightarrow \quad C(\sigma_1^2\mathbf{V}_1\mathbf{S} + \sigma_2^2\mathbf{V}_2\mathbf{S}) \subset C(\mathbf{X}) \text{ for all } \sigma_1^2, \sigma_2^2 \\
&\Leftrightarrow \quad C((\sigma_1^2\mathbf{V}_1 + \sigma_2^2\mathbf{V}_2)\mathbf{S}) \subset C(\mathbf{X}) \quad \text{for all } \sigma_1^2, \sigma_2^2 \\
&\Leftrightarrow \quad C(\mathbf{VS}) \subset C(\mathbf{X}) \\
&\Leftrightarrow \quad \mathbf{S}'\mathbf{y} \text{ BLUE for } \mathbf{S}'\mathbf{X}\beta \\
&\qquad\quad \text{in the } LM(\mathbf{y},\mathbf{X}\beta,\mathbf{V}=\sigma_1^2\mathbf{V}_1 + \sigma_2^2\mathbf{V}_2). \quad \text{(Zyskind, 1967)} \\
&\Leftrightarrow \quad \ell'(\mathbf{S}'\mathbf{X})^-\mathbf{S}'\mathbf{y} \text{ BLUE for } \ell'(\mathbf{S}'\mathbf{X})^-(\mathbf{S}'\mathbf{X})\beta \text{ for all } \ell \\
&\Leftrightarrow \quad \ell'(\mathbf{S}'\mathbf{X})^-\mathbf{S}'\mathbf{y} \text{ BLUE for } \ell'\beta \text{ for all } \ell \in R(\mathbf{S}'\mathbf{X}) \,, \\
&\qquad\quad \text{since } \ell'(\mathbf{S}'\mathbf{X})^-(\mathbf{S}'\mathbf{X}) = \ell' \text{ for all } \ell \in R(\mathbf{S}'\mathbf{X}) \,.
\end{aligned}$$

But ℓ,$\mathbf{S}$ and $\mathbf{X}$ are independent of (σ_1^2, σ_2^2), and thus $\ell'\hat{\beta} = \ell'(\mathbf{S}'\mathbf{X})^-\mathbf{S}'\mathbf{y}$ is independent of (σ_1^2, σ_2^2).

Necessity:

The BLUE $\ell'\hat{\beta}$ of an estimable linear function $\ell'\beta$ of β is independent of (σ_1^2, σ_2^2) in the $LM(\mathbf{y},\mathbf{X}\beta,\mathbf{V}=\sigma_1^2\mathbf{V}_1 + \sigma_2^2\mathbf{V}_2)$. Thus $\ell'\hat{\beta}$ is BLUE for $\ell'\beta$ particularly in the models

(1.48) (i) $LM(\mathbf{y},\mathbf{X}\beta,\mathbf{V}_1+\mathbf{V}_2)$, i.e. $\sigma_1^2 = \sigma_2^2 = 1$

(ii) $LM(\mathbf{y},\mathbf{X}\beta,2\mathbf{V}_1+\mathbf{V}_2)$, i.e. $\sigma_1^2 = 2, \sigma_2^2 = 1$

(iii) $LM(\mathbf{y},\mathbf{X}\beta,\mathbf{V}_1+2\mathbf{V}_2)$, i.e. $\sigma_1^2 = 1, \sigma_2^2 = 2$.

Further, $\ell'\hat{\beta}$ can be written as $\ell'\hat{\beta} = \ell'\mathbf{X}^-\mathbf{y}$, $\mathbf{X}^-$ being the g-inverse of $\mathbf{X}$ constrained by $\mathbf{VZ}$ (Rao, 1976). Hence

(1.49) (i) $(\mathbf{V}_1 + \mathbf{V}_2)\mathbf{X}^{-'}\ell \in C(\mathbf{X})$

(ii) $(2\mathbf{V}_1 + \mathbf{V}_2)\mathbf{X}^{-'}\ell \in C(\mathbf{X})$

(iii) $(\mathbf{V}_1 + 2\mathbf{V}_2)\mathbf{X}^{-'}\ell \in C(\mathbf{X})$ (Zyskind, 1967)

Subtracting (i) from (ii) and (iii) in (1.49) yields

(1.50) (ii) $-$(i): $\mathbf{V}_1\mathbf{X}^{-'}\ell \in C(\mathbf{X})$

(iii)$-$(i): $\mathbf{V}_2\mathbf{X}^{-'}\ell \in C(\mathbf{X})$

$\Leftrightarrow$ "by definition": $\mathbf{X}^{-'}\ell \in C(\mathbf{S})$, where $\mathbf{S}$ a matrix of maximum rank such that $C(\mathbf{V}_1\mathbf{S}) \subset C(\mathbf{X})$ and $C(\mathbf{V}_2\mathbf{S}) \subset C(\mathbf{X})$

$\Leftrightarrow$ $\mathbf{X}'\mathbf{X}^{-'}\ell \in C(\mathbf{X}'\mathbf{S})$

$\Leftrightarrow$ $\mathbf{X}'\mathbf{X}^{-'}\ell \in R(\mathbf{S}'\mathbf{X})$

$\Leftrightarrow$ $\ell \in R(\mathbf{S}'\mathbf{X})$, since $\ell'\beta$ is an estimable linear function of β, thus $\ell \in R(\mathbf{X})$, which in turn implies $\mathbf{X}'\mathbf{X}^{-'}\ell = \ell$.

□

To compute a matrix $\mathbf{S}$ such that $C(\mathbf{V}_i\mathbf{S}) \subset C(\mathbf{X})$, $i = 1, \ldots, k$, thus enabling us to check the condition (1.46) for a given linear function $\ell'\beta$ of β, or to determine the space of estimable linear functions $\ell'\beta$ of β whose BLUE $\ell'\hat{\beta}$ is independent of $(\sigma_1^2, \ldots, \sigma_k^2)$ in the $LM(\mathbf{y}, \mathbf{X}\beta, \Sigma\sigma_i^2\mathbf{V}_i)$, we solve for $\mathbf{S}$ in

(1.51)
$$\begin{bmatrix} \mathbf{Z}'\mathbf{V}_1 \\ \vdots \\ \mathbf{Z}'\mathbf{V}_k \end{bmatrix} \mathbf{S} = \begin{bmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}$$

where $\mathbf{Z}$ is as above any matrix of maximum rank such that $\mathbf{X}'\mathbf{Z} = \mathbf{0}$.

Example 1.8

Consider the linear model

(1.52)
$$\begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_k \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_k \end{bmatrix} \beta + \mathbf{e}\ ; \quad cov(\mathbf{e}) = \begin{bmatrix} \sigma_1^2\mathbf{I}_1 & & \\ & \ddots & \\ & & \sigma_k^2\mathbf{I}_k \end{bmatrix}$$

conformably partitioned, where $\mathbf{I}_1, \ldots, \mathbf{I}_k$ are identity matrices of dimension $n_1, \ldots, n_k$.

As a corollary to Theorem 1.7 it can be shown that the BLUE $\mathbf{X}\hat{\beta}$ for $\mathbf{X}\beta$ is independent of the variance components $(\sigma_1^2, \ldots, \sigma_k^2)$ if and only if

(1.53)
$$R(\mathbf{X}_i) \cap \sum_{\substack{j=1 \\ j \neq i}}^{k} R(\mathbf{X}_j) = \{\mathbf{0}\}, \; i = 1, \ldots, k$$

i.e. if and only if the row spaces of $\mathbf{X}_1, \ldots, \mathbf{X}_k$ are disconnected (the same holds if we take arbitrary but nonsingular $\mathbf{V}_1, \ldots, \mathbf{V}_k$ instead of $\mathbf{I}_1, \ldots, \mathbf{I}_k$). Clearly, in such a case the ordinary least squares estimator (OLSE) for $\mathbf{X}\beta$ is BLUE for any set of variance components $(\sigma_1^2, \ldots, \sigma_k^2) \neq 0$. Thus, if groups of observations correspond to disconnected row spaces in the design matrix, the OLSE is robust against the violation of homoscedasticity between those groups thus retaining its minimum variance property in the class of linear unbiased estimators.

We note that it is easily verified that under (1.53) the natural estimators

(1.54)
$$\hat{\sigma}_i^2 = \mathbf{y}_i' (\mathbf{I}_i - \mathbf{X}_i(\mathbf{X}_i'\mathbf{X}_i)^- \mathbf{X}_i')\mathbf{y}_i/(n_i - p_i)$$

for σ_i^2 $(p_i = rank(\mathbf{X}_i))$, $i = 1, \ldots, k$ are MINQUE.

□

Kendall and Stuart (1973) state the well-known result that optimal designs in polynomial and trigonometric regression are characterized by the following property: if k parameters are to be estimated, i.e. we fit a polynomial of order $k-1$ to the data, and $N = n \cdot k$ observations can be taken, it is optimal that the design matrix consists of precisely k different linearly independent predictors $\mathbf{x}_1, \ldots, \mathbf{x}_k$, which appear n times each in the design. If

(1.55)
$$\mathbf{X}_i = \left.\begin{bmatrix} \mathbf{x}_i \\ \vdots \\ \mathbf{x}_i \end{bmatrix}\right\} \; n \text{ times}$$

then the design matrix $\mathbf{X}$ for the polynomial regression can be written as

(1.56)
$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_k \end{bmatrix}$$

where the matrices $\mathbf{X}_1, \ldots, \mathbf{X}_k$ satisfy the condition (1.53). Thus optimal designs in polynomial and trigonometric regression are robust against heteroscedasticity between groups of observations corresponding to different predictors, that is, the OLSE $\hat{\beta}$ for the parameter vector β remains BLUE even under heteroscedasticity. This is not to say that the design remains optimal under heteroscedasticity. However, it is reasonable in a practical situation, doing polynomial or trigonometric regression, to assume homoscedasticity within groups of observations corresponding to a given predictor, and to allow for heteroscedasticity between groups. Then, in the absence of any prior information on the variances $\sigma_1^2, \ldots, \sigma_k^2$, a design of the form (1.55), (1.56) is optimal.

CHAPTER 2

The Linear Model

2.1 INTRODUCTION

Through most of this thesis, with the exception of the later chapters where a multivariate structure is considered, we are concerned with the well-known linear regression model

$$\underset{(n\times 1)}{\mathbf{y}} = \underset{(n\times p)}{\mathbf{X}} \cdot \underset{(p\times 1)}{\beta} + \underset{(n\times 1)}{\mathbf{e}} \quad ; \quad cov(\mathbf{e}) = \sigma^2 \underset{(n\times n)}{\mathbf{V}} \tag{2.1}$$

where $\mathbf{y}$ is a vector of n observations, $\mathbf{X}$ is the known design matrix, β is an unknown vector of regression parameters and $\mathbf{e}$ is an unobservable random variate.

It is assumed in (2.1) that

$$E(\mathbf{e}) = \mathbf{0}\,, \quad \text{and} \tag{2.2}$$

$$cov(\mathbf{e}) = E(\mathbf{e}\mathbf{e}') = \sigma^2\mathbf{V} \tag{2.3}$$

Equations (2.2) and (2.3) imply that

$$E(\mathbf{y}) = \mathbf{X}\beta \quad \text{and} \tag{2.4}$$

$$cov(\mathbf{y}) = cov(\mathbf{e}) = \sigma^2\mathbf{V}. \tag{2.5}$$

The matrix $\mathbf{V}$ gives the variance-covariance structure of $\mathbf{e}$ and of $\mathbf{y}$, which is known up to an unknown scale factor σ^2. Being a covariance matrix, $\mathbf{V}$ is symmetric and at least nonnegative-definite. The rank of $\mathbf{V}$ is possibly smaller than n, that is, we allow for singular $\mathbf{V}$.

The design matrix $\mathbf{X}$ is by assumption non-stochastic. To avoid an overspecification of the model we take $p \leq n$, which results in no loss of generality. However, we allow for

$$r = rank(\mathbf{X}) < p \leq n, \tag{2.6}$$

i.e. $\mathbf{X}$ is possibly not of full rank.

When the dimensions of the vectors and matrices involved in (2.1) are clear in the context, we simply write

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}; \quad cov(\mathbf{e}) = \sigma^2\mathbf{V}, \tag{2.1a}$$

or equivalently we denote (2.1) by $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$.

The model (2.1) reflects essentially the assumption that the observed variate values $\mathbf{y}$ can be explained or modelled by a systematic part represented by $\mathbf{X}\beta$ and a purely random part represented by $\mathbf{e}$. The columns of $\mathbf{X}$ are commonly called (explanatory) variables, and the form of the model implies that the relationship between $\mathbf{y}$ and the variables is linear, apart from the noise $\mathbf{e}$.

Equations (2.1) through (2.3) describe the linear model, and in the special case of $\mathbf{y}$ (and consequently $\mathbf{e}$) being normally distributed, they uniquely define the respective distributions of the variates $\mathbf{y}$ and $\mathbf{e}$ in the model, since a normal distribution is uniquely defined by specifying its first and second moments. Thus, in the normal case, we have $\mathbf{y} \sim N_n(\mathbf{X}\beta, \sigma^2\mathbf{V})$ and $\mathbf{e} \sim N_n(\mathbf{0}, \sigma^2\mathbf{V})$.

Allowing $\mathbf{V}$ to be any symmetric nonnegative-definite matrix, possibly not of full rank, is the most general way to set up a linear model of the form (2.1). The "linear model under arbitrary known variance-covariance structure" was considered by such authors as Goldman and Zelen (1964), Zyskind and Martin (1969) and Rao (1971, 1972), who examined the problem of best linear unbiased estimation of the parameters in the model, and the related problem of testing a linear hypothesis about the parameter vector β. Their results have been surveyed, and relationships between their approaches described, in Dunne (1982). In the following sections on best linear unbiased estimation and tests of hypotheses in a linear model we will provide the necessary insights and methods for a later chapter on outliers and influential observations in the linear model, but we will not elaborate on the wide body of the theory.

A model slightly less general than (2.1) with $\mathbf{V}$ being nonsingular and thus positive-definite and symmetric (pds) was first considered by Aitken (1935). We write

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}; \quad cov(\mathbf{e}) = \sigma^2\mathbf{V}, \quad \mathbf{V} \text{ pds.} \tag{2.7}$$

The mathematical treatment of this model is considerably simplified by the fact that it can be transformed by a nonsingular (linear) transformation

$$\mathbf{T} = \mathbf{V}^{-1/2} \quad \text{(say)} \tag{2.8}$$

to obtain the model

$$\mathbf{Ty} = \mathbf{TX}\beta + \mathbf{Te}; \quad cov(\mathbf{Te}) = \sigma^2\mathbf{TVT}' \tag{2.9}$$

$$\Leftrightarrow \quad \tilde{\mathbf{y}} = \tilde{\mathbf{X}}\beta + \tilde{\mathbf{e}}; \quad cov(\tilde{\mathbf{e}}) = \sigma^2\mathbf{I}_n\,. \tag{2.10}$$

The model (2.10) is the classical Gauß-model. Gauß (1809) laid the foundations of the *Theory of Least Squares,* which was independently reinvented by Markoff (1900) and was subsequently generalized by Aitken (1935) for the model (2.7), and by Zyskind and Martin (1969) and Rao (1971) for the general model (2.1). In some texts the Gauß-model is described as the Gauß-Markoff-model.

In model (2.10), which is a special case of (2.7) and (2.1), the error terms are uncorrelated and have identical variance σ^2. In the normal case they are independent and identically distributed (iid) as $N(0, \sigma^2)$, since normally distributed variates are independent if they are uncorrelated.

The Gauß-model, with and without the normality assumption, is certainly a widely used statistical model and is of great practical importance. This is much more than can be said about a model like (2.7) or even (2.1), when $\mathbf{V}$ is singular. Whereas every statistician has performed an analysis of data under a Gauß-model, especially since it includes all analysis of variance (ANOVA) and analysis of covariance (ANACOVA) problems, it is a science in itself to find a practical and nontrivial example for a linear model with singular variance-covariance structure.

In the following, however, we will present all results, where possible, in terms of the general model (2.1). Specifically, we allow $\mathbf{V}$ and $\mathbf{X}$ to be possibly not of full rank. It is firstly of mathematical interest to treat the theory in this way, to develop it in its most general form and to obtain practical results, if there are any, as special case of the general results. Secondly, the theory of the linear model under arbitrary known variance-covariance structure will provide a theoretical framework for examining the cases where $\mathbf{V}$ is either known only up to an additive structure as in the variance components model, or where $\mathbf{V}$ is completely unknown and has to be estimated from a sample of observations $\mathbf{y}$ as in the multivariate models.

2.1.1 Decomposition of the sure equations

When the variance-covariance structure $\mathbf{V}$ of a linear model is singular, there will, in general, exist sure equations in the model which restrict the space of observations $\mathbf{y}$ and the space of parameters β.

Lemma 2.1

With respect to the $LM(\mathbf{y},\mathbf{X}\beta, \sigma^2\mathbf{V})$, let $\mathbf{N}$ be a matrix of maximum and full rank such that

(2.11) $$\mathbf{N}'\mathbf{V} = \mathbf{0} \quad .$$

Then

(2.12) $$\mathbf{y} \in C([\mathbf{X}:\mathbf{V}]) \text{ , with probability 1 (w.p.1)}$$

(2.13) $$\mathbf{N}'\mathbf{y} =: \mathbf{d} = \mathbf{N}'\mathbf{X}\beta \text{ , w.p.1.}$$

Proof: (Rao, 1973)
Without loss of generality let $\mathbf{N}$ be partitioned as $\mathbf{N} = [\mathbf{N}_1 : \mathbf{N}_2]$, where $\mathbf{N}_2$ is a matrix of maximum rank such that

(2.14) $$\mathbf{N}_2' [\mathbf{X}:\mathbf{V}] = [\mathbf{0}:\mathbf{0}] .$$

Then we have that

(2.15) $$E(\mathbf{N}_2'\mathbf{y}) = \mathbf{N}_2'\mathbf{X}\beta = \mathbf{0} , \quad \text{and}$$

(2.16) $$E(\mathbf{N}_2'\mathbf{y}\mathbf{y}'\mathbf{N}_2) = cov(\mathbf{N}_2'\mathbf{V}\mathbf{N}_2) = \mathbf{0} .$$

$$\Rightarrow \quad \mathbf{N}_2'\mathbf{y} = \mathbf{0} , \quad \text{w.p.1}$$

$$\Rightarrow \quad \mathbf{y} \in C([\mathbf{X}:\mathbf{V}]) , \quad \text{w.p.1} \quad \text{from (2.14)} .$$

From the definition of the model we have for any observation $\mathbf{y}$

(2.17) $$(\mathbf{y}-\mathbf{X}\beta) = \mathbf{e} \in C(\mathbf{V})$$

and for any two observations $\mathbf{y}_1$, $\mathbf{y}_2$

(2.18) $$(\mathbf{y}_1-\mathbf{y}_2) = (\mathbf{e}_1-\mathbf{e}_2) \in C(\mathbf{V}) .$$

$$\Rightarrow \quad \mathbf{N}'\mathbf{y}_1-\mathbf{N}'\mathbf{y}_2 = \mathbf{N}'(\mathbf{e}_1-\mathbf{e}_2) = \mathbf{0} , \quad \text{w.p.1} \quad \text{for all } \mathbf{y}_1, \mathbf{y}_2$$

$$\Rightarrow \quad \mathbf{N}'\mathbf{y} = const =: \mathbf{d} = \mathbf{N}'\mathbf{X}\beta , \quad \text{w.p.1}$$

□

Thus all observations $\mathbf{y}$ in a linear model come from the space spanned by the columns of $\mathbf{X}$ and $\mathbf{V}$. A linear model is called consistent, when $\mathbf{y} \in C([\mathbf{X}:\mathbf{V}])$ is satisfied (Rao, 1973), which is equivalent with the consistency of the sure equations $\mathbf{d} = \mathbf{N}'\mathbf{X}\beta$. Of course, $\mathbf{d}$ is in general fixed only for fixed $\mathbf{N}$, and any nonsingular transformation $\mathbf{T}$ yields an equivalent set of sure equations $\mathbf{T}\mathbf{d} = \mathbf{T}\mathbf{N}'\mathbf{X}\beta$, w.p.1. If $\mathbf{V}$ is singular, we call the linear model (2.1) singular, and if $C([\mathbf{X}:\mathbf{V}]) \neq \mathbb{R}^n$, we say the linear model is not of full rank. A linear model which is not of full rank is necessarily singular.

The following Lemma gives a decomposition of the sure equations of a singular linear model and presents them in a canonical form.

Lemma 2.2 (Schall)

In the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$,

(i) the maximum number of linearly independent linear contrasts of $\mathbf{y}$ with zero variance is $n - rank(\mathbf{V})$, yielding the $n - rank(\mathbf{V})$ sure equations

(2.19) $$\mathbf{d} = \mathbf{N}'\mathbf{X}\beta \text{ (say)} , \quad \text{w.p.1}$$

where $\mathbf{N}$ as in (2.11).

(ii) Those $n - rank(\mathbf{V})$ equations can be decomposed into $s = rank(\mathbf{N}_1) = rank([\mathbf{X}:\mathbf{V}]) - rank(\mathbf{V})$ linearly independent sure equations

(2.20) $$\mathbf{d}_1 = \mathbf{N}_1'\mathbf{X}\beta \text{ (say)} , \quad \text{w.p.1}$$

and $t = rank(\mathbf{N}_2) = n - rank([\mathbf{X}:\mathbf{V}])$ trivial or redundant sure equations

(2.21) $$\mathbf{0} = \mathbf{0}\beta , \quad \text{w.p.1}$$

where $\mathbf{d}' = [\mathbf{d}_1' : \mathbf{d}_2']$ is conformably partitioned with $\mathbf{N} = [\mathbf{N}_1 : \mathbf{N}_2]$ as in (2.14). When t is zero, the linear model is of full rank.

(iii) The s linearly independent sure equations can be decomposed into at least $(s-1)$ sure equations with a zero left hand side

(2.22) $$\mathbf{0} = \mathbf{S}\mathbf{N}_1'\mathbf{X}\beta \text{ (say)}, \quad \text{w.p.1}$$

and one sure equation with possibly a nonzero left hand side

(2.23) $$\mathbf{d}_1'\mathbf{d}_1 = \mathbf{d}_1'\mathbf{N}_1\mathbf{X}\beta \text{ (say)}, \quad \text{w.p.1}\ .$$

Proof:
(i) This follows directly from Lemma 2.1.

(ii)
$$\mathbf{d} = \mathbf{N}'\mathbf{X}\beta$$
$$\Leftrightarrow \begin{bmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{N}_1' \\ \mathbf{N}_2' \end{bmatrix} \mathbf{X}\beta$$
$$\Leftrightarrow \begin{bmatrix} \mathbf{d}_1 \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{N}_1'\mathbf{X}\beta \\ \mathbf{0}\beta \end{bmatrix}, \quad \text{from (2.14)}\ .$$

(iii) Let $\mathbf{S}$ be a matrix of maximum rank such that $\mathbf{S}\mathbf{d} = \mathbf{0}$. Then $rank(\mathbf{S}) = s-1$ if $\mathbf{d} \neq \mathbf{0}$, otherwise $rank(\mathbf{S}) = s$. In either case, premultiplying (2.20) by $\mathbf{S}$ yields (2.22), and if $\mathbf{d}_1 \neq \mathbf{0}$ we obtain (2.23) by premultiplying (2.20) by $\mathbf{d}_1'$.

□

We note that equation (2.22) restricts the space for the parameter vector β, and consequently restricts further the space of admissible observations.

Corollary 2.2.1 (Rao, 1973)

With the notation as in Lemma 2.2 we have

(2.24) $$\beta \in R(\mathbf{S}\mathbf{N}_1'\mathbf{X})^{\perp}, \quad \text{w.p.1 and thus}$$

(2.25) $$\mathbf{y} \in C([\mathbf{X}\mathbf{B} : \mathbf{V}]), \quad \text{w.p.1}$$

where the columns of $\mathbf{B}$ span the space $R(\mathbf{S}\mathbf{N}_1'\mathbf{X})^{\perp}$.

□

Of course $\mathbf{B}$ is only known when $\mathbf{d}_1$ is known, and that is the case only when some observation vector $\mathbf{y}$ is available as data. The nontrivial sure equations (2.20) are known only after some observation has been made, and the restriction $\mathbf{y} \in C([\mathbf{X}\mathbf{B} : \mathbf{V}])$ is an *a-posteriori* one. Before making an observation, *a-priori*, we only know $\mathbf{y} \in C([\mathbf{X} : \mathbf{V}])$, w.p.1.

In fact, once an observation vector $\mathbf{y}$ is known, the space given in (2.25) can be expressed in an alternative form.

Lemma 2.3 (Schall)

Let $\mathbf{M}$ be given as

(2.26) $$\mathbf{M} = \mathbf{N}(\mathbf{N}'\mathbf{N})^{-1}\mathbf{N}'$$

and $\mathbf{z} := \mathbf{My}$. Then

(2.27) $$\mathbf{y} \in C([\mathbf{z}:\mathbf{V}]) \ , \quad \text{w.p.1}$$

Proof:

$$\mathbf{My} = \mathbf{N}(\mathbf{N}'\mathbf{N})^{-1}\mathbf{d} = \mathbf{z} \ , \quad \text{w.p.1}$$

$$\Rightarrow \ \mathbf{M}(\mathbf{y}-\mathbf{z}) = \mathbf{0} \ , \quad \text{w.p.1 since } \mathbf{M} \text{ is idempotent}$$

$$\Rightarrow \ \mathbf{y}-\mathbf{z} \in C(\mathbf{V}) \ , \quad \text{w.p.1 from (2.11)}$$

$$\Rightarrow \ \mathbf{y} \in C([\mathbf{z}:\mathbf{V}]) \ , \quad \text{w.p.1} \ .$$

□

Corollary 2.3.1

The space of admissible observations, *a-posteriori*, is the affine space $\{\mathbf{z} + C(\mathbf{V})\}$.

□

2.1.2 Decomposition of the space R^n.

In the previous section we have given a decomposition of the sure equations in a linear model into redundancies (2.21), sure equations with zero left hand side (2.22) and possibly one sure equation with a nonzero left hand side (2.23). The remainder, of course, are the stochastic equations in the model. That development essentially constitutes a decomposition of the row space $R([\mathbf{y}:\mathbf{X}:\mathbf{V}])$. In this section we will give a corresponding decomposition of the space of admissible observations $C([\mathbf{X}:\mathbf{V}])$ which for admissible $\mathbf{y}$ coincides with the column space $C([\mathbf{y}:\mathbf{X}:\mathbf{V}])$.

To begin with we require

Lemma 2.4 (Rao, 1974)

Let $\mathbf{Z}$ be a matrix of maximum rank such that $\mathbf{Z}'\mathbf{X} = \mathbf{0}$. Then

(2.28) $$\text{(i)} \quad C(\mathbf{X}) \cap C(\mathbf{VZ}) = \{\mathbf{0}\}$$

$$\text{(ii)} \quad C([\mathbf{X}:\mathbf{VZ}]) = C([\mathbf{X}:\mathbf{V}]) \ .$$

Proof:
(i) Assume the contrary, let be $\mathbf{VZ}\lambda \in C(\mathbf{X})$ for some λ, then $\mathbf{Z}'\mathbf{VZ}\lambda = \mathbf{0}$ implies $\mathbf{VZ}\lambda = \mathbf{0}$ which proves (i).

(ii) We have only to show that $C([\mathbf{X}:\mathbf{V}]) \subset C([\mathbf{X}:\mathbf{VZ}])$ since $C([\mathbf{X}:\mathbf{VZ}]) \subset C([\mathbf{X}:\mathbf{V}])$ is trivially satisfied.
Assume the contrary, let $\mathbf{x}$ be a vector such that $\mathbf{x} \in C[\mathbf{X}:\mathbf{V}]$ but $\mathbf{x} \notin C([\mathbf{X}:\mathbf{VZ}])$. Without loss of generality take $\mathbf{x} \in C([\mathbf{X}:\mathbf{VZ}])^{\perp}$. Then $\mathbf{x}$ can be written as $\mathbf{x} = \mathbf{Z}\lambda$ for some λ. By assumption we have $\lambda'\mathbf{Z}'\mathbf{VZ} = \mathbf{0}$ which implies $\lambda'\mathbf{Z}'\mathbf{V} = \mathbf{0}$. Thus $\mathbf{x} = \mathbf{Z}\lambda \in C([\mathbf{X}:\mathbf{V}])^{\perp}$ which is a contradiction to $\mathbf{x} \in C([\mathbf{X}:\mathbf{V}])$ unless $\mathbf{x} = \mathbf{0}$.

□

The lemma states that $C([\mathbf{X}:\mathbf{V}])$ can be decomposed as

$$C([\mathbf{X}:\mathbf{VZ}]) = C(\mathbf{X}) \oplus C(\mathbf{VZ}) \tag{2.29}$$

where $\oplus$ denotes the direct sum of vector spaces.

Calling a matrix $\mathbf{B}$ a base of a vector space S if $\mathbf{B}$ is a matrix of basis vectors of S, we proceed by letting $\mathbf{X}_2$ be a base of $C(\mathbf{X}) \cap C(\mathbf{V})$ and $\mathbf{X}_1$ be an extension of $\mathbf{X}_2$ to a base of $C(\mathbf{X})$. Then

Corollary 2.4.1

The space $C([\mathbf{X}:\mathbf{V}])$ can be decomposed as

$$C([\mathbf{X}:\mathbf{V}]) = C(\mathbf{X}_1) \oplus C(\mathbf{X}_2) \oplus C(\mathbf{VZ}) \tag{2.30}$$

where $C(\mathbf{X}) = C(\mathbf{X}_1) \oplus C(\mathbf{X}_2)$, with $C(\mathbf{X}_2) \subset C(\mathbf{V})$, $C(\mathbf{X}_1) \cap C(\mathbf{V}) = \{\mathbf{0}\}$.

□

As a consequence of (2.30), any admissible observation $\mathbf{y}$ in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ can be decomposed as

$$\mathbf{y} = \mathbf{y}_1 + \mathbf{y}_2 + \mathbf{y}_3 \text{ , where} \tag{2.31}$$

$$\mathbf{y}_1 \in C(\mathbf{X}_1) \text{ , } \mathbf{y}_2 \in C(\mathbf{X}_2) \text{ and } \mathbf{y}_3 \in C(\mathbf{VZ}) \text{ .}$$

For any two admissible observations $\mathbf{y}$ and $\mathbf{z}$ and their corresponding decompositions similar to (2.31) we have $\mathbf{y}_1 = \mathbf{z}_1$, w.p.1, which follows directly from $(\mathbf{y}-\mathbf{z}) \in C(\mathbf{V})$.

We may rephrase the import of Lemma 2.3 and its corollary as

Lemma 2.5 (Schall)

Let $\mathbf{z}$ be an admissible observation under the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$, and $\mathbf{z} = \mathbf{z}_1 + \mathbf{z}_2 + \mathbf{z}_3$ a decomposition of $\mathbf{z}$ similar to (2.31). Then

$$\mathbf{y} \in C([\mathbf{z}_1:\mathbf{V}]) \text{ , w.p.1}$$

for any admissible observation $\mathbf{y}$.

□

Corollary 2.5.1

The space of admissible observations, *a-posteriori*, is the affine space $\{\mathbf{z}_1 + C(\mathbf{V})\}$.

□

Finally, we can exhibit the correspondences between the decomposition of the equations in a linear model and the decomposition of $C([\mathbf{X}:\mathbf{V}]) \subset \mathbb{R}^n$. Denoting by $\mathbf{N}_2$ a base of $C([\mathbf{X}:\mathbf{V}])^{\perp}$ and by $\mathbf{X}_S$ an extension of $\{\mathbf{z}_1\}$ to a base of $C(\mathbf{X}_1)$, we can say:

(i) $C(\mathbf{N}_2)$ corresponds to t redundancies

(ii) $C(\mathbf{X}_S)$ corresponds to $rank(\mathbf{X}_S) \in \{s-1, s\}$ sure equations with zero left hand side

(iii) $\{\mathbf{z}_1\}$ corresponds to $a \in \{0, 1\}$ sure equation with nonzero left hand side.

(iv) $C(\mathbf{V}) = C(\mathbf{X}_2) \oplus C(\mathbf{VZ})$ corresponds to the $rank(\mathbf{V})$ stochastic equations in the model.

2.1.3 Reduction of a linear model

In the previous sections we have considered the case of a linear model with singular variance-covariance structure, that is $C(\mathbf{V}) \neq \mathbb{R}^n$. Then a matrix $\mathbf{N}$ exists such that

$$\mathbf{N}'\mathbf{V} = \mathbf{0} . \tag{2.11}$$

Now we investigate a linear model where even

$$C([\mathbf{X}:\mathbf{V}]) \neq \mathbb{R}^n , \tag{2.32}$$

i.e. where the space of admissible observations is not the whole space $\mathbb{R}^n$.

With the notation as in Lemma 2.1, there exists a nontrivial submatrix $\mathbf{N}_2$ of $\mathbf{N} = [\mathbf{N}_1:\mathbf{N}_2]$, possibly after a rearrangement of the columns of $\mathbf{N}$, such that

$$\mathbf{N}_2'[\mathbf{X}:\mathbf{V}] = [\mathbf{0}:\mathbf{0}] . \tag{2.14}$$

We will show a well-known fact that such a model can always be reduced, by dropping $t = rank(\mathbf{N}_2)$ components of $\mathbf{y}$ and corresponding rows of $\mathbf{X}$ and rows and columns of $\mathbf{V}$, to a $LM(\mathbf{y}_1, \mathbf{X}_1\beta, \mathbf{V}_{11})$ such that

$$C([\mathbf{X}_1:\mathbf{V}_{11}]) = \mathbb{R}^{n-t} , \tag{2.33}$$

where the reduced model is equivalent to the original one. This amounts to discarding the redundancies.

Lemma 2.6

A linear model $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$, which is not of full rank, that is, $rank([\mathbf{X}:\mathbf{V}]) = n-t < n$, can be reduced to a $LM(\mathbf{y}_1, \mathbf{X}_1\beta, \sigma^2\mathbf{V}_{11})$ by dropping t model equations from the original model, so that the reduced model is stochastically equivalent to the original one.

Proof: Let $\mathbf{N}_2$, as given in (2.14), be partitioned as

$$\mathbf{N}_2' = [\mathbf{N}_{21}' : \mathbf{N}_{22}'] \tag{2.34}$$

such that $\mathbf{N}_{22}$ is a square matrix. Without loss of generality we take $\mathbf{N}_{22}$ to be nonsingular, since this can always be achieved by a rearrangement of the rows of $\mathbf{N}_2$. (If the rows of $\mathbf{N}_2$ were rearranged, we make a corresponding rearrangement of the model equations in the original model $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$).

Partitioning the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ conformably with $\mathbf{N}_2' = [\mathbf{N}_{21}' : \mathbf{N}_{22}']$ we obtain

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \beta + \mathbf{e}, \quad cov(\mathbf{e}) = \sigma^2 \begin{bmatrix} \mathbf{V}_{11} : \mathbf{V}_{12} \\ \mathbf{V}_{21} : \mathbf{V}_{22} \end{bmatrix} \tag{2.35}$$

Clearly, the columns of $\mathbf{N}_2$ span the space orthogonal to $C([\mathbf{X}:\mathbf{V}])$, and thus $C([\mathbf{N}_2:\mathbf{X}:\mathbf{V}]) = \mathbb{R}^n$.

From equation (2.14) we have

$$\mathbf{N}_2'[\mathbf{y}:\mathbf{X}:\mathbf{V}] = [\mathbf{0}:\mathbf{0}:\mathbf{0}] \tag{2.36}$$

$$\Leftrightarrow \quad [\mathbf{y}_2:\mathbf{X}_2:\mathbf{V}_{21}:\mathbf{V}_{22}] = -\mathbf{N}_{22}'^{-1}\mathbf{N}_{21}' [\mathbf{y}_1:\mathbf{X}_1:\mathbf{V}_{11}:\mathbf{V}_{12}]$$

which states precisely that the last t equations of the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ are linear combinations of the first $(n-t)$ equations.

The reduced model

$$\mathbf{y}_1 = \mathbf{X}_1\beta + \mathbf{e}_1 ; \quad cov(\mathbf{e}_1) = \sigma^2\mathbf{V}_{11} \tag{2.37}$$

is thus equivalent to the original one.

Further, the columns of $\mathbf{X}_1$ and $\mathbf{V}_{11}$ span the whole space $\mathbb{R}^{n-t}$, since $\ell'[\mathbf{X}_1:\mathbf{V}_{11}] = \mathbf{0}$ implies that $[\ell':\mathbf{0}] \in R(\mathbf{N}_2') = R[\mathbf{N}_{21}':\mathbf{N}_{22}']$, which is a contradiction to the nonsingularity of $\mathbf{N}_{22}$ unless $\ell = \mathbf{0}$. The reduced model is thus of full rank.

□

A linear model which is not of full rank may be construed as adding, in effect, a number of redundant sure equations $\mathbf{0} = \mathbf{0}\beta$, and performing linear combinations on the model observations at these annihilated equations, as can be seen from Lemma 2.2. A reduction of this model to a full rank model does therefore not change the statistical character or the statistical information contained in the model. Any meaningful statistical procedure should be invariant under a reduction of a model given by Lemma 2.6, and we will presently see that this is the case with linear estimation and tests of linear hypotheses.

Any analysis of a model which is not of full rank should be preceeded by a reduction to a full rank model. This could be done in a manner that is computationally stable, since we need not compute accurately the linear combination which would give a redundant model equation, but we would only have to ascertain that there is such a linear combination, and omit rows after rearrangement.

2.2 BEST LINEAR UNBIASED ESTIMATION

In this section we will treat the problem of best linear unbiased (BLU) estimation in the linear model. Beginning with the well-known Gauß-model $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{I})$, via the $LM(\mathbf{y},\mathbf{X}\beta, \sigma^2\mathbf{V})$ with $\mathbf{V}$ positive definite to the general case where $\mathbf{V}$ may be singular, our *Leitmotiv* will be to see the BLUE $\mathbf{X}\hat{\beta}$ for $\mathbf{X}\beta$ as the image of $\mathbf{y}$ under a projection onto the space $C(\mathbf{X})$.

This geometric interpretation of BLU-estimation will be useful in simplifying the notation as well as the algebra, and insights may be gained in a direct manner which otherwise could be difficult to obtain.

2.2.1 Estimation in the Gauß-model

Carl Friedrich Gauß (1777–1855), in the first part of his *Theoria Combinationis Observationum Erroribus Minimus Obnoxiae* (1821) essentially proposed, as a method to estimate the unknown parameter vector β in a linear model

(2.38) $$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}\,, \quad cov(\mathbf{e}) = \sigma^2\mathbf{I}$$

that a vector $\hat{\beta}$ should be taken such that

(2.39) $$(\mathbf{y} - \mathbf{X}\beta)'\,(\mathbf{y} - \mathbf{X}\beta)$$

is minimized at $\beta = \hat{\beta}$.

Of course, Gauß did not use the "modern" matrix notation as in (2.38) and (2.39), which was introduced much later by Aitken (1935).

To find the minimum of (2.39), or equivalently to solve the approximation problem (2.39), we take the derivative with respect to β, which, set equal to zero, yields the equation

(2.40) $$\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{y}\,,$$

whose solution is labelled $\hat{\beta}$. These are the well-known normal equations (NE's). Whereas the so-called model equations

(2.41) $$\mathbf{y} = \mathbf{X}\beta$$

are inconsistent in general, i.e. $\mathbf{y} \notin C(\mathbf{X})$, the NE's are always consistent since $C(\mathbf{X}'\mathbf{X}) = C(\mathbf{X}')$ and thus a solution to (2.40) does always exist. Such a solution $\hat{\beta}$ is called an ordinary least-squares (OLS) solution, and the method to estimate β by an OLS-solution $\hat{\beta}$ of the NE's is called the least-squares method.

If $\mathbf{X}$ has full rank p then $\mathbf{X}'\mathbf{X}$ is nonsingular and premultiplying (2.40) by $(\mathbf{X}'\mathbf{X})^{-1}$ yields

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\,\mathbf{y} \tag{2.42}$$

and $\hat{\beta}$ is obviously the unique solution of the NE's.

However, if $\mathbf{X}$ has rank $r < p$, then the solution of the NE's is not unique. It is easy to show that a general solution is given by

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-}\,\mathbf{X}'\mathbf{y} + (\mathbf{I} - (\mathbf{X}'\mathbf{X})^{-}\,\mathbf{X}'\mathbf{X})\,\lambda \tag{2.43}$$

where $\lambda \in \mathbb{R}^p$ is an arbitrary vector.

But while $\hat{\beta}$ is not unique, perhaps some linear functions $\ell'\hat{\beta}$ of $\hat{\beta}$ are unique over all choices $\hat{\beta}$ as given in (2.43)? In other words, even though β cannot be estimated uniquely by the method of least squares, perhaps a linear function $\ell'\beta$ of β can uniquely be estimated by $\ell'\hat{\beta}$, where $\hat{\beta}$ is any solution to the NE's. This gives rise to the following definition:

Definition 2.7 (Bose, 1944)

In the $LM(\mathbf{y},\mathbf{X}\beta, \sigma^2\mathbf{I})$ a linear function $\ell'\beta$ of β is called estimable if and only if $\ell'\hat{\beta}$ is unique over all solutions $\hat{\beta}$ to the normal equations $\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{y}$.

□

The question of which linear functions $\ell'\beta$ of β are estimable is readily answered by the

Theorem 2.8 (Bose, 1944)

In the $LM(\mathbf{y},\mathbf{X}\beta, \sigma^2\mathbf{I})$, a linear function $\ell'\beta$ of β is estimable if and only if $\ell \in R(\mathbf{X})$.

Proof: Let be $\ell \in R(\mathbf{X})$, which is equivalent with $\ell \in R(\mathbf{X}'\mathbf{X})$, which in turn is equivalent with $\ell = \mathbf{X}'\mathbf{X}\lambda$ for some λ. Then, using (2.43), $\ell'\hat{\beta} = \lambda'\mathbf{X}'\mathbf{X}\hat{\beta} = \lambda'\mathbf{X}'\mathbf{y}$ and the sufficiency of $\ell \in R(\mathbf{X})$ is established.

Now let $\ell'\hat{\beta}$ be unique. Then $\ell'(\mathbf{I} - (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X})\lambda = 0$, for all λ and $\mathbf{X}'\mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-}\ell = \ell$, so that $\ell \in C(\mathbf{X}')$ and $\ell \in R(\mathbf{X})$.

□

Clearly, $\ell'\hat{\beta} = \ell'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}$ for an estimable linear function $\ell'\hat{\beta}$ of β is a linear estimator and, taking expectation,

$$E(\ell'\hat{\beta}) = \ell'\,(\mathbf{X}'\mathbf{X})^{-}\,\mathbf{X}'\mathbf{X}\beta = \ell'\beta \tag{2.44}$$

we observe that $\ell'\hat{\beta}$ is unbiased for $\ell'\beta$, that is, the expectation of the estimator is the quantity to be estimated. Estimable linear functions of β in a linear model can also be characterized using the concept of unbiasedness, thus vindicating the definition above:

Theorem 2.9

In the $LM(\mathbf{y},\mathbf{X}\beta, \sigma^2\mathbf{I})$, a linear function $\ell'\beta$ of β is estimable if and only if there exists a linear unbiased estimator $\mathbf{m}'\mathbf{y}$ for $\ell'\beta$.

Proof:

$$\begin{aligned} & \mathbf{m}'\mathbf{y} \text{ unbiased for } \ell'\beta \\ \Leftrightarrow\ & \mathbf{m}'\mathbf{X}\beta = \ell'\beta \text{ , for all } \beta \\ \Leftrightarrow\ & \mathbf{X}'\mathbf{m} = \ell \\ \Leftrightarrow\ & \ell \in R(\mathbf{X}) \text{ .} \end{aligned}$$

□

We can also determine the variance of $\ell'\hat{\beta}$ when $\ell'\beta$ is an estimable linear function of β as

$$var\ (\ell'\hat{\beta}) = \ell'\ (\mathbf{X}'\mathbf{X})^{-}\ \ell \tag{2.45}$$

which is unique over all choices of a g-inverse $(\mathbf{X}'\mathbf{X})^{-}$ of $(\mathbf{X}'\mathbf{X})$. A good estimator should have small variance, and for linear unbiased estimators (LUE's) we have a best estimator in this class in view of

Theorem 2.10 (Gauß, 1821/1823)

In the $LM(\mathbf{y},\mathbf{X}\beta, \sigma^2\mathbf{I})$, $\ell'\hat{\beta}$ is the minimum variance estimator for an estimable linear function $\ell'\beta$ of β, in the class of linear unbiased estimators for $\ell'\beta$, where $\hat{\beta}$ is a solution to the NE's.

Proof: Since $\ell'\beta$ is estimable, ℓ can be written as $\ell = \mathbf{X}'\mathbf{X}\,\lambda$ for some λ. Let $\mathbf{m}'\mathbf{y}$ be any unbiased estimator for $\ell'\beta$, i.e. $\mathbf{m}'\mathbf{X} = \ell'$.

The variance of $\mathbf{m}'\mathbf{y}$ is

$$\begin{aligned} var\ (\mathbf{m}'\mathbf{y}) &= var(\mathbf{m}'\mathbf{y} - \lambda'\mathbf{X}'\mathbf{y} + \lambda'\mathbf{X}'\mathbf{y}) \\ &= var(\mathbf{m}'\mathbf{y} - \lambda'\mathbf{X}'\mathbf{y}) + var\ (\lambda'\mathbf{X}'\mathbf{y}) \\ &= var(\mathbf{m}'\mathbf{y} - \lambda'\mathbf{X}'\mathbf{y}) + var\ (\lambda'\hat{\beta}) \text{ .} \end{aligned}$$

This equation holds, and proves $var(\mathbf{m}'\mathbf{y}) \geq var(\ell'\hat{\beta})$, since

$$\begin{aligned} cov(\mathbf{m}'\mathbf{y} - \lambda'\mathbf{X}'\mathbf{y}, \lambda'\mathbf{X}'\mathbf{y}) &= \sigma^2\ (\mathbf{m}' - \lambda'\mathbf{X}')\ \mathbf{X}\lambda \\ &= \sigma^2\ (\mathbf{m}'\mathbf{X} - \lambda'\mathbf{X}'\mathbf{X})\ \lambda \\ &= \sigma^2\ (\ell' - \ell')\ \lambda \\ &= 0 \text{ .} \end{aligned}$$

□

This theorem, often called the Gauß-Markoff-theorem, states that in the $LM(\mathbf{y},\mathbf{X}\beta, \sigma^2\mathbf{I})$ the OLS-estimator $\ell'\hat{\beta}$ for an estimable linear function $\ell'\beta$ of β is the best linear unbiased estimator (BLUE) for $\ell'\beta$. This result holds independently of any assumption about the distribution of $\mathbf{y}$, but Gauß also showed that the BLUE $\ell'\hat{\beta}$ for an estimable linear function $\ell'\beta$ of β is the maximum-likelihood estimate (MLE) for $\ell'\beta$ if $\mathbf{y}$ follows a normal distribution.

Whether or not all linear functions $\ell'\beta$ of β are estimable, or equivalently $\mathbf{X}$ is of full rank, the vector of means $\mathbf{X}\beta$ is always estimable, and the so-called fitted value vector

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y} = \mathbf{M}\mathbf{y} \tag{2.46}$$

is its estimate. That is, the BLUE $\mathbf{X}\hat{\beta}$ of $\mathbf{X}\beta$ always exists. The matrix $\mathbf{M}$ is idempotent, and clearly

$$\mathbf{M}\mathbf{X} = \mathbf{X} \tag{2.47}$$

Thus $\mathbf{M}$ is the orthogonal projection operator onto the space $C(\mathbf{X})$ along $C(\mathbf{X})^{\perp}$, and the fitted values $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ are therefore the orthogonal projection of $\mathbf{y}$ onto the space $C(\mathbf{X})$. This gives a well-known geometric interpretation of the method of least-squares as presented e.g. by Rao (1973).

If $\mathbf{Z}$ is a matrix of maximum rank such that $\mathbf{Z}'\mathbf{X} = \mathbf{0}$, then $C(\mathbf{Z}) = C(\mathbf{X})^{\perp}$, and we denote the orthogonal projection operator onto $C(\mathbf{X})$ along $C(\mathbf{Z})$ by $P_{\mathbf{X}|\mathbf{Z}}$. The BLUE $\mathbf{X}\hat{\beta}$ for $\mathbf{X}\beta$, and thus the BLUE $\ell'\hat{\beta}$ for an estimable linear function $\ell'\beta$ of β can now be expressed in terms of the projection operator $P_{\mathbf{X}|\mathbf{Z}}$

Theorem 2.11 (Rao, 1974)

(i) In the $LM(\mathbf{y},\mathbf{X}\beta, \sigma^2\mathbf{I})$, the BLUE $\mathbf{X}\hat{\beta}$ for $\mathbf{X}\beta$ is given by

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = P_{\mathbf{X}|\mathbf{Z}}\,\mathbf{y} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}\,. \tag{2.48}$$

(ii) The BLUE $\ell'\hat{\beta}$ for any estimable linear function $\ell'\beta$ of β is given by

$$\ell'\hat{\beta} = \mathbf{m}'P_{\mathbf{X}|\mathbf{Z}}\,\mathbf{y} \tag{2.49}$$

where $\mathbf{m} \in R^n$ is any vector such that $\mathbf{m}'\mathbf{y}$ is an unbiased estimator for $\ell'\beta$.

Proof: (i) is obvious from what has been said above in equations (2.46) and (2.47).

(ii) $\mathbf{m}'P_{\mathbf{X}|\mathbf{Z}}\,\mathbf{y} = \mathbf{m}'\mathbf{X}\hat{\beta}$, from (i)
$= \ell'\hat{\beta}$, from the unbiasedness of $\mathbf{m}'\mathbf{y}$.

□

2.2.2 Estimation in a linear model with arbitrary but nonsingular variance-covariance structure

The linear model

(2.7) $$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}\ ;\quad cov(\mathbf{e}) = \sigma^2\mathbf{V}\ ,\quad \mathbf{V}\ \text{pds}$$

where $\mathbf{V}$ is arbitrary but nonsingular can be transformed by the nonsingular transformation

(2.8) $$\mathbf{T} = \mathbf{V}^{-1/2}\quad \text{(say)}$$

to obtain the model

(2.9) $$\mathbf{Ty} = \mathbf{TX}\beta + \mathbf{Te}\ ,\quad cov(\mathbf{Te}) = \sigma^2\mathbf{TVT}'$$

(2.10) $$\Leftrightarrow\quad \tilde{\mathbf{y}} = \tilde{\mathbf{X}}\beta + \tilde{\mathbf{e}}\ ;\quad cov(\tilde{\mathbf{e}}) = \sigma^2\mathbf{I}\ .$$

The transformed model (2.10) is the Gauß-model treated in the previous section, whose results can now directly be generalized in the following theorem.

Theorem 2.12 (Aitken, 1935)

In the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$, $\mathbf{V}$ pds the BLUE $\ell'\hat{\beta}$ for an estimable linear function $\ell'\beta$ of β is given by

(2.50) $$\ell'\hat{\beta} = \ell'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\,\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}\ .$$

Proof: To prove unbiasedness, we only have to note that $\ell \in R(\mathbf{X})$ implies that $\ell \in R(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})$, which holds due to the nonsingularity of $\mathbf{V}$. Minimum variance is shown along the lines of the proof of Theorem 2.10, simply replacing $\mathbf{X}$ by $\tilde{\mathbf{X}} = \mathbf{V}^{-1/2}\mathbf{X}$ and $\mathbf{y}$ by $\tilde{\mathbf{y}} = \mathbf{V}^{-1/2}\mathbf{y}$.

□

In general, the BLUE (2.50) and the OLS-estimate for $\ell'\beta$ will not coincide, but the idea of least-squares can be generalized. Writing the approximation problem (2.39) in terms of the transformed model (2.10), we obtain with

(2.51) $$(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\beta)'(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\beta) \equiv (\mathbf{y} - \mathbf{X}\beta)'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\beta)$$

a generalized least-squares problem, leading to the generalized normal equations (GNE's)

(2.52) $$\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\beta = \mathbf{X}'\mathbf{V}^{-1}\mathbf{y}\ .$$

Clearly, the BLUE $\ell'\hat{\beta}$ for an estimable linear function $\ell'\beta$ of β as in (2.50) is given by $\ell'\hat{\beta}$ where $\hat{\beta}$ is any solution to the GNE's (2.52). Thus the development of the previous section can be paralleled.

Examining (2.52) indicates why we could take over the concept of estimability unchanged from the Gauß-model. Clearly, $\ell \in R(\mathbf{X})$ if and only if $\ell \in R(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})$, and thus the same linear functions $\ell'\hat{\beta}$ of $\hat{\beta}$ are unique over all solutions $\hat{\beta}$ to (2.52) and (2.40) respectively.

The fitted values $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$, $\mathbf{V}$ pds are given by

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \tag{2.53}$$

which relationship enables us to generalize Theorem 2.11.

Theorem 2.13 (Rao, 1974)

In the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$, $\mathbf{V}$ pds the BLUE $\mathbf{X}\hat{\beta}$ for $\mathbf{X}\beta$ is given by

$$\mathbf{X}\hat{\beta} = P_{\mathbf{X}|\mathbf{VZ}}\mathbf{y} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} , \tag{2.54}$$

where $P_{\mathbf{X}|\mathbf{VZ}}$ denotes the projection operator onto $C(\mathbf{X})$ along $C(\mathbf{VZ})$.

The BLUE $\ell'\hat{\beta}$ for an estimable linear function $\ell'\beta$ of β is given by

$$\ell'\hat{\beta} = \mathbf{m}'P_{\mathbf{X}|\mathbf{VZ}}\mathbf{y} , \tag{2.55}$$

where $\mathbf{m} \in \mathbb{R}^n$ is any vector such that $\mathbf{m}'\mathbf{y}$ is unbiased for $\ell'\beta$.

Proof: Let be $\mathbf{M} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}$. Then $\mathbf{MX} = \mathbf{X}$ and $\mathbf{MVZ} = \mathbf{0}$. Thus $\mathbf{X}\hat{\beta} = \mathbf{My} = P_{\mathbf{X}|\mathbf{VZ}}\mathbf{y}$. The rest follows similar to the proof of Theorem 2.11. □

The parallels of BLU-estimation in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$, $\mathbf{V}$ pds to BLU-estimation in the Gauß-model are now quite clear: in both cases we can see BLU-estimates as the solution to a (generalized) least-squares problem, and in both cases the fitted values $\hat{\mathbf{y}}$ in the respective models, or the BLUE $\mathbf{X}\hat{\beta}$ for $\mathbf{X}\beta$, are the images of $\mathbf{y}$ under a projection onto $C(\mathbf{X})$, namely $P_{\mathbf{X}|\mathbf{Z}}$ or $P_{\mathbf{X}|\mathbf{VZ}}$ respectively.

2.2.3 Estimation in a linear model with possibly singular variance-covariance structure

Seeing that the generalization of BLU-estimation from the Gauß-model to the linear model with arbitrary but nonsingular variance-covariance structure was quite a straightforward affair, we might be led to consider, in analogy to the generalized least-squares problem (2.51), the approximation problem

$$(\mathbf{y} - \mathbf{X}\beta)'\mathbf{V}^{-}(\mathbf{y} - \mathbf{X}\beta) , \tag{2.56}$$

where $\mathbf{V}^{-}$ is a g-inverse of $\mathbf{V}$. We would hope that a solution $\mathbf{X}\hat{\beta}$ minimizing (2.56) would yield the BLUE for $\mathbf{X}\beta$ in the linear model

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e} \; ; \quad \mathrm{cov}(\mathbf{e}) = \sigma^2\mathbf{V} \tag{2.1}$$

Unfortunately, this is not the case in general (at least not for an arbitrary g-inverse $\mathbf{V}^{-}$).

Clearly, $\mathbf{X}\hat{\beta}$ minimizing (2.56) would be of the form

$$\mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-}\mathbf{X})\mathbf{X}'\mathbf{V}^{-}\mathbf{y} \tag{2.57}$$

but in the case of a singular g-inverse $\mathbf{V}^-$ of $\mathbf{V}$ not even the unbiasedness of $\mathbf{X}\hat{\beta}$ as in (2.57) for $\mathbf{X}\beta$ is guaranteed. Also, since a g-inverse $\mathbf{V}^-$ of $\mathbf{V}$ is not unique when $\mathbf{V}$ is singular, $\mathbf{X}\hat{\beta}$ as given by (2.57) might not be unique over all choices of a g-inverse $\mathbf{V}^-$ of $\mathbf{V}$. Thus a generalization of the least-squares approach, using an arbitrary g-inverse $\mathbf{V}^-$ of $\mathbf{V}$ in (2.56) must fail.

But apart from the characterization of the BLUE $\mathbf{X}\hat{\beta}$ for $\mathbf{X}\beta$ as the solution of a generalized least-squares problem in a linear model with arbitrary but nonsingular variance-covariance structure, we had the characterization of $\mathbf{X}\hat{\beta}$ as the image of $\mathbf{y}$ under the projection operator $P_{\mathbf{X}|\mathbf{VZ}}$. In fact, this is also true when $\mathbf{V}$ is singular.

Theorem 2.14 (Rao, 1974)

In the $LM(\mathbf{y},\mathbf{X}\beta, \sigma^2\mathbf{V})$, where $\mathbf{V}$ is arbitrary nonnegative-definite and symmetric, the BLUE $\mathbf{X}\hat{\beta}$ for $\mathbf{X}\beta$ is given by

$$\mathbf{X}\hat{\beta} = P_{\mathbf{X}|\mathbf{VZ}}\,\mathbf{y}\,, \tag{2.58}$$

and the BLUE $\ell'\hat{\beta}$ of an estimable linear function $\ell'\beta$ of β is given by

$$\ell'\hat{\beta} = \mathbf{m}'P_{\mathbf{X}|\mathbf{VZ}}\,\mathbf{y} \tag{2.59}$$

where $\mathbf{m} \in \mathbb{R}^n$ is any vector such that $\mathbf{m}'\mathbf{y}$ is unbiased for $\ell'\beta$.

Proof: We prove only the relationship (2.59), the rest follows accordingly.

Let $\mathbf{m}'\mathbf{y}$ be any unbiased estimator for $\ell'\beta$, i.e. $\mathbf{m}'\mathbf{X} = \ell'$. Consider

$$\begin{aligned} var\,(\mathbf{m}'\mathbf{y}) &= var\,(\mathbf{m}'\mathbf{y} - \mathbf{m}'\,P_{\mathbf{X}|\mathbf{VZ}}\,\mathbf{y} + \mathbf{m}'P_{\mathbf{X}|\mathbf{VZ}}\,\mathbf{y}) \\ &= var\,(\mathbf{m}'\mathbf{y} - \mathbf{m}'\,P_{\mathbf{X}|\mathbf{VZ}}\,\mathbf{y}) + var\,(\ell'\hat{\beta}). \end{aligned}$$

This proves minimum variance of $\ell'\hat{\beta} = \mathbf{m}'P_{\mathbf{X}|\mathbf{VZ}}\,\mathbf{y}$ if we can show that

$$\begin{aligned} &cov\,(\mathbf{m}'\mathbf{y} - \mathbf{m}'P_{\mathbf{X}|\mathbf{VZ}}\,\mathbf{y},\ \mathbf{m}'P_{\mathbf{X}|\mathbf{VZ}}\,\mathbf{y}) \\ &= \sigma^2\,\mathbf{m}'\,(\mathbf{I} - P_{\mathbf{X}|\mathbf{VZ}})\,\mathbf{V}P'_{\mathbf{X}|\mathbf{VZ}}\,\mathbf{m} \\ &= 0\,. \end{aligned}$$

But

$$\begin{aligned} &(\mathbf{I} - P_{\mathbf{X}|\mathbf{VZ}})\,\mathbf{X} = \mathbf{0} \\ \Rightarrow\ &R(\mathbf{I} - P_{\mathbf{X}|\mathbf{VZ}}) \subset R(\mathbf{Z}') \\ \Rightarrow\ &(\mathbf{I} - P_{\mathbf{X}|\mathbf{VZ}})\,\mathbf{V}\,P'_{\mathbf{X}|\mathbf{VZ}} = \mathbf{0}\,. \end{aligned}$$

The unbiasedness of $\ell' \hat{\beta} = \mathbf{m}' P_{\mathbf{X}|\mathbf{VZ}} \mathbf{y}$ we obtain as

$$\begin{aligned} E(\mathbf{m}' P_{\mathbf{X}|\mathbf{VZ}} \mathbf{y}) &= \mathbf{m}' P_{\mathbf{X}|\mathbf{VZ}} \mathbf{X}\beta \\ &= \mathbf{m}' \mathbf{X}\beta \\ &= \ell' \beta . \end{aligned}$$

□

Note that the operator $P_{\mathbf{X}|\mathbf{VZ}}$ is not unique when $C([\mathbf{X}:\mathbf{V}]) = C([\mathbf{X}:\mathbf{VZ}])$ is not the whole space $\mathbb{R}^n$. But we have

Corollary 2.14.1

The BLUE $\mathbf{X}\hat{\beta}$ for $\mathbf{X}\beta$ is unique.

Proof: It is sufficient to show that $P_{\mathbf{X}|\mathbf{VZ}} \mathbf{y}$ is unique over all choices of $P_{\mathbf{X}|\mathbf{VZ}}$, for arbitrary but fixed $\mathbf{y} \in C([\mathbf{X}:\mathbf{VZ}])$. But by definition, $P_{\mathbf{X}|\mathbf{VZ}} [\mathbf{X}:\mathbf{VZ}] = [\mathbf{X}:\mathbf{0}]$ independently of the choice of $P_{\mathbf{X}|\mathbf{VZ}}$.

□

Using the uniqueness of the fitted values, which implies the uniqueness of the BLUE of every estimable linear function $\ell' \beta$ of β, an earlier result on BLUE's in a linear model can be proved:

Theorem 2.15 (Zyskind, 1967)

In the $LM(\mathbf{y},\mathbf{X}\beta, \sigma^2\mathbf{V})$, a linear function $\mathbf{m}'\mathbf{y}$ of $\mathbf{y}$ is BLUE for its expectation $E(\mathbf{m}'\mathbf{y}) = \mathbf{m}'\mathbf{X}\beta$ if and only if $\mathbf{Vm} \in C(\mathbf{X})$.

Proof:

$$\begin{aligned} & \mathbf{m}'\mathbf{y} \text{ BLUE for } \mathbf{m}'\mathbf{X}\beta \\ \Leftrightarrow\ & \mathbf{m}'\mathbf{y} = \mathbf{m}' P_{\mathbf{X}|\mathbf{VZ}} \mathbf{y}, \text{ for all } \mathbf{y} \in C([\mathbf{X}:\mathbf{V}]) \\ \Leftrightarrow\ & \mathbf{m}'[\mathbf{X}:\mathbf{VZ}] = \mathbf{m}' P_{\mathbf{X}|\mathbf{VZ}} [\mathbf{X}:\mathbf{VZ}] \\ \Leftrightarrow\ & \mathbf{m}'\mathbf{VZ} = \mathbf{0} \\ \Leftrightarrow\ & \mathbf{Vm} \in C(\mathbf{X}). \end{aligned}$$

□

At the beginning of this section we began the search for a BLUE $\mathbf{X}\hat{\beta}$ for $\mathbf{X}\beta$ by considering the approximation problem (2.56) but observed that a generalization of the least-squares approach in a linear model with nonsingular variance-covariance structure failed, when we replaced the unique inverse $\mathbf{V}^{-1}$ by an arbitrary choice of a g-inverse $\mathbf{V}^-$ of $\mathbf{V}$, in the case of singular $\mathbf{V}$. Modifying this approach we could now ask the question whether there is at least a class of g-inverses $\mathbf{V}^*$ of $\mathbf{V}$ such that a solution $\mathbf{X}\hat{\beta}$ to the approximation problem

$$(\mathbf{y} - \mathbf{X}\beta)' \mathbf{V}^* (\mathbf{y} - \mathbf{X}\beta) = min \tag{2.60}$$

would yield the BLUE for $\mathbf{X}\beta$.

If such a g-inverse $\mathbf{V}^*$ exists, or a class of g-inverses $\mathbf{V}^*$ of $\mathbf{V}$, then the theory of least squares could be unified for all linear models. Zyskind and Martin (1969) and Rao (1971) achieved this *Unified Theory of Linear Estimation.*

Theorem 2.16 (Rao, 1971)

A solution $\mathbf{X}\hat{\beta}$ to the approximation problem (2.60), or equivalently a solution $\mathbf{X}\hat{\beta}$ to the GNE's

(2.61) $$\mathbf{X}'\mathbf{V}^*\mathbf{X}\beta = \mathbf{X}'\mathbf{V}^*\mathbf{y}, \quad \mathbf{V}^* \text{ a g-inverse of } \mathbf{V},$$

is the BLUE for $\mathbf{X}\beta$ in the $LM(\mathbf{y},\mathbf{X}\beta, \sigma^2\mathbf{V})$, if and only if

(2.62) $$rank\ (\mathbf{X}'\mathbf{V}^*\mathbf{X}) = rank\ (\mathbf{X}) \text{ , and}$$

(2.63) $$\mathbf{V}^* = (\mathbf{V} + \mathbf{X}\mathbf{U}\mathbf{X}')^-$$

where $\mathbf{U}$ is an arbitrary matrix such that $C(\mathbf{V}) \cap C(\mathbf{X}\mathbf{U}\mathbf{X}') = \{\mathbf{0}\}$ and (2.62) is satisfied.

The minimum variances may be obtained from

(2.64) $$var\ (\mathbf{X}\hat{\beta}) = \sigma^2\ (\mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^-\mathbf{X}' - \mathbf{X}\mathbf{U}\mathbf{X}')$$

Proof: (Dunne, 1982)

In the light of Theorems 2.14 and 2.15 we must show that

(i) $$\mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^-\ \mathbf{X}'\mathbf{V}^*\mathbf{X} = \mathbf{X} \text{ , and}$$

(ii) $$\mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^-\ \mathbf{X}'\mathbf{V}^*\mathbf{V}\mathbf{Z} = \mathbf{0}$$

if and only if (2.62) and (2.63) hold:

(i) $$\mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^-\ \mathbf{X}'\mathbf{V}^*\ \mathbf{X} = \mathbf{X}$$

$$\Leftrightarrow \quad R(\mathbf{X}) = R(\mathbf{X}'\mathbf{V}^*\mathbf{X})$$

$$\Leftrightarrow \quad rank(\mathbf{X}) = rank\ (\mathbf{X}'\mathbf{V}^*\mathbf{X})$$

(ii) Using (i) we conclude

(2.65) $$\mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^-\ \mathbf{X}'\mathbf{V}^*\mathbf{V}\mathbf{Z} = \mathbf{0}$$

$$\Leftrightarrow \quad \mathbf{X}'\mathbf{V}^*\mathbf{V}\mathbf{Z} = \mathbf{0}$$

Let $\mathbf{V}^* = (\mathbf{V} + \mathbf{X}\mathbf{U}\mathbf{X}')^-$ be given as in (2.63), and let $[\mathbf{X}_1 : \mathbf{X}_2]$ be any base of $\mathbf{X}$ such that $C(\mathbf{X}_1) \cap C(\mathbf{V}) = \{\mathbf{0}\}$ and $C(\mathbf{X}_2) \subset C(\mathbf{V})$. Then $\mathbf{X}$ can be written as $\mathbf{X} = \mathbf{X}_1\mathbf{C} + \mathbf{X}_2\mathbf{D}$, for some $\mathbf{C}$, $\mathbf{D}$. But

(2.66) $$\mathbf{X}'\mathbf{V}^*\mathbf{V}\mathbf{Z} = (\mathbf{C}'\mathbf{X}_1' + \mathbf{D}'\mathbf{X}_2')\ (\mathbf{V} + \mathbf{X}\mathbf{U}\mathbf{X}')^-\ \mathbf{V}\mathbf{Z}$$

$$= \mathbf{D}'\mathbf{X}_2'\mathbf{Z} \text{ , from Theorem 1.6}$$

$$= \mathbf{0} \text{ .}$$

This proves the sufficiency of (2.63). To exhibit a choice of $\mathbf{U}$, now let

$$
\begin{aligned}
(2.67) \qquad & \mathbf{X}'\mathbf{V}^*\mathbf{V}\mathbf{Z} = \mathbf{0} \\
\Leftrightarrow \quad & C(\mathbf{V}\mathbf{V}^{*\prime}\mathbf{X}) \subset C(\mathbf{X}) \\
\Leftrightarrow \quad & \mathbf{V}\mathbf{V}^{*\prime}\mathbf{X} = \mathbf{X}\mathbf{B}, \text{ for some } \mathbf{B} \\
\Leftrightarrow \quad & \mathbf{V}\mathbf{V}^{*\prime}(\mathbf{X}_1\mathbf{C} + \mathbf{X}_2\mathbf{D}) = \mathbf{X}_2\mathbf{E}\,, \quad \text{for some } \mathbf{C}, \mathbf{D}, \mathbf{E} \\
\Leftrightarrow \quad & \mathbf{V}\mathbf{V}^{*\prime}\mathbf{X}_1\mathbf{C} = \mathbf{X}_2(\mathbf{E} - \mathbf{D}) \\
& \qquad\qquad\quad = \mathbf{V}\mathbf{V}^{*\prime}\mathbf{X}_2(\mathbf{E} - \mathbf{D}) \\
& \qquad\qquad\quad = \mathbf{V}\mathbf{V}^{*\prime}\mathbf{X}_2\mathbf{F}\,.
\end{aligned}
$$

If $\mathbf{X}_2\mathbf{F} = \mathbf{0}$ then write $\mathbf{V}^*$ as

$$
(2.68) \qquad \mathbf{V}^* = (\mathbf{V} + \mathbf{X}_1\mathbf{C}\mathbf{C}'\mathbf{X}_1)^-\,,
$$

otherwise write $\mathbf{V}^*$ as

$$
(2.69) \qquad \mathbf{V}^* = (\mathbf{V} + (\mathbf{X}_1\mathbf{C} - \mathbf{X}_2\mathbf{F})(\mathbf{X}_1\mathbf{C} - \mathbf{X}_2\mathbf{F})')^-\,.
$$

By assumption we have $C(\mathbf{X}_1\mathbf{C}) \cap C(\mathbf{V}) = \{\mathbf{0}\}$ and $\mathbf{V}\mathbf{V}^{*\prime}(\mathbf{X}_1\mathbf{C} - \mathbf{X}_2\mathbf{F}) = \mathbf{0}$, from (2.67), which implies $C(\mathbf{X}_1\mathbf{C} - \mathbf{X}_2\mathbf{F}) \cap C(\mathbf{V}) = \{\mathbf{0}\}$.

The minimum variances in (2.64) are obtained by writing $\mathbf{V} = (\mathbf{V} + \mathbf{X}\mathbf{U}\mathbf{X}') - \mathbf{X}\mathbf{U}\mathbf{X}'$. □

Neither $(\mathbf{V} + \mathbf{X}\mathbf{U}\mathbf{X}')$ nor $\mathbf{V}^* = (\mathbf{V} + \mathbf{X}\mathbf{U}\mathbf{X})^-$ need be symmetric, but they can always be taken to be symmetric. For the rest of this thesis, $\mathbf{V}^*$ will denote a g-inverse of $\mathbf{V}$ as given by (2.62) and (2.63), and unless otherwise specified we will take $\mathbf{V}^*$ to be symmetric.

With Theorem 2.16 we have now a method to explicitly compute a projection operator $P_{\mathbf{X}|\mathbf{V}\mathbf{Z}}$ yielding the BLUE $\mathbf{X}\hat{\beta} = P_{\mathbf{X}|\mathbf{V}\mathbf{Z}}\,\mathbf{y}$ for $\mathbf{X}\beta$. One choice for $P_{\mathbf{X}|\mathbf{V}\mathbf{Z}}$ is $P_{\mathbf{X}|\mathbf{V}\mathbf{Z}} = \mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^-\mathbf{X}'\mathbf{V}^*$, and one choice for $P_{\mathbf{V}\mathbf{Z}|\mathbf{X}}$ is $P_{\mathbf{V}\mathbf{Z}|\mathbf{X}} = \mathbf{I} - P_{\mathbf{X}|\mathbf{V}\mathbf{Z}} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^-\mathbf{X}'\mathbf{V}^*$. As noted below Theorem 2.14, $P_{\mathbf{X}|\mathbf{V}\mathbf{Z}}$ is unique if and only if $C([\mathbf{X}:\mathbf{V}]) = \mathbb{R}^n$, in which case $\mathbf{V}^*$ is the unique inverse $(\mathbf{V} + \mathbf{X}\mathbf{U}\mathbf{X}')^{-1}$ of $(\mathbf{V} + \mathbf{X}\mathbf{U}\mathbf{X}')$.

Another method for computing $P_{\mathbf{X}|\mathbf{V}\mathbf{Z}}$ is the IPM (inverse partitioned matrix) method of Rao (1971). Efficient algorithms to compute $P_{\mathbf{X}|\mathbf{V}\mathbf{Z}}$ are presented in Section 2.6.

As in the case where $\mathbf{V}$ is arbitrary but nonsingular the concept of estimability need not be changed. We need only note that (2.62) implies that $R(\mathbf{X}) = R(\mathbf{X}'\mathbf{V}^*\mathbf{X})$, and thus estimable linear functions are the same in the Gauß-model $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{I})$ and in the general linear model $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$.

In summary, and paralleling the concluding remarks of the previous sections, we point out that in the $LM(\mathbf{y},\mathbf{X}\beta, \sigma^2\mathbf{V})$, where $\mathbf{V}$ is arbitrary and possibly singular, the BLUE $\mathbf{X}\hat{\beta}$ for $\mathbf{X}\beta$ is given by the image of $\mathbf{y}$ under a projection operator $P_{\mathbf{X}|\mathbf{VZ}}$. Alternatively, it can be characterized as the solution of a (generalized) least-squares problem of the form (2.60).

We conclude this section by commenting on the notion of unbiasedness of linear estimators in a linear model.

A linear function $\mathbf{m}'\mathbf{y}$ of $\mathbf{y}$ is called an unbiased estimator for a linear function $\ell'\beta$ of β if and only if $E(\mathbf{m}'\mathbf{y}) = \mathbf{m}'\mathbf{X}\beta = \ell'\beta$, for all β. Of course, this is equivalent with

$$\mathbf{m}'\mathbf{X} = \ell' \iff \mathbf{X}'\mathbf{m} = \ell \tag{2.70}$$

when the variance-covariance structure of the model is nonsingular. This follows essentially from the fact that the parameter space for the p-vector β is $\mathbb{R}^p$, the whole space. In a linear model with singular variance-covariance structure, however, the parameter space is not the whole space $\mathbb{R}^p$, in general, due to the possible presence of sure equations in the model. Rao (1972) points out that the class of unbiased linear estimators can be extended in this case, since $\mathbf{m}'\mathbf{X}\beta = \ell'\beta$ need only hold for the smaller set of β which satisfy the sure equations in the model, to render $\mathbf{m}'\mathbf{y}$ an unbiased estimator for $\ell'\beta$.

Lemma 2.17 (Rao, 1972)

In the $LM(\mathbf{y},\mathbf{X}\beta, \sigma^2\mathbf{V})$, a linear function $\mathbf{m}'\mathbf{y}$ of $\mathbf{y}$ is an unbiased estimator for the linear function $\ell'\beta$ of β if and only if

$$\begin{aligned} & \mathbf{m}'\mathbf{X}\beta = \ell'\beta \,, \quad \text{w.p.1} \\ \iff \; & (\mathbf{m}'\mathbf{X} - \ell')\beta = \mathbf{0} \,, \quad \text{w.p.1} \\ \iff \; & (\mathbf{X}'\mathbf{m} - \ell) \in C(\mathbf{X}'\mathbf{N}_1\mathbf{S}) \end{aligned} \tag{2.71}$$

where $\mathbf{N}_1$, $\mathbf{S}$ are as given in Lemma 2.2.

Proof: The result follows directly from Lemma 2.2 (iii).

□

Of course, we do not know the sure equations in a model before taking an observation, and thus (2.71) can not be checked *a-priori*, when $\mathbf{X}'\mathbf{m} \neq \ell$. In view of that fact it remains a matter of taste whether we should not define unbiasedness as the relationship (2.70), even more so since for any $\mathbf{m}$ such that $\mathbf{m}'\mathbf{y}$ is unbiased for $\ell'\beta$ in the sense of (2.71), there exists a vector $\mathbf{k}$ such that $\mathbf{k}'\mathbf{y} = \mathbf{m}'\mathbf{y}$, w.p.1 and $\mathbf{k}'\mathbf{y}$ is unbiased in the sense of (2.70) (Rao, 1972).

2.2.4 The class of operators yielding BLUE's

In Section 2.1.2 we have observed that the vector space $\mathbb{R}^n$ can be decomposed as

$$\mathbb{R}^n = C(\mathbf{X}) \oplus C(\mathbf{VZ}) \oplus C(\mathbf{N}) \tag{2.72}$$

where $\mathbf{N}$ is a base extending a base of $C([\mathbf{X}:\mathbf{V}])$ to a base of $\mathbb{R}^n$. Without loss of generality $\mathbf{N}$ can be taken to be a base of $C([\mathbf{X}:\mathbf{V}])^{\perp}$.

The BLUE $\mathbf{X}\hat{\beta}$ for $\mathbf{X}\beta$ in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ is given by

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = P_{\mathbf{X}|\mathbf{VZ}}\mathbf{y} \tag{2.58}$$

where $P_{\mathbf{X}|\mathbf{VZ}}$ is a projection operator onto $C(\mathbf{X})$ along $C(\mathbf{VZ})$. $P_{\mathbf{X}|\mathbf{VZ}}$ is not unique, and a wide class of such operators exists, when $C([\mathbf{X}:\mathbf{VZ}]) \neq \mathbb{R}^n$ or equivalently when $\mathbf{N} \neq \mathbf{0}$ in (2.72). This is so because $P_{\mathbf{X}|\mathbf{VZ}}\lambda$ for $\lambda \in C(\mathbf{N})$ can be assigned an arbitrary value leaving BLUE's invariant, since $\mathbf{y} \in C([\mathbf{X}:\mathbf{V}])$, w.p.1.

We can, with respect to the decomposition (2.72) of the vector space $\mathbb{R}^n$, characterize all linear operators P such that $P\mathbf{y}$ yields the BLUE $\mathbf{X}\hat{\beta}$ for $\mathbf{X}\beta$ in the $LM(\mathbf{X}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ as follows:

Lemma 2.18

A linear operator $P:\mathbb{R}^n \to \mathbb{R}^n$ yields the BLUE $\mathbf{X}\hat{\beta}$ for $\mathbf{X}\beta$ in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ if and only if

$$P[\mathbf{X}:\mathbf{VZ}:\mathbf{N}] = [\mathbf{X}:\mathbf{0}:\mathbf{A}] \tag{2.73}$$

where $\mathbf{A}$ is an arbitrary conformable matrix.

□

Obviously the class of operators P given by Lemma 2.18 is precisely the class of projection operators $P_{\mathbf{X}|\mathbf{VZ}}$ given by Theorem 2.14, and there are $rank(\mathbf{N})+1 = n+1-rank([\mathbf{X}:\mathbf{V}])$ linearly independent operators. Lemma 2.18 is thus just a rephrasing of the result of Theorem 2.14. But now we assume that the sure equations in the model are known. Then the class of linear operators leading to BLUE's can be extended. This is essentially due to the fact that space of admissible observations can be further restricted.

Lemma 2.19

Let $\mathbf{z}$ be as in Lemma 2.3, that is $C([\mathbf{z}:\mathbf{V}])$ is the smallest vector space containing all admissible observations in the model. Further let $[\mathbf{X}_1:\mathbf{X}_2]$ be a base such that $[\mathbf{z}:\mathbf{X}_1:\mathbf{X}_2]$ is a base of $\mathbf{X}$, with $C(\mathbf{X}_1) \cap C(\mathbf{V}) = \{\mathbf{0}\}$ and $C(\mathbf{X}_2) \subset C(\mathbf{V})$. Then a linear operator $P:\mathbb{R}^n \to \mathbb{R}^n$ yields the BLUE $\mathbf{X}\hat{\beta}$ for $\mathbf{X}\beta$ in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ if and only if

$$P[\mathbf{z}:\mathbf{X}_1:\mathbf{X}_2:\mathbf{VZ}:\mathbf{N}] = [\mathbf{z}:\mathbf{A}_1:\mathbf{X}_2:\mathbf{0}:\mathbf{A}] \tag{2.74}$$

where $\mathbf{A}_1$ and $\mathbf{A}$ are arbitrary conformable matrices.

□

In general the class of operators P given by Lemma 2.19 will be greater than the class given by Lemma 2.18, since (2.74) yields $rank(\mathbf{N}) + rank(\mathbf{A}_1) + 1 = n+1 - rank([\mathbf{z}:\mathbf{V}])$ linearly independent operators.

Noting that the space of admissible observations is actually an affine space (Corollary 2.3.1), even a wider class of operators can be found.

Lemma 2.20 (Schall)

Let $\mathbf{z}$, $\mathbf{X}_1$, $\mathbf{X}_2$ be as in Lemma 2.19. Then a linear operator $Q:\mathbb{R}^n \to \mathbb{R}^n$ of the form

$$Q\mathbf{y} = \mathbf{z} + P(\mathbf{y}-\mathbf{z}) \tag{2.75}$$

yields the BLUE $\mathbf{X}\hat{\beta}$ for $\mathbf{X}\beta$ in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ if and only if

$$P[\mathbf{z}:\mathbf{X}_1:\mathbf{X}_2:\mathbf{VZ}:\mathbf{N}] = [\mathbf{a}:\mathbf{A}_1:\mathbf{X}_2:\mathbf{0}:\mathbf{A}] \tag{2.76}$$

where $\mathbf{a}$ is an arbitrary vector and $\mathbf{A}_1$, $\mathbf{A}$ are arbitrary conformable matrices.

□

The class of operators yielding BLUE's is extended from Lemma 2.19 to Lemma 2.20 if $\mathbf{z} \neq \mathbf{0}$.

2.3 ANALYSIS OF VARIANCE

2.3.1 Estimation of the scale parameter σ^2; the sum of squares for error.

In the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ the BLUE $\mathbf{X}\hat{\beta}$ for $\mathbf{X}\beta$, or the fitted values $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ are given by

$$\begin{aligned} \hat{\mathbf{y}} = \mathbf{X}\hat{\beta} &= \mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^-\mathbf{X}'\mathbf{V}^*\mathbf{y} \\ &= \mathbf{M}\mathbf{y} \end{aligned} \tag{2.78}$$

where $\mathbf{V}^*$ is a (symmetric) g-inverse of $\mathbf{V}$ in the manner of Theorem 2.16.

The estimated error term $\hat{\mathbf{e}}$ is given by

$$\begin{aligned} \hat{\mathbf{e}} = \mathbf{y}-\hat{\mathbf{y}} &= (\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^-\mathbf{X}'\mathbf{V}^*)\mathbf{y} \\ &= (\mathbf{I}-\mathbf{M})\mathbf{y} \,. \end{aligned} \tag{2.79}$$

The operator $(\mathbf{I}-\mathbf{M})$ is the projection operator $P_{\mathbf{VZ}|\mathbf{X}}$ onto $C(\mathbf{VZ})$ along $C(\mathbf{X})$, and thus

$$E(\hat{\mathbf{e}}) = (\mathbf{I}-\mathbf{M})\mathbf{X}\beta = \mathbf{0}\,, \quad \text{and} \tag{2.80}$$

$$\hat{\mathbf{e}} \in C(\mathbf{VZ}) \subset C(\mathbf{V})\,, \quad \text{w.p.1.} \tag{2.81}$$

The variance-covariance matrix of $\hat{\mathbf{e}}$ is given by

$$
\begin{aligned}
(2.82) \qquad cov(\hat{\mathbf{e}}) &= \sigma^2(\mathbf{I}-\mathbf{M})\mathbf{V}(\mathbf{I}-\mathbf{M})' \\
&= \sigma^2(\mathbf{V}-\mathbf{MV}-\mathbf{VM}'+\mathbf{MVM}') \\
&= \sigma^2(\mathbf{V}+\mathbf{XUX}'-\mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^-\mathbf{X}') \\
&= \sigma^2\mathbf{N} ,
\end{aligned}
$$

which is verified by writing $\mathbf{V}$ as $\mathbf{V} = (\mathbf{V}+\mathbf{XUX}') - \mathbf{XUX}'$.

Since $\hat{\mathbf{e}} \in C(\mathbf{V})$, w.p.1, the quadratic form $\hat{\mathbf{e}}'\mathbf{V}^-\hat{\mathbf{e}}$ is invariant (w.p.1) over all choices of a g-inverse $\mathbf{V}^-$ of $\mathbf{V}$ and we may therefore, without loss of generality, write

$$
\begin{aligned}
(2.83) \qquad \hat{\mathbf{e}}'\mathbf{V}^*\hat{\mathbf{e}} &= \mathbf{y}'(\mathbf{I}-\mathbf{M})'\mathbf{V}^*(\mathbf{I}-\mathbf{M})\mathbf{y} \\
&= \mathbf{y}'\mathbf{Q}\mathbf{y} .
\end{aligned}
$$

The following theorem is easily verified by checking the conditions in (1.1) through (1.3) of Khatri (1962, 1963).

Theorem 2.21 (Zyskind and Martin, 1969)

The quadratic form $\hat{\mathbf{e}}'\mathbf{V}^-\hat{\mathbf{e}}$ is invariant (w.p.1) over all choices of a g-inverse $\mathbf{V}^-$ of $\mathbf{V}$, and

$$
(2.84) \qquad E(\hat{\mathbf{e}}'\mathbf{V}^-\hat{\mathbf{e}}) = s\cdot\sigma^2
$$

where $s = rank([\mathbf{X}:\mathbf{V}]) - rank(\mathbf{X})$.

Under the assumption of normality, $\hat{\mathbf{e}}'\mathbf{V}^-\hat{\mathbf{e}}$ follows a central $\sigma^2\chi^2_s$ distribution. □

Theorem 2.21 implies, that

$$
(2.85) \qquad \hat{\sigma}^2 = \hat{\mathbf{e}}'\mathbf{V}^-\hat{\mathbf{e}} \div s
$$

is an unbiased estimator for σ^2, and Rao (1973, p.319) points out that $\hat{\sigma}^2$ as in (2.85) is the minimum variance unbiased estimator for σ^2 under the assumption of normality.

The quadratic form $\hat{\mathbf{e}}'\mathbf{V}^-\hat{\mathbf{e}}$ is commonly called the sum of squares for error (SSE) in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$. Writing the total sum of squares (SS) in the model as

$$
(2.86) \qquad SS = \mathbf{y}'\mathbf{V}^*\mathbf{y}
$$

it can be decomposed into uncorrelated sums of squares as

$$
\begin{aligned}
(2.87) \qquad SS &= \mathbf{y}'\mathbf{V}^*\mathbf{y} \\
&= (\hat{\mathbf{y}}+\hat{\mathbf{e}})'\mathbf{V}^*(\hat{\mathbf{y}}+\hat{\mathbf{e}}) \\
&= \hat{\mathbf{y}}'\mathbf{V}^*\hat{\mathbf{y}}+\hat{\mathbf{e}}'\mathbf{V}^*\hat{\mathbf{e}} \\
&= SSR+SSE ,
\end{aligned}
$$

where $SSR = \hat{\mathbf{y}}'\mathbf{V}^*\hat{\mathbf{y}}$ denotes the sum of squares for regression.

The decomposition (2.87) follows from $\mathbf{M}'\mathbf{V}^*(\mathbf{I}-\mathbf{M}) = \mathbf{0}$, and

(2.88) $$cov(\hat{\mathbf{e}}, \hat{\mathbf{y}}) = \sigma^2 P_{\mathbf{VZ}|\mathbf{X}} \mathbf{V} P'_{\mathbf{X}|\mathbf{VZ}} = \mathbf{0}$$

yields the zero correlation of $\hat{\mathbf{y}}'\mathbf{V}^*\hat{\mathbf{y}}$ and $\hat{\mathbf{e}}'\mathbf{V}^*\hat{\mathbf{e}}$.

The decomposition (2.87) and the distributional result of Theorem 2.21 are important when linear hypotheses are tested in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$. The ANOVA-table corresponding to (2.87) is

Table 2.22

ANOVA for the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$.

SS	df	statistic	expectation
SSR	$r(\mathbf{V})-s = r(\mathbf{X})-r([\mathbf{X}:\mathbf{V}])+r(\mathbf{V})$	$\hat{\mathbf{y}}'\mathbf{V}^*\hat{\mathbf{y}}$	$\beta'\mathbf{X}'\mathbf{V}^*\mathbf{X}\beta + (r(\mathbf{V})-s)\cdot\sigma^2$
SSE	$s = r([\mathbf{X}:\mathbf{V}])-r(\mathbf{X})$	$\hat{\mathbf{e}}'\mathbf{V}^*\hat{\mathbf{e}}$	$s\cdot\sigma^2$
SS	$r(\mathbf{V})$	$\mathbf{y}'\mathbf{V}^*\mathbf{y}$	$\beta'\mathbf{X}'\mathbf{V}^*\mathbf{X}\beta + r(\mathbf{V})\cdot\sigma^2$

□

Note that in general $\hat{\mathbf{y}}'\mathbf{V}^*\hat{\mathbf{y}}$ and thus $\mathbf{y}'\mathbf{V}^*\mathbf{y}$ are not distributed chi-squared, even though $\hat{\mathbf{e}}'\mathbf{V}^*\hat{\mathbf{e}}$ satisfies the requisite conditions. Conditions for the chi-squaredness of $\hat{\mathbf{y}}'\mathbf{V}^*\hat{\mathbf{y}}$ and $\mathbf{y}'\mathbf{V}^*\mathbf{y}$ are given in Section 1.2.

2.3.2 Tests of linear hypotheses

Suppose, in the linear model

(2.1) $$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}\ , \quad cov(\mathbf{e}) = \sigma^2\mathbf{V}$$

we wish to test the linear hypothesis

(2.89) $$H_0 : \mathbf{L}\beta = \mathbf{c}\ .$$

A linear hypothesis is called consistent, when the equations (2.89) are consistent, i.e. $\mathbf{c} \in C(\mathbf{L})$, and when the sure equations in the model (2.1) are consistent with $\mathbf{L}\beta = \mathbf{c}$ (Rao, 1972). In the following, unless specified otherwise, we will only consider consistent hypotheses.

The usual method to test the hypothesis (2.89) is to compare the SSE under model (2.1) with the SSE under the model

(2.90) $$\begin{cases} \mathbf{y} = \mathbf{X}\beta + \mathbf{e}\ ; \quad cov(\mathbf{e}) = \sigma^2\mathbf{V} \\ \mathbf{L}\beta = \mathbf{c} \end{cases}$$

with the additional restrictions $\mathbf{L}\beta = \mathbf{c}$, or equivalently under the model

(2.91) $$\begin{bmatrix} \mathbf{y} \\ \mathbf{c} \end{bmatrix} = \begin{bmatrix} \mathbf{X} \\ \mathbf{L} \end{bmatrix} \beta + \mathbf{e}\,; \quad cov(\mathbf{e}) = \sigma^2 \begin{bmatrix} \mathbf{V} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

This leads to the usual F-test, well-known from ANOVA and ANACOVA problems in the Gauß-model, which has many optimal properties (see e.g. Kendall and Stuart, 1973).

Theorem 2.23 (Zyskind and Martin, 1969)

Let $\hat{\mathbf{e}}$ and $\tilde{\mathbf{e}}$ be the residual vectors under models (2.1) and (2.90/2.91) respectively. The F-statistic

(2.92) $$F = \frac{\tilde{\mathbf{e}}'\mathbf{V}^-\tilde{\mathbf{e}} - \hat{\mathbf{e}}'\mathbf{V}^-\hat{\mathbf{e}}}{\hat{\mathbf{e}}'\mathbf{V}^-\hat{\mathbf{e}}} \cdot \frac{s}{h}$$

has, under the assumption of normality, a central $F_{h,s}$-distribution under the null hypothesis (2.89). The respective degrees of freedom are

(2.93) $$s = rank([\mathbf{X}:\mathbf{V}]) - rank\ (\mathbf{X})\ , \text{ and}$$

(2.94) $$h = rank(\mathbf{X}) - rank\left(\begin{bmatrix} \mathbf{X} \\ \mathbf{L} \end{bmatrix}\right) + rank\left(\begin{bmatrix} \mathbf{N}'\mathbf{X} \\ \mathbf{L} \end{bmatrix}\right) - rank(\mathbf{N}'\mathbf{X})$$

where $\mathbf{N}$ is a matrix of maximum rank such that $\mathbf{N}'\mathbf{V} = \mathbf{0}$.

Proof: We show that

(2.95) $$cov(\tilde{\mathbf{e}} - \hat{\mathbf{e}},\ \hat{\mathbf{e}}) = \mathbf{0}\ , \quad \text{and}$$

(2.96) $$(\tilde{\mathbf{e}} - \hat{\mathbf{e}})'\mathbf{V}^-(\tilde{\mathbf{e}} - \hat{\mathbf{e}}) = \tilde{\mathbf{e}}'\mathbf{V}^-\tilde{\mathbf{e}} - \hat{\mathbf{e}}'\mathbf{V}^-\hat{\mathbf{e}}\ .$$

Then, under normality, the numerator and denominator in (2.12) are independenly distributed.

Further, the numerator $(\tilde{\mathbf{e}} - \hat{\mathbf{e}})'\mathbf{V}^-(\tilde{\mathbf{e}} - \hat{\mathbf{e}})$ as the difference of two $\sigma^2\chi^2$ variates is independent of one of the variates, namely $\hat{\mathbf{e}}'\mathbf{V}^-\hat{\mathbf{e}}$, and thus distributed $\sigma^2\chi^2$. Finally, the ratio of two independent χ^2-variates follows a F-distribution.

The sure equations in model (2.1) are

(2.97) $$\mathbf{N}'\mathbf{X}\beta = \mathbf{N}'\mathbf{y}\ , \quad \text{from Lemma 2.1}\ .$$

Of course, in model (2.90) we need only consider those constraints $\mathbf{L}_0\beta - \mathbf{c}_0$ which are not already contained in (2.97), or more precisely if $\mathbf{L}_1$ is a base of $R(\mathbf{N}'\mathbf{X}) \cap R(\mathbf{L})$, then extend this base by $\mathbf{L}_0$ to a base of $\mathbf{L}$, yielding $R(\mathbf{L}_0) \cap R(\mathbf{N}'\mathbf{X}) = \{\mathbf{0}\}$. Thus we consider without loss of generality the model (2.90) with $R(\mathbf{L}) \cap R(\mathbf{N}'\mathbf{X}) = \{\mathbf{0}\}$.

Writing $\mathbf{X}$ as

(2.98) $$\begin{aligned} \mathbf{X} &= \mathbf{X}\,(\mathbf{I} - \mathbf{L}'(\mathbf{L}\mathbf{L}')^-\mathbf{L}) + \mathbf{X}\mathbf{L}'(\mathbf{L}\mathbf{L}')^-\mathbf{L} \\ &= \mathbf{X}_1 + \mathbf{X}_2\mathbf{L} \end{aligned}$$

it is clear that the fitted values $\tilde{\mathbf{y}}$ in model (2.90) are given by

(2.99) $$\tilde{\mathbf{y}} = P_{\mathbf{X}_1|\mathbf{V}\mathbf{Z}_1}(\mathbf{y}-\mathbf{X}_2\mathbf{c}) + \mathbf{X}_2\mathbf{c}$$
$$= P_{\mathbf{X}_1|\mathbf{V}\mathbf{Z}_1}\mathbf{y} + P_{\mathbf{V}\mathbf{Z}_1|\mathbf{X}_1}\mathbf{X}_2\mathbf{c} ,$$

where $\mathbf{Z}_1$ is a matrix of maximum rank such that $\mathbf{Z}_1'\mathbf{X}_1 = \mathbf{0}$.

Noting that

(2.100) $$\tilde{\mathbf{e}}-\hat{\mathbf{e}} = \tilde{\mathbf{e}}-\mathbf{y}+\mathbf{y}-\hat{\mathbf{e}}$$
$$= \hat{\mathbf{y}}-\tilde{\mathbf{y}}$$

we can write

(2.101) $$cov(\tilde{\mathbf{e}}-\hat{\mathbf{e}}, \hat{\mathbf{e}})$$
$$= cov(\hat{\mathbf{y}}-\tilde{\mathbf{y}}, \hat{\mathbf{e}})$$
$$= \sigma^2(P_{\mathbf{X}|\mathbf{V}\mathbf{Z}}-P_{\mathbf{X}_1|\mathbf{V}\mathbf{Z}_1})\mathbf{V}P'_{\mathbf{V}\mathbf{Z}|\mathbf{X}}$$
$$= \sigma^2(P_{\mathbf{X}|\mathbf{V}\mathbf{Z}}-P_{\mathbf{X}_1|\mathbf{V}\mathbf{Z}_1})\mathbf{V}\mathbf{Z}\mathbf{C} , \quad \text{for some } \mathbf{C}$$
$$= \mathbf{0} , \quad \text{since } C(\mathbf{V}\mathbf{Z}) \subset C(\mathbf{V}\mathbf{Z}_1) , \quad \text{from } C(\mathbf{X}_1) \subset C(\mathbf{X}) .$$

Next we show that $C(\mathbf{X}_2) \subset C([\mathbf{X}_1:\mathbf{V}]) = C([\mathbf{X}_1:\mathbf{V}\mathbf{Z}_1])$. Assume the contrary. Then there exists a matrix $\mathbf{N}_2$ such that $\mathbf{N}_2'[\mathbf{X}_1:\mathbf{V}] = [\mathbf{0}:\mathbf{0}]$ but $\mathbf{N}_2'\mathbf{X}_2 \neq \mathbf{0}$. This implies

(2.102) $$\mathbf{N}_2'\mathbf{y} = \mathbf{N}_2'\mathbf{X}_2\mathbf{L}\beta , \quad \text{w.p.1}$$

is a subset of the sure equations in model (2.1), which is a contradiction to the assumption $R(\mathbf{L}) \cap R(\mathbf{N}'\mathbf{X}) = \{\mathbf{0}\}$.

Thus we can write $\mathbf{X}_2$ as

(2.103) $$\mathbf{X}_2 = \mathbf{X}_1\mathbf{B} + \mathbf{V}\mathbf{Z}_1\mathbf{C} , \quad \text{for some } \mathbf{B} \text{ and } \mathbf{C}$$

and $C(\mathbf{V}\mathbf{Z}_1\mathbf{C}) \subset C(\mathbf{X})$ from (2.103).

Any admissible observation $\mathbf{y}$ can be written as

(2.104) $$\mathbf{y} = \mathbf{X}_1\lambda_1 + \mathbf{X}_2\mathbf{L}\lambda_2 + \mathbf{V}\mathbf{Z}\gamma$$
$$= \mathbf{X}_1\lambda_1 + \mathbf{X}_1\mathbf{B}\mathbf{L}\lambda_2 + \mathbf{V}\mathbf{Z}_1\mathbf{C}\mathbf{L}\lambda_2 + \mathbf{V}\mathbf{Z}\gamma$$

for some $\lambda_1, \lambda_2, \gamma$.

By (2.99) and using $\hat{\mathbf{y}} = P_{\mathbf{X}|\mathbf{V}\mathbf{Z}}\mathbf{y}$ we obtain

(2.105) $$\hat{\mathbf{e}} = \mathbf{V}\mathbf{Z}\gamma$$
$$\tilde{\mathbf{e}} = \mathbf{V}\mathbf{Z}\gamma + \mathbf{V}\mathbf{Z}_1\mathbf{C}\mathbf{L}(\lambda_2-\mathbf{c}) .$$

To prove (2.96) we must only show that

(2.106) $$\hat{\mathbf{e}}'\mathbf{V}^-\tilde{\mathbf{e}} = \hat{\mathbf{e}}'\mathbf{V}^-\hat{\mathbf{e}} ,$$

or, in view of (2.105),

(2.107)
$$\begin{aligned} & \gamma' \mathbf{Z}' \mathbf{V} \mathbf{V}^{-} \mathbf{V} \mathbf{Z}_1 \mathbf{C} \mathbf{L} (\lambda_2 - \mathbf{c}) \\ &= \gamma' \mathbf{Z}' \mathbf{V} \mathbf{Z}_1 \mathbf{C} \mathbf{L} (\lambda_2 - \mathbf{c}) \\ &= 0 \, . \end{aligned}$$

This is clearly the case from our earlier remark below (2.103), that $C(\mathbf{V}\mathbf{Z}_1\mathbf{C}) \subset C(\mathbf{X})$.

The degrees of freedom for the numerator are clearly given by

(2.94a)
$$\begin{aligned} h &= rank(\mathbf{X}) - rank(\mathbf{X}_1) \\ &= rank(\mathbf{L}) - rank\left(\begin{bmatrix} \mathbf{X} \\ \mathbf{L} \end{bmatrix} \right) + rank(\mathbf{X}) \end{aligned}$$

since the rows of $\mathbf{X}_1$ are the projections of the rows of $\mathbf{X}$ onto $R(\mathbf{L})^{\perp}$.

Generally, when $R(\mathbf{L}) \cap R(\mathbf{N}'\mathbf{X}) = \{\mathbf{0}\}$ is not satisfied, the degrees of freedom for the numerator are given by (2.94), where we adjust, in
$rank\left(\begin{bmatrix} \mathbf{N}'\mathbf{X} \\ \mathbf{L} \end{bmatrix} \right) - rank(\mathbf{N}'\mathbf{X})$, for the subspace of $R(\mathbf{L})$ which is in $R(\mathbf{N}'\mathbf{X})$.

□

Using Theorem 2.23 we can test any consistent linear hypothesis $\mathbf{L}\beta = \mathbf{c}$, whether or not $\mathbf{L}$ is estimable. However, if $\mathbf{L}$ is estimable, i.e. $R(\mathbf{L}) \subset R(\mathbf{X})$, then the degrees of freedom for the numerator are given by

(2.94b)
$$h = rank\left(\begin{bmatrix} \mathbf{N}'\mathbf{X} \\ \mathbf{L} \end{bmatrix} \right) - rank(\mathbf{N}'\mathbf{X})$$

and when in addition $R(\mathbf{N}'\mathbf{X})$ and $R(\mathbf{L})$ are disconnected then

(2.94c)
$$h = rank(\mathbf{L}) \, .$$

An apparent disadvantage of the F-statistic (2.92) is that we must fit two models to compute it, namely models (2.1) and (2.91) respectively. But we will presently see that (2.92) can be written in an alternative form avoiding the fit of the second model (2.91).

Implicit in the development of the proof of Theorem 2.23 is the decomposition of the sum of squares for regression in model (2.1) into the sum of squares for regression in model (2.91) and the sum of squares for the hypothesis (SSH) in the numerator of the F-statitic (2.92). We obtain

(2.108)
$$\begin{aligned} SS &= SSR + SSE \\ &= \hat{\mathbf{y}}' \mathbf{V}^* \hat{\mathbf{y}} + \hat{\mathbf{e}}' \mathbf{V}^* \hat{\mathbf{e}} \\ &= (SSR - SSH) + SSH + SSE \\ &= (SSR - SSH) + (\tilde{\mathbf{e}} - \hat{\mathbf{e}})' \mathbf{V}^* (\tilde{\mathbf{e}} - \hat{\mathbf{e}}) + \hat{\mathbf{e}}' \mathbf{V}^* \hat{\mathbf{e}} \, , \end{aligned}$$

where $(SSR - SSH)$, SSH and SSE are mutually uncorrelated and $SSH = (\tilde{\mathbf{e}} - \hat{\mathbf{e}})' \mathbf{V}^* (\tilde{\mathbf{e}} - \hat{\mathbf{e}})$ is the numerator in the F-statistic (2.92).

The corresponding ANOVA-table is

Table 2.24

ANOVA for the hypothesis H_0: $\mathbf{L}\beta = \mathbf{c}$ in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$.

SS	df	statistic
$SSR - SSH$	$r(\mathbf{V}) - h - s$	$SS - SSE - SSH$
SSH	h	$(\tilde{\mathbf{e}} - \hat{\mathbf{e}})' \mathbf{V}^* (\tilde{\mathbf{e}} - \hat{\mathbf{e}})$
SSE	s	$\hat{\mathbf{e}}' \mathbf{V}^* \hat{\mathbf{e}}$
SS	$r(\mathbf{V})$	$\mathbf{y}' \mathbf{V}^* \mathbf{y}$

□

We note that $\hat{\mathbf{y}}' \mathbf{V}^* \hat{\mathbf{y}}$ and $\tilde{\mathbf{y}}' \mathbf{V}^* \tilde{\mathbf{y}}$ are not in general distributed chi-squared as pointed out in Section 1.2. To obtain chi-squared variates an adjustment as in Table 1.3 must be made.

2.4 REPARAMETRIZATIONS AND TRANSFORMATIONS

2.4.1 Reparametrization

Definition 2.25

A linear model

(2.109) $$\mathbf{y} = \mathbf{X}^* \beta^* + \mathbf{e} \; ; \quad cov(\mathbf{e}) = \sigma^2 \mathbf{V}$$

is said to be a reparametrization of the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ if and only if there exist conformable matrices $\mathbf{U}$ and $\mathbf{T}$ such that

(2.110) $$\mathbf{X}^* = \mathbf{X}\mathbf{U} \; ,$$

$$\beta^* = \mathbf{T}\beta \; , \quad \text{and}$$

$$\mathbf{X}^* \beta^* = \mathbf{X}\mathbf{U}\mathbf{T}\beta \; , \quad \text{for all } \beta, \beta^* \; .$$

□

From this definition it is obvious that $C(\mathbf{X}) = C(\mathbf{X}^*)$ which implies the following

Theorem 2.26 (Pringle and Rayner, 1971)

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

Proof: The fitted values in the original model $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ and the reparametrized model $(\mathbf{y}, \mathbf{X}^*\beta^*, \sigma^2\mathbf{V})$ coincide:

$$\hat{\mathbf{y}} = P_{\mathbf{X}|\mathbf{V}\mathbf{Z}}\mathbf{y} = P_{\mathbf{X}^*|\mathbf{V}\mathbf{Z}^*}\mathbf{y} = \hat{\mathbf{y}}^* \tag{2.111}$$

since $C(\mathbf{X}) = C(\mathbf{X}^*)$ if and only if $C(\mathbf{Z}) = C(\mathbf{X})^{\perp} = C(\mathbf{Z}^*) = C(\mathbf{X}^*)^{\perp}$.

□

But testing of a linear hypothesis is also invariant under a reparametrization of a linear model.

Corollary 2.26.1

The F-statistic (2.92) associated with testing the linear hypothesis $H_0: \mathbf{L}\beta = \mathbf{c}$ in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ is invariant under a reparametrization of the model, that is, it coincides with the F-statistic associated with testing the linear hypothesis $H_0: \mathbf{L}^*\beta^* = \mathbf{c}$ in the reparametrized model $(\mathbf{y}, \mathbf{X}^*\beta^*, \sigma^2\mathbf{V})$, where $\mathbf{L}^* = \mathbf{L}\mathbf{U}$.

Proof: By Theorem 2.26, the fitted values in model (2.91) and model

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{c} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^* \\ \mathbf{L}^* \end{bmatrix} \beta^* + \mathbf{e} \; ; \quad cov(\mathbf{e}) = \sigma^2 \begin{bmatrix} \mathbf{V} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \tag{2.112}$$

coincide, thus the respective residual vectors, and finally the respective F-statistics coincide.

□

When the variance-covariance structure $\mathbf{V}$ of a linear model is singular, the class of reparametrizations of this model can be extended. Instead of requiring $\mathbf{X}^*\beta^* = \mathbf{X}\mathbf{U}\mathbf{T}\beta$ for all β ans β^* as in (2.110), we would only require $\mathbf{X}^*\beta^* = \mathbf{X}\mathbf{U}\mathbf{T}\beta$ for all β and β^* which satisfy the sure equations in the models $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ and $(\mathbf{y}, \mathbf{X}^*\beta^*, \sigma^2\mathbf{V})$ respectively.

If the columns of $\mathbf{B}$ and $\mathbf{B}^*$ (say) respectively span the spaces of admissible parameter vectors β and β^* (see Corollary 2.2.1), then clearly a $LM(\mathbf{y}, \mathbf{X}^*\beta^*, \sigma^2\mathbf{V})$ is a reparametrization of the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ if and only if $C(\mathbf{X}\mathbf{B}) = C(\mathbf{X}^*\mathbf{B}^*)$.

In Corollary 2.26.1 we have used the fact that the estimated residual vectors are invariant under a reparametrization of a model. This leads us to the following algorithm for the reduction of a linear model with singular variance-covariance structure to a linear model with nonsingular variance-covariance structure, leaving estimated residuals invariant, and leading to a linear model which is statistically equivalent to the original one.

Algorithm 2.27 (Schall)

Reduction of a singular linear model

Step 1: Reparametrize the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ to obtain the model

(2.113)
$$\mathbf{y} = [\mathbf{X}^* : \mathbf{X}_2^*] \begin{bmatrix} \beta^* \\ \beta_2^* \end{bmatrix} + \mathbf{e} \; ; \quad cov(\mathbf{e}) = \sigma^2\mathbf{V}$$

such that $C(\mathbf{X}^*) \subset C(\mathbf{V})$ and $C(\mathbf{X}_2^*) \cap C(\mathbf{V}) = \{\mathbf{0}\}$. Then $\mathbf{X}_2^*\beta_2^* = const$, w.p.1 and $\mathbf{X}^*\beta^*$ is unaffected by the sure equations in the model.

Step 2: Subtract $\mathbf{X}_2^*\beta_2^*$ on both sides of the equation (2.113), to obtain with $\mathbf{y}^* = \mathbf{y} - \mathbf{X}_2^*\beta^*$ the model

(2.114)
$$\mathbf{y}^* = \mathbf{X}^*\beta^* + \mathbf{e} \; ; \quad cov(\mathbf{e}) = \sigma^2\mathbf{V}$$

Step 3: Reduce the model (2.114) to a model of full rank as in Lemma 2.6, to obtain

$$\mathbf{y}_1^* = \mathbf{X}_1^*\beta^* + \mathbf{e}_1 \; ; \quad cov(\mathbf{e}_1) = \sigma^2\mathbf{V}_{11}$$

□

The model (2.114) has a nonsingular variance-covariance structure $\mathbf{V}_{11}$, which is a submatrix of $\mathbf{V}$ (possibly after a rearrangement of the model). By preserving the error term $\mathbf{e}_1$ it is statistically equivalent to the original model $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$, since $\mathbf{e}_2$ is a linear function of $\mathbf{e}_1$ (w.p.1).

From

(2.115)
$$\begin{aligned} & \hat{\mathbf{y}} = \mathbf{X}^*\hat{\beta}^* + \mathbf{X}_2^*\hat{\beta}_2^* \\ \Leftrightarrow \quad & \hat{\mathbf{y}} - \mathbf{X}_2^*\hat{\beta}_2^* = \mathbf{X}^*\hat{\beta}^* \\ \Leftrightarrow \quad & \hat{\mathbf{y}}^* = \mathbf{X}\hat{\beta}^* \\ \Leftrightarrow \quad & \hat{\mathbf{y}}_1^* = \mathbf{X}_1\hat{\beta}^* \end{aligned}$$

we can conclude that $\hat{\mathbf{e}}_1 = \mathbf{y}_1 - \hat{\mathbf{y}}_1 = \mathbf{y}_1^* - \hat{\mathbf{y}}_1^* = \hat{\mathbf{e}}_1^*$ and the estimated residual vectors in the original model $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ and the reduced model (2.114) are equivalent in the sense that $(n-t)$ components (say) are identical and the remaining t components in the original model are a linear function of the first $(n-t)$ components.

The reduction of a $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ to the form (2.114) is useful when a residual analysis of this model is performed, while testing for outliers and influential observations in the model. Most residual based procedures and statistics are invariant under a reduction of a model as given by Algorithm 2.27, and by performing the reduction we avoid problems arising out of a singular variance-covariance structure.

Computationally not much extra effort is required to perform the reduction, since essentially it involves, in step 1 and step 2 of the algorithm, the determination of the sure equations in the original model, which is explicit in the estimation in the original model.

2.4.2 Transformation

Definition 2.28

A linear model

$$\mathbf{Ty} = \mathbf{TX}\beta + \mathbf{Te}\ ;\quad cov(\mathbf{Te}) = \sigma^2\mathbf{TVT}' \tag{2.116}$$

$$\Leftrightarrow:\quad \tilde{\mathbf{y}} = \tilde{\mathbf{X}}\beta + \tilde{\mathbf{e}}\ ;\quad cov(\tilde{\mathbf{e}}) = \sigma^2\tilde{\mathbf{V}}$$

is said to be a transformation of the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ if and only if $\mathbf{T}$ is nonsingular. □

Similar to the previous section we show that the BLUE $\mathbf{X}\hat{\beta}$ for $\mathbf{X}\beta$ and the F-statistic associated with the hypothesis H_0: $\mathbf{L}\beta = \mathbf{c}$ are invariant under a transformation of a linear model.

Theorem 2.29 (Mitra and Rao, 1968)

The BLUE $\ell'\hat{\beta}$ for any estimable linear function $\ell'\beta$ of β is invariant under a transformation of the model.

Proof: Noting that $R(\mathbf{X}) = R(\mathbf{TX}) = R(\tilde{\mathbf{X}})$ for nonsingular $\mathbf{T}$, we must only show that $\mathbf{X}\hat{\beta} = \mathbf{X}\tilde{\beta}$, where $\mathbf{X}\hat{\beta}$ and $\mathbf{X}\tilde{\beta}$ are respectively the BLUE's for $\mathbf{X}\beta$ in the original model $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ and the transformed model $(\tilde{\mathbf{y}}, \tilde{\mathbf{X}}\beta, \sigma^2\tilde{\mathbf{V}})$:

$$\begin{aligned} & \mathbf{X}\hat{\beta} = P_{\mathbf{X}|\mathbf{VZ}}\mathbf{y} \\ \Leftrightarrow\quad & \mathbf{TX}\hat{\beta} = \mathbf{T}P_{\mathbf{X}|\mathbf{VZ}}\mathbf{y} \\ & \qquad = P_{\mathbf{TX}|\mathbf{TVZ}}\mathbf{Ty}\ ,\quad \text{for all } \mathbf{y} \in C([\mathbf{X}:\mathbf{VZ}]) \\ & \qquad = \mathbf{TX}\tilde{\beta}\ . \\ \Leftrightarrow\quad & \mathbf{X}\hat{\beta} = \mathbf{X}\tilde{\beta}\ . \end{aligned} \tag{2.117}$$

□

Corollary 2.29.1

The F-statistic (2.92) associated with testing the linear hypothesis H_0: $\mathbf{L}\beta = \mathbf{c}$ in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ is invariant under a transformation of the model.

Proof: Let $\hat{\mathbf{e}}$ and $\tilde{\mathbf{e}}$ be respectively the residual vectors under models (2.1) and (2.90). Then by Theorem 2.29 we have that $\mathbf{T}\hat{\mathbf{e}}$ and $\mathbf{T}\tilde{\mathbf{e}}$ are the corresponding residual vectors under the transformed models.

But if $\mathbf{V}^-$ is a g-inverse of $\mathbf{V}$, then $\mathbf{T}^{-1\prime}\mathbf{V}^-\mathbf{T}^{-1}$ is a g-inverse of $\tilde{\mathbf{V}} = \mathbf{T}\mathbf{V}\mathbf{T}'$, and thus the respective F-statistics coincide.

We may note that $\mathbf{y}$ is normally distributed if $\tilde{\mathbf{y}} = \mathbf{T}\mathbf{y}$ is normally distributed.

□

With Theorem 2.29 and its corollary we can now prove our claim made at the end of Section 2.1.3 that BLU-estimation and testing of a linear hypothesis are invariant under a reduction of a linear model, which is not of full rank, to a reduced model of full rank:

Let $\mathbf{N}_2' = [\mathbf{N}_{21}' : \mathbf{N}_{22}']$ be as in Lemma 2.6, i.e. $\mathbf{N}_2'[\mathbf{X}:\mathbf{V}] = [\mathbf{0}:\mathbf{0}]$ and $\mathbf{N}_{22}$ is a nonsingular $t \times t$ matrix. Then

(2.118)
$$\mathbf{T}: = \begin{bmatrix} \mathbf{I}_{n-t} : \mathbf{0} \\ \mathbf{N}_{21}' : \mathbf{N}_{22}' \end{bmatrix}$$

is a nonsingular matrix, and transforming the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ by $\mathbf{T}$ we obtain

(2.119)
$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{0} \end{bmatrix}\beta + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{0} \end{bmatrix}; \quad cov(\mathbf{e}_1) = \sigma^2\mathbf{V}_{11}\,.$$

Except for the redundant zero's this is precisely the reduced model (2.37), and as an obvious consequence of Theorem 2.29 and Corollary 2.29.1 we obtain the desired

Corollary 2.29.2

BLU-estimation and F-statistics are invariant under a reduction of a linear model to a linear model of full rank.

□

When the space $C([\mathbf{X}:\mathbf{V}])$ of admissible observations in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ is not the whole space $\mathbb{R}^n$ (see Lemma 2.3), we can generalize the notion of transformation of a linear model.

Instead of the nonsingularity of $\mathbf{T}$ in Definition 2.28 we require only that a matrix $\mathbf{N}$ exists such that

(2.120)
(i) $\mathbf{T}+\mathbf{N}$ is nonsingular, and

(ii) $\mathbf{N}'[\mathbf{X}:\mathbf{V}] = [\mathbf{0}:\mathbf{0}]$.

A reduction of a linear model would then be performed by a transformation $\mathbf{T}$ of the form

(2.121)
$$\mathbf{T} = \begin{bmatrix} \mathbf{I}_{n-t} : \mathbf{0} \\ \mathbf{0} \quad : \mathbf{0} \end{bmatrix}$$

after a suitable rearrangement of the model equations.

2.5 AUGMENTING AND PARTITIONING A LINEAR MODEL

2.5.1 Augmenting a linear model

An analysis of the data arising from a proposed $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ might lead to the conclusion that the proposed model is inadequate and does not fit the data well.

To improve the model, we attempt to fit additional variables as the conformable matrix $\mathbf{A}$ (say), that is, we augment the original model (2.1) by $\mathbf{A}$ to obtain

$$\mathbf{y} = [\mathbf{X}:\mathbf{A}] \begin{bmatrix} \beta \\ \lambda \end{bmatrix} + \mathbf{e} \; ; \quad cov(\mathbf{e}) = \sigma^2\mathbf{V} \quad . \tag{2.122}$$

These additional variables could be concomitant variables to adjust for concomitant variation, or some dummy variables to adjust the original model for possible outliers or missing observations in the data. The latter application, due to Draper (1961), will play an important role in the following chapter on outliers.

The first question which could be asked, in the case of singular $\mathbf{V}$, is what happens to the sure equations in the model.

Certainly, in some situations such as testing hypotheses on β, we might say that the sure equations in the original model (2.1) should also hold in the augmented model (2.122), i.e. the process of fitting the new variables $\mathbf{A}$ should not lead to the contradiction of the sure equations in (2.1). This is the case when in fitting the new variables we wish to preserve the sure equations. The statistical process of fitting a new variable is not supposed to interfere with nonstatistical sure information on the parameters prior to introducing new concomitant information and new variables.

In one sense this is the converse problem to the testability of the hypothesis H_0: $\mathbf{L}\beta = \mathbf{c}$ in a linear model with singular variance-covariance structure, where H_0 must be consistent with the sure equations to be testable (Rao, 1972). Whereas testing a hypothesis inherently means a reduction of the number of variables in the model (without contradicting the sure equations), we are now confronted with the problem of increasing the number of variables without contradicting the sure equations in the model. Consequently we define

Definition 2.30 (Schall)

A set of variables $\mathbf{A}$ (say) is called admissible to augment the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ if and only if the sure equations in the original model still hold in the augmented model.

□

Even in situations where we would allow for rewriting the sure equations in a model by augmenting it, it will be of interest to know when in fact we do rewrite the sure equations, whether we are altering nonstochastic information.

The following lemma allows us to ascertain whether a new variable is admissible to augment a given linear model.

Lemma 2.31 (Schall)

The variables $\mathbf{A}$ are admissible to augment the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ if and only if

$$C(\mathbf{A}) \cap C([\mathbf{X}:\mathbf{V}]) \subset C(\mathbf{V}) \; . \tag{2.123}$$

Proof: Let $\mathbf{N} = [\mathbf{N}_1:\mathbf{N}_2]$ be a matrix of maximum and full rank such that

$$\mathbf{N}'\mathbf{V} = \mathbf{0} \tag{2.11}$$

where $\mathbf{N}_2$ is a matrix of maximum rank such that

$$\mathbf{N}_2'[\mathbf{X}:\mathbf{V}] = [\mathbf{0}:\mathbf{0}] \tag{2.14}$$

Then the sure equations in the original model $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ are (Lemma 2.2)

$$\mathbf{N}_1'\mathbf{y} = \mathbf{N}_1'\mathbf{X}\beta \; , \quad \text{w.p.1} \; . \tag{2.20}$$

Similarly in the augmented model the set of sure equations is

$$\begin{bmatrix} \mathbf{N}_1' \\ \mathbf{N}_2' \end{bmatrix} \mathbf{y} = \begin{bmatrix} \mathbf{N}_1' \\ \mathbf{N}_2' \end{bmatrix} [\mathbf{X}:\mathbf{A}] \begin{bmatrix} \beta \\ \lambda \end{bmatrix} \; , \quad \text{w.p.1} \tag{2.124}$$

$$\Leftrightarrow \quad \begin{cases} \mathbf{N}_1'\mathbf{y} = \mathbf{N}_1'\mathbf{X}\beta + \mathbf{N}_1'\mathbf{A}\lambda \\ \mathbf{0} = \mathbf{N}_2'\mathbf{A}\lambda \; , \quad \text{w.p.1} \end{cases} \; , \quad \text{w.p.1}$$

Now the following equivalences hold:

$$\begin{aligned} & \{(2.124) \Rightarrow (2.20)\} \qquad (2.125) \\ \Leftrightarrow \quad & \{\mathbf{N}_2'\mathbf{A}\lambda = \mathbf{0} \Rightarrow \mathbf{N}_1'\mathbf{A}\lambda = \mathbf{0}\} \\ \Leftrightarrow \quad & \{\mathbf{N}_2'\mathbf{A}\lambda = \mathbf{0} \Rightarrow \mathbf{N}'\mathbf{A}\lambda = \mathbf{0}\} \\ \Leftrightarrow \quad & \{\mathbf{A}\lambda \in C([\mathbf{X}:\mathbf{V}]) \Rightarrow \mathbf{A}\lambda \in C(\mathbf{V})\} \; , \text{ from (2.11), (2.14)} \\ \Leftrightarrow \quad & \{C(\mathbf{A}) \cap C([\mathbf{X}:\mathbf{V}]) \subset C(\mathbf{V})\} \; . \end{aligned}$$

□

We can write any new variable $\mathbf{A}$, which is not necessarily admissible in the sense of Lemma 2.31, as

$$\begin{aligned} \mathbf{A} &= [\mathbf{X}:\mathbf{V}]\mathbf{B} + \mathbf{N}\mathbf{C} \; , \quad \text{for conformable } \mathbf{B} \text{ and } \mathbf{C} \qquad (2.126) \\ &= \mathbf{A}_1 + \mathbf{A}_2 \; , \end{aligned}$$

where $\mathbf{N}$ is a base of $C([\mathbf{X}:\mathbf{V}])^{\perp}$ (any base $\mathbf{N}$ extending a base of $C([\mathbf{X}:\mathbf{V}])$ to a base of the space $\mathbb{R}^n$ will suffice). The following lemma is of interest when a consistent linear model is augmented.

Lemma 2.32 (Schall)

With respect to the augmented $LM(2.122)$ let $\mathbf{A}$ be written as $\mathbf{A} = \mathbf{A}_1 + \mathbf{A}_2$, where $C(\mathbf{A}_1) \subset C([\mathbf{X}:\mathbf{V}])$ and $C(\mathbf{A}_2) \cap C([\mathbf{X}:\mathbf{V}]) = \{\mathbf{0}\}$. If the original model $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ was consistent, then

$$\mathbf{A}_2\hat{\lambda} = \mathbf{0}\,, \quad \text{w.p.1} \tag{2.127}$$

for any $\hat{\lambda}$ such that $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} + \mathbf{A}\hat{\lambda}$ is the BLUE for $\mathbf{X}\beta + \mathbf{A}\lambda$ in the augmented model (2.122).

Proof: If the original model was consistent, then $\mathbf{y} \in C([\mathbf{X}:\mathbf{V}])$, w.p.1 and $\hat{\mathbf{e}} \in C(\mathbf{V})$ in the augmented model (2.122). Thus in the augmented model

$$\begin{aligned} & \hat{\mathbf{y}} = (\mathbf{y} - \hat{\mathbf{e}}) \in C([\mathbf{X}:\mathbf{V}])\,, \quad \text{w.p.1} \\ \Rightarrow\ & \mathbf{X}\hat{\beta} + \mathbf{A}\hat{\lambda} = (\mathbf{X}\hat{\beta} + \mathbf{A}_1\hat{\lambda} + \mathbf{A}_2\hat{\lambda}) \in C([\mathbf{X}:\mathbf{V}])\,, \quad \text{w.p.1} \\ \Rightarrow\ & \mathbf{A}_2\hat{\lambda} \in C([\mathbf{X}:\mathbf{V}])\,, \quad \text{w.p.1} \\ \Rightarrow\ & \mathbf{A}_2\hat{\lambda} = \mathbf{0}\,, \quad \text{w.p.1}\,. \end{aligned} \tag{2.128}$$

□

Lemma 2.32 implies that, without loss of extra fit, we can always take the variables $\mathbf{A}$ to augment a consistent model from the space $C([\mathbf{X}:\mathbf{V}])$, and all admissible variables from the space $C(\mathbf{V})$. Equivalent to (2.127) is that $\mathbf{A}\hat{\lambda} \in C([\mathbf{X}:\mathbf{V}])$, w.p.1.

The following theorem, which generalizes a similar result for the case $\mathbf{V} = \mathbf{I}$ (see e.g. Searle, 1971), will play an important role in the analysis of augmented linear models. To clarify the notation, we label a set of linear models, all of them under common variance-covariance structure $\sigma^2\mathbf{V}$.

(2.129)

(1)	$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$	: the original model
(2)	$\mathbf{y} = [\mathbf{X}:\mathbf{A}]\begin{bmatrix}\beta\\ \lambda\end{bmatrix} + \mathbf{e}$	: the augmented model, with $C(\mathbf{A}) \subset C([\mathbf{X}:\mathbf{V}])$
(3)	$\mathbf{y} = [\mathbf{X}:P_{\mathbf{VZ}\|\mathbf{X}}\mathbf{A}]\begin{bmatrix}\beta\\ \lambda\end{bmatrix} + \mathbf{e}$	: a reparametrization of (2)
(4)	$\mathbf{y} = P_{\mathbf{VZ}\|\mathbf{X}}\mathbf{A}\lambda + \mathbf{e}$	: a reduction of (3)

Theorem 2.33 (Schall and Dunne, 1986d)

Let $\hat{\mathbf{y}}^{(i)}$, $\mathbf{X}\hat{\beta}^{(i)}$ and $P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\hat{\lambda}^{(i)}$ respectively denote the BLU-estimates for the fitted values, $\mathbf{X}\beta$ (if estimable in (2)) and $P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\lambda$ in the models in question. Then the following relationships hold, when $C(\mathbf{A}) \subset C([\mathbf{X}:\mathbf{V}])$ is satisfied.

(a) $\hat{\mathbf{y}}^{(2)} = \hat{\mathbf{y}}^{(3)}$

(b) $\mathbf{X}\hat{\beta}^{(1)} = \mathbf{X}\hat{\beta}^{(3)}$

(c) $P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\hat{\lambda}^{(2)} = P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\hat{\lambda}^{(3)} = P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\hat{\lambda}^{(4)} = P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\hat{\lambda}$. If model (4) is not consistent, $\mathbf{y}$ can be replaced by $\hat{\mathbf{e}}^{(1)} = \mathbf{y} - \mathbf{X}\hat{\beta}^{(1)}$.

(d) $\mathbf{X}\hat{\beta}^{(2)} = \mathbf{X}\hat{\beta}^{(1)} - P_{\mathbf{X}|\mathbf{VZ}}\mathbf{A}\hat{\lambda}$, if $\mathbf{X}\beta$ is estimable in (2), otherwise (d) yields the BLUE $\ell'\hat{\beta}^{(2)}$ for any estimable linear function $\ell'\beta$ of β.

(e) $P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\hat{\lambda}$ is uncorrelated with $\mathbf{X}\hat{\beta}^{(3)} = \mathbf{X}\hat{\beta}^{(1)}$.

(f) The additional sum of squares due to fitting $\mathbf{A}$ in (2) is

$$SSA = (P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\hat{\lambda})'\mathbf{V}^{-}P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\hat{\lambda}$$

(g) The total sum of sqaures in (2) and (3) can be decomposed into uncorrelated sums of squares as

$$\begin{aligned} SS &= SSR^{(2)} + SSE^{(2)} \\ &= SSR^{(1)} + SSA + SSE^{(2)} \\ &= SSR^{(1)} + SSA + (SSE^{(1)} - SSA) \ . \end{aligned}$$

(h) The F-statistic associated with the hypothesis H_0: $\mathbf{A}\lambda = \mathbf{0}$, or more precisely with H_0: $P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\lambda = \mathbf{0}$, is given by

$$F = \frac{SSA}{SSE^{(2)}} \cdot \frac{s-a}{a} = \frac{SSA}{SSE^{(1)} - SSA} \cdot \frac{s-a}{a} \quad , \text{ where } a = rank(P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}) \ ,$$

which under normality follows a $F_{a,\,s-a}$-distribution.

Proof:
(a) We need only to show that $C([\mathbf{X}:\mathbf{A}]) = C([\mathbf{X}:P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}])$, which is clearly the case since for $C(\mathbf{A}) \subset C([\mathbf{X}:\mathbf{V}])$ we can write

$$\mathbf{A} = P_{\mathbf{X}|\mathbf{VZ}}\mathbf{A} + P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A} \ . \tag{2.130}$$

Thus the column spaces of the design matrices in (2) and (3) are identical and the result follows from Theorem 2.14.

(b) Let $\tilde{\mathbf{Z}}$ be a matrix of maximum rank such that $\tilde{\mathbf{Z}}'[\mathbf{X}:\mathbf{A}] = [\mathbf{0}:\mathbf{0}]$. Then any admissible observation $\mathbf{y} \in C([\mathbf{X}:\mathbf{V}])$ can be written as

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\xi + \mathbf{A}\alpha + \mathbf{V}\tilde{\mathbf{Z}}\gamma \ , \quad \text{for some } \xi, \alpha, \gamma \\ &= \mathbf{X}\xi + P_{\mathbf{X}|\mathbf{VZ}}\mathbf{A}\alpha + P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\alpha + \mathbf{V}\tilde{\mathbf{Z}}\gamma \ . \end{aligned} \tag{2.131}$$

Clearly,

$$\mathbf{X}\hat{\beta}^{(1)} = \mathbf{X}\xi + P_{\mathbf{X}|\mathbf{VZ}}\mathbf{A}\alpha = \mathbf{X}\hat{\beta}^{(3)}. \tag{2.132}$$

(c)
$$\begin{aligned} \hat{\mathbf{y}}^{(3)} &= \mathbf{X}\hat{\beta}^{(3)} + P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\hat{\lambda}^{(3)} \\ &= \hat{\mathbf{y}}^{(2)} \ , \quad \text{from (a)} \\ &= \mathbf{X}\hat{\beta}^{(2)} + \mathbf{A}\hat{\lambda}^{(2)} \\ &= \mathbf{X}\hat{\beta}^{(2)} + P_{\mathbf{X}|\mathbf{VZ}}\mathbf{A}\hat{\lambda}^{(2)} + P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\hat{\lambda}^{(2)} \ , \quad \text{from (2.130).} \end{aligned}$$

Now the first equality follows from $C(\mathbf{X}) \cap C(\mathbf{VZ}) = \{\mathbf{0}\}$ (Lemma 2.4). The second equality is proved, provided that model (4) is consistent, using (2.131):

$$P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\hat{\lambda}^{(3)} = P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\alpha = P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\hat{\lambda}^{(4)} . \tag{2.133}$$

If (4) is not consistent, we can replace $\mathbf{y}$ by $\hat{\mathbf{e}}^{(1)}$.

(d) From (c) we have

$$\begin{aligned} \mathbf{X}\hat{\beta}^{(2)} &= \hat{\mathbf{y}}^{(3)} - \mathbf{A}\hat{\lambda} \\ &= \mathbf{X}\hat{\beta}^{(1)} + P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\hat{\lambda} - \mathbf{A}\hat{\lambda} , \quad \text{from (b)} \\ &= \mathbf{X}\hat{\beta}^{(1)} + P_{\mathbf{X}|\mathbf{VZ}}\mathbf{A}\hat{\lambda} . \end{aligned} \tag{2.134}$$

(e)
$$\begin{aligned} & cov(P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\hat{\lambda}, \mathbf{X}\hat{\beta}^{(3)}) \\ &= cov(P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\hat{\lambda}, \mathbf{X}\hat{\beta}^{(1)}) , \quad \text{from (b)} \\ &= cov(P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\hat{\lambda}, P_{\mathbf{X}|\mathbf{VZ}}\mathbf{y}) \\ &= \mathbf{0} . \end{aligned}$$

(f)
$$\begin{aligned} SSR^{(2)} &= \hat{\mathbf{y}}^{(2)'}\mathbf{V}^*\hat{\mathbf{y}}^{(2)} \\ &= (\mathbf{X}\hat{\beta}^{(2)} + \mathbf{A}\hat{\lambda})'\mathbf{V}^*(\mathbf{X}\hat{\beta}^{(2)} + \mathbf{A}\hat{\lambda}) \\ &= (\mathbf{X}\hat{\beta}^{(1)} + P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\hat{\lambda})'\mathbf{V}^*(\mathbf{X}\hat{\beta}^{(1)} + P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\hat{\lambda}) \\ &= (\mathbf{X}\hat{\beta}^{(1)})'\mathbf{V}^*\mathbf{X}\hat{\beta}^{(1)} + (P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\hat{\lambda})'\mathbf{V}^*P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\hat{\lambda} \\ &= SSR^{(1)} + SSA , \end{aligned}$$

since $\mathbf{X}'\mathbf{V}^*P_{\mathbf{VZ}|\mathbf{X}} = \mathbf{0}$.

(g) and (h) are direct consequences of (f).

□

Useful in practical situations, when the additional sum of squares SSA and adjusted estimates in model (2) have to be computed, is the following corollary.

Corollary 2.33.1

Let the BLUE $\mathbf{X}\hat{\beta}$ for $\mathbf{X}\beta$ in (1) be given by $\mathbf{X}\hat{\beta} = P_{\mathbf{X}|\mathbf{VZ}}\mathbf{y} = \mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^-\mathbf{X}'\mathbf{V}^*\mathbf{y}$, and let $\mathbf{M}$ denote the matrix

$$\mathbf{M} = \mathbf{V}^* - \mathbf{V}^*\mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^-\mathbf{X}'\mathbf{V}^* \tag{2.135}$$

Then

(a) $P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\hat{\lambda} = (\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^-\,\mathbf{X}'\mathbf{V}^*)\mathbf{A}\,(\mathbf{A}'\mathbf{M}\mathbf{A})^-\,\mathbf{A}'\mathbf{V}^*\hat{\mathbf{e}}$

(b) $SSA = \hat{\mathbf{e}}'\mathbf{V}^*\mathbf{A}(\mathbf{A}'\mathbf{M}\mathbf{A})^-\,\mathbf{A}'\mathbf{V}^*\hat{\mathbf{e}}$

(c) $\mathbf{X}\hat{\beta}^{(2)} = \mathbf{X}\hat{\beta}^{(1)}-\mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^-\,\mathbf{X}'\mathbf{V}^*\mathbf{A}\,(\mathbf{A}'\mathbf{M}\mathbf{A})^-\,\mathbf{A}'\mathbf{V}^*\hat{\mathbf{e}}$, if $\mathbf{X}\hat{\beta}^{(2)}$ is estimable, otherwise the BLUE $\ell'\hat{\beta}$ for any estimable linear function $\ell'\beta$ of β is given by (c).

Proof:

Let $\mathbf{N} = \mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^-\,\mathbf{X}'\mathbf{V}^* = P_{\mathbf{VZ}|\mathbf{X}}$, then $\mathbf{M} = \mathbf{N}'\mathbf{V}^*\mathbf{N}$.

(a) Using Theorem 2.33 (c)

$$\begin{aligned} P_{\mathbf{VZ}|\mathbf{X}}\hat{\lambda} &= \mathbf{NA}(\mathbf{A}'\mathbf{N}'\mathbf{V}^*\mathbf{NA})^-\,\mathbf{A}'\mathbf{N}'\mathbf{V}^*\hat{\mathbf{e}} \\ &= \mathbf{NA}(\mathbf{A}'\mathbf{M}\mathbf{A})^-\,\mathbf{A}'\mathbf{V}^*\mathbf{N}\hat{\mathbf{e}} \\ &= \mathbf{NA}(\mathbf{A}'\mathbf{M}\mathbf{A})^-\,\mathbf{A}'\mathbf{V}^*\hat{\mathbf{e}} \,. \end{aligned}$$

(b) Using (a) and Theorem 2.33 (f):

$$\begin{aligned} SSA &= \hat{\mathbf{e}}'\mathbf{V}^*\mathbf{A}(\mathbf{A}'\mathbf{M}\mathbf{A})^-\mathbf{A}'\mathbf{N}'\mathbf{V}^*\mathbf{NA}(\mathbf{A}'\mathbf{M}\mathbf{A})^-\mathbf{A}'\mathbf{V}^*\hat{\mathbf{e}} \\ &= \hat{\mathbf{e}}'\mathbf{V}^*\mathbf{A}(\mathbf{A}'\mathbf{M}\mathbf{A})^-\mathbf{A}'\mathbf{V}^*\hat{\mathbf{e}} \,. \end{aligned}$$

(c) Follows directly from (a) and Theorem 2.33 (d).

□

For the formulation of Theorem 2.33 and its corollary we required that the additional variables $\mathbf{A}$ satisfy the condition $C(\mathbf{A}) \subset C([\mathbf{X}:\mathbf{V}])$. If this is not case, quantities like $P_{\mathbf{X}|\mathbf{VZ}}\mathbf{A}$ and $P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}$ are not invariant over the special choice of the projection operator in question, and the models (3) and (4) are not well-defined. However, since we have that $\mathbf{A}\hat{\lambda} \in C([\mathbf{X}:\mathbf{V}])$, w.p.1 as noted below Lemma 2.32, the results (d) through (h) of Theorem 2.33 hold for any $\mathbf{A}$.

If $C(\mathbf{A}) \subset C([\mathbf{X}:\mathbf{V}])$ is not satisfied, a possible course of action is to reparametrize the $\mathbf{A}\lambda$-part of the augmented model (2.122) as

(2.136)
$$\mathbf{A}\lambda = [\mathbf{A}_1^*:\mathbf{A}_2^*]\begin{bmatrix}\lambda_1^*\\ \lambda_2^*\end{bmatrix}, \quad \text{for all } \lambda,\, \lambda^*$$

such that $C(\mathbf{A}_1^*) \subset C([\mathbf{X}:\mathbf{V}])$ and $C(\mathbf{A}_2^*) \cap C([\mathbf{X}:\mathbf{V}]) = \{\mathbf{0}\}$. Then, using Lemma 2.32 we conclude that the BLUE $\mathbf{A}_2^*\hat{\lambda}_2^*$ for $\mathbf{A}_2^*\lambda_2^*$ is zero, w.p.1 . Thus it is sufficient to fit the variables $\mathbf{A}_1^*$ only, for which Theorem 2.33 can be applied.

Theorem 2.33 and its corollary allow a complete treatment of an augmented linear model, BLU-estimation and tests for additional fit. In the light of results (b), (c), (d) and (f) of the theorem and related results of the corollary, the augmented model need not actually explicitly be fitted to compute adjusted parameter estimates and the test-statistic for additional fit.

Especially if $\mathbf{A}$ is a single column vector, or if it contains only few columns compared with $\mathbf{X}$, it is much more economical to compute the F-statistic (h) by fitting model (4) rather than (2).

A further important application of the theorem is its use when a model is downdated, i.e. a model is preserved but the data is reduced. It is well-known that the removal of the i-th observation from a linear model is equivalent to fitting a dummy variable $\mathbf{u}_i = (0, \ldots, 1, \ldots, 0)'$ where the 1 appears at the i-th component of $\mathbf{u}_i$. Other types of data reduction will be treated in the next chapter. In any case, however, the theorem allows us to compute the parameter estimates in the reduced data model without actually fitting the dummy variable or removing the observation in question, in an economical manner.

Finally, the principle that the additional sum of squares SSA due to fitting $\mathbf{A}$ after $\mathbf{X}$ is uncorrelated with (and, under normality, independent of) the sum of squares for regression due to fitting $\mathbf{X}$ alone, we will call the additional sum of squares principle, in the manner of Searle (1971).

The corresponding ANOVA table, as indicated by the decomposition (g) of Theorem 2.33, may be given as below:

Table 2.34

ANOVA for the augmented model (2.122)

SS	df	statistic	expectation
$SSR^{(1)}$	$r(\mathbf{V})-s$	$\hat{\mathbf{y}}^{(1)\prime}\mathbf{V}^*\hat{\mathbf{y}}^{(1)}$	$\beta'\mathbf{X}'\mathbf{V}^*\mathbf{X}\beta + (r(\mathbf{V})-s)\cdot\sigma^2$
SSA	a	$(P_{\mathbf{VZ}\mid\mathbf{X}}\mathbf{A}\hat{\lambda})'\mathbf{V}^*P_{\mathbf{VZ}\mid\mathbf{X}}\mathbf{A}\hat{\lambda}$	$(P_{\mathbf{VZ}\mid\mathbf{X}}\mathbf{A}\lambda)'\mathbf{V}^*(P_{\mathbf{VZ}\mid\mathbf{X}}\mathbf{A}\lambda) + a\cdot\sigma^2$
$SSE^{(1)}-SSA$	$s-a$	$\hat{\mathbf{e}}^{(1)\prime}\mathbf{V}^*\hat{\mathbf{e}}^{(1)}-SSA$	$(s-a)\cdot\sigma^2$
SS	$r(\mathbf{V})$	$\mathbf{y}'\mathbf{V}^*\mathbf{y}$	$(\mathbf{X}\beta + P_{\mathbf{VZ}\mid\mathbf{X}}\mathbf{A}\lambda)'\mathbf{V}^*(\mathbf{X}\beta + P_{\mathbf{VZ}\mid\mathbf{X}}\mathbf{A}\lambda) + r(\mathbf{V})\cdot\sigma^2$

□

2.5.2 Partitioning a linear model

We partition the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ conformably as

(2.137)
$$\mathbf{y} = [\mathbf{X}_1:\mathbf{X}_2]\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix} + \mathbf{e}\ ;\quad cov(\mathbf{e}) = \sigma^2\mathbf{V}\ .$$

If $\mathbf{y} \in C([\mathbf{X}_1:\mathbf{V}])$, we can treat this model precisely as we have treated the augmented model in the previous section. Estimates and a decomposition of the sum of squares can be obtained in a similar way, with $\mathbf{X} = \mathbf{X}_1$ and $\mathbf{A} = \mathbf{X}_2$ or *vice versa*, when $\mathbf{y} \in C([\mathbf{X}_2:\mathbf{V}])$ is satisfied.

By an interesting application of Theorem 2.33 we can also write the F-statistic (2.92) of Theorem 2.23 in an alternative way.

Recall that we wish to test the (consistent) linear hypotheses H_0: $\mathbf{L}\beta = \mathbf{c}$ in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$. By the invariance of the test-statistic (2.92) under a reparametrization of the model (Corollary 2.26.1), we assume without loss of generality that the hypothesis is of the form H_0: $[\mathbf{0}:\mathbf{L}_2]\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix} = \mathbf{c}$, where $\mathbf{L}_2$ is of full rank and $R([\mathbf{0}:\mathbf{L}_2]) \subset R(\mathbf{X})$.

Further, since the test-statistic (2.92) is also invariant under a reduction of a linear model to full rank, we can assume without loss of generality that $\mathbf{L}_2$ is a square matrix and thus nonsingular. This can be achieved by reducing the model (2.91) by dropping all redundant equations in $\mathbf{L}_2\beta_2 = \mathbf{c}$. With these assumptions we can prove

Theorem 2.35 (Zyskind and Martin, 1969)

The F-statistic

(2.138) $$F = \frac{(\mathbf{L}\hat{\beta} - \mathbf{c})'(cov(\mathbf{L}\hat{\beta}))^-(\mathbf{L}\hat{\beta} - \mathbf{c})}{\hat{\mathbf{e}}'\mathbf{V}^-\hat{\mathbf{e}}} \cdot \frac{s}{h}$$

is associated with the hypothesis H_0: $\mathbf{L}\beta = \mathbf{c}$ in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$.

Proof: (Schall)

Without loss of generality we assume that H_0 is of the form

(2.139) $$H_0: \quad [\mathbf{0}:\mathbf{L}_2]\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix} = \mathbf{c}$$

where $\mathbf{L}_2$ is a nonsingular matrix such that $R([\mathbf{0}:\mathbf{L}_2]) \subset R(\mathbf{X})$.

Let the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ be conformably partitioned as (2.137) and subsequently reparametrized as

(2.140) $$\mathbf{y} = [\mathbf{X}_1:P_{\mathbf{V}\mathbf{Z}_1|\mathbf{X}_1}\mathbf{X}_2]\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix} + \mathbf{e}\ ; \quad cov(\mathbf{e}) = \sigma^2\mathbf{V}$$

The BLUE $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ for $\mathbf{X}\beta$ in (2.140) can be written as

(2.141) $$\hat{\mathbf{y}} = \mathbf{X}_1\hat{\beta}_1 + P_{\mathbf{V}\mathbf{Z}_1|\mathbf{X}_1}\mathbf{X}_2\hat{\beta}_2$$

and with Theorem 2.33 and using $\mathbf{L}_2\beta_2 = \mathbf{c}$, i.e. $\beta_2 = \mathbf{L}_2^{-1}\mathbf{c}$, we can write the BLUE $\tilde{\mathbf{y}} = \mathbf{X}\tilde{\beta}$ in (2.140) under the additional restrictions $\mathbf{L}_2\beta_2 = \mathbf{c}$ as

(2.142) $$\tilde{\mathbf{y}} = \mathbf{X}_1\hat{\beta}_1 + P_{\mathbf{V}\mathbf{Z}_1|\mathbf{X}_1}\mathbf{X}_2\mathbf{L}_2^{-1}\mathbf{c}$$

Clearly,

(2.143) $$(\hat{\mathbf{e}} - \tilde{\mathbf{e}})'\mathbf{V}^-(\hat{\mathbf{e}} - \tilde{\mathbf{e}})$$
$$= (\mathbf{X}_2\hat{\beta}_2 - \mathbf{X}_2\mathbf{L}_2^{-1}\mathbf{c})'P'_{\mathbf{V}\mathbf{Z}_1|\mathbf{X}_1}\mathbf{V}^-P_{\mathbf{V}\mathbf{Z}_1|\mathbf{X}_1}(\mathbf{X}_2\hat{\beta}_2 - \mathbf{X}_2\mathbf{L}_2^{-1}\mathbf{c}) .$$

But

(2.144)
$$cov([\mathbf{0}:\mathbf{L}_2]\begin{bmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{bmatrix})$$
$$= [\mathbf{0}:\mathbf{L}_2](\begin{bmatrix}\mathbf{X}_1'\\ \mathbf{X}_2'P'_{\mathbf{VZ}_1|\mathbf{X}_1}\end{bmatrix}\mathbf{V}^*[\mathbf{X}_1:P_{\mathbf{VZ}_1|\mathbf{X}_1}\mathbf{X}_2])^-\begin{bmatrix}\mathbf{0}\\ \mathbf{L}_2'\end{bmatrix}$$
$$= \mathbf{L}_2(\mathbf{X}_2'P'_{\mathbf{VZ}_1|\mathbf{X}_1}\mathbf{V}^*P_{\mathbf{VZ}_1|\mathbf{X}_1}\mathbf{X}_2)^-\mathbf{L}_2'$$
$$= \mathbf{L}_2(\mathbf{X}_2'P'_{\mathbf{VZ}_1|\mathbf{X}_1}\mathbf{V}^-P_{\mathbf{VZ}_1|\mathbf{X}_1}\mathbf{X}_2)^-\mathbf{L}_2' .$$

This implies

(2.145)
$$(\mathbf{L}_2\hat{\beta}_2-\mathbf{c})'cov(\mathbf{L}_2\hat{\beta}_2)^-(\mathbf{L}_2\hat{\beta}_2-\mathbf{c})$$
$$= (\hat{\beta}_2-\mathbf{L}_2^{-1}\mathbf{c})'\mathbf{X}_2'P'_{\mathbf{VZ}_1|\mathbf{X}_1}\mathbf{V}^-P_{\mathbf{VZ}_1|\mathbf{X}_1}\mathbf{X}_2(\hat{\beta}_2-\mathbf{L}_2^{-1}\mathbf{c})$$

and this is by (2.143) precisely the numerator of the F-statistic (2.92).

□

In the implication from (2.141) to (2.142), using Theorem 2.33, we actually require that $\mathbf{y} \in C([\mathbf{X}_1:\mathbf{V}])$ to be able to use Theorem 2.33. If this is not the case, this condition can be relaxed somewhat by requiring $(\mathbf{y}-P_{\mathbf{VZ}_1|\mathbf{X}_1}\mathbf{X}_2\mathbf{L}_2^{-1}\mathbf{c}) \in C([\mathbf{X}_1:\mathbf{V}])$. But this is always the case if the hypothesis $\mathbf{L}_2\beta_2 = \mathbf{c}$ is consistent. We formulate this as a lemma and complete thus the proof of Theorem 2.35.

Lemma 2.36 (Schall)

The hypothesis $\mathbf{H}_0$: $\mathbf{L}\beta = \mathbf{c}$ in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ is consistent if and only if

(2.146)
$$(\mathbf{y}-\mathbf{X}_2\mathbf{c}) \in C([\mathbf{X}_1:\mathbf{V}])$$

where $\mathbf{X}_1 = \mathbf{X}(\mathbf{I}-\mathbf{L}'(\mathbf{L}\mathbf{L}')^-\mathbf{L})$ and $\mathbf{X}_2 = \mathbf{X}\mathbf{L}'(\mathbf{L}\mathbf{L}')^-$ as given in (2.98).

Proof: Let $\mathbf{N}$ be a matrix of maximum rank such that $\mathbf{N}'[\mathbf{X}_1:\mathbf{V}] = [\mathbf{0}:\mathbf{0}]$. Then

(2.147)
$$\mathbf{N}'\mathbf{y} = \mathbf{N}'\mathbf{X}_2\mathbf{L}\beta \ , \quad \text{w.p.1}$$

is a subset of the sure equations in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$. But

$$(\mathbf{y}-\mathbf{X}_2\mathbf{c}) \in C([\mathbf{X}_1:\mathbf{V}])$$
$$\Leftrightarrow \quad \mathbf{N}'(\mathbf{y}-\mathbf{X}_2\mathbf{c}) = \mathbf{0}$$
$$\Leftrightarrow \quad \mathbf{N}'(\mathbf{y}-\mathbf{X}_2\mathbf{L}\beta) = \mathbf{0}$$
$$\Leftrightarrow \quad \mathbf{N}'\mathbf{y} = \mathbf{N}'\mathbf{X}_2\mathbf{L}\beta$$
$$\Leftrightarrow \quad \mathbf{L}\beta = \mathbf{c} \quad \text{is consistent,}$$

since $\mathbf{X}_1\beta$ is unaffected by the hypothesis $\mathbf{L}\beta = \mathbf{c}$.

□

2.6 COMPUTATIONAL ISSUES

The most efficient and numerically stable methods to solve least-squares problems, or equivalently to perform the BLU-estimation of parameters in the classical Gauß-model, are generally considered to be those which apply a QR-factorization or a singular value decomposition (SVD) of the design matrix $\mathbf{X}$.

Wilkinson and Reinsch (1971) provide an excellent collection of algorithms useful in the statistical linear algebra. Of special interest here are contributions I/8 by Businger and Golub (1965) and I/10 by Golub and Reinsch (1970), respectively concerning the solution of least-squares problems by Householder-transformations and by a singular value decomposition. In the following we generalize these methods to be applicable in a $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ where $\mathbf{X}$ and $\mathbf{V}$ are possibly not of full rank. In addition we apply these methods to the related problem of testing a linear hypothesis in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$.

Firstly, we propose an algorithm for the solution of the generalized least-squares problem in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ which is useful when only estimation of the parameters β and σ^2 is of interest.

Thereafter we present two approaches to the problem of BLU-estimation and testing of a linear hypothesis which lead directly to the use of efficient and numerically stable algorithms provided elsewhere in the literature (see e.g. Wilkinson and Reinsch, 1971). In the problem of testing a linear hypothesis, they allow for the check of consistency of a linear hypothesis, check for testability of the hypothesis and for the computation of degrees of freedom of the associated statistics. These approaches are related to the work of Goldman and Zelen (1964), Zyskind and Martin (1969) and Rao (1971). The approaches are unified here in the sense that they include the well-known Gauß-model as a special case, as well as best linear constrained unbiased estimation (BLICUE).

Insight into the nature of a linear model under arbitrary variance-covariance structure is provided, and it is shown that every linear model under arbitrary variance can be reduced through a transformation and a subsequent reparametrization to a model with full rank design matrix and variance-covariance structure $\sigma^2\mathbf{I}$, which is statistically equivalent to the original model.

2.6.1 Solving generalized least squares problems

Observing that any admissible observation $\mathbf{y} \in C([\mathbf{X}:\mathbf{V}])$ under the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ can be written as

$$\mathbf{y} = \mathbf{y}_1 + \mathbf{y}_2 \tag{2.148}$$

where $\mathbf{y}_1 \in C(\mathbf{X})$ and $\mathbf{y}_2 \in C(\mathbf{VZ})$ (Lemma 2.4), and noting that the BLUE $\mathbf{X}\hat{\beta}$ for $\mathbf{X}\beta$ under the model (2.1) is given by $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = P_{\mathbf{X}|\mathbf{VZ}}\mathbf{y} = \mathbf{y}_1$, we observe that the computation of the BLUE $\mathbf{X}\hat{\beta}$ for $\mathbf{X}\beta$ amounts to solving the system of linear equations

(2.149) $$\mathbf{y} = [\mathbf{X}:\mathbf{VZ}] \begin{bmatrix} \beta \\ \lambda \end{bmatrix} .$$

This system is consistent for admissible $\mathbf{y}$, and $\mathbf{X}\hat{\beta}$ is BLUE for $\mathbf{X}\beta$ for any solution $\begin{bmatrix} \hat{\beta} \\ \hat{\lambda} \end{bmatrix}$ to (2.149).

If only estimation in a linear model is of interest, an efficient and numerically stable procedure to find the BLUE $\mathbf{X}\hat{\beta}$ for $\mathbf{X}\beta$ is to solve (2.149) using Householder-tranformations.

We propose the following algorithm:

Algorithm 2.37 (Schall)

BLU-estimation in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$

Step 0: Start with the system linear of equations

(2.150) $$\mathbf{y} = [\mathbf{X}:\mathbf{V}] \begin{bmatrix} \beta \\ \lambda \end{bmatrix}$$

Step 1: Compute the QR-factorization of $\mathbf{X}$ (see appendix), i.e. $\mathbf{X} = \mathbf{Q}_1 \begin{bmatrix} \mathbf{R}_1:\mathbf{D}_1 \\ \mathbf{0} : \mathbf{0} \end{bmatrix} \mathbf{M}_1$ and set

$$\mathbf{y}^{(1)} := \mathbf{Q}_1'\mathbf{y}$$

$$\mathbf{X}^{(1)} := \mathbf{Q}_1'\mathbf{X}\mathbf{M}_1 = \begin{bmatrix} \mathbf{R}_1:\mathbf{D}_1 \\ \mathbf{0} : \mathbf{0} \end{bmatrix}$$

$$\mathbf{V}^{(1)} := \mathbf{Q}_1'\mathbf{V}\mathbf{Q}_1$$

$$\beta^{(1)} := \mathbf{M}_1\beta$$

$$\lambda^{(1)} := \mathbf{Q}_1'\lambda$$

to obtain the system

(2.151) $$\mathbf{y}^{(1)} = \begin{bmatrix} \mathbf{R}_1:\mathbf{D}_1:\mathbf{V}_{11}^{(1)}:\mathbf{V}_{12}^{(1)} \\ \mathbf{0} : \mathbf{0} :\mathbf{V}_{21}^{(1)}:\mathbf{V}_{22}^{(1)} \end{bmatrix} \begin{bmatrix} \beta_1^{(1)} \\ \beta_2^{(1)} \\ \lambda_1^{(1)} \\ \lambda_2^{(1)} \end{bmatrix} ,$$

where $\mathbf{V}^{(1)}$ is partitioned conformably to the partitioning of $\mathbf{X}^{(1)}$ such that $\mathbf{V}_{22}^{(1)}$ is a square matrix of order $(n-r)\times(n-r)$, $r = rank(\mathbf{X})$.

Step 2: Drop the redundant unknowns $\beta_2^{(1)}$ and $\lambda_1^{(1)}$ and the corresponding variables $\begin{bmatrix} \mathbf{D}_1 \\ \mathbf{0} \end{bmatrix}$ and $\begin{bmatrix} \mathbf{V}_{11}^{(1)} \\ \mathbf{V}_{21}^{(1)} \end{bmatrix}$ to obtain the system

(2.152) $$\begin{bmatrix} \mathbf{y}_1^{(1)} \\ \mathbf{y}_2^{(1)} \end{bmatrix} = \begin{bmatrix} \mathbf{R}_1:\mathbf{V}_{12}^{(1)} \\ \mathbf{0} :\mathbf{V}_{22}^{(1)} \end{bmatrix} \begin{bmatrix} \beta_1^{(1)} \\ \lambda_2^{(1)} \end{bmatrix} ,$$

where $\mathbf{y}^{(1)}$ is conformably partitioned.

Step 3: Compute the QR-factorization of $\mathbf{V}_{22}^{(1)}$, i.e. $\mathbf{V}_{22}^{(1)} = \mathbf{Q}_2 \begin{bmatrix} \mathbf{R}_2 : \mathbf{D}_2 \\ \mathbf{0} : \mathbf{0} \end{bmatrix} \mathbf{M}_2$ and set

$$\mathbf{y}_1^{(2)} := \mathbf{y}_1^{(1)}$$

$$\mathbf{y}_2^{(2)} = \begin{bmatrix} \mathbf{y}_{21}^{(2)} \\ \mathbf{y}_{22}^{(2)} \end{bmatrix} := \mathbf{Q}_2' \mathbf{y}_2^{(1)}$$

$$\mathbf{V}_{12}^{(2)} = [\mathbf{V}_{121}^{(2)} : \mathbf{V}_{122}^{(2)}] := \mathbf{V}_{12}^{(1)}$$

$$\mathbf{V}_{22}^{(2)} := \mathbf{Q}_2' \mathbf{V}_{22}^{(1)} \mathbf{M}_2 = \begin{bmatrix} \mathbf{R}_2 : \mathbf{D}_2 \\ \mathbf{0} : \mathbf{0} \end{bmatrix}$$

$$\beta_1^{(2)} := \beta_1^{(1)}$$

$$\lambda_2^{(2)} = \begin{bmatrix} \lambda_{21}^{(2)} \\ \lambda_{22}^{(2)} \end{bmatrix} := \mathbf{M}_2 \lambda_2^{(1)}$$

to obtain the system

$$\begin{bmatrix} \mathbf{y}_1^{(2)} \\ \mathbf{y}_{21}^{(2)} \\ \mathbf{y}_{22}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{R}_1 : \mathbf{V}_{121}^{(2)} : \mathbf{V}_{122}^{(2)} \\ \mathbf{0} : \mathbf{R}_2 : \mathbf{D}_2 \\ \mathbf{0} : \mathbf{0} : \mathbf{0} \end{bmatrix} \begin{bmatrix} \beta_1^{(2)} \\ \lambda_{21}^{(2)} \\ \lambda_{22}^{(2)} \end{bmatrix} \tag{2.153}$$

Step 4: Drop the redundant unknows $\lambda_{22}^{(2)}$ and the corresponding variable, as well as the spurious equations $\mathbf{y}_{22}^{(2)} = \mathbf{0}$ to obtain the system

$$\begin{bmatrix} \mathbf{y}_1^{(2)} \\ \mathbf{y}_{21}^{(1)} \end{bmatrix} = \begin{bmatrix} \mathbf{R}_1 : \mathbf{V}_{121}^{(2)} \\ \mathbf{0} : \mathbf{R}_2 \end{bmatrix} \begin{bmatrix} \beta_1^{(2)} \\ \lambda_{21}^{(2)} \end{bmatrix} \tag{2.154}$$

Step 5: Compute the solution $\begin{bmatrix} \hat{\beta}_1^{(2)} \\ \hat{\lambda}_{21}^{(2)} \end{bmatrix}$ of the system (2.154).

A solution $\begin{bmatrix} \hat{\beta} \\ \hat{\lambda} \end{bmatrix}$ to the system (2.149) is given by

$$\hat{\beta} = \mathbf{M}_1 \begin{bmatrix} \hat{\beta}_1^{(2)} \\ \mathbf{0} \end{bmatrix}, \text{ and} \tag{2.155}$$

$$\hat{\lambda} = \mathbf{Q}_1 \mathbf{M}_2 \begin{bmatrix} \hat{\lambda}_{21}^{(2)} \\ \mathbf{0} \end{bmatrix}. \tag{2.156}$$

The BLUE for $\mathbf{X}\beta$ is given by $\mathbf{X}\hat{\beta}$, and an unbiased estimate for σ^2 is

$$\hat{\sigma}^2 = \hat{\lambda}_{21}^{(2)\prime} [\mathbf{V}_{121}^{(2)\prime} : \mathbf{R}_2'] \begin{bmatrix} \mathbf{V}_{121}^{(2)} \\ \mathbf{R}_2 \end{bmatrix} \hat{\lambda}_{21}^{(2)} \div s \quad , \tag{2.157}$$

where $s := rank\ (\mathbf{R}_2)$.

□

Notes on the steps of the algorithm:

Steps 1 and 2:

By multiplying $\mathbf{V}$ in $\mathbf{V}^{(1)}$: $= \mathbf{Q}'_1\mathbf{V}\mathbf{Q}_1$ from the right by $\mathbf{Q}_1$ in step 1 and by dropping $\begin{bmatrix} \mathbf{V}_{11}^{(1)} \\ \mathbf{V}_{21}^{(1)} \end{bmatrix}$ in step 2 we essentially compute $\mathbf{VZ}$ so that the system (2.152) is equivalent to the system (2.149) which the algorithm solves.

Step 3: The original system (2.150) is consistent if and only if $\mathbf{y}_2^{(2)} = \mathbf{0}$.

Step 4: The system (2.154) is upper triangular and this facilitates the computation of a solution in step 5.

Step 5: Usually, $\hat{\lambda}$ need not be computed since σ^2 can be estimated by (2.157) directly using $\hat{\lambda}_{21}{}^{(2)}$.

2.6.2 Transformation and Reparametrization for best linear unbiased estimation

In this section we propose two algorithms for BLU-estimation in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ which are especially useful when tests of linear hypotheses in the model are also of interest. In the first algorithm, BLU-estimation is performed by applying two transformations of the linear model, whereas in the second algorithm the BLU-estimation is performed through a transformation and a subsequent reparametrization of the model. We will call those approaches the transformation (T-) and reparametrization (R-) method respectively. A third algorithm allows for the testing of linear hypotheses, including a check of consistency and testability of the hypothesis, as well as the computation of the degrees of freedom of the F-statistic in question.

In the process, we will rely heavily on the results of Theorems 2.26 and 2.29 and associated Corollaries 2.26.1 and 2.29.1, to the effect that BLUE's and F-statistics are invariant under transformations and reparametrizations of a linear model.

The T-method

The BLUE $\mathbf{X}\hat{\beta}$ for $\mathbf{X}\beta$ in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ can be obtained by two subsequent transformations of the model.

Algorithm 2.38 (Schall and Dunne, 1986a)

BLU-estimation by transformation

Step 1: Choose a transformation $\mathbf{T}_1$ such that

(2.152) $$\mathbf{T}_1\mathbf{V}\mathbf{T}_1' = \begin{bmatrix} \mathbf{I}_k : \mathbf{0} \\ \mathbf{0} : \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{I} : \mathbf{0} \\ \mathbf{0} : \mathbf{0} \end{bmatrix}, \quad k = rank(\mathbf{V})$$

and set

(2.153) $$\mathbf{y}^{(1)} = \begin{bmatrix} \mathbf{y}_1^{(1)} \\ \mathbf{y}_2^{(1)} \end{bmatrix} := \mathbf{T}_1\mathbf{y}$$

(2.154) $$\mathbf{X}^{(1)} = \begin{bmatrix} \mathbf{X}_1^{(1)} \\ \mathbf{X}_2^{(1)} \end{bmatrix} := \mathbf{T}_1\mathbf{X} \quad \text{conformably partitioned,}$$

to obtain the model

(2.155) $$\begin{bmatrix} \mathbf{y}_1^{(1)} \\ \mathbf{y}_2^{(1)} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^{(1)} \\ \mathbf{X}_2^{(1)} \end{bmatrix} \beta + \mathbf{e}^{(1)}; \quad cov(\mathbf{e}^{(1)}) = \sigma^2 \begin{bmatrix} \mathbf{I} : \mathbf{0} \\ \mathbf{0} : \mathbf{0} \end{bmatrix}$$

Step 2: Choose a transformation

(2.156) $$\mathbf{T}_2 = \begin{bmatrix} \mathbf{I} : -\mathbf{A} \\ \mathbf{0} : \mathbf{I} \end{bmatrix}$$

and set

(2.157) $$\mathbf{y}^{(2)} = \begin{bmatrix} \mathbf{y}_1^{(2)} \\ \mathbf{y}_2^{(2)} \end{bmatrix} := \mathbf{T}_2 \begin{bmatrix} \mathbf{y}_1^{(1)} \\ \mathbf{y}_2^{(1)} \end{bmatrix} = \begin{bmatrix} \mathbf{y}_1^{(1)} - \mathbf{A}\mathbf{y}_2^{(1)} \\ \mathbf{y}_2^{(1)} \end{bmatrix}$$

(2.158) $$\mathbf{X}^{(2)} = \begin{bmatrix} \mathbf{X}_1^{(2)} \\ \mathbf{X}_2^{(2)} \end{bmatrix} := \mathbf{T}_2 \begin{bmatrix} \mathbf{X}_1^{(1)} \\ \mathbf{X}_2^{(1)} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^{(1)} - \mathbf{A}\mathbf{X}_2^{(1)} \\ \mathbf{X}_2^{(1)} \end{bmatrix}$$

such that $R(\mathbf{X}_1^{(2)}) \cap R(\mathbf{X}_2^{(2)}) = \{\mathbf{0}\}$. With

(2.159) $$\mathbf{T}_2 \begin{bmatrix} \mathbf{I} : \mathbf{0} \\ \mathbf{0} : \mathbf{0} \end{bmatrix} \mathbf{T}_2' = \begin{bmatrix} \mathbf{I} : \mathbf{0} \\ \mathbf{0} : \mathbf{0} \end{bmatrix}$$

we obtain model

(2.160) $$\begin{bmatrix} \mathbf{y}_1^{(2)} \\ \mathbf{y}_2^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^{(2)} \\ \mathbf{X}_2^{(2)} \end{bmatrix} \beta + \mathbf{e}^{(2)}; \quad cov(\mathbf{e}^{(2)}) = \sigma^2 \begin{bmatrix} \mathbf{I} : \mathbf{0} \\ \mathbf{0} : \mathbf{0} \end{bmatrix}$$

Step 3: The BLUE $\mathbf{X}^{(2)}\hat{\beta}$ for $\mathbf{X}^{(2)}\beta$ in model (2.160) is the straightforward least-squares-solution, i.e.

(2.161) $$\mathbf{X}^{(2)}\hat{\beta} = \mathbf{X}^{(2)}(\mathbf{X}^{(2)'}\mathbf{X}^{(2)})^{-}\mathbf{X}^{(2)'}\mathbf{y}$$

and thus the BLUE $\mathbf{X}\hat{\beta}$ for $\mathbf{X}\beta$ in the original model (2.1) is

(2.162) $$\mathbf{X}\hat{\beta} = \mathbf{T}_1^{-1}\mathbf{T}_2^{-1}\mathbf{X}^{(2)}\hat{\beta} = \mathbf{T}^{-1}\mathbf{X}^{(2)}\hat{\beta}, \quad \text{where } \mathbf{T} := \mathbf{T}_2\mathbf{T}_1.$$

□

Notes on the steps of the algorithm:

Step 1: $\mathbf{T}_1$ can be obtained by standard methods, and we name but one in particular: let

(2.163) $$\mathbf{V} = \mathbf{P}\begin{bmatrix} \triangle : \mathbf{0} \\ \mathbf{0} : \mathbf{0} \end{bmatrix}\mathbf{P}'$$

be the complete singular value decomposition (SVD) of $\mathbf{V}$, i.e. $\mathbf{P}$ is an orthogonal matrix of the same order as $\mathbf{V}$ and $\triangle$ is a $k \times k$ diagonal matrix containing the positive eigenvalues of $\mathbf{V}$. Then

(2.164) $$\mathbf{T}_1 = \begin{bmatrix} \triangle^{-1/2} : \mathbf{0} \\ \mathbf{0} \quad : \mathbf{I} \end{bmatrix}\mathbf{P}' \quad \text{(say)}.$$

Program libraries like NAG and IMSL provide the SVD of a matrix, using the algorithm of Golub and Reinsch (1970) which is efficient and numerically stable.

Step 2: Without loss of generality, $\mathbf{A}$ can be taken as

(2.165) $$\mathbf{A} = \mathbf{X}_1^{(1)}\mathbf{X}_2^{(1)\prime}(\mathbf{X}_2^{(1)}\mathbf{X}_2^{(1)\prime})^- .$$

Then

(2.166) $$\mathbf{X}_1^{(2)} = \mathbf{X}_1^{(1)} - \mathbf{A}\mathbf{X}_2^{(1)} = \mathbf{X}_1^{(1)}(\mathbf{I}_p - \mathbf{X}_2^{(1)\prime}(\mathbf{X}_2^{(1)}\mathbf{X}_2^{(1)\prime})^-\mathbf{X}_2^{(1)}) .$$

Thus $R(\mathbf{X}_1^{(2)}) \perp R(\mathbf{X}_2^{(2)})$ which implies of course $R(\mathbf{X}_1^{(2)}) \cap R(\mathbf{X}_2^{(2)}) = \{\mathbf{0}\}$.

The matrix $\mathbf{A}$ is most effectively computed using the QR-factorization of $\mathbf{X}_2^{(1)\prime}$ (see appendix). If

(2.167) $$\mathbf{X}_2^{(1)\prime} = \mathbf{Q}_1[\mathbf{R}_1 : \mathbf{D}]\mathbf{M}$$

as in (2.190), then

(2.168) $$\mathbf{A} = \mathbf{X}_1^{(1)}\mathbf{Q}_1[(\mathbf{R}_1^{-1})' : \mathbf{0}']\mathbf{M} \quad \text{(say), and}$$

(2.169) $$\mathbf{X}_1^{(2)} = \mathbf{X}_1^{(1)}(\mathbf{I}_p - \mathbf{Q}_1\mathbf{Q}_1') , \quad \text{from (2.191).}$$

Step 3: No further computations are necessary to obtain the inverses $\mathbf{T}_1^{-1}$ and $\mathbf{T}_2^{-1}$, since if

(2.170) $$\mathbf{T}_1 = \begin{bmatrix} \triangle^{-1/2} : \mathbf{0} \\ \mathbf{0} \quad : \mathbf{I} \end{bmatrix}\mathbf{P}' \text{ then } \mathbf{T}_1^{-1} = \mathbf{P}\begin{bmatrix} \triangle^{1/2} : \mathbf{0} \\ \mathbf{0} \quad : \mathbf{I} \end{bmatrix},$$

and if

(2.171) $$\mathbf{T}_2 = \begin{bmatrix} \mathbf{I} : -\mathbf{A} \\ \mathbf{0} : \mathbf{I} \end{bmatrix} \text{ then } \mathbf{T}_2^{-1} = \begin{bmatrix} \mathbf{I} : \mathbf{A} \\ \mathbf{0} : \mathbf{I} \end{bmatrix} .$$

We note that $\mathbf{X}^{(2)}\hat{\beta}$ in (2.161) can be computed using the QR-factorization of $\mathbf{X}^{(2)}$, as described by Businger and Golub (1965).

Linking this approach to known theory, we note that Goldman and Zelen (1964) obtained the transformed model (2.155), and in effect proceeded by minimizing

(2.172) $$(\mathbf{y}_1^{(1)} - \mathbf{X}_1^{(1)}\beta)'(\mathbf{y}_1^{(1)} - \mathbf{X}_1^{(1)}\beta) \quad \text{subject to}$$

(2.173) $$\mathbf{y}_2^{(1)} = \mathbf{X}_2^{(1)}\beta \ ,$$

i.e. by solving constrained least squares, whereas we proceed in step 2 of the algorithm by disconnecting the constraints from the stochastic part of the data.

It is in step 2 of the algorithm, that BLICU-estimation may be effected. If we have the constraints $\mathbf{c} = \mathbf{C}\beta$, we augment $\mathbf{y}_2^{(1)}$ by $\mathbf{c}$ and $\mathbf{X}_2^{(1)}$ by $\mathbf{C}$ and proceed with

(2.174) $$\mathbf{y}_2^{(c)} = \begin{bmatrix} \mathbf{y}_2^{(1)} \\ \mathbf{c} \end{bmatrix} \text{ and } \mathbf{X}_2^{(c)} = \begin{bmatrix} \mathbf{X}_2^{(1)} \\ \mathbf{C} \end{bmatrix} .$$

Further, we note that in (2.162) $\mathbf{X}\hat{\beta}$ can be written as

(2.175) $$\begin{aligned} \mathbf{X}\hat{\beta} &= \mathbf{T}^{-1}\mathbf{X}^{(2)}\hat{\beta} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{T}'\mathbf{T}\mathbf{X})^{-}\mathbf{X}'\mathbf{T}'\mathbf{T}\mathbf{y} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{V}^{*}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{*}\mathbf{y} \ , \quad \mathbf{V}^{*} = \mathbf{T}'\mathbf{T} \ . \end{aligned}$$

$\mathbf{V}^*$ is easily shown to be a g-inverse of $\mathbf{V}$, and it can be shown that $\mathbf{V}\mathbf{V}^*\mathbf{X} = \mathbf{X}\mathbf{B}$, for some $\mathbf{B}$, and that $rank(\mathbf{X}'\mathbf{V}^*\mathbf{X}) = rank(\mathbf{X})$. Thus $\mathbf{V}^*$ is a g-inverse of $\mathbf{V}$ as given by Theorem 2.16. (Algorithm 2.38 can then be used for the computation of such a g-inverse.)

That (2.161) gives the BLUE $\mathbf{X}^{(2)}\hat{\beta}$ for $\mathbf{X}^{(2)}\beta$ in model (2.160) is obvious if $R(\mathbf{X}_1^{(2)}) \perp R(\mathbf{X}_2^{(2)})$. But $R(\mathbf{X}_1^{(2)}) \cap R(\mathbf{X}_2^{(2)}) = \{\mathbf{0}\}$ is sufficient indeed for (2.161) to be true, which follows from Theorem 1.6.

Zyskind and Martin (1969), Rao (1971) and Dunne (1982) characterized the g-inverses $\mathbf{V}^*$ of $\mathbf{V}$ which yield the BLUE $\mathbf{X}\hat{\beta}$ for $\mathbf{X}\beta$ in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ as the solution of the GNE's (2.61), even though $\mathbf{V}$ is singular. In contrast, Algorithm 2.38 characterizes a class of transformations $\mathbf{T} = \mathbf{T}_2\mathbf{T}_1$ of the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ as given in (2.162) such that in the transformed model $(\tilde{\mathbf{y}}, \tilde{\mathbf{X}}\beta, \sigma^2\tilde{\mathbf{V}})$ the BLUE $\tilde{\mathbf{X}}\hat{\beta}$ for $\tilde{\mathbf{X}}\beta$ is obtained as the simple least-squares estimate. (This construction is a parallel to the idea of Aitken (1935), for the case of singular $\mathbf{V}$.) The remarks above however, notably the relationship (2.175), show that the two approaches are closely connected.

The R-method

The BLUE $\mathbf{X}\hat{\beta}$ for $\mathbf{X}\beta$ in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ can be obtained by a transformation and subsequent reparametrization of the model:

Algorithm 2.39 (Schall and Dunne, 1986a)

BLU-estimation by reparametrization

Step 1: Same as step 1 in Algorithm 2.38.

Step 2: Compute the QR-factorization of $\mathbf{X}_2^{(1)\prime}$, i.e. $\mathbf{X}_2^{(1)} = \mathbf{M}[\mathbf{R}' : \mathbf{0}']\mathbf{Q}'$ as in (2.190), and reparametrize the model (2.155) as

$$\mathbf{X}^* := \mathbf{X}^{(1)}\mathbf{Q} \; ; \tag{2.176}$$

$$\beta^* := \mathbf{Q}'\beta$$

to obtain the model

$$\begin{bmatrix} \mathbf{y}_1^{(1)} \\ \mathbf{y}_2^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{11}^* : \mathbf{X}_{12}^* \\ \mathbf{R}_1' : \mathbf{0}' \end{bmatrix} \begin{bmatrix} \beta_1^* \\ \beta_2^* \end{bmatrix} + \mathbf{e}^{(1)}; \; cov(\mathbf{e}^{(1)}) = \sigma^2 \begin{bmatrix} \mathbf{I} : \mathbf{0} \\ \mathbf{0} : \mathbf{0} \end{bmatrix} \tag{2.177}$$

where $\mathbf{y}_2^{(2)} := \mathbf{M}\mathbf{y}_2^{(1)}$, and β_1^* is a $r_1 \times 1$-vector, $r_1 = rank(\mathbf{X}_2^{(1)})$.

Step 3: Compute $\beta_1^* = \hat{\beta}_1^*$ from

$$\mathbf{y}_2^{(2)} = \begin{bmatrix} \mathbf{R}_1' \\ \mathbf{D}' \end{bmatrix} \beta_1^* \text{ (}\mathbf{D}, \mathbf{R}_1 \text{ as in the appendix).} \tag{2.178}$$

Step 4: The BLUE $\mathbf{X}_{12}^* \hat{\beta}_2^*$ for $\mathbf{X}_{12}^* \beta_2^*$ is

$$\mathbf{X}_{12}^* \hat{\beta}_2^* = \mathbf{X}_{12}^* (\mathbf{X}_{12}^{*\prime}\mathbf{X}_{12}^*)^- \mathbf{X}_{12}^{*\prime}(\mathbf{y}_1^{(1)} - \mathbf{X}_{11}^* \hat{\beta}_1^*) \tag{2.179}$$

and the BLUE $\mathbf{X}\hat{\beta}$ for $\mathbf{X}\beta$ is

$$\hat{\beta} = \mathbf{T}_1^{-1} \begin{bmatrix} \mathbf{X}_{11}^* : \mathbf{X}_{12}^* \\ \mathbf{R}_1' : \mathbf{0}' \end{bmatrix} \hat{\beta}^* \tag{2.180}$$

□

Notes on the steps of the algorithm:

Step 1: Same as for step 1 of Algorithm 2.38.

Step 2: Of course, in a computer application the components of $\mathbf{y}_2^{(1)}$ and the rows of $\mathbf{X}_2^{(1)}$ need not actually explicity be permuted by $\mathbf{M}$ to obtain $\mathbf{R}_1$ upper triangular. Only a vector $\mathbf{m}$ of the same dimension $(n-k)$ as $\mathbf{y}_2^{(1)}$ is needed to store the appropriate permutation, which will later be used in the computation of $\hat{\beta}_1^*$ in step 3.

Step 3: If $\mathbf{y}_2^{(2)\prime} = [\mathbf{y}_{21}^{(2)\prime} : \mathbf{y}_{22}^{(2)\prime}]$ is partitioned conformably with $[\mathbf{R}_1 : \mathbf{D}]$, then of course $\beta_1^* = \hat{\beta}_1^*$ is computed from

$$\mathbf{y}_{21}^{(2)} = \mathbf{R}_1' \beta_1^* \quad , \tag{2.181}$$

which is particularly easy and stable since $\mathbf{R}_1'$ is triangular.

Step 4: $\mathbf{X}_{12}^* \hat{\beta}_2^*$ is again computed using the QR-factorization of $\mathbf{X}_{12}^*$.

This approach can also be linked to the work of Zyskind and Martin (1969), Rao (1971) and Dunne (1982): since in model (2.177) we have

$$(2.182) \qquad C\begin{pmatrix}\mathbf{X}_{12}^{*}\\ \mathbf{0}\end{pmatrix} \subset C(\mathbf{V}) \text{ and } C\begin{pmatrix}\mathbf{X}_{11}^{*}\\ \mathbf{R}'\end{pmatrix} \cap C(\mathbf{V}) = \{\mathbf{0}\},\ \mathbf{V} = \begin{bmatrix}\mathbf{I} : \mathbf{0}\\ \mathbf{0} : \mathbf{0}\end{bmatrix}$$

we conclude that

$$(2.183) \qquad \mathbf{V}^{*} = \left(\mathbf{V} + \begin{bmatrix}\mathbf{X}_{11}^{*}\\ \mathbf{R}'\end{bmatrix}[\mathbf{X}_{11}^{*\prime} : \mathbf{R}]\right)^{-}$$

is a g-inverse of $\mathbf{V}$ as given by the above authors in model (2.177), i.e. $\mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{+}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{+}\mathbf{y}$ is the BLUE for $\mathbf{X}\beta$ in model (2.1), where $\mathbf{V}^{+} = \mathbf{T}_1'\mathbf{V}^{*}\mathbf{T}_1$. Thus Algorithm 2.39 is also a method to compute such a g-inverse.

Reduction of a linear model and testing of a linear hypothesis

With steps 1 and 2 of Algorithm 2.39 we have reduced the linear model (2.1) to the form (2.177). Since $\beta_1^{*} = \mathbf{b}_1 := (\mathbf{R}_1')^{-1}\mathbf{y}_{21}^{(2)}$ with probability 1, the model

$$(2.184) \qquad (\mathbf{y}_1^{(1)} - \mathbf{X}_{11}^{*}\mathbf{b}_1) = \mathbf{X}_{12}^{*}\beta_2^{*} + \mathbf{e}_1^{(1)};\quad cov(\mathbf{e}_1^{(1)}) = \sigma^2\mathbf{I}$$

is statistically equivalent to (2.177), under normal theory.

By a further reparametrization of (2.184), we can obtain a model with full rank design matrix (if $\mathbf{X}_{12}^{*}$ is not of full rank in the first place).

If we wish to test the linear hypothesis H_0: $\mathbf{L}\beta = \mathbf{c}$ we can proceed as follows:

Algorithm 2.40 (Schall and Dunne, 1986a)

Testing of the hypothesis $H_0 : \mathbf{L}\beta = \mathbf{c}$

Step 1: Reparametrize H_0 as $\mathbf{L}^{*} = \mathbf{L}\mathbf{Q}$, $\beta^{*} = \mathbf{Q}'\beta$ where $\mathbf{Q}$ follows from step 2 of Algorithm 2.39. Now, conformably partitioned, H_0 can be written as

$$(2.185) \qquad H_0\colon [\mathbf{L}_1^{*} : \mathbf{L}_2^{*}]\begin{bmatrix}\beta_1^{*}\\ \beta_2^{*}\end{bmatrix} = \mathbf{c}\ .$$

Step 2: Since $\mathbf{L}_1^{*}\beta_1^{*}$ is constant and known (β_1^{*} from step 3 of Algorithm 2.39), H_0 can be written as

$$(2.186) \qquad H_0\colon \mathbf{L}_2^{*}\beta_2^{*} = (\mathbf{c} - \mathbf{L}_1^{*}\beta_1^{*})\ .$$

Step 3: (i) check for consistency:

$$(2.187) \qquad H_0 \text{ consistent} \Leftrightarrow (\mathbf{c} - \mathbf{L}_1^{*}\beta_1^{*}) \in C(\mathbf{L}_2^{*})\ .$$

(ii) check for testability:

(2.188) $$H_0 \text{ testable} \iff R(\mathbf{L}_2^*) \subset R(\mathbf{X}_{12}^*) .$$

(iii) degrees of freedom for hypothesis:

(2.189) $$df = rank(\mathbf{L}_1^*) .$$

(iv) test statistic:

will be provided by any standard regression program: simply test H_0: $\mathbf{L}_2^*\beta_2^* = (\mathbf{c} - \mathbf{L}_1^*\beta_1^*)$ in the model (2.184) with variance-covariance structure $\sigma^2\mathbf{I}$.

□

2.6.3 APPENDIX

Let $\mathbf{X}$ be a matrix of order $n \times p$ with rank $r \leq min(n, p)$. Then by applying r Householder-transformations, $\mathbf{X}$ can be decomposed as

(2.190) $$\mathbf{X} = \mathbf{Q} \begin{bmatrix} \mathbf{R}_1 : \mathbf{D} \\ \mathbf{0}_1 : \mathbf{0}_2 \end{bmatrix} \mathbf{M} = \mathbf{Q} \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix} \mathbf{M}$$

where $\mathbf{Q}$ orthogonal of order $n \times n$
$\mathbf{R}_1$ upper triangular of order $r \times r$
$\mathbf{D}$ some matrix of order $r \times (p-r)$
$\mathbf{0}_1$ a null-matrix of order $(n-r) \times r$
$\mathbf{0}_2$ a null-matrix of order $(n-r) \times (p-r)$
$\mathbf{R} = [\mathbf{R}_1 : \mathbf{D}]$, $\mathbf{0} = [\mathbf{0}_1 : \mathbf{0}_2]$
$\mathbf{M}$ a permutation matrix of order $p \times p$, permuting the columns of $\mathbf{Q}'\mathbf{X}$ to ensure that $\mathbf{R}_1$ is upper triangular.

Some or all of the matrices $\mathbf{D}$, $\mathbf{0}_1$, $\mathbf{0}_2$ may vanish:

$\mathbf{D}$ vanishes if $r = p$, i.e. $\mathbf{X}$ is of full column rank
$\mathbf{0}_1$ vanishes if $r = n$, i.e. $\mathbf{X}$ is of full row rank
$\mathbf{0}_2$ vanishes if $\mathbf{D}$ or $\mathbf{0}_1$ vanishes.

$\mathbf{M}$ may be taken as $\mathbf{M} = \mathbf{I}_p$ if the first r columns of $\mathbf{X}$ are linearly independent.

If $\mathbf{Q} = [\mathbf{Q}_1 : \mathbf{Q}_2]$ is conformably partitioned, i.e. $\mathbf{Q}_1$ is of order $n \times r$ and $\mathbf{Q}_2$ is of order $n \times (n-r)$, then $\mathbf{X} = \mathbf{Q}_1[\mathbf{R}_1 : \mathbf{D}]\mathbf{M}$.

Important are the following identities:

(2.191) (i) $\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}' = \mathbf{Q}_1\mathbf{Q}_1'$ and

(ii) $$(\mathbf{X}'\mathbf{X})^-\mathbf{X}' = \mathbf{M}' \begin{bmatrix} \mathbf{R}_1^{-1} \\ \mathbf{0} \end{bmatrix} \mathbf{Q}_1' = \mathbf{M}' \begin{bmatrix} \mathbf{R}_1^{-1}\mathbf{Q}_1' \\ \mathbf{0} \end{bmatrix} \text{ (say)} .$$

CHAPTER 3

Outliers and Influence under Arbitrary Known Variance

The problem of outliers in a linear model under arbitrary known variance-covariance structure was first considered by Dunne (1982) who showed how to adjust the model for what we will call distributional outliers in the data. By fitting as additional variables to the model those columns of the variance-covariance matrix $\mathbf{V}$ of the observed variate $\mathbf{y}$ which correspond to the data points suspected to be outlying, one obtains an F-statistic providing a test for additional fit due to the new variables. This test was shown to be a generalization of the usual F-test proposed by many authors to be applied when testing for outliers in the classical Gauß-model with $\mathbf{V} = \mathbf{I}$, thus including the case of arbitrary known variance-covariance structure. Similarly, the proposed adjusted model was observed to be a generalization of a formulation by John and Draper (1978) and Cook and Weisberg (1979).

In the development, a different type of outlier, by "additive shift", was briefly mentioned and relationships to missing data estimation were examined.

In the following, we will rest substantially upon insights provided by Dunne (1982) when we distinguish three types of outlier: *Distributional Outliers*, *Outliers by Additive Shifts* and *Transformational Outliers*. These three types of outlier can only be distinguished when the variance-covariance structure $\mathbf{V}$ of the model is not a diagonal matrix.

For each type of outlier, we will specify the appropriate reduced data model, relate it to missing data problems and generalize the recursive residuals approach by John and Draper (1978) originally applied in the $\mathbf{V} = \mathbf{I}$ case.

Using the additional sum of squares principle, the tests in question can be carried out without actually explicitly fitting the corresponding adjusted model, and estimates adjusted for outlier effects as well as recursive residuals can be computed in an economical manner.

A unified approach to outliers in a linear model is achieved by showing that the testing for outliers can be carried out involving the linear model with the vector $\hat{\mathbf{e}}$ of estimated residuals from the original linear model as observed variate, zero mean, and the variance-covariance structure $\mathbf{N}$ of $\hat{\mathbf{e}}$. The matrix $\mathbf{N}$ is singular in general, whether or not $\mathbf{V}$ is singular, and thus there is no basic distinction between linear models with singular and nonsingular variance-covariance struture $\mathbf{V}$ respectively.

In the second part of this chapter, the statistics proposed by Cook (1977) and Andrews and Pregibon (1978) for the detection of influential observations in a Gauß-model are generalized for the linear model under arbitrary known variance-covariance structure, and corresponding to three types of outlier we obtain three different versions in general of each of the above named statistics.

A third part will be devoted to a brief discussion of outliers and influence in the normal variance components model, extending the ideas from the case where $\mathbf{V}$ is known to the case where $\mathbf{V}$ is only known up to some additive structure.

Generally, it will transpire that the notion of outliers by additive shifts is strongly related to classical missing data estimation, whereas transformational outliers are related to principal component analysis. Distributional outliers, however, excite the mathematical mind by the beauty of the results obtained, as they yield the true counterparts of statistics and estimates known from the treatment of outliers in the Gauß-model.

The development that follows is essentially possible without any distributional assumptions. The main thrust of the work, however, is for the normal linear model, and whenever distributional statements are made the assumption of the normality of the observed variate $\mathbf{y}$ is implicit.

Since the test-statistics relating to different types of outlier will in general yield different results when the observations are correlated, it will be here, in the linear model with nondiagonal variance-covariance structure that an outlier from the tail of the underlying probability distribution (but an otherwise valid observation) and an outlier by an invasive mean shift can be distinguished for the first time. This distinction is by assumption impossible when the observations are independently distributed, and thus the work by Dunne (1982) marks a turning point in the statistical theory of outliers.

3.1 OUTLIERS IN THE GENERAL LINEAR MODEL

We consider the linear model (2.1) conformably partitioned as

$$(3.1) \qquad \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \beta + \mathbf{e} \; ; \quad cov(\mathbf{e}) = \sigma^2 \begin{bmatrix} \mathbf{V}_{11} : \mathbf{V}_{12} \\ \mathbf{V}_{21} : \mathbf{V}_{22} \end{bmatrix}$$

where $\mathbf{y}_2$ and $\mathbf{e}_2$ are $k \times 1$-vectors, $\mathbf{X}_2$ is $k \times p$ and $\mathbf{V}_{22}$ is a $k \times k$-matrix.

We assume that the model is consistent, that is, $\mathbf{y} \in C([\mathbf{X}:\mathbf{V}])$. Equivalently the sure equations in the model are consistent with the data.

The observation(s) $\mathbf{y}_2$ are suspected to be outlying, in some sense which we will specify below, and the remainder of the data, $\mathbf{y}_1$, is considered to be "clean". Naturally, in general some rearrangement of the components of $\mathbf{y}$, and a corresponding rearrangement of the rows of $\mathbf{X}$ and the rows and columns of $\mathbf{V}$ is needed to achieve the division in (3.1).

Implicit in the division in (3.1) is the assumption that we know the number of outliers in the data, k, and that we know further where these outliers are in the data, if there are any. Essentially we will not address the problem of determining the number of outliers in the model, and locating them in the data, if k and the location are unknown. These problems are handled, using the body of theory provided by this chapter, completely analogous to the methods used in the case where $\mathbf{V} = \mathbf{I}$. However, by generalizing the idea of recursive residuals we provide a very useful tool for this task.

Distributional Outliers

Suppose initially that the observations $\mathbf{y}_2$ in (3.1) are rare but otherwise valid observations from the tail of the underlying probability distribution of the variate $\mathbf{y}_2$. Then we assume for $\mathbf{y}_2$ the linear model

$$\mathbf{y}_2 = \mathbf{X}_2\beta + \mathbf{e}_2 \; ; \quad cov(\mathbf{e}_2) = \sigma^2\mathbf{V}_{22} \tag{3.2}$$

where the error term $\mathbf{e}_2$ is considered to be significantly large.

In the situation (3.2), where $E(\mathbf{y}_2)$ is still assumed to be correctly specified by $\mathbf{X}_2\beta$ but $\mathbf{e}_2$ is considered to be significantly large or outlying, we call $\mathbf{y}_2$ a *Distributional Outlier*, or *D*-outlier.

Outliers by Additive Shifts

In a different situation we allow for an additive shift in the mean of $\mathbf{y}_2$, causing the outlier, which is invasive and unrelated to the presumed linear model (3.1).

Then the linear model for $\mathbf{y}_2$, adjusted for the mean shift is

$$\mathbf{y}_2 = \mathbf{X}_2\beta + \lambda + \mathbf{e}_2 \; ; \quad cov(\mathbf{e}_2) = \sigma^2\mathbf{V}_{22} \tag{3.3}$$

where λ is an arbitrary and generally unknown and unobservable vector. In the situation (3.3) we call $\mathbf{y}_2$ an *Outlier by Additive Shift*, or *A*-outlier.

Transformational Outliers

Finally, departing in a way from the model (3.1) we consider the linear model for the principal components (PC's) $\mathbf{y}^*$ of $\mathbf{y}$: if

$$\mathbf{V} = \mathbf{P} \triangle \mathbf{P}' = [\mathbf{P}_1 : \mathbf{P}_2] \begin{bmatrix} \triangle_1 & \\ & \triangle_2 \end{bmatrix} \begin{bmatrix} \mathbf{P}_1' \\ \mathbf{P}_2' \end{bmatrix} \tag{3.4}$$

is the singular value decomposition (SVD) of $\mathbf{V}$, then the PC's $\mathbf{y}^*$ of $\mathbf{y}$ are defined as $\mathbf{y}^* = \mathbf{P}'\mathbf{y}$ and the corresponding linear model is obtained by transforming (3.1) by $\mathbf{P}'$. Conformably partitioned we write

(3.5) $$\begin{bmatrix} \mathbf{y}_1^* \\ \mathbf{y}_2^* \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^* \\ \mathbf{X}_2^* \end{bmatrix} \beta + \mathbf{e}^* \; ; \quad cov(\mathbf{e}^*) = \sigma^2 \begin{bmatrix} \triangle_1 : \mathbf{0} \\ \mathbf{0} : \triangle_2 \end{bmatrix}$$

If $\mathbf{y}_2^*$ is suspected to be outlying under the model (3.5), in the usual sense since $cov(\mathbf{y}^*)$ is diagonal and D- and A-outliers are undistinguishable, we call $\mathbf{y}_2^*$, or equivalently $\mathbf{P}_2'\mathbf{y}$, a *Transformational Outlier*, or T-outlier, since $\mathbf{y}^*$ has been obtained by a transformation of $\mathbf{y}$. Clearly the effect of such outliers on the data vector $\mathbf{y}$ is smeared across all the observations and not just a corresponding or conformable subvector of $\mathbf{y}$.

3.1.1 Distributional Outliers

The adjusted model

For the D-outlier $\mathbf{y}_2$ we assumed the linear model

(3.2) $$\mathbf{y}_2 = \mathbf{X}_2\beta + \mathbf{e}_2 \; ; \quad cov(\mathbf{e}_2) = \sigma^2\mathbf{V}_{22}$$

where $\mathbf{e}_2$ is considered to be significantly large.

The linear model for $\mathbf{y}_1$ can be written as

(3.6) $$\begin{aligned} \mathbf{y}_1 &= \mathbf{X}_1\beta + \mathbf{e}_1 \\ &= \mathbf{X}_1\beta + \mathbf{V}_{12}\mathbf{V}_{22}^{-}\mathbf{e}_2 + \mathbf{e}_1 - \mathbf{V}_{12}\mathbf{V}_{22}^{-}\mathbf{e}_2 \\ &= \mathbf{X}_1\beta + \mathbf{V}_{12}\mathbf{V}_{22}^{-}\mathbf{e}_2 + \tilde{\mathbf{e}}_1 \; ; \end{aligned}$$

Clearly, $\tilde{\mathbf{e}}_1$ and $\mathbf{e}_2$ are uncorrelated and

(3.7) $$cov(\tilde{\mathbf{e}}_1) = \sigma^2(\mathbf{V}_{11} - \mathbf{V}_{12}\mathbf{V}_{22}^{-}\mathbf{V}_{21}) = \sigma^2\tilde{\mathbf{V}}_{11} \; .$$

This can be verified by writing $\tilde{\mathbf{e}}_1$ as

(3.8) $$\tilde{\mathbf{e}}_1 = [\mathbf{I} : -\mathbf{V}_{12}\mathbf{V}_{22}^{-}] \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix} = \mathbf{T}_1\mathbf{e}$$

and observing that

(3.9) $$cov(\tilde{\mathbf{e}}_1) = \sigma^2\mathbf{T}_1\mathbf{V}\mathbf{T}_1' = \sigma^2\tilde{\mathbf{V}}_{11} \; , \quad \text{and}$$

(3.10) $$cov(\tilde{\mathbf{e}}_1, \mathbf{e}_2) = \sigma^2[\mathbf{I} : -\mathbf{V}_{12}\mathbf{V}_{22}^{-}] \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} = \mathbf{0} \; .$$

We can interpret $\tilde{\mathbf{e}}_1$ as the variate $\mathbf{e}_1$ adjusted for the covariate $\mathbf{e}_2$, and $\mathbf{X}_1\beta + \mathbf{V}_{12}\mathbf{V}_{22}^{-}\mathbf{e}_2$ can be interpreted as the conditional expectation of $\mathbf{y}_1 \mid \mathbf{e}_2$.

Combining the model part (3.6) and (3.2) to a common linear model for $\mathbf{y}$, we obtain (3.1) in the alternative form

(3.11) $$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \beta + \begin{bmatrix} \mathbf{V}_{12} \\ \mathbf{V}_{22} \end{bmatrix} \mathbf{V}_{22}^{-}\mathbf{e}_2 + \begin{bmatrix} \tilde{\mathbf{e}}_1 \\ \mathbf{0} \end{bmatrix} .$$

If $\mathbf{e}_2$ is assumed to be significantly large, an improvement to the model (3.1) would be to fit the additional variables $\mathbf{V}_2 = [\mathbf{V}_{21} : \mathbf{V}_{22}]'$ to obtain the augmented model (Dunne, 1982)

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 : \mathbf{V}_{12} \\ \mathbf{X}_2 : \mathbf{V}_{22} \end{bmatrix} \begin{bmatrix} \beta \\ \lambda \end{bmatrix} + \mathbf{e} \; ; \quad cov(\mathbf{e}) = \sigma^2 \mathbf{V} \tag{3.12}$$

We observe that this model constitutes a generalization of the approach by John and Draper (1978) and Cook and Weisberg (1979), who fitted the additional variables $\mathbf{V}_2 = [\mathbf{0} : \mathbf{I}]'$ to adjust for the possible outliers $\mathbf{y}_2$ in the case $\mathbf{V} = \mathbf{I}$. Of course, the variables $[\mathbf{0} : \mathbf{I}_k]'$ form the last k columns of the variance-covariance structure $\mathbf{V} = \mathbf{I}_n$ of a Gauß-model, whereas the additional variables fitted in (3.12) form the last k columns of the variance-covariance structure $\mathbf{V}$ of the general linear model (3.1).

We note that all the sure information contained in model (3.1) is preserved in model (3.12) (Lemma 2.31), since trivially $C(\mathbf{V}_2) \subset C(\mathbf{V})$. Thus fitting the additional variables $\mathbf{V}_2$ as in model (3.12) is always admissible in the sense of Lemma 2.31, for any subset $\mathbf{y}_2$ of the data. The method amounts to an assumption that $\mathbf{y}_2$ is anomalous, but legitimate, and that we wish to adjust the model to compensate for the anomaly.

Testing and adjusted estimates

With respect to the linear model (3.12) we test the hypothesis $\mathbf{H_0}$: $\lambda = \mathbf{0}$, which constitutes the test that $\mathbf{y}_2$ is not a D-outlier against the alternative that $\mathbf{y}_2$ is a D-outlier. The hypothesis is consistent since $C(\mathbf{V}_2) \subset C(\mathbf{V}) \subset C([\mathbf{X} : \mathbf{V}])$ (Lemma 2.36).

Theorem 3.1 (Dunne, 1982)

The F-statistic associated with the hypothesis $\mathbf{H_0}$: $\lambda = \mathbf{0}$ in the linear model (3.12) is

$$F = \frac{\hat{\mathbf{e}}_2' \mathbf{N}_{22}^- \hat{\mathbf{e}}_2}{\hat{\mathbf{e}}' \mathbf{V}^- \hat{\mathbf{e}} - \hat{\mathbf{e}}_2' \mathbf{N}_{22}^- \hat{\mathbf{e}}_2} \cdot \frac{r([\mathbf{X} : \mathbf{V}]) - r(\mathbf{X}) - r(\mathbf{N}_{22})}{r(\mathbf{N}_{22})} \tag{3.13}$$

where $\mathbf{N}_{22}$ is the trailing principal $k \times k$-submatrix of

$$\mathbf{N} = \mathbf{V} + \mathbf{X}\mathbf{U}\mathbf{X}' - \mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^- \mathbf{X}' \tag{3.14}$$

which, up to the scale factor σ^2, is the variance-covariance matrix of $\hat{\mathbf{e}}$.

Proof: (Schall)

Using Theorem 2.33, we must only show that the additional sum of squares SSA due to fitting the new variables $\mathbf{A} = \mathbf{V}_2$ in the model (3.12) is given by

$$SSA = \hat{\mathbf{e}}_2' \mathbf{N}_{22}^- \hat{\mathbf{e}}_2 \; . \tag{3.15}$$

But with Corollary 2.33.1 (b) we have

$$SSA = \hat{\mathbf{e}}' \mathbf{V}^* \mathbf{V}_2 (\mathbf{V}_2' \mathbf{M} \mathbf{V}_2)^- \mathbf{V}_2' \mathbf{V}^* \hat{\mathbf{e}} \; , \quad \text{where} \tag{3.16}$$

$$\mathbf{M} = \mathbf{V}^* - \mathbf{V}^* \mathbf{X} (\mathbf{X}' \mathbf{V}^* \mathbf{X})^- \mathbf{X}' \mathbf{V}^* \; . \tag{3.17}$$

Noting that

$$\hat{\mathbf{e}}'\mathbf{V}^*\mathbf{V}_2 = \hat{\mathbf{e}}_2', \quad \text{w.p.1} \tag{3.18}$$

since $\hat{\mathbf{e}} \in C(\mathbf{V})$, w.p.1 and writing $\mathbf{V}_2$ as $\mathbf{V}_2 + \mathbf{X}\mathbf{U}\mathbf{X}_2' - \mathbf{X}\mathbf{U}\mathbf{X}_2'$ we obtain

$$\begin{aligned} \mathbf{V}_2'\mathbf{M}\mathbf{V}_2 &= \mathbf{V}_2'\mathbf{V}^*\mathbf{V}_2 - \mathbf{V}_2'\mathbf{V}^*\mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^-\mathbf{X}'\mathbf{V}^*\mathbf{V}_2 \\ &= \mathbf{V}_{22} - \mathbf{X}_2(\mathbf{X}'\mathbf{V}^*\mathbf{X})^-\mathbf{X}_2' + \mathbf{X}_2\mathbf{U}\mathbf{X}_2' \end{aligned} \tag{3.19}$$

and the result is established.

□

The F-statistic (3.13), a generalization of a statistic proposed by Gentleman and Wilk (1975), permits the testing of the hypothesis H_0: $\lambda = \mathbf{0}$ without actually explicitly fitting the entire model (3.12). The terms $\hat{\mathbf{e}}$ and $\mathbf{N}$ are known from the initial analysis of the original model (3.1), and thus the F-statistic (3.13) is computed in a very economical manner, possibly for various subsets of the data.

If tests show the variables $\mathbf{V}_2$ to be significant, we assume thereafter that the model (3.12) is the correct one. We will now compute the adjusted estimate $\mathbf{X}\tilde{\beta}$ for $\mathbf{X}\beta$ under model (3.12), noting that $\mathbf{X}\beta$ will be estimable under (3.12) if and only if

$$C(\mathbf{V}_2) \cap C(\mathbf{X}) = \{\mathbf{0}\} . \tag{3.20}$$

If this condition is not satisfied, we have introduced at least one superfluous variable, which can be removed. Thus we can, without loss of generality, and without loss of extra fit, assume that (3.20) is satisfied, in much the same way that $R(\mathbf{X}_2) \subset R(\mathbf{X}_1)$ is usually required in the literature while testing for $\mathbf{y}_2$ to be outlying in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{I})$, an equivalent condition to (3.20) when $\mathbf{V}$ is at least diagonal. However, the following corollary gives a general result for the adjusted estimates under model (3.12), including the case where (3.20) is not satisfied.

Corollary 3.1.1 (Schall)

If $\mathbf{X}\beta$ and $\mathbf{V}_2\lambda$ are estimable under model (3.12), that is, $C(\mathbf{V}_2) \cap C(\mathbf{X}) = \{\mathbf{0}\}$, the BLUE $\mathbf{X}\tilde{\beta}$ for $\mathbf{X}\beta$ under the model (3.12) is given by

$$\begin{aligned} \mathbf{X}\tilde{\beta} &= P_{\mathbf{X}|\mathbf{V}\mathbf{Z}}(\mathbf{y} - \mathbf{V}_2\mathbf{N}_{22}^-\hat{\mathbf{e}}_2) \\ &= \mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^-\mathbf{X}'\mathbf{V}^*(\mathbf{y} - \mathbf{V}_2\mathbf{N}_{22}^-\hat{\mathbf{e}}_2) \\ &= \mathbf{X}\hat{\beta} - \mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^-\mathbf{X}'\mathbf{V}^*\mathbf{V}_2\tilde{\lambda} \end{aligned} \tag{3.21}$$

where $\mathbf{X}\hat{\beta}$ is the BLUE for $\mathbf{X}\beta$ under the original model (3.1) and $\mathbf{V}_2\tilde{\lambda} = \mathbf{V}_2\mathbf{N}_{22}^-\hat{\mathbf{e}}_2$ is the BLUE for $\mathbf{V}_2\lambda$ under model (3.12).

In general, (3.21) yields the BLUE $\ell'\tilde{\beta}$ for any estimable linear function $\ell'\beta$ of β under (3.12), and $P_{\mathbf{V}\mathbf{Z}|\mathbf{X}}\mathbf{V}_2\mathbf{N}_{22}^-\hat{\mathbf{e}}_2$ is the BLUE for $P_{\mathbf{V}\mathbf{Z}|\mathbf{X}}\mathbf{V}_2\lambda$ under (3.12), yielding the BLUE $\mathbf{m}'\tilde{\lambda}$ of any estimable linear function $\mathbf{m}'\lambda$ of λ.

Proof: Using Theorem 2.33 we need only show that $P_{\mathbf{VZ}|\mathbf{X}}\mathbf{V}_2\tilde{\lambda} = P_{\mathbf{VZ}|\mathbf{X}}\mathbf{V}_2\mathbf{N}_{22}^{-}\hat{\mathbf{e}}_2$ in model (3.12). But

$$(3.22) \qquad P_{\mathbf{VZ}|\mathbf{X}}\mathbf{V}_2\tilde{\lambda} = P_{\mathbf{VZ}|\mathbf{X}}\mathbf{V}_2(\mathbf{V}_2'\mathbf{M}\mathbf{V}_2)^{-}\mathbf{V}_2'\mathbf{V}^{*}\hat{\mathbf{e}} \quad \text{(Theorem 2.33 (c))}$$

$$= P_{\mathbf{VZ}|\mathbf{X}}\mathbf{V}_2\mathbf{N}_{22}^{-}\hat{\mathbf{e}}_2 \,.$$

□

Reduced data model

In general the BLUE $\mathbf{X}\tilde{\beta}$ for $\mathbf{X}\beta$ under model (3.12) will be different from the BLUE $\mathbf{X}\hat{\beta}$ for $\mathbf{X}\beta$ under model (3.1), or more generally, when $\mathbf{X}\beta$ is not estimable in (3.12), the BLUE $\ell'\tilde{\beta}$ for an estimable linear function $\ell'\beta$ of β under (3.12) will differ from the corresponding BLUE $\ell'\hat{\beta}$ under (3.1).

The following problem arises: in which sort of reduced data model (with the outlying observation $\mathbf{y}_2$ 'removed' in some sense) would we equivalently obtain the BLUE $\mathbf{X}\tilde{\beta}$ for $\mathbf{X}\beta$ (or the BLUE $\ell'\tilde{\beta}$ for any estimable linear function $\ell'\beta$ of β)? The answer is given by defining a transformation $\mathbf{T}$ as

$$(3.23) \qquad \mathbf{T} = \begin{bmatrix} \mathbf{I} : -\mathbf{V}_{12}\mathbf{V}_{22}^{-} \\ \mathbf{0} : \quad \mathbf{I} \end{bmatrix},$$

where $\mathbf{T}$ is clearly nonsingular, and transforming the model (3.12) by $\mathbf{T}$ into the linear model

$$(3.24) \qquad \mathbf{T}\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \mathbf{T}\begin{bmatrix} \mathbf{X}_1 : \mathbf{V}_{12} \\ \mathbf{X}_2 : \mathbf{V}_{22} \end{bmatrix}\begin{bmatrix} \beta \\ \lambda \end{bmatrix} + \mathbf{T}\mathbf{e} \; ; \; \tilde{\mathbf{V}} = \mathbf{T}\mathbf{V}\mathbf{T}'$$

$$\Leftrightarrow \begin{bmatrix} \mathbf{y}_1 - \mathbf{V}_{12}\mathbf{V}_{22}^{-}\mathbf{y}_2 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 - \mathbf{V}_{12}\mathbf{V}_{22}^{-}\mathbf{X}_2 \; : \; \mathbf{0} \\ \mathbf{X}_2 \qquad\qquad\quad : \mathbf{V}_{22} \end{bmatrix}\begin{bmatrix} \beta \\ \lambda \end{bmatrix} + \begin{bmatrix} \mathbf{e}_1 - \mathbf{V}_{12}\mathbf{V}_{22}^{-}\mathbf{e}_2 \\ \mathbf{e}_2 \end{bmatrix}$$

$$\tilde{\mathbf{V}} = \begin{bmatrix} \mathbf{V}_{11} - \mathbf{V}_{12}\mathbf{V}_{22}^{-}\mathbf{V}_{21} : \mathbf{0} \\ \mathbf{0} \qquad\qquad\quad : \mathbf{V}_{22} \end{bmatrix}$$

By the invariance of BLUE's under a nonsingular transformation (Theorem 2.29), we have that the BLUE's under models (3.24) and (3.12) are identical.

Now, when $\mathbf{V}_{22}$ is nonsingular, we obtain the BLUE $(\mathbf{X}_1 - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{X}_2)\tilde{\beta}$ for $(\mathbf{X}_1 - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{X}_2)\beta$ under the model (3.24) also in the reduced model

$$(3.25) \qquad (\mathbf{y}_1 - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{y}_2) = (\mathbf{X}_1 - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{X}_2)\beta + \tilde{\mathbf{e}}_1 \; ;$$

$$cov(\tilde{\mathbf{e}}_1) = (\mathbf{V}_{11} - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21})$$

because model (3.24) is essentially disconnected.

We call model (3.25) the reduced data model, and the space of estimable linear functions $\ell'\beta$ of β in model (3.12) can now be seen to be $R(\mathbf{X}_1 - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{X}_2)$ when $\mathbf{V}_{22}$ is nonsingular.

If $\mathbf{V}_{22}$ is singular, let without loss of generality $\mathbf{W} = \mathbf{V}_{22}$ be partitioned as

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_{11} : \mathbf{W}_{12} \\ \mathbf{W}_{21} : \mathbf{W}_{22} \end{bmatrix} \tag{3.26}$$

such that $\mathbf{W}_{22}$ is nonsingular and $rank(\mathbf{V}_{22}) = rank(\mathbf{W}_{22})$.

Now, in (3.24) we partition the model part associated with $\mathbf{y}_2$ conformably with $\mathbf{W}$ as

$$\begin{bmatrix} \mathbf{y}_{21} \\ \mathbf{y}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{21} : \mathbf{W}_{11} : \mathbf{W}_{12} \\ \mathbf{X}_{22} : \mathbf{W}_{21} : \mathbf{W}_{22} \end{bmatrix} \begin{bmatrix} \beta \\ \lambda_1 \\ \lambda_2 \end{bmatrix} + \mathbf{e}_2 \tag{3.27}$$

and transform by

$$\mathbf{T}_2 = \begin{bmatrix} \mathbf{I} & : & -\mathbf{W}_{12}\mathbf{W}_{22}^{-1} \\ \mathbf{0} & : & \mathbf{I} \end{bmatrix} \tag{3.28}$$

to obtain

$$\begin{bmatrix} \mathbf{y}_1 - \mathbf{V}_{12}\mathbf{V}_{22}^{-}\mathbf{y}_2 \\ \mathbf{y}_{21} - \mathbf{W}_{12}\mathbf{W}_{22}^{-1}\mathbf{y}_{22} \\ \mathbf{y}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 - \mathbf{V}_{12}\mathbf{V}_{22}^{-}\mathbf{X}_2 & : \mathbf{0} & : \mathbf{0} \\ \mathbf{X}_{21} - \mathbf{W}_{12}\mathbf{W}_{22}^{-1}\mathbf{X}_{22} & : \mathbf{0} & : \mathbf{0} \\ \mathbf{X}_{22} & : \mathbf{W}_{21} & : \mathbf{W}_{22} \end{bmatrix} \begin{bmatrix} \beta \\ \lambda_1 \\ \lambda_2 \end{bmatrix} + \tilde{\mathbf{e}} \tag{3.29}$$

$$\tilde{\tilde{\mathbf{V}}} = \begin{bmatrix} \mathbf{V}_{11} - \mathbf{V}_{12}\mathbf{V}_{22}^{-}\mathbf{V}_{21} & : \mathbf{0} & : \mathbf{0} \\ \mathbf{0} & : \mathbf{0} & : \mathbf{0} \\ \mathbf{0} & : \mathbf{0} & : \mathbf{W}_{22} \end{bmatrix}$$

The reduced data model is then given by dropping the equations $\mathbf{y}_{22} = \mathbf{X}_{22}\beta + \mathbf{W}_{21}\lambda_1 + \mathbf{W}_{22}\lambda_2 + \mathbf{e}_{22}$ from (3.29). The motivation for the application of the second transformation $\mathbf{T}_2$ was to preserve the sure equations

$$(\mathbf{y}_{21} - \mathbf{W}_{12}\mathbf{W}_{22}^{-1}\mathbf{y}_{22}) = (\mathbf{X}_{21} - \mathbf{W}_{12}\mathbf{W}_{22}^{-1}\mathbf{X}_{22})\beta \ , \quad \text{w.p.1} \tag{3.30}$$

in the reduced data model. Generally, $rank(\mathbf{V}_{22})$ equations are dropped from the model (3.12) to obtain the reduced data model.

From one point of view it makes sense to assume $\mathbf{V}_{22}$ to be nonsingular. In the case of a single outlier it means that $\mathbf{V}_{22} \neq 0$ and we would certainly not test a sure equation as possibly (stochastically) outlying (and fit a column of zeros to adjust for such an outlier). For multiple outliers the singularity of $\mathbf{V}_{22}$ means that the space $C(\mathbf{V}_2)$ is spanned by a subset of l (say) linear independent columns of $\mathbf{V}_2$ ($l < k$), and we have fitted $k - l$ redundant columns, and inherently specified the same number of sure equations as outlying. Thus, without loss of extra fit, we can always take $\mathbf{V}_{22}$ to be nonsingular.

But whether or not this condition is satisfied, we observe in model (3.24) that

$$\mathbf{y}_2 = \mathbf{X}_2\tilde{\beta} + \mathbf{V}_{22}\tilde{\lambda} = \tilde{\mathbf{y}}_2 \,, \tag{3.31}$$

that is, the fitted value $\tilde{\mathbf{y}}_2$ for $\mathbf{y}_2$ under model (3.24) and consequently under model (3.12) is $\mathbf{y}_2$ itself, as it is the case when $\mathbf{V} = \mathbf{I}$ (John and Draper, 1978).

If $\mathbf{V}_2\lambda$ is estimable under the models (3.12) and (3.24), i.e. $C(\mathbf{V}_2) \cap C(\mathbf{X}) = \{\mathbf{0}\}$, and $\mathbf{V}_{22}$ is nonsingular, then using (3.31) we can write (3.24) as

$$\mathbf{y}_1 - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}(\mathbf{X}_2\tilde{\beta} + \mathbf{V}_{22}\tilde{\lambda}) = (\mathbf{X}_1 - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{X}_2)\beta + \tilde{\mathbf{e}} \tag{3.32}$$

$$\tilde{\mathbf{V}}_{11} = (\mathbf{V}_{11} - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}) \,.$$

We obtain the same BLUE $\mathbf{X}\tilde{\beta}$ for $\mathbf{X}\beta$ in

$$\mathbf{y}_1 - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}(\mathbf{X}_2\beta + \mathbf{V}_{22}\tilde{\lambda}) = (\mathbf{X}_1 - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{X}_2)\beta + \tilde{\mathbf{e}}_1 \tag{3.33}$$

$$\Leftrightarrow \quad (\mathbf{y}_1 - \mathbf{V}_{12}\tilde{\lambda}) = \mathbf{X}_1\beta + \tilde{\mathbf{e}}_1 \,;\; \tilde{\mathbf{V}} = (\mathbf{V}_{11} - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21})$$

If $\mathbf{V}_{22}$ is singular, similar expressions for (3.32) and (3.33) can be found involving the reduced data model obtained from (3.29).

Hence: If we detect observation $\mathbf{y}_2$ as an D-outlier, we obtain the reduced data model not by simply removing $\mathbf{y}_2$ and corresponding components from model (3.1). We must also remove the estimated correlated error effect transmitted as $\mathbf{V}_{12}\tilde{\lambda}$ in models (3.12) through (3.33) to the other equations, and adjust the covariance structure to that of the conditional variance for $\mathbf{y}_1 \mid \mathbf{y}_2$, precisely because those terms were in fact observed though anomalous. Note that although we adjust $\mathbf{y}_1$ by the *estimated* outlier effect $\mathbf{V}_{12}\tilde{\lambda}$, the error term $\mathbf{e}_1$ is adjusted by the *true* outlier effect $\mathbf{e}_2$, as can be seen in (3.24) where $\tilde{\mathbf{e}}_1 = \mathbf{e}_1 - \mathbf{V}_{12}\mathbf{V}_{22}^{-}\mathbf{e}_2$. Thus the effect of fitting the additional variables $\mathbf{A} = [\mathbf{V}_{21} : \mathbf{V}_{22}]'$ to the model (3.1) is to remove (w.p.1) the outlying unobserved variate $\mathbf{e}_2$ from the estimation of $\mathbf{X}\beta$. In fact, $\mathbf{e}_2$ and not necessarily $\mathbf{y}_2$ is removed from the model when we adopt the reduced data model. This process reduces the data to a set of plausible unobserved adjusted $\mathbf{y}$, which might have been observed had the offending term $\mathbf{e}_2$ in the observations $\mathbf{y}_2$ not occurred and not affected all the other observations through correlation. Of course, for $\mathbf{V}$ at least diagonal, as in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{I})$, $\mathbf{e}_2$ is removed together with $\mathbf{y}_2$.

Missing data estimation

From a missing value point of view, in (3.12) we obtain $\mathbf{V}_{22}\tilde{\lambda}$ (the estimate of the distributional outlier effect) by first estimating $\mathbf{X}_2\beta$ by $\mathbf{X}_2\tilde{\beta}$ in the reduced data model (3.24), and then $\mathbf{V}_{22}\tilde{\lambda} = \mathbf{y}_2 - \mathbf{X}_2\tilde{\beta}$. That is, we see the difference between the observed value $\mathbf{y}_2$ and the missing plot estimate $\mathbf{X}_2\tilde{\beta}$ as the outlier effect. This is the same procedure as in the familiar case $\mathbf{V} = \mathbf{I}$, where we also have that the fitted value $\tilde{\mathbf{y}}_2$ in the adjusted model is precisely the observed value $\mathbf{y}_2$. Of course, missing data estimation of $\mathbf{y}_2$ in the case $\mathbf{V} = \mathbf{I}$ is only possible if $R(\mathbf{X}_2) \subset R(\mathbf{X}_1)$, and similarly the missing data interpretation for arbitrary $\mathbf{V}$ as above is only possible if $C(\mathbf{X}) \cap C(\mathbf{V}_2) = \{\mathbf{0}\}$.

Recursive residuals

John and Draper (1978) show that the statistic Q_k of Gentleman and Wilk (1975), "can be written in general as the sum of squares of k successive revised normalized uncorrelated residuals". Since Gentleman and Wilks' Q_k is a special case, for $\mathbf{V} = \mathbf{I}$, of the additional sum of squares $\hat{\mathbf{e}}_2' \mathbf{N}_{22}^{-} \hat{\mathbf{e}}_2$ in (3.13), we can generalize the idea of recursive residuals for the linear model under arbitrary known variance-covariance structure.

In the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{I})$ the computation of recursive residuals is equivalent to fitting new vectors $\mathbf{u}_n = (0, \ldots, 0, 1)'$, then $\mathbf{u}_{n-1}$ after $\mathbf{u}_n$, and so on, to the model, provided that $\mathbf{u}_i \notin C([\mathbf{X}:\mathbf{U}_2])$, where $\mathbf{U}_2$ denotes the new variables already in the model. We obtain the models

$$\mathbf{y} = \begin{bmatrix} \mathbf{X}_{(-n)} & : \mathbf{0} \\ \mathbf{x}_n & : 1 \end{bmatrix} \begin{bmatrix} \beta \\ \lambda_1 \end{bmatrix} + \mathbf{e} \tag{3.34}$$

through

$$\mathbf{y} = \begin{bmatrix} \mathbf{X}_{(-R)} & : \mathbf{0} \\ \mathbf{X}_R & : \mathbf{I}_R \end{bmatrix} \begin{bmatrix} \beta \\ \lambda_R \\ \vdots \\ \lambda_1 \end{bmatrix} + \mathbf{e} \tag{3.35}$$

where $rank(\mathbf{X}_{(-R)}) = rank(\mathbf{X})$ and $R = n - rank(\mathbf{X})$.

If Q_k is the additional sum of squares due to fitting $[\mathbf{0}:\mathbf{I}_k]'$ at the k-th step, then the k-th recursive residual is defined as the positive square root of

$$q_k = Q_k - Q_{k-1} \,. \tag{3.36}$$

Clearly, q_k is uncorrelated with Q_{k-1}, since it is the additional sum of squares due to fitting $\mathbf{u}_{n-k+1}$ after $[\mathbf{X}:\mathbf{u}_n : \ldots : \mathbf{u}_{n-k+2}]$ (Theorem 2.33). Since this reasoning can be applied at any stage $i = 1, \ldots, k$, Q_k can be written as

$$Q_k = \sum_{i=1}^{k} q_i \tag{3.37}$$

where the q_i are mutually uncorrelated $(i = 1, \ldots, k)$.

Similarly, in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ with $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_n]$ we add successively the variables $\mathbf{v}_n$, then $\mathbf{v}_{n-1}$ after $\mathbf{v}_n$, and so on, provided that $\mathbf{v}_i \notin C([\mathbf{X}:\mathbf{V}_2])$ where $\mathbf{V}_2$ denotes here the new variables already in the model at the $(i-1)$-th step. This condition results in no loss of generality, since with $\mathbf{v}_i \in C([\mathbf{X}:\mathbf{V}_2])$ the fitting of $\mathbf{v}_i$ provides no extra fit.

As in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{I})$ we define the k-th recursive residual as the positive square root of

$$q_k = Q_k - Q_{k-1} \,, \tag{3.38}$$

where Q_k is the additional sum of squares due to fitting $[\mathbf{v}_n, \ldots, \mathbf{v}_{n-k+1}]$ after $\mathbf{X}$, and similarly Q_i, $i = 1, \ldots, k-1$. The columns $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_n]$ are possibly renumbered to ensure that the condition $\mathbf{v}_i \notin C([\mathbf{X}:\mathbf{V}_2])$ is satisfied at each step. Thus $rank([\mathbf{X}:\mathbf{V}]) - rank(\mathbf{X})$ recursive residuals are obtained in total.

From Theorem 3.1 we know that

$$Q_k = \hat{\mathbf{e}}_2' \mathbf{N}_{22}^{-1} \hat{\mathbf{e}}_2 \tag{3.39}$$

where the use of the true inverse $\mathbf{N}_{22}^{-1}$ of $\mathbf{N}_{22}$ is justified by the condition $\mathbf{v}_i \notin C([\mathbf{X}:\mathbf{V}_2])$, $i = 1, \ldots, k$. But Q_k can be written as

$$Q_k = [\hat{e}_k : \hat{\mathbf{e}}_{(-k)}'] \begin{bmatrix} d : \mathbf{c}' \\ \mathbf{c} : \mathbf{N}_{(-k)} \end{bmatrix}^{-1} \begin{bmatrix} \hat{e}_k \\ \hat{\mathbf{e}}_{(-k)} \end{bmatrix} \tag{3.40}$$

where $\hat{\mathbf{e}}_2'$ is partitioned as $\hat{\mathbf{e}}_2' = [\hat{e}_k : \hat{\mathbf{e}}_{(-k)}']$, with $\hat{e}_k$ being a number, and $\mathbf{N}_{(-k)}$ denotes the "$\mathbf{N}_{22}$-matrix" at the $(k-1)$-th step. Then, by applying a standard result given by Rao (1973, p. 33) for the inversion of partitioned matrices

$$Q_k = (\hat{e}_k - \mathbf{c}' \mathbf{N}_{(-k)}^{-1} \hat{\mathbf{e}}_{(-k)})^2 / (d - \mathbf{c}' \mathbf{N}_{(-k)}^{-1} \mathbf{c}) + \hat{\mathbf{e}}_{(-k)}' \mathbf{N}_{(-k)}^{-1} \hat{\mathbf{e}}_{(-k)} \tag{3.41}$$

and thus

$$q_k = (\hat{e}_k - \mathbf{c}' \mathbf{N}_{(-k)}^{-1} \hat{\mathbf{e}}_{(-k)})^2 / (d - \mathbf{c}' \mathbf{N}_{(-k)}^{-1} \mathbf{c}) \,. \tag{3.42}$$

This way, the q_i, $i = 1, \ldots, (rank([\mathbf{X}:\mathbf{V}]) - rank(\mathbf{X}))$ can be computed recursively using (3.42), with

$$q_1 = \hat{e}_n^2 / N_{nn} \tag{3.43}$$

in the first step.

Since q_k is distributed $\sigma^2 \chi_1^2$ under normality, being the additional sum of squares due to fitting $\mathbf{v}_{n-k+1}$, the recursive residuals all have common variance σ^2 (independent of the normality assumption), that is, they are normalized.

3.1.2 Outliers by additive shifts

The adjusted model

To adjust for a shift in the mean of the A-outlier $\mathbf{y}_2$, we fitted the linear model

$$\mathbf{y}_2 = \mathbf{X}_2 \beta + \lambda + \mathbf{e}_2 \,; \quad cov(\mathbf{e}_2) = \sigma^2 \mathbf{V}_{22} \tag{3.3}$$

for $\mathbf{y}_2$. The common linear model for $\mathbf{y}$, adjusted for the A-outlier $\mathbf{y}_2$, is then

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 : \mathbf{0} \\ \mathbf{X}_2 : \mathbf{I} \end{bmatrix} \begin{bmatrix} \beta \\ \lambda \end{bmatrix} + \mathbf{e} \,; \quad cov(\mathbf{e}) = \sigma^2 \mathbf{V} \,. \tag{3.44}$$

This model can also be seen as a generalization of the formulation by John and Draper (1978) and Cook and Weisberg (1979).

While fitting the additional variables $\mathbf{A} = [\mathbf{0}:\mathbf{I}]'$ in model (3.44), it may happen that $C(\mathbf{A}) \cap C([\mathbf{X}:\mathbf{V}]) \not\subset C(\mathbf{V})$, and thus fitting $\mathbf{A}$ is not necessarily admissible in the sense of Lemma 2.31. If we are convinced that the sure equations in the original model (3.1) are not contaminated by an additive outlier, $C(\mathbf{A}) \cap C([\mathbf{X}:\mathbf{V}]) \not\subset C(\mathbf{V})$ indicates that we are specifying at least one component of $\mathbf{y}_2$ incorrectly as an A-outlier. In this case we would have to drop a suitable number of columns from $\mathbf{A}$ to ensure that fitting the new variables is admissible.

Further, if an additive shift has occured in the mean of $\mathbf{y}_2$, the original model (3.1) might be inconsistent due to the A-outlier $\mathbf{y}_2$, and fitting $\mathbf{A} = [\mathbf{0}:\mathbf{I}]'$ is meant to restore the consistency in the adjusted model (3.44). However, in this case the hypothesis H_0: $\lambda = \mathbf{0}$ in model (3.44) can not be tested since it is inconsistent (Lemma 2.36). But from the inconsistency of the original model (3.1), and its restoration by fitting $\mathbf{A} = [\mathbf{0}:\mathbf{I}]'$ in (3.44) we know, with probability 1, that $\mathbf{y}_2$ is outlying, and no test is needed. In a practical situation we would fit the minimum number of columns $\mathbf{u}_i$, $\mathbf{u}_j$, ... which are needed to restore the consistency of the model. We shall presently see that this amounts to removing the corresponding observations $\mathbf{y}_i$, $\mathbf{y}_j$, ... , the "sure" outliers, from the model. In the reduced model, which is consistent, the search for stochastic outliers might continue. In the following we assume thus, without loss of generality, the consistency of the model.

If the model (3.1) is not of full rank, that is, $C([\mathbf{X}:\mathbf{V}]) \neq \mathbb{R}^n$, it is possible that $C(\mathbf{A}) \not\subset C([\mathbf{X}:\mathbf{V}])$, where $\mathbf{A} = [\mathbf{0}:\mathbf{I}]'$. In this case, being confronted with a consistent model, we would perform a reduction of model (3.1) to a model of full rank, as given by Lemma 2.6. This reduction facilitates considerably the computation of adjusted estimates and test-statistics in the augmented model (3.44), since Theorem 2.33 and its corollary can then be applied without reparametrizing the $\mathbf{A}\lambda$-part of (3.44) as $\mathbf{A}\lambda = \mathbf{A}_1^*\lambda_1 + \mathbf{A}_2^*\lambda_2$ such that $C(\mathbf{A}_1^*) \subset C([\mathbf{X}:\mathbf{V}])$ and $C(\mathbf{A}_2^*) \cap C([\mathbf{X}:\mathbf{V}]) = \{\mathbf{0}\}$, and a subsequent fit of $\mathbf{A}_1^*$ only. We assume thus that $C(\mathbf{A}) \subset C([\mathbf{X}:\mathbf{V}])$ is satisfied in model (3.44).

Testing and adjusted estimates

In the adjusted model (3.44) we test the hypothesis H_0: $\lambda = \mathbf{0}$ against the alternative H_1: $\lambda \neq \mathbf{0}$, which constitutes the test for the hypothesis that $\mathbf{y}_2$ is not an A-outlier against the alternative that $\mathbf{y}_2$ is an A-outlier.

Theorem 3.2 (Schall)

The F-statistic associated with the hypothesis H_0: $\lambda = \mathbf{0}$ in the linear model (3.44) is

$$F = \frac{\hat{\mathbf{e}}'\mathbf{M}_2'\mathbf{M}_{22}^-\mathbf{M}_2\hat{\mathbf{e}}}{\hat{\mathbf{e}}'\mathbf{V}^-\hat{\mathbf{e}} - \hat{\mathbf{e}}'\mathbf{M}_2'\mathbf{M}_{22}^-\mathbf{M}_2\hat{\mathbf{e}}} \cdot \frac{r([\mathbf{X}:\mathbf{V}]) - r(\mathbf{X}) - r(\mathbf{M}_{22})}{r(\mathbf{M}_{22})} \tag{3.45}$$

provided that $C(\mathbf{A}) \subset C([\mathbf{X}:\mathbf{V}])$, $\mathbf{A} = [\mathbf{0}:\mathbf{I}]'$, where

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_1 \\ \mathbf{M}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{M}_{11}\ \mathbf{M}_{12} \\ \mathbf{M}_{21}\ \mathbf{M}_{22} \end{bmatrix} = (\mathbf{V}^* - \mathbf{V}^*\mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^-\mathbf{X}'\mathbf{V}^*) \,. \tag{3.17}$$

Proof: Similarly to the proof of Theorem 3.1 we must show that the additional sum of squares due to fitting $\mathbf{A} = [\mathbf{0}:\mathbf{I}]'$ in model (3.44) is

$$SSA = \hat{\mathbf{e}}'\mathbf{M}_2'\mathbf{M}_{22}^{-}\mathbf{M}_2\hat{\mathbf{e}} \,. \tag{3.46}$$

Again, using Corollary 2.33.1 (b), we have

$$\begin{aligned} SSA &= \hat{\mathbf{e}}'\mathbf{V}^*\mathbf{N}\mathbf{A}(\mathbf{A}'\mathbf{M}\mathbf{A})^{-}\mathbf{A}'\mathbf{N}\mathbf{V}^*\hat{\mathbf{e}} \\ &= \hat{\mathbf{e}}'\mathbf{N}\mathbf{V}^*\mathbf{N}\mathbf{A}(\mathbf{A}'\mathbf{M}\mathbf{A})^{-}\mathbf{A}\mathbf{N}\mathbf{V}^*\mathbf{N}\hat{\mathbf{e}} \\ &= \hat{\mathbf{e}}'\mathbf{M}\mathbf{A}(\mathbf{A}'\mathbf{M}\mathbf{A})^{-}\mathbf{A}'\mathbf{M}\hat{\mathbf{e}} \\ &= \hat{\mathbf{e}}'\mathbf{M}_2'\mathbf{M}_{22}^{-}\mathbf{M}_2\hat{\mathbf{e}} \,, \quad \text{with } \mathbf{A} = [\mathbf{0}:\mathbf{I}] \,. \end{aligned} \tag{3.47}$$

□

Setting $\mathbf{V} = \mathbf{V}^* = \mathbf{I}$, the statistic (3.45) reduces also in this case to the statistic proposed by Gentlemen and Wilk (1975), and here we obtain Q_k as $Q_k = \hat{\mathbf{e}}'\mathbf{M}_2'\mathbf{M}_{22}^{-}\mathbf{M}_2\hat{\mathbf{e}}$.

If tests show the variables $\mathbf{A} = [\mathbf{0}:\mathbf{I}]'$ to be significant in model (3.44), we compute the adjusted parameter estimates under model (3.44). The proof of the following corollary follows the lines of the proof of Corollary 3.1.1.

Corollary 3.2.1

If $\mathbf{X}\beta$ is estimable under model (3.44), that is, $C(\mathbf{A}) \cap C(\mathbf{X}) = \{\mathbf{0}\}$, $\mathbf{A} = [\mathbf{0}:\mathbf{I}]'$, or equivalently $R(\mathbf{X}_2) \subset R(\mathbf{X}_1)$, the BLUE $\mathbf{X}\tilde{\beta}$ for $\mathbf{X}\beta$ under the model (3.44) is given by

$$\begin{aligned} \mathbf{X}\tilde{\beta} &= \mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^*\left(\mathbf{y} - \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} \mathbf{M}_{22}^{-}\mathbf{M}_2\hat{\mathbf{e}}\right) \\ &= \mathbf{X}\hat{\beta} - \mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^* \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} \mathbf{M}_{22}^{-}\mathbf{M}_2\hat{\mathbf{e}} \,, \end{aligned} \tag{3.48}$$

provided that $C(\mathbf{A}) \subset C([\mathbf{X}:\mathbf{V}])$, where $\mathbf{X}\hat{\beta}$ is the BLUE for $\mathbf{X}\beta$ under the original model (3.1). If $\mathbf{X}\beta$ is not estimable under (3.44), then (3.48) yields the BLUE $\ell'\tilde{\beta}$ of any estimable linear function $\ell'\beta$ of β. The space of estimable linear functions $\ell'\beta$ of β is given by $R(\mathbf{X}_1)$.

□

Reduced data model

We now seek to specify the reduced data model where we would equivalently obtain $\mathbf{X}\tilde{\beta}$ from Corollary 3.2.1 as the BLUE for $\mathbf{X}\beta$.

The BLUE $\mathbf{X}_1\tilde{\beta}$ for $\mathbf{X}_1\beta$ in the model

$$\mathbf{y}_1 = \mathbf{X}_1\beta + \mathbf{e}_1 \,; \quad cov(\mathbf{e}_1) = \sigma^2\mathbf{V}_{11} \tag{3.49}$$

is given by

(3.50) $$\mathbf{X}_1\tilde{\beta} = \mathbf{X}_1(\mathbf{X}_1'\mathbf{V}_{11}^*\mathbf{X}_1)^-\mathbf{X}_1'\mathbf{V}_{11}^*\mathbf{y}_1 .$$

Now $\mathbf{X}_1\tilde{\beta}$ as in (3.50) is BLUE for $\mathbf{X}_1\beta$ under the model (3.44) if and only if

(3.51) $$C\left(\begin{bmatrix}\mathbf{V}_{11}:\mathbf{V}_{12}\\ \mathbf{V}_{21}:\mathbf{V}_{22}\end{bmatrix}\begin{bmatrix}\mathbf{V}_{11}^*\mathbf{X}_1(\mathbf{X}_1'\mathbf{V}_{11}^*\mathbf{X}_1)^-\mathbf{X}_1'\\ \mathbf{0}\end{bmatrix}\right) \subset C\left(\begin{bmatrix}\mathbf{X}_1:\mathbf{0}\\ \mathbf{X}_2:\mathbf{I}\end{bmatrix}\right)$$

$$\Leftrightarrow\quad C\left(\begin{bmatrix}\mathbf{V}_{11}\mathbf{V}_{11}^*\mathbf{X}_1(\mathbf{X}_1'\mathbf{V}_{11}^*\mathbf{X}_1)^-\mathbf{X}_1'\\ \mathbf{V}_{21}\mathbf{V}_{11}^*\mathbf{X}_1(\mathbf{X}_1'\mathbf{V}_{11}^*\mathbf{X}_1)^-\mathbf{X}_1'\end{bmatrix}\right) \subset C\left(\begin{bmatrix}\mathbf{X}_1:\mathbf{0}\\ \mathbf{X}_2:\mathbf{I}\end{bmatrix}\right)$$

using Theorem 2.15 (Zyskind, 1967). But this is clearly the case. Hence the BLUE's $\mathbf{X}_1\tilde{\beta}$ for $\mathbf{X}_1\beta$ in the models (3.44) and (3.49) are identical, and we can therefore interpret (3.49) as the reduced data model. Note that the effect of fitting $\mathbf{A} = [\mathbf{0}:\mathbf{I}]'$ is to remove the observations $\mathbf{y}_2$ and corresponding rows of $\mathbf{X}$ and rows and columns of $\mathbf{V}$ from the model, without any adjustment of $\mathbf{y}_1$ and the variance-covariance structure of $\mathbf{y}_1$.

It is also possible to write the subvector $\tilde{\mathbf{e}}_2$ of the estimated residuals $\tilde{\mathbf{e}}$ under (3.44) in an interesting way.

Lemma 3.3 (Schall)

Let

(3.52) $$\tilde{\mathbf{e}}_1 = \begin{bmatrix}\tilde{\mathbf{e}}_1\\ \tilde{\mathbf{e}}_2\end{bmatrix} = \begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix} - \begin{bmatrix}\mathbf{X}_1 : \mathbf{0}\\ \mathbf{X}_2 : \mathbf{I}\end{bmatrix}\begin{bmatrix}\tilde{\beta}\\ \tilde{\lambda}\end{bmatrix}$$

be the estimated residuals under the linear model (3.44). Then

(3.53) $$\tilde{\mathbf{e}}_2 = \mathbf{V}_{21}\mathbf{V}_{11}^-\tilde{\mathbf{e}}_1 .$$

Proof: Using the reduced data model (3.49), $\tilde{\mathbf{e}}_1$ can be written as

(3.54) $$\tilde{\mathbf{e}}_1 = (\mathbf{I}-\mathbf{X}_1(\mathbf{X}_1'\mathbf{V}_{11}^*\mathbf{X}_1)^-\mathbf{X}_1'\mathbf{V}_{11}^*)\mathbf{y}_1$$

Clearly, $E(\mathbf{y}_2-\tilde{\mathbf{e}}_2) = \mathbf{X}_2\beta + \lambda$, and $\mathbf{y}_2-\mathbf{V}_{21}\mathbf{V}_{11}^-\tilde{\mathbf{e}}_1$ is then the BLUE for $\mathbf{X}_2\beta+\lambda$ if and only if (Theorem 2.15)

(3.55) $$C\left(\begin{bmatrix}\mathbf{V}_{11} : \mathbf{V}_{12}\\ \mathbf{V}_{21} : \mathbf{V}_{22}\end{bmatrix}\begin{bmatrix}(\mathbf{V}_{11}^*\mathbf{X}_1(\mathbf{X}_1'\mathbf{V}_{11}^*\mathbf{X}_1)^-\mathbf{X}_1'-\mathbf{I})\mathbf{V}_{11}^-\mathbf{V}_{12}\\ \mathbf{I}\end{bmatrix}\right) \subset C\left(\begin{bmatrix}\mathbf{X}_1 : \mathbf{0}\\ \mathbf{X}_2 : \mathbf{I}\end{bmatrix}\right)$$

$$\Leftrightarrow\quad C(\mathbf{V}_{11}\mathbf{V}_{11}^*\mathbf{X}_1'(\mathbf{X}_1'\mathbf{V}_{11}^*\mathbf{X}_1)^-\mathbf{X}_1'\mathbf{V}_{11}^-\mathbf{V}_{12}-\mathbf{V}_{12}+\mathbf{V}_{12}) \subset C(\mathbf{X}_1)$$

which is clearly the case, since $\mathbf{V}_{11}\mathbf{V}_{11}^*\mathbf{X}_1' = \mathbf{X}_1\mathbf{B}$ for some $\mathbf{B}$ (Theorem 2.16). Thus $\mathbf{y}_2-\tilde{\mathbf{e}}_2$ $= \mathbf{y}_2-\mathbf{V}_{21}\mathbf{V}_{11}^-\tilde{\mathbf{e}}_1$, and the lemma is proved.

□

We note that both the result on the reduced data model (3.49) and the result given in Lemma 3.3 do not depend on the assumption that $C(\mathbf{A}) \subset C([\mathbf{X} : \mathbf{V}])$.

Missing data estimation

We can interpret $\tilde{\mathbf{e}}_2$ as a missing data estimate of the error $\mathbf{e}_2$, since we obtain $-\hat{\lambda} = \tilde{\mathbf{e}}_2$ in the model

$$\begin{bmatrix} \tilde{\mathbf{e}}_1 \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} \lambda + \mathbf{e} \; ; \mathbf{V} = \begin{bmatrix} \mathbf{V}_{11} : \mathbf{V}_{12} \\ \mathbf{V}_{21} : \mathbf{V}_{22} \end{bmatrix} \tag{3.56}$$

Equivalently we have a missing data estimate $\mathbf{y}_2^*$ of $\mathbf{y}_2$, provided that $\mathbf{X}_2\beta$ is estimable under model (3.44), given by

$$\mathbf{y}_2^* = \mathbf{X}_2\tilde{\beta} + \tilde{\mathbf{e}}_2 \tag{3.57}$$

where $\mathbf{X}_2\tilde{\beta}$ and $\tilde{\mathbf{e}}_2$ are estimates from model (3.44).

The estimated outlier effect $\tilde{\lambda}$ under model (3.44) is then the difference between the observation $\mathbf{y}_2$ and the missing data estimate $\mathbf{y}_2^*$, i.e.

$$\tilde{\lambda} = \mathbf{y}_2 - \mathbf{X}_2\tilde{\beta} - \mathbf{V}_{21}\mathbf{V}_{11}^*\tilde{\mathbf{e}}_1 = \mathbf{y}_2 - \mathbf{y}_2^* \; . \tag{3.58}$$

We observe that the notion of an additive outlier is closely related to classical missing data estimation. The missing data estimate $\mathbf{y}_2^*$ is obtained not only by computing $\mathbf{X}_2\tilde{\beta}$ from the reduced model, but by adding to this quantity the term $\tilde{\mathbf{e}}_2 = \mathbf{V}_{21}\mathbf{V}_{11}^-\tilde{\mathbf{e}}_1$, thus obtaining the conditional expectation for $\mathbf{y}_2 \mid \mathbf{y}_1$.

Recursive residuals

In a manner similar to the method outlined in section 3.1.1, we can generalize the idea of recursive residuals for the case of A-outliers.

The k-th recursive residual is defined as the positive square root of

$$q_k = Q_k - Q_{k-1} \tag{3.59}$$

where Q_k and Q_{k-1} are respectively the additional sums of squares due to fitting $[\mathbf{u}_n : \ldots : \mathbf{u}_{n-k+1}]$ and $[\mathbf{u}_n : \ldots : \mathbf{u}_{n-k+2}]$ as new variables to the model (3.1), and $\mathbf{u}_i$ is the vector of zeros except for a 1 in the i-th component. We again require that $\mathbf{u}_i \notin C([\mathbf{X} : \mathbf{U}_2])$, $i = 1, \ldots, k$ where $\mathbf{U}_2$ denotes the new variables already in the model.

Clearly, Q_k can be written as

$$Q_k = \sum_{i=1}^{k} q_i \tag{3.60}$$

where the q_i are mutually uncorrelated, and similarily to the development leading to (3.42), we can write q_k as

$$q_k = (\mathbf{m}_k\hat{\mathbf{e}} - \mathbf{c}'\mathbf{M}_{(-)}^{-1}\mathbf{M}_{(-k)}\hat{\mathbf{e}})^2/(\mathrm{d} - \mathbf{c}'\mathbf{M}_{(-)}^{-1}\mathbf{c}) \tag{3.61}$$

where $\mathbf{M}_{22}$ is partitioned as

(3.62)
$$\mathbf{M}_{22} = \begin{bmatrix} \mathrm{d} : \mathbf{c}' \\ \mathbf{c} : \mathbf{M}_{(-)} \end{bmatrix}$$

and $\mathbf{M}$ is partitioned as

(3.63)
$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_1 \\ \mathbf{M}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{M}_1 \\ \mathbf{m}_k \\ \mathbf{M}_{(-k)} \end{bmatrix}$$

Thus the statistic Q_k can again be written as the sum of k uncorrelated normalized squared residuals.

3.1.3 A comparison of D- and A-outliers

In comparing the respective approaches to outliers in a linear model under arbitrary variance we have considered so far, it is instructive to begin with the missing data interpretations of the respective approaches.

In the case of a D-outlier $\mathbf{y}_2$ we regard the difference between $\mathbf{y}_2$ and the adjusted estimate $\mathbf{X}_2\tilde{\beta}$ for $\mathbf{X}_2\beta$ as the outlier effect, which is consistent with the definition of a D-outlier as an observation $\mathbf{y}_2$ with

(3.2)
$$\mathbf{y}_2 = \mathbf{X}_2\beta + \mathbf{e}_2$$

where $\mathbf{e}_2$ was considered significantly large. Equivalently

(3.64)
$$\mathbf{e}_2 = \mathbf{y}_2 - \mathbf{X}_2\beta$$

is the true outlier effect, and in performing the estimation in the adjusted model (3.12) we do nothing else than replace $\mathbf{X}_2\beta$ in (3.64) by the estimate $\mathbf{X}_2\tilde{\beta}$ from the other observations, and consequently attribute the remainder $\mathbf{y}_2 - \mathbf{X}_2\tilde{\beta}$ to the estimated outlier effect $\mathbf{V}_{22}\tilde{\lambda}$, leaving $\tilde{\mathbf{e}}_2$ to be zero.

As an A-outlier $\mathbf{y}_2$ we defined

(3.3)
$$\mathbf{y}_2 = \mathbf{X}_2\beta + \lambda + \mathbf{e}_2$$

where λ represented an arbitrary mean shift causing $\mathbf{y}_2$ to be an outlier. Equivalently,

(3.65)
$$\lambda = \mathbf{y}_2 - \mathbf{X}_2\beta - \mathbf{e}_2$$

and it makes perfect sense that in the adjusted model (3.44) not only the estimate $\mathbf{X}_2\tilde{\beta}$ for $\mathbf{X}_2\beta$ but also some estimate $\tilde{\mathbf{e}}_2$ for $\mathbf{e}_2$ is subtracted from $\mathbf{y}_2$ to estimate the outlier effect λ. In fact, $\tilde{\mathbf{e}}_2 = \mathbf{V}_{21}\mathbf{V}_{11}^{-}\tilde{\mathbf{e}}_1$ is the conditional expectation of $\mathbf{e}_2 \mid \tilde{\mathbf{e}}_1$, and this is a sensible estimate since $\mathbf{e}_1$ is not contaminated by the mean shift λ which is independent of $\mathbf{e}_1$ by assumption.

Downdating

The two approaches to outliers also offer two different methods to downdate a linear model.

From the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{I})$ it is well-known that fitting the variable $\mathbf{u}_n = (0, \ldots, 0, 1)'$ (say) to the model is equivalent to removing the n-th observation from the model, that is, the estimate $\mathbf{X}\tilde{\beta}$ for $\mathbf{X}\beta$ in the augmented model is computed using the first $(n-1)$ observations only. This process of adjusting the estimate $\mathbf{X}\hat{\beta}$ for $\mathbf{X}\beta$ in the full model $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{I})$ to the estimate $\mathbf{X}\tilde{\beta}$ in the model with an observation removed, Cook and Weisberg (1982) call downdating of the model. Corresponding to two types of outlier in a linear model under arbitrary variance we can downdate a model by either fitting $\mathbf{v}_i$ or $\mathbf{u}_i$ (say). In the latter case, the observation y_i is actually removed from the model, as was shown in the previous section through the discussion of the reduced data model associated with fitting $\mathbf{u}_i$. Fitting $\mathbf{v}_i$, however, removes the error term e_i, and we may call this stochastic downdating of the model, as opposed to deterministic downdating which is effected by fitting $\mathbf{u}_i$. In either case, Corollary 3.1.1 and Corollary 3.1.2 offer efficient methods to downdate a model without actually recomputing the estimates in the reduced models, but rather by adjusting in a simple and economical manner the estimate in the full model.

Duality between D- and A-outliers

Finally, we may note a duality relationship between D- and A-outliers in a linear model, where we assume for simplicity that the variance-covariance structure $\mathbf{V}$ is arbitrary but nonsingular.

Transforming the adjusted models (3.12) and (3.44) by $\mathbf{T} = \mathbf{V}^{-1}$ we obtain respectively

$$\begin{bmatrix} \tilde{\mathbf{y}}_1 \\ \tilde{\mathbf{y}}_2 \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{X}}_1 : \mathbf{0} \\ \tilde{\mathbf{X}}_2 : \mathbf{I} \end{bmatrix} \begin{bmatrix} \beta \\ \lambda \end{bmatrix} + \tilde{\mathbf{e}}\,, \quad cov(\tilde{\mathbf{e}}) = \sigma^2\mathbf{V}^{-1} \tag{3.66}$$

and

$$\begin{bmatrix} \tilde{\mathbf{y}}_1 \\ \tilde{\mathbf{y}}_2 \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{X}}_1 : \mathbf{V}^{12} \\ \tilde{\mathbf{X}}_2 : \mathbf{V}^{22} \end{bmatrix} \begin{bmatrix} \beta \\ \lambda \end{bmatrix} + \tilde{\mathbf{e}}\,; \quad cov(\tilde{\mathbf{e}}) = \sigma^2 \begin{bmatrix} \mathbf{V}^{11} : \mathbf{V}^{12} \\ \mathbf{V}^{21} : \mathbf{V}^{22} \end{bmatrix} \tag{3.67}$$

where $\tilde{\mathbf{y}} = \mathbf{V}^{-1}\mathbf{y}$, $\tilde{\mathbf{X}} = \mathbf{V}^{-1}\mathbf{X}$ and $\tilde{\mathbf{e}} = \mathbf{V}^{-1}\mathbf{e}$. We observe that the D-outlier in (3.12) has turned into an A-outlier in the transformed model (3.66), and *vice versa*, the A-outlier in (3.44) has turned into a D-outlier in (3.67).

This duality relationship can also be observed by considering the reduced data model (3.25) in the case of D-outliers, and the model

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 - \mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{y}_1 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & : \mathbf{0} \\ \mathbf{X}_2 - \mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{X}_1 & : \mathbf{I} \end{bmatrix} \begin{bmatrix} \beta \\ \lambda \end{bmatrix} + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 - \mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{e}_1 \end{bmatrix} \tag{3.68}$$

in the case of A-outliers. The model (3.68) is obtained from (3.44) by applying the transformation

(3.69) $$\mathbf{T} = \begin{bmatrix} \mathbf{I} & : \mathbf{0} \\ -\mathbf{V}_{21}\mathbf{V}_{11}^{-1} & : \mathbf{I} \end{bmatrix}.$$

The reduced data model (3.25) can be seen as the linear model for the variate $\mathbf{y}_1 \mid \mathbf{y}_2$, and the missing data estimate $\mathbf{y}_2^* = \mathbf{X}_2\tilde{\beta} + \mathbf{V}_{22}\mathbf{V}_{11}^{-1}\tilde{\mathbf{e}}_1$ for $\mathbf{y}_2$ in the model (3.68) is the estimate for the conditional expectation of $\mathbf{y}_2 \mid \mathbf{y}_1$.

3.1.4 The problem of bias when outliers are rejected

An old problem in the treatment of outliers is the question whether the removal of an observation, which is considered to be significantly outlying, from a sample introduces bias into the analysis of the corresponding model. Certainly, rejecting an observation which is contaminated by an invasive mean shift would remove bias from the model rather than introduce it, but it can be argued that the removal of what we called a *D*-outlier, an anomalous but valid observation from the tail of the underlying distribution, introduces bias. Of course, when the observations in a sample are independently distributed, those two types of outlier can not be distinguished statistically. The estimated outlier effect, the difference between the observation in question and the estimate of its mean, can be attributed to a mean shift as well as to a large error term.

When the data are correlated, however, a distinction is possible, and the test-statistics (3.13) and (3.45) will in general lead to different results which could be interpreted in a way that a discordant observation is more likely to be an outlier of one type rather than the other.

3.1.5 Transformational outliers

The adjusted model

For the third approach to outliers in a linear model with arbitrary known variance-covariance structure we consider the model of the principal components (PC's) $\mathbf{y}^*$ of $\mathbf{y}$, which is given by

(3.5) $$\begin{bmatrix} \mathbf{y}_1^* \\ \mathbf{y}_2^* \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^* \\ \mathbf{X}_2^* \end{bmatrix} \beta + \mathbf{e}^*, \quad cov(\mathbf{e}^*) = \sigma^2\triangle = \sigma^2 \begin{bmatrix} \triangle_1 & \\ & \triangle_2 \end{bmatrix}$$

The model (3.5) is obtained by a transformation of the model (3.1) by $\mathbf{T} = \mathbf{P}'$, where

(3.4) $$\mathbf{V} = \mathbf{P}\triangle\mathbf{P}' = [\mathbf{P}_1{}'\mathbf{P}_2] \begin{bmatrix} \triangle_1 & \\ & \triangle_2 \end{bmatrix} \begin{bmatrix} \mathbf{P}_1' \\ \mathbf{P}_2' \end{bmatrix}$$

is the (complete) singular value decomposition of $\mathbf{V}$, i.e. $\mathbf{P}$ is an orthogonal matrix of order $n \times n$ and $\triangle \geq \mathbf{0}$ is a diagonal matrix.

Principal Component Analysis (PCA) is a well-known and established technique of multivariate statistical analysis, and this section on transformational outliers, or equivalently outliers in PC's, in the case where the variance-covariance structure $\mathbf{V}$ is known, serves also to prepare the ground for the treatment of outliers in PC's in multivariate samples, when $\mathbf{V}$ is unknown and has to be estimated from a sample of observations $\mathbf{y}$.

If $\mathbf{y}_2^*$ in (3.5) is suspected to be outlying, we fit the adjusted model

$$(3.70) \qquad \begin{bmatrix} \mathbf{y}_1^* \\ \mathbf{y}_2^* \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^* : \mathbf{0} \\ \mathbf{X}_2^* : \mathbf{I} \end{bmatrix} \begin{bmatrix} \beta \\ \lambda \end{bmatrix} + \mathbf{e}^* , \quad cov(\mathbf{e}^*) = \sigma^2 \begin{bmatrix} \triangle_1 & \\ & \triangle_2 \end{bmatrix}$$

and we call $\mathbf{y}_2^* = \mathbf{P}_2'\mathbf{y}$ a transformational outlier, or T-outlier.

Of course, fitting $\mathbf{A} = [\mathbf{0}:\mathbf{I}]'$ makes only sense if $\triangle_2 > \mathbf{0}$ is nonsingular, if we assume that the sure equations in the model are not contaminated. In this case, $\triangle_2 > \mathbf{0}$, fitting $\mathbf{A} = [\mathbf{0}:\mathbf{I}]'$ is equivalent to fitting $\mathbf{A} = [\mathbf{0}:\triangle_2]'$, and we may note that in the model (3.70) A- and D-outliers can not be distinguished. In the following we will assume that $\triangle_2 > \mathbf{0}$.

Rewriting (3.70) in terms of the original model, we obtain the model adjusted for a T-outlier as

$$(3.71) \qquad \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 : \mathbf{P}_{12} \\ \mathbf{X}_2 : \mathbf{P}_{22} \end{bmatrix} \begin{bmatrix} \beta \\ \lambda \end{bmatrix} + \mathbf{e} , \quad cov(\mathbf{e}) = \sigma^2\mathbf{V} ,$$

which, as for the case of D- and A-outliers, can be seen as a generalization of the formulation by John and Draper (1978) and Cook and Weisberg (1979).

Testing, adjusted estimates, the reduced data model and recursive residuals

The test-statistic for the hypothesis H_0: $\lambda = \mathbf{0}$ in the model (3.71), as well as adjusted parameter estimates can be obtained directly in the transformed model (3.70), using the invariance of BLU-estimation and tests of linear hypotheses under a transformation of the model (Theorem 2.29 and Corollary 2.29.1). The methods of the section on either A- or D-outliers can be applied, or, if $\triangle$ is nonsingular, the well-known methods from the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{I})$, since (3.70) can then be transformed by $\mathbf{T} = \triangle^{-1/2}$ to have variance-covariance structure $\sigma^2\mathbf{I}$.

Trivially, the reduced data model is

$$(3.72) \qquad \mathbf{y}_1^* = \mathbf{X}_1^*\beta + \mathbf{e}_1^* ; \quad cov(\mathbf{e}_1^*) = \sigma^2\triangle_1$$

and recursive residuals in model (3.70) can be computed using the results of the previous sections.

Downdating

As a third type of downdating we may consider principal components downdating, which is effected by fitting $\mathbf{p}_i$ (say), the i-th column of $\mathbf{P}$, to the model (2.1). Then the i-th principal component $\mathbf{y}_i^*$ is removed from the model.

3.1.6 May the matrix $\mathbf{N}_{22}$ be singular?

When the additional variables $\mathbf{A} = \mathbf{V}_2 = [\mathbf{V}_{21} : \mathbf{V}_{22}]'$ are fitted to the model (3.1), and $\mathbf{V}_{22}$ is nonsingular, then the matrix $\mathbf{N}_{22}$ appearing in the corresponding outlier sum of squares $Q_k = \hat{\mathbf{e}}_2' \mathbf{N}_{22}^{-} \hat{\mathbf{e}}_2$ is nonsingular if and only if $C(\mathbf{V}_2) \cap C(\mathbf{X}) = \{\mathbf{0}\}$. (The question of singular $\mathbf{V}_{22}$ we have treated elsewhere, observing that $\mathbf{V}_{22}$ singular implies that we have introduced at least one superfluous column in $\mathbf{V}_2$, and that at least one estimable linear function $\ell'\beta$ in $\mathbf{y}_2 = \mathbf{X}_2\beta$ is sure.)

An equivalent condition to the nonsingularity of $\mathbf{N}_{22}$ is that $\mathbf{X}\beta$ and λ are estimable in the adjusted model (3.12). Clearly, the outlier sum of squares is well defined even when $\mathbf{N}_{22}$ is singular, and so is the resulting test-statistic (3.13). This follows from $\hat{\mathbf{e}}_2 \in C(\mathbf{N}_{22})$, w.p.1. But we might still pose the question whether it makes any sense, statistically, to fit a variable $\mathbf{V}_2$ to adjust for outliers in the data, such that $C(\mathbf{V}_2) \cap C(\mathbf{X}) \neq \{\mathbf{0}\}$. This question is equivalent with a problem posed by Dunne (1982, p. 6.17) who asked whether "a subset (could) be deemed to be outlying if k degrees of freedom are not essential for their explanation?" (A similar problem can be posed in the context of A-outliers, regarding the nonsingularity of $\mathbf{M}_{22}$.)

We think that the matrix $\mathbf{N}_{22}$ (or $\mathbf{M}_{22}$) may not be singular, believing that, essentially, no inference can be drawn, as far as the hypothesis is concerned that the observations $\mathbf{y}_2$ are not outlying, from an associated F-statistic involving Q_k with singular $\mathbf{N}_{22}$. We list a battery of arguments:

1. Consider a single column $\mathbf{v}_n$ (say), and the corresponding observation y_n. Obviously, it is statistically meaningless to fit the additional variable $\mathbf{v}_n$ if $\mathbf{v}_n \in C(\mathbf{X})$.

2. The F-statistic (3.13), involving the outlier sum of squares Q_k, purports to allow to test the hypothesis H_0: $\lambda = \mathbf{0}$ in the model (3.12). But if $\mathbf{N}_{22}$ is singular, this hypothesis is not strictly testable since λ is not estimable in the model (3.12). Essentially, we do not test the hypothesis H_0: $\lambda = \mathbf{0}$, but H_0: $P_{\mathbf{VZ}|\mathbf{X}}\mathbf{V}_2\lambda = \mathbf{0}$, and the two hypotheses are not equivalent if $C(\mathbf{V}_2) \cap C(\mathbf{X}) \neq \{\mathbf{0}\}$.

3. The usual method to test whether $\mathbf{y}_2$ is outlying is to compare $\mathbf{y}_2$ with an estimate of the mean of $\mathbf{y}_2$ after the removal of the contaminant $\mathbf{y}_2$, or $\mathbf{e}_2$ in the case of D-outliers, from the estimation of the mean of $\mathbf{y}_2$. But if $\mathbf{N}_{22}$ is singular, the mean of $\mathbf{y}_2$ is not estimable after the removal of $\mathbf{e}_2$. Thus we have no uncontaminated estimate of $E(\mathbf{y}_2)$ to compare the possible outlier $\mathbf{y}_2$ with.

At least on one linear function of $E(\mathbf{y}_2)$ we have no statistical information whatsoever if we declare $\mathbf{y}_2$ as an outlier, and thus it is actually impossible to determine whether $\mathbf{y}_2$ is an outlier. There is no good data to compare the bad data with.

4. Consider as example the *LM*

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \beta + \mathbf{e} \; ; \quad cov(\mathbf{e}) = \sigma^2\mathbf{I} \tag{3.73}$$

where $R(\mathbf{X}_1) \cap R(\mathbf{X}_2) = \{\mathbf{0}\}$, and we suspect $\mathbf{y}_2$ to be outlying.

It is clear that all the statistical information on $E(\mathbf{y}_2)$ is contained in $\mathbf{y}_2$, since $\mathbf{X}_2\hat{\beta} = \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^-\mathbf{X}_2'\mathbf{y}_2$ (and $\mathbf{X}_1\hat{\beta} = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^-\mathbf{X}_1'\mathbf{y}_1$) . The model is disconnected, and if σ^2 was known, assuming a normal distribution for $\mathbf{y}$, any search for outliers in (3.73) could be performed in the model parts $\mathbf{y}_1 = \mathbf{X}_1\beta + \mathbf{e}_1$ and $\mathbf{y}_2 = \mathbf{X}_2\beta + \mathbf{e}_2$ separately. Now, specifying $\mathbf{y}_2$ as outlier means that we specify a whole sample as outlying, which is not necessarily wrong but certainly unverifiable by any statistical method, without further assumptions.

We maintain that, in the general case, when $\mathbf{N}_{22}$ is singular, the same happens in a less obvious but still as severe a way: as $\mathbf{X}_2\beta$ is no longer estimable when $\mathbf{y}_2$ is specified as an outlier, at least one linear contrast $\ell'\mathbf{X}_2\beta$ of $\mathbf{X}_2\beta$ is not estimable from the remaining data. Thus we inherently specify the whole sample on $\ell'\mathbf{X}_2\beta$ as outlying.

3.1.7 A unified approach to outliers in a linear model

Hawkins (1980, p. 22) notes that optimal tests to locate outliers among the estimated residuals $\hat{\mathbf{e}}$ from a $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{I})$ are not known, and somewhat later (pp. 86–87) he remarks that "even though all information about outliers is contained in $\hat{\mathbf{e}}$, it does not follow that the individual elements of $\hat{\mathbf{e}}$ are themselves much use in detecting outliers ... (among the observations $\mathbf{y}$)".

It does follow, providing that we know which subset to examine:

The linear model for the estimated residuals $\hat{\mathbf{e}}$ from the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{I})$ can formally be written as

$$\hat{\mathbf{e}} = \mathbf{0}\beta + \hat{\mathbf{e}} \; ; \quad cov(\hat{\mathbf{e}}) = \sigma^2\mathbf{N} \tag{3.74}$$

where $\mathbf{N} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}')$.

If we suspect the residuals $\hat{\mathbf{e}}_2$ to be *D*-outliers, we fit the adjusted model

$$\begin{bmatrix} \hat{\mathbf{e}}_1 \\ \hat{\mathbf{e}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{0} : \mathbf{N}_{12} \\ \mathbf{0} : \mathbf{N}_{22} \end{bmatrix} \begin{bmatrix} \beta \\ \lambda \end{bmatrix} + \mathbf{e} \; , \quad cov(\mathbf{e}) = \sigma^2 \begin{bmatrix} \mathbf{N}_{11} : \mathbf{N}_{12} \\ \mathbf{N}_{21} : \mathbf{N}_{22} \end{bmatrix} \tag{3.75}$$

Clearly, testing the hypothesis H_0: $\lambda = \mathbf{0}$ in (3.75) leads to the same F-statistic which is obtained when H_0: $\lambda = \mathbf{0}$ is tested in the underlying $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{I})$ which is adjusted for the outliers $\mathbf{y}_2$, namely

(3.76) $$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 : \mathbf{0} \\ \mathbf{X}_2 : \mathbf{I} \end{bmatrix} \begin{bmatrix} \beta \\ \lambda \end{bmatrix} + \mathbf{e}\,, \quad cov(\mathbf{e}) = \sigma^2\mathbf{I}\,.$$

Thus testing for outliers in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{I})$ is equivalent to testing for D-outliers in the $LM(\hat{\mathbf{e}}, \mathbf{0}, \sigma^2\mathbf{N})$, where $\hat{\mathbf{e}}$ is the vector of estimated residuals from the model $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{I})$, and $\sigma^2\mathbf{N} = \sigma^2(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}')$ is the variance-covariance matrix of $\hat{\mathbf{e}}$. Thus, knowing $\hat{\mathbf{e}}$ and $\mathbf{N}$ we can test for outliers among the observations $\mathbf{y}$, without knowing $\mathbf{y}$ and $\mathbf{X}$ (of course, $C(\mathbf{X})$ is known if $\mathbf{N}$ is known).

Similarly, in the general linear model $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$, testing for D-outliers in this model is equivalent to testing for D-outliers in the $LM(\hat{\mathbf{e}}, \mathbf{0}, \sigma^2\mathbf{N})$, where $\hat{\mathbf{e}}$ is the vector of estimated residuals from the model $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ and $\sigma^2\mathbf{N} = \sigma^2(\mathbf{V} + \mathbf{X}\mathbf{U}\mathbf{X}' - \mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^-\mathbf{X}')$ is the variance-covariance matrix of $\hat{\mathbf{e}}$.

Observing that the variance-covariance matrix $\sigma^2\mathbf{N}$ of the vector of estimated residuals $\hat{\mathbf{e}}$ from a $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ is always singular (except for the trivial case when $\mathbf{X} = \mathbf{0}$ and $\mathbf{V}$ nonsingular), whether or not $\mathbf{V} = \mathbf{I}$, $\mathbf{V}$ is nonsingular or $\mathbf{V}$ is singular, we realize that there is no fundamental distinction between testing for outliers in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{I})$ and testing for D-outliers in any general model $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$. Thus we have a unified approach to outlier-testing in the general linear model, which involves the vector $\hat{\mathbf{e}}$ of estimated residuals from the model and its variance-covariance matrix $\sigma^2\mathbf{N}$.

With respect to A- and T-outliers in a general linear model, testing the null hypothesis H_0: $\lambda = \mathbf{0}$ in the adjusted models (3.44) and (3.71) is equivalent to testing the hypothesis H_0: $\lambda = \mathbf{0}$ in the models $(\hat{\mathbf{e}}, P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\lambda, \sigma^2\mathbf{N})$ where $\mathbf{A} = [\mathbf{0}:\mathbf{I}]'$ and $\mathbf{A} = \mathbf{P}_2$ respectively.

What we essentially do is that we transform the respective adjusted models (3.12), (3.44) and (3.71) by the *singular* transformation $\mathbf{T} = P_{\mathbf{VZ}|\mathbf{X}}$. If the adjusted model is

(3.77) $$\mathbf{y} = [\mathbf{X}:\mathbf{A}] \begin{bmatrix} \beta \\ \lambda \end{bmatrix} + \mathbf{e}\,; \quad cov(\mathbf{e}) = \sigma^2\mathbf{V}$$

with $\mathbf{A} = \mathbf{V}_2$, $\mathbf{A} = [\mathbf{0}:\mathbf{I}]'$ and $\mathbf{A} = \mathbf{P}_2$ respectively, the respective model transformed by $\mathbf{T} = P_{\mathbf{VZ}|\mathbf{X}}$ is

(3.78) $$\hat{\mathbf{e}} = [\mathbf{0}:P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}] \begin{bmatrix} \beta \\ \lambda \end{bmatrix} + \tilde{\mathbf{e}}\,; \quad cov(\tilde{\mathbf{e}}) = \sigma^2\mathbf{N}\,.$$

Clearly, for $\mathbf{A} = \mathbf{V}_2$ we have

(3.79) $$\begin{aligned} P_{\mathbf{VZ}|\mathbf{X}}\mathbf{V}_2 &= (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^-\mathbf{X}'\mathbf{V}^*)\mathbf{V}_2 \\ &= \mathbf{V}_2 + \mathbf{X}\mathbf{U}\mathbf{X}_2' - \mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^-\mathbf{X}'\mathbf{V}^*(\mathbf{V}_2 + \mathbf{X}\mathbf{U}\mathbf{X}_2') \\ &= \mathbf{V}_2 + \mathbf{X}\mathbf{U}\mathbf{X}_2' - \mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^-\mathbf{X}_2' \\ &= \mathbf{N}_2\,. \end{aligned}$$

This proves our claim above that testing for D-outliers in the models $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ and $(\hat{\mathbf{e}}, \mathbf{0}, \sigma^2\mathbf{N})$ respectively is equivalent.

Summing up, testing respectively for D-, A- and T-outliers in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ and fitting in the process the additional variables $\mathbf{A} = \mathbf{V}_2$, $\mathbf{A} = [\mathbf{0}:\mathbf{I}]'$ and $\mathbf{A} = \mathbf{P}_2$, is equivalent to testing for outliers in the $LM(\hat{\mathbf{e}}, \mathbf{0}\beta, \sigma^2\mathbf{N})$ and fitting respectively the additional variables

$$\mathbf{A} = P_{\mathbf{VZ}|\mathbf{X}}\mathbf{V}_2 \quad = P_{\mathbf{VZ}|\mathbf{X}}\mathbf{V}^1 \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix},$$

$$\mathbf{A} = P_{\mathbf{VZ}|\mathbf{X}}\mathbf{P}_2\triangle_2^{1/2} = P_{\mathbf{VZ}|\mathbf{X}}\mathbf{V}^{1/2} \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix}, \quad \text{and}$$

$$\mathbf{A} = P_{\mathbf{VZ}|\mathbf{X}} \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} = P_{\mathbf{VZ}|\mathbf{X}}\mathbf{V}^0 \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix}.$$

Here $\mathbf{P}\triangle^{1/2}$ is denoted somewhat loosely by $\mathbf{V}^{1/2}$, where $\triangle^{1/2}$ denotes the diagonal matrix satisfying $\triangle^{1/2} \cdot \triangle^{1/2} = \triangle$.

3.2 INFLUENTIAL OBSERVATIONS

In this section we generalize statistics proposed by Cook (1977) and Andrews and Pregibon (1978) to detect influential observations in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{I})$ for the general linear model $(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ with arbitrary known variance-covariance structure.

It is well-known that those statistics, in the case $\mathbf{V} = \sigma^2\mathbf{I}$, are combined measures for outliers and influence. Consequently, and corresponding to three types of outlier distinguishable in the general linear model, each statistic, *Cook's distance* and the *AP-statistic*, can be generalized in three ways, to detect observations which are respectively outlying and/or influential with respect to D-, A- and T-outliers.

We show that those statistics are invariant under a reparametrization of the underlying linear model, and for the AP-statistic we consider the case when the influence of certain observations on the estimate of only a specified subset of linear contrasts of the regression parameters is under investigation, rather than on the estimate of all the regression parameters.

3.2.1 Cook's distance

To detect influential observations in the Gauß-model

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \beta + \mathbf{e}\,; \quad cov(\mathbf{e}) = \sigma^2 \begin{bmatrix} \mathbf{I}_{n-k} : \mathbf{0} \\ \mathbf{0} \quad : \mathbf{I}_k \end{bmatrix} \tag{2.38}$$

Cook (1977) proposed the statistic

(3.80) $$C_{(2)} = \frac{(\hat\beta - \tilde\beta)'\mathbf{X}'\mathbf{X}(\hat\beta - \tilde\beta)}{\hat{\mathbf{e}}'\hat{\mathbf{e}}} \cdot \frac{n-p}{p}$$

where $\mathbf{X}$ was assumed to have full rank p and $\hat\beta$ and $\tilde\beta$ are respectively the BLUE for β in the full model (2.38) and in the model with the observations $\mathbf{y}_2$ (say) deleted, whose influence on the estimate $\hat\beta$ for β should be determined. Equivalently, $\tilde\beta$ is the BLUE for β in the adjusted model

(3.81) $$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 : \mathbf{0} \\ \mathbf{X}_2 : \mathbf{I} \end{bmatrix} \begin{bmatrix} \beta \\ \lambda \end{bmatrix} + \mathbf{e}\,, \quad cov(\mathbf{e}) = \sigma^2\mathbf{I}\,.$$

Note that $\{\tilde\beta \mid C_{(2)}(\tilde\beta) \leq F_{p,n-p}\,(1-\alpha)\}$ is a $(1-\alpha)\cdot 100\%$ confidence ellipsoid for β in the model (2.38), and thus $C_{(2)}$ can be interpreted as detecting those observations to be influential whose removal from the model would yield the adjusted estimate $\tilde\beta$ for β far out from the center $\hat\beta$ of the confidence ellipsoid for β.

Generalizing for the linear model under arbitrary variance-covariance structure, and corresponding to three types of outlier, we define

(3.82) $$C_{(2)} = \frac{(\hat\beta - \tilde\beta)'\mathbf{X}'\mathbf{V}^*\mathbf{X}(\hat\beta - \tilde\beta)}{\hat{\mathbf{e}}'\mathbf{V}^-\hat{\mathbf{e}}} \cdot \frac{r([\mathbf{X}:\mathbf{V}]) - r(\mathbf{X})}{r(\mathbf{V}) - r([\mathbf{X}:\mathbf{V}]) + r(\mathbf{X})}$$

where $\mathbf{X}\hat\beta$ is the BLUE for $\mathbf{X}\beta$ in the original model (3.1), and $\mathbf{X}\tilde\beta$ is respectively the BLUE for $\mathbf{X}\beta$ in the adjusted models (3.12), (3.44) and (3.71). (We require that $\mathbf{X}\beta$ is estimable in the adjusted models, i.e. $C(\mathbf{A}) \cap C(\mathbf{X}) = \{\mathbf{0}\}$, where $\mathbf{A}$ denotes the new variable in the model.) Noting that

(3.83) $$C = \{\mathbf{X}\tilde\beta \mid C_{(2)}(\mathbf{X}\tilde\beta) \leq F_{r,s}(1-\alpha)\}$$

is a $(1-\alpha)\cdot 100\%$ confidence ellipsoid for $\mathbf{X}\beta$ in the original model (3.1), and recalling that fitting the corresponding adjusted models is equivalent to the removal of the error term $\mathbf{e}_2$, the observation $\mathbf{y}_2$ or the PC $\mathbf{y}_2^*$ from the model, the interpretation of $C_{(2)}$ in the Gauß-model as given above can also be applied in the linear model under arbitrary variance.

We denote by $C_{(2)}^D$, $C_{(2)}^A$ and $C_{(2)}^T$ the statistics $C_{(2)}$ as given by (3.82) which are computed with respect to the adjusted models (3.12), (3.44) and (3.71).

Using Theorem 2.33 we can write

(3.84) $$\begin{aligned} \mathbf{X}\hat\beta - \mathbf{X}\tilde\beta &= -P_{\mathbf{X}|\mathbf{VZ}}\mathbf{A}\tilde\lambda \\ &= P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\tilde\lambda - \mathbf{A}\tilde\lambda \end{aligned}$$

where $\mathbf{A}$ is the corresponding new variable (If $\mathbf{A} = [\mathbf{0} : \mathbf{I}]'$, and $C(\mathbf{A}) \not\subset C([\mathbf{X}:\mathbf{V}])$, $\mathbf{A}$ must be replaced, here and in the following, by $\mathbf{A}_1^*$ such that $\mathbf{A}\lambda = \mathbf{A}_1^*\lambda_1 + \mathbf{A}_2^*\lambda_2$ is a reparametrization of $\mathbf{A}\lambda$ with $C(\mathbf{A}_1^*) \subset C([\mathbf{X}:\mathbf{V}])$ and $C(\mathbf{A}_2^*) \cap C([\mathbf{X}:\mathbf{V}]) = \{\mathbf{0}\}$) .

Clearly, $P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\tilde\lambda \in C(\mathbf{V})$, and when $\mathbf{A} = \mathbf{V}_2$ or $\mathbf{A} = \mathbf{P}_2$ we have $\mathbf{A}\tilde\lambda \in C(\mathbf{V})$. Thus $C_{(2)}^D$ and $C_{(2)}^T$ are invariant over all choices of a g-inverse $\mathbf{V}^*$ of $\mathbf{V}$ in (3.82).

If $\mathbf{A} = [\mathbf{0}:\mathbf{I}]'$, then $\mathbf{A}\tilde{\lambda} \in C(\mathbf{V})$ if and only if $C(\mathbf{A}) \cap C([\mathbf{X}:\mathbf{V}]) \subset C(\mathbf{V})$, i.e. if and only if fitting $\mathbf{A}$ leaves the sure equations invariant. Now, if fitting $\mathbf{A}$ changes the sure equations in the model, then the corresponding observations $\mathbf{y}_2$ are influential with probability 1, and we do not need the statistic $C_{(2)}^{A}$ to ascertain that fact. If, however, fitting $\mathbf{A}$ does not change the sure equations, then $\mathbf{A}\tilde{\lambda} \in C(\mathbf{V})$, and we have also in this case that $C_{(2)}^{A}$ is invariant over all choices of a g-inverse $\mathbf{V}^*$ of $\mathbf{V}$ in (3.82).

In the following we may assume that $\mathbf{A}\tilde{\lambda} \in C(\mathbf{V})$, but this assumption which ensures the invariance of $C_{(2)}^{A}$ over all choices of a g-inverse $\mathbf{V}^*$ of $\mathbf{V}$ is not necessary for the development.

Similar to the development in Draper and John (1981), we write the numerator in (3.82) in an alternative form:

$$
\begin{aligned}
(3.85) \qquad & (\hat{\beta}-\tilde{\beta})'\mathbf{X}'\mathbf{V}^*\mathbf{X}(\hat{\beta}-\tilde{\beta}) \\
&= (P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\tilde{\lambda}-\mathbf{A}\tilde{\lambda})'\mathbf{V}^*(P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\tilde{\lambda}-\mathbf{A}\tilde{\lambda}) \ , \quad \text{from (3.84)} \\
&= (P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\tilde{\lambda})'\mathbf{V}^*(P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\tilde{\lambda}) - 2(P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\tilde{\lambda})'\mathbf{V}^*\mathbf{A}\tilde{\lambda} + (\mathbf{A}\tilde{\lambda})'\mathbf{V}^*\mathbf{A}\tilde{\lambda} \\
&= Q_k - 2(P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\tilde{\lambda})'\mathbf{V}^*(P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\tilde{\lambda} + P_{\mathbf{X}|\mathbf{VZ}}\mathbf{A}\tilde{\lambda}) + (\mathbf{A}\tilde{\lambda})'\mathbf{V}^*\mathbf{A}\tilde{\lambda} \\
&= Q_k - 2Q_k + (\mathbf{A}\tilde{\lambda})'\mathbf{V}^*\mathbf{A}\tilde{\lambda} \\
&= (\mathbf{A}\tilde{\lambda})'\mathbf{V}^*\mathbf{A}\tilde{\lambda} - Q_k \ ,
\end{aligned}
$$

where Q_k denotes the additional sum of squares due to fitting $\mathbf{A}$.

Thus $C_{(2)}$ can be written as

$$
(3.86) \qquad C_{(2)} = \frac{Q_k}{\hat{\mathbf{e}}'\mathbf{V}^-\hat{\mathbf{e}}} \cdot \left(\frac{(\mathbf{A}\tilde{\lambda})'\mathbf{V}^*\mathbf{A}\tilde{\lambda}}{Q_k} - 1 \right) \cdot \frac{r([\mathbf{X}:\mathbf{V}]) - r(\mathbf{X})}{r(\mathbf{V}) - r([\mathbf{X}:\mathbf{V}]) + r(\mathbf{X})} \ .
$$

This factorization of $C_{(2)}$ proves the invariance of $C_{(2)}$ under a reparametrization of the underlying linear model (3.1), since clearly Q_k, $\hat{\mathbf{e}}'\mathbf{V}^-\hat{\mathbf{e}}$ and $\mathbf{A}\tilde{\lambda}$ are invariant under a reparametrization of the model.

The first component $Q_k/\hat{\mathbf{e}}'\mathbf{V}^-\hat{\mathbf{e}}$ in (3.86) is an outlier measure as interpreted by Draper and John (1981), since it is large when Q_k, "the outlier sum of squares", is large.

To investigate the second component more closely, we compute $(\mathbf{A}\tilde{\lambda})'\mathbf{V}^*\mathbf{A}\tilde{\lambda}$ for $\mathbf{A} = \mathbf{V}_2$ and $\mathbf{A} = [\mathbf{0}:\mathbf{I}]'$ (for $\mathbf{A} = \mathbf{P}_2$ we are, in the transformed model (3.70), back at the case $\mathbf{A} = [\mathbf{0}:\mathbf{I}]'$):

$$
\begin{aligned}
(3.87) \qquad & \tilde{\lambda}'\mathbf{V}_2'\mathbf{V}^*\mathbf{V}_2\tilde{\lambda} \\
&= \tilde{\lambda}'\mathbf{V}_{22}\tilde{\lambda} \\
&= \hat{\mathbf{e}}_2'\mathbf{N}_{22}^-\mathbf{V}_{22}\mathbf{N}_{22}^-\hat{\mathbf{e}}_2 \ .
\end{aligned}
$$

(3.88)
$$\tilde{\lambda}'[\mathbf{0}:\mathbf{I}]\mathbf{V}^* \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} \tilde{\lambda}$$
$$= \hat{\mathbf{e}}'\mathbf{M}_2'\mathbf{M}_{22}^-\mathbf{V}_{22}^*\mathbf{M}_{22}^-\mathbf{M}_2\hat{\mathbf{e}} \ .$$

Now, if $\mathbf{A} = \mathbf{V}_2 = \mathbf{v}_n$ is a single column in (3.87), then

(3.89)
$$\frac{(\mathbf{A}\tilde{\lambda})'\mathbf{V}^*\mathbf{A}\tilde{\lambda}}{Q_k^D} - 1$$
$$= \frac{\hat{\mathbf{e}}_n^2 \cdot \mathrm{V}_{nn} \cdot \mathrm{N}_{nn}^{-2}}{\hat{\mathbf{e}}_n^2 \cdot \mathrm{N}_{nn}^{-1}} - 1$$
$$= \frac{\mathrm{V}_{nn}}{\mathrm{N}_{nn}} - 1$$
$$= \frac{\mathrm{V}_{nn} - \mathrm{N}_{nn}}{\mathrm{N}_{nn}} \ .$$

Similarly, if $\mathbf{A} = \mathbf{u}_n = (0, \ldots, 0, 1)'$ is a single column in (3.88), we obtain

(3.90)
$$\frac{(\mathbf{A}\tilde{\lambda})'\mathbf{V}^*\mathbf{A}\tilde{\lambda}}{Q_k^A} - 1$$
$$= \frac{\hat{\mathbf{e}}'\mathbf{m}_n'\mathrm{M}_{nn}^{-1}\mathrm{V}_{nn}^*\mathrm{M}_{nn}^{-1}\mathbf{m}_n\hat{\mathbf{e}}}{\hat{\mathbf{e}}'\mathbf{m}_n'\mathrm{M}_{nn}^{-1}\mathbf{m}_n\hat{\mathbf{e}}} - 1$$
$$= \frac{\mathrm{V}_{nn}^* - \mathrm{M}_{nn}}{\mathrm{M}_{nn}} \ .$$

For $\mathbf{V} = \mathbf{I}$, (3.89) and (3.90) are identical, and the numerator $\mathrm{V}_{nn} - \mathrm{N}_{nn} = \mathrm{V}_{nn}^* - \mathrm{M}_{nn} = \mathbf{x}_n'(\mathbf{X}'\mathbf{X})^-\mathbf{x}_n$ is the leverage of the data point y_n (Hoaglin and Welsh, 1978), and thus in the case of a single observation the second term in (3.86) is large if the leverage of the observation in question is large.

For $\mathbf{V} \neq \mathbf{I}$ we can also interpret the diagonal elements of the matrices

(3.91)
$$\mathbf{V} - \mathbf{N} = \mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^-\mathbf{X}' - \mathbf{X}\mathbf{U}\mathbf{X}' \ , \quad \text{and}$$

(3.92)
$$\mathbf{V}^* - \mathbf{M} = \mathbf{V}^*\mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^-\mathbf{X}'\mathbf{V}^*$$

as the leverage of the corresponding observation, generalized for arbitrary $\mathbf{V}$ and with respect to D- and A-outliers. The i-th leverage $\mathrm{V}_{ii}^* - \mathrm{M}_{ii}$ is unique over all possible choices of $\mathbf{V}^*$ if $\mathbf{u}_i \in C([\mathbf{X} : \mathbf{V}])$.

The different values for the leverage of an observation with respect to different types of outlier result from the fact that with the i-th diagonal element of (3.91) we have the leverage of observation i via the error term e_i, i.e. via a vector $\mathbf{v}_i\mathrm{V}_{ii}^{-1}\mathrm{e}_i$, whereas with the i-th diagonal element of (3.92) we have the leverage of observation y_i via a mean shift $(0, \ldots, \lambda_i, 0, \ldots, 0)'$.

If $\mathbf{A}$ in (3.86) consists of more than one column, or equivalently if more than one observation is under investigation by $C_{(2)}$, the terms involving the residuals $\hat{\mathbf{e}}_2$ and $\mathbf{M}_2\hat{\mathbf{e}}$ in (3.89) and (3.90) respectively do not cancel out, and Draper and John (1981) admit that in this case they "do not see a meaningful interpretation for (the second component in (3.86)) in terms of the influence of a set of points".

But noting that

$$(3.93) \qquad \frac{(\mathbf{A}\tilde{\lambda})'\mathbf{V}^*\mathbf{A}\tilde{\lambda}}{Q_k} = \frac{(\mathbf{A}\tilde{\lambda})'\mathbf{V}^*\mathbf{A}\tilde{\lambda}}{(P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\tilde{\lambda})'\mathbf{V}^*(P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\tilde{\lambda})} \geq 1 ,$$

the second component in (3.86) will be large if (3.93) is large. (3.93) equals 1 and thus the second component in (3.86) will be zero if and only if

$$(3.94) \qquad \begin{aligned} & P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\tilde{\lambda} = \mathbf{A}\tilde{\lambda} \\ \Leftrightarrow \quad & \mathbf{A}\tilde{\lambda} \in C(\mathbf{VZ}) \\ \Leftrightarrow \quad & \mathbf{X}\hat{\beta} = \mathbf{X}\tilde{\beta} . \end{aligned}$$

Thus the second term in (3.86) measures the actual influence of a set of data points, since the difference between $\mathbf{A}\tilde{\lambda}$ and $P_{\mathbf{VZ}|\mathbf{X}}\mathbf{A}\tilde{\lambda}$ is $P_{\mathbf{X}|\mathbf{VZ}}\mathbf{A}\tilde{\lambda}$ which is precisely the difference between $\mathbf{X}\hat{\beta}$ and $\mathbf{X}\tilde{\beta}$.

This actual influence itself is influenced by two factors: firstly the potential influence of the observations $\mathbf{y}_2$, as reflected by the matrix $\mathbf{A}'P'_{\mathbf{X}|\mathbf{VZ}}\mathbf{V}^*P_{\mathbf{X}|\mathbf{VZ}}\mathbf{A}$, or some function of it (e.g. its eigenvalues). Clearly, large eigenvalues (say) of $\mathbf{A}'P'_{\mathbf{X}|\mathbf{VZ}}\mathbf{V}^*P_{\mathbf{X}|\mathbf{VZ}}\mathbf{A}$ will cause the second component in (3.86) to be large, for fixed $\tilde{\lambda}$, and *vice versa*. Secondly, the actual influence of $\mathbf{y}_2$ will crucially depend on which eigenvalues of $\mathbf{A}'P'_{\mathbf{X}|\mathbf{VZ}}\mathbf{V}^*P_{\mathbf{X}|\mathbf{VZ}}\mathbf{A}$ correspond to $\tilde{\lambda}$. If $\tilde{\lambda}$ corresponds to a zero eigenvalue, i.e. $\mathbf{A}\tilde{\lambda} \in C(\mathbf{VZ})$, then the second component in (3.86) is zero even if $\tilde{\lambda}$ is large. In this context, see also the work by Cook and Weisberg (1982, pp. 137–141).

In summary, the first component of (3.86) measures how much $\mathbf{y}_2$ is outlying, and the second component of (3.86) measures the actual influence of $\mathbf{y}_2$ on the estimate $\mathbf{X}\hat{\beta}$ for $\mathbf{X}\beta$, combining potential influence and the relative actual situation of the residual $\hat{\mathbf{e}}$ via $\mathbf{A}\tilde{\lambda}$.

In the important special case of $\mathbf{V} = \mathbf{I}$, the condition (3.94) amounts to

$$(3.95) \qquad \mathbf{X}\hat{\beta} = \mathbf{X}\tilde{\beta} \quad \Leftrightarrow \quad \tilde{\lambda} \in C(\mathbf{X}_2)^{\perp} .$$

For a single observation, (3.95) can not be satisfied except in the trivial case $\mathbf{x}_2 = \mathbf{0}$. If two observations are considered, (3.95) can only be satisfied when $\mathbf{X}_2$ consists of two identical predictors or row vectors. Thus the effect of two outliers on the estimation of $\mathbf{X}\beta$ can only cancel out if they occur at the same point in the design space. Similarly, if k outliers are considered, their effect on the estimation of $\mathbf{X}\beta$ can only cancel out if $rank(\mathbf{X}_2) < k$.

If $\mathbf{V} \neq \mathbf{I}$ but nonsingular, we have in the case of D-outliers that $\mathbf{V}_2\tilde{\lambda} \in C(\mathbf{VZ})$ is equivalent with $\tilde{\lambda} \in C(\mathbf{X}_2)^\perp$, thus obtaining the same condition as in the case $\mathbf{V} = \mathbf{I}$. In the case of A-outliers, we arrive at the condition $\tilde{\lambda} \in C(\tilde{\mathbf{X}}_2)^\perp$, where $\tilde{\mathbf{X}} = \mathbf{V}^{-1}\mathbf{X}$.

We consider as a simple example the estimation of a regression line in a $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{I})$, where the effect of two specified observations is under investigation.

For the actual influence of two given data points to be zero we need, as pointed out above, that

(3.96)
$$\tilde{\lambda} \in C(\mathbf{X}_2)^\perp, \quad \text{where } rank(\mathbf{X}_2) = 1 \iff C(\mathbf{X}_2) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$
$$\iff \mathbf{N}_{22}^{-1}\hat{\mathbf{e}}_2 \perp \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$
$$\iff \hat{\lambda} = \begin{bmatrix} \hat{\lambda}_1 \\ \hat{\lambda}_2 \end{bmatrix} \text{ such that } \hat{\lambda}_1 = -\hat{\lambda}_2 .$$

But $\mathbf{N}_{22}$ is of the form

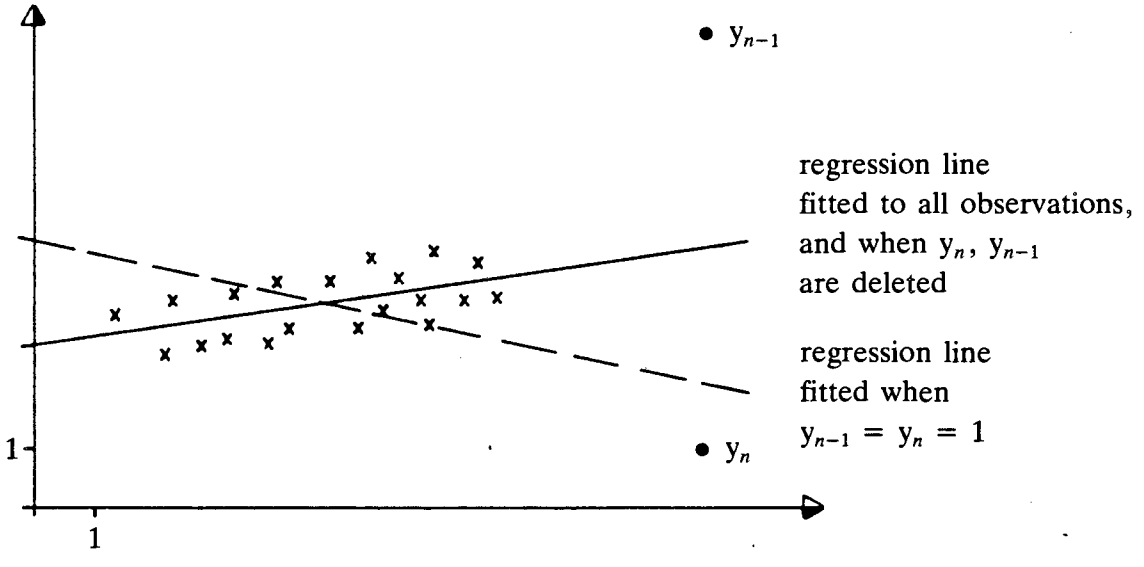
(3.97)
$$\mathbf{N}_{22} = \begin{bmatrix} 1-a & : & -a \\ -a & : & 1-a \end{bmatrix}, \quad \text{for some } 0 < a < 1$$

and similarly $\mathbf{N}_{22}^{-1}$ is of the form

(3.98)
$$\mathbf{N}_{22}^{-1} = \begin{bmatrix} 1-b & : & -b \\ -b & : & 1-b \end{bmatrix}, \quad b = \frac{a}{a+b-1} .$$

Thus $\hat{\lambda}_1 = -\hat{\lambda}_2$ if and only if $\hat{\mathbf{e}}_n = -\hat{\mathbf{e}}_{n-1}$, and only in this case the combined effect of two given residuals does cancel out. Further, we note that the potential influence of $\mathbf{y}_2$ is large if a is large, i.e. if y_n and y_{n-1} lie far out in the design space. We illustrate these effects by the following figure.

Figure 3.4



The data points y_n and y_{n-1} denoted by • clearly have the largest Q_2 in the sample, for all combinations of two observations, but with $\hat{e}_{n-1} = -\hat{e}_n = 3.5$ their combined influence is zero.

We note that y_n and y_{n-1} do not only have the largest Q_2, but they also have the largest potential influence of all pairs of observations in the sample, since they lie furthest out from the bulk of the data. Thus it may happen that a pair of observations has the largest Q_2 and the largest potential influence of all pairs of observations in the sample, but their combined actual influence might be zero, or close to zero.

When $y_{n-1} = y_n = 1$ and consequently $\hat{e}_{n-1} = -\hat{e}_n$, we obtain the same value for Q_2 as before, but a large influence of y_n and y_{n-1} on the regression line fitted is given.

Finally, and for completeness, we may generalize in a straightforward way some alternative choices to (3.80) of normed influence measures, as given by Cook and Weisberg (1982, p. 124) for the case $\mathbf{V} = \mathbf{I}$.

Alternatively to (3.80) the distance measure may be scaled by $\hat{\mathbf{e}}'\hat{\mathbf{e}} - Q_k$ rather than by $\hat{\mathbf{e}}'\hat{\mathbf{e}}$, and the generalization for arbitrary $\mathbf{V}$ obviously involves the scaling of (3.82) by $\hat{\mathbf{e}}'\mathbf{V}^{-}\hat{\mathbf{e}} - Q_k^D$ with respect to $C_{(2)}^D$, and by $\hat{\mathbf{e}}'\mathbf{V}^{-}\hat{\mathbf{e}} - Q_k^A$ with respect to $C_{(2)}^A$, instead of scaling by $\hat{\mathbf{e}}'\mathbf{V}^{-}\hat{\mathbf{e}}$.

Another modification of (3.80) is possible by replacing the matrix $\mathbf{X}'\mathbf{X}$ in the numerator by $\mathbf{X}_1'\mathbf{X}_1$, i.e. $\mathbf{X}_2$ is removed from $\mathbf{X}$. Similarly, for $C_{(2)}^A$ we would replace $\mathbf{X}'\mathbf{V}^*\mathbf{X}$ by $\mathbf{X}_1'\mathbf{V}_{11}^*\mathbf{X}_1$, and for $C_{(2)}^D$ we would replace $\mathbf{X}'\mathbf{V}^*\mathbf{X}$ by $\tilde{\mathbf{X}}_1'\tilde{\mathbf{V}}_{11}^*\tilde{\mathbf{X}}_1$, where

$$\tilde{\mathbf{X}}_1 = \mathbf{X}_1 - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{X}_2 , \quad \text{and} \tag{3.99}$$

$$\tilde{\mathbf{V}}_{11} = \mathbf{V}_{11} - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21} . \tag{3.100}$$

Those modified distance measures use the design matrices and variance-covariance structures from the corresponding reduced data models (3.49) and (3.25).

3.2.2 The Andrews-Pregibon statistic

With respect to the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{I})$, Andrews and Pregibon (1978), to "find the outliers that matter", proposed the statistic

$$AP_{(2)} = \frac{|\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2|}{|\tilde{\mathbf{X}}_1'\tilde{\mathbf{X}}_1|} \tag{3.101}$$

where $\tilde{\mathbf{X}}_1 = [\mathbf{X}:\mathbf{y}]$, $\tilde{\mathbf{X}}_2 = [\mathbf{X}:\mathbf{A}:\mathbf{y}]$, and with $\mathbf{A} = [\mathbf{0}:\mathbf{I}_k]$ (say) the influence of the last k observations $\mathbf{y}_2$ in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{I})$ on the least-squares estimate of β is measured by the "AP-statistic" $AP_{(2)}$. "Small values of $AP_{(2)}$ are associated with deviant and/or influential observations" (Andrews and Pregibon, 1978).

We note that $AP_{(2)}$ as in (3.101) is only defined if $\mathbf{X}$ is of full rank, since otherwise $|\tilde{\mathbf{X}}_1'\tilde{\mathbf{X}}_1| = 0$. Further, $C(\mathbf{A}) \cap C(\mathbf{X}) = \{\mathbf{0}\}$ is required for $AP_{(2)}$ to be nonzero.

In this section we generalize the statistic $AP_{(2)}$ for the linear model under arbitrary variance-covariance structure, with respect to three types of outlier. The invariance of $AP_{(2)}$ under a rank preserving reparametrization of the underlying linear model is shown, and thus a definition of $AP_{(2)}$ using determinants as in (3.101) is always possible. Finally, the statistic is generalized to be applicable when the influence of a group of observations on the estimate of only a specified subset of linear contrasts of the regression parameters is under investigation, rather than on the estimate of all regression parameters.

With respect to the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$, where $\mathbf{X}$ is assumed to be of full rank for the present, we define the statistics

(3.102)
$$AP_{(2)} = \frac{|\tilde{\mathbf{X}}_2'\mathbf{V}^*\tilde{\mathbf{X}}_2|}{|\tilde{\mathbf{X}}_1'\mathbf{V}^*\tilde{\mathbf{X}}_1|}$$

where $\tilde{\mathbf{X}}_1 = [\mathbf{X}:\mathbf{y}]$ as before, $\tilde{\mathbf{X}}_2 = [\mathbf{X}:\mathbf{A}:\mathbf{y}]$, and with $\mathbf{A}$ respectively set equal to $\mathbf{A} = \mathbf{V}_2$, $\mathbf{A} = [\mathbf{0}:\mathbf{I}]'$ and $\mathbf{A} = \mathbf{P}_2$ we obtain the statistics $AP_{(2)}^D$, $AP_{(2)}^A$ and $AP_{(2)}^T$, corresponding to three types of outlier.

Along the lines of the development in Draper and John (1981) we write $AP_{(2)}$ in an equivalent but more revealing form, initially assuming that $C(\mathbf{A}) \cap C(\mathbf{X}) = \{\mathbf{0}\}$, and $\mathbf{A}$ is of full rank. (As before, if $C(\mathbf{A}) \subset C([\mathbf{X}:\mathbf{V}])$ is not satisfied, $\mathbf{A}$ must be replaced by $\mathbf{A}_1^*$ as given in the previous section.) The denominator and the numerator in (3.102) can respectively be written as

(3.103)
$$\begin{aligned}
&|\tilde{\mathbf{X}}_1'\mathbf{V}^*\tilde{\mathbf{X}}_1| \\
&= \begin{vmatrix} \mathbf{X}'\mathbf{V}^*\mathbf{X} : \mathbf{X}'\mathbf{V}^*\mathbf{y} \\ \mathbf{y}'\mathbf{V}^*\mathbf{X} : \mathbf{y}'\mathbf{V}^*\mathbf{y} \end{vmatrix} \\
&= |\mathbf{X}'\mathbf{V}^*\mathbf{X}| \cdot (\mathbf{y}'\mathbf{V}^*\mathbf{y} - \mathbf{y}'\mathbf{V}^*\mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^*\mathbf{y}) \\
&= |\mathbf{X}'\mathbf{V}^*\mathbf{X}| \cdot \hat{\mathbf{e}}'\mathbf{V}^*\hat{\mathbf{e}} \,.
\end{aligned}$$

(3.104)
$$\begin{aligned}
&|\tilde{\mathbf{X}}_2'\mathbf{V}^*\tilde{\mathbf{X}}_2| \\
&= \begin{vmatrix} \mathbf{X}'\mathbf{V}^*\mathbf{X} : \mathbf{X}'\mathbf{V}^*\mathbf{A} : \mathbf{X}'\mathbf{V}^*\mathbf{y} \\ \mathbf{A}'\mathbf{V}^*\mathbf{X} : \mathbf{A}'\mathbf{V}^*\mathbf{A} : \mathbf{A}'\mathbf{V}^*\mathbf{y} \\ \mathbf{y}'\mathbf{V}^*\mathbf{X} : \mathbf{y}'\mathbf{V}^*\mathbf{A} : \mathbf{y}'\mathbf{V}^*\mathbf{y} \end{vmatrix} \\
&= |\mathbf{X}'\mathbf{V}^*\mathbf{X}| \cdot \left| \begin{bmatrix} \mathbf{A}'\mathbf{V}^*\mathbf{A} : \mathbf{A}'\mathbf{V}^*\mathbf{y} \\ \mathbf{y}'\mathbf{V}^*\mathbf{A} : \mathbf{y}'\mathbf{V}^*\mathbf{y} \end{bmatrix} - \begin{bmatrix} \mathbf{A}'\mathbf{V}^*\mathbf{X} \\ \mathbf{y}'\mathbf{V}^*\mathbf{X} \end{bmatrix} (\mathbf{X}'\mathbf{V}^*\mathbf{X})^{-1}[\mathbf{X}'\mathbf{V}^*\mathbf{A} : \mathbf{X}'\mathbf{V}^*\mathbf{y}] \right| \\
&= |\mathbf{X}'\mathbf{V}^*\mathbf{X}| \cdot \begin{vmatrix} \mathbf{A}'\mathbf{M}\mathbf{A} : \mathbf{A}'\mathbf{V}^*\hat{\mathbf{e}} \\ \hat{\mathbf{e}}'\mathbf{V}^*\mathbf{A} : \hat{\mathbf{e}}'\mathbf{V}^*\hat{\mathbf{e}} \end{vmatrix} \\
&= |\mathbf{X}'\mathbf{V}^*\mathbf{X}| \cdot |\mathbf{A}'\mathbf{M}\mathbf{A}| \, (\hat{\mathbf{e}}'\mathbf{V}^*\hat{\mathbf{e}} - \hat{\mathbf{e}}'\mathbf{V}^*\mathbf{A}(\mathbf{A}'\mathbf{M}\mathbf{A})^{-1}\mathbf{A}'\mathbf{V}^*\hat{\mathbf{e}}) \\
&= |\mathbf{X}'\mathbf{V}^*\mathbf{X}| \cdot |\mathbf{A}'\mathbf{M}\mathbf{A}| \, (\hat{\mathbf{e}}'\mathbf{V}^*\hat{\mathbf{e}} - Q_k) \,.
\end{aligned}$$

Combining (3.103) and (3.104) we obtain $AP_{(2)}$ factorized as

$$AP_{(2)} = \frac{\hat{\mathbf{e}}'\mathbf{V}^{-}\hat{\mathbf{e}} - Q_k}{\hat{\mathbf{e}}'\mathbf{V}^{-}\hat{\mathbf{e}}} \cdot |\mathbf{A}'\mathbf{M}\mathbf{A}| \,. \tag{3.105}$$

For $\mathbf{A} = \mathbf{V}_2$ and $\mathbf{A} = [\mathbf{0}:\mathbf{I}_k]'$ we compute $AP_{(2)}$:

$$AP^{D}_{(2)} = \frac{\hat{\mathbf{e}}'\mathbf{V}^{-}\hat{\mathbf{e}} - Q^{D}_k}{\hat{\mathbf{e}}'\mathbf{V}^{-}\hat{\mathbf{e}}} |\mathbf{N}_{22}| \,, \quad \text{and} \tag{3.106}$$

$$AP^{A}_{(2)} = \frac{\hat{\mathbf{e}}'\mathbf{V}^{-}\hat{\mathbf{e}} - Q^{A}_k}{\hat{\mathbf{e}}'\mathbf{V}^{-}\hat{\mathbf{e}}} |\mathbf{M}_{22}| \,. \tag{3.107}$$

The first component in $AP_{(2)}$ is obviously an outlier measure, since it will be small when Q_k, the outlier sum of squares, is large. The second component measures the potential influence of the observations $\mathbf{y}_2$. Draper and John (1981), for the case $\mathbf{V} = \mathbf{I}$, write that it "provides a measure of the remoteness of the set of observations in the predictor space, smaller values of $|\mathbf{A}'\mathbf{M}\mathbf{A}|$ indicating 'more remote' points".

Clearly, for $k = 1$ the numbers $|\mathrm{N}_{ii}| = \mathrm{N}_{ii}$ and $|\mathrm{M}_{ii}| = \mathrm{M}_{ii}$ are decreasing functions of the leverage $(\mathrm{V}_{ii} - \mathrm{N}_{ii})$ and $(\mathrm{V}_{ii} - \mathrm{M}_{ii})$ of the observation in question, high leverage resulting in small values of $|\mathrm{N}_{ii}|$ and $|\mathrm{M}_{ii}|$ respectively.

We observe that the condition $C(\mathbf{A}) \cap C(\mathbf{X}) = \{\mathbf{0}\}$ results in no loss of generality. Similar to the argument in the section dealing with outliers, we note that $AP_{(2)}$ is associated with removing the observations $\mathbf{y}_2$ (or the error term $\mathbf{e}_2$, the PC's $\mathbf{y}^*_2$) from the estimation of the regression parameters β, which is equivalent to fitting the new variables $\mathbf{A}$ to the original model. If $C(\mathbf{A}) \cap C(\mathbf{X}) = \{\mathbf{0}\}$ is not satisfied, at least one linear function $\ell'\beta$ of β is not estimable in the augmented model (or after the removal of $\mathbf{y}_2$, $\mathbf{e}_2$ or $\mathbf{y}^*_2$), and the notion of influence of these removed terms on the estimate of the parameter vector β makes no sense since these terms completely determine the estimate $\ell'\hat{\beta}$ of $\ell'\beta$.

However, the condition that $\mathbf{X}$ be of full rank can be dropped. The representation (3.105) of $AP_{(2)}$ is defined whether or not $\mathbf{X}$ is of full rank, and thus $AP_{(2)}$ can be defined using (3.105) in a linear model with non-full rank design matrix $\mathbf{X}$.

We note that $AP_{(2)}$ is invariant under a rank preserving reparametrization of the underlying linear model. Clearly, $\hat{\mathbf{e}}'\mathbf{V}^{-}\hat{\mathbf{e}}$ and Q_k in (3.105) are invariant under a reparametrization, and so is $\mathbf{M} = \mathbf{V}^*(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^*)$.

Thus, in the case where $\mathbf{X}$ is not of full rank, we can perform a full rank reparametrization of the model and subsequently define $AP_{(2)}$ using determinants as in (3.102). Equivalent to this procedure is a definition of $AP_{(2)}$ as the ratio of the products of the nonzero eigenroots of $|\tilde{\mathbf{X}}_2'\mathbf{V}^*\tilde{\mathbf{X}}_2|$ and $|\tilde{\mathbf{X}}_1'\mathbf{V}^*\tilde{\mathbf{X}}_1|$ respectively. Then the condition $C(\mathbf{A}) \cap C(\mathbf{X}) = \{\mathbf{0}\}$ and the condition that $\mathbf{A}$ is of full rank can also be dropped, achieving the

full generality as in the cases where we obtained the F-statistics for outliers irrespective of any rank relations and column space conditions concerning the matrices $\mathbf{X}$ and $\mathbf{A}$.

If $C(\mathbf{A}) \subset C(\mathbf{V})$ is satisfied, then $AP_{(2)}$ is also invariant over all choices of a g-inverse $\mathbf{V}^*$ of $\mathbf{V}$. In general, this condition can be assumed without loss of generality, since $AP_{(2)}$ will only be applied in a consistent model, which can be reduced to a model of full rank. Now, $\mathbf{A} = \mathbf{V}_2$ and $\mathbf{A} = \mathbf{P}_2$ trivially imply $C(\mathbf{A}) \subset C(\mathbf{V})$, and if $\mathbf{A} = [\mathbf{0}:\mathbf{I}]'$ we have $C(\mathbf{A}) \subset C(\mathbf{V})$ if and only if the sure equations are not changed by fitting $\mathbf{A}$. But only in this case does it make sense to compute $AP_{(2)}$.

We now turn to the problem of assessing the influence of a group of data points on only a set of specified linear contrasts of the regression parameters. A similar generalization of Cook's distance was given by Cook (1979). See also Cook and Weisberg (1982, pp. 124–126). For simplicity, we initially take $\mathbf{X}$ of full rank and $\mathbf{V} = \mathbf{I}$, but the generalization for $\mathbf{X}$ possibly not of full rank and arbitrary $\mathbf{V}$ will follow directly.

To measure the influence of the data points $\mathbf{y}_2$ on the least-squares estimate $\mathbf{L}'\hat{\beta}$ of a set of linear functions $\mathbf{L}'\beta$ of β in the $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{I})$, where $\mathbf{X}$ and $\mathbf{L}$ are of full rank, we propose the statistic

$$AP_{(2)}(\mathbf{L}) = \frac{|\mathbf{L}_2'\,\mathbf{L}_2|}{|\mathbf{L}_1'\,\mathbf{L}_1|} \tag{3.108}$$

where $\mathbf{L}_1 = [\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}:\mathbf{y}]$ and $\mathbf{L}_2 = [\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}:\mathbf{A}:\mathbf{y}]$, with $\mathbf{A} = [\mathbf{0}:\mathbf{I}]'$.

We note that from the invariance of $AP_{(2)}$ over a reparametrization we have that $AP_{(2)}(\mathbf{L})$ is invariant over all matrix representations of the subset of linear contrasts in question, that is, $AP_{(2)}(\mathbf{L}) = AP_{(2)}(\mathbf{L}^*)$ where $\mathbf{L}^* = \mathbf{LB}$ for some nonsingular $\mathbf{B}$, or equivalently where $C(\mathbf{L}) = C(\mathbf{L}^*)$.

Considering the extreme case of all linear contrasts of β we may take $\mathbf{L} = \mathbf{I}_p$ or $\mathbf{L} = \mathbf{X}'\mathbf{X}$, thus obtaining

$$AP_{(2)}(\mathbf{I}) = AP_{(2)}(\mathbf{X}'\mathbf{X}) = AP_{(2)}\,, \tag{3.109}$$

where $AP_{(2)}$ is from (3.102). Thus the definition (3.108) is consistent with (3.102).

The special choice of the statistic $AP_{(2)}(\mathbf{L})$ as in (3.108) can be motivated as follows:

Suppose the columns of $\mathbf{X}$ are orthogonal or equivalently $\mathbf{X}'\mathbf{X} = \mathbf{I}_p$. Then clearly the influence of a group of observations on the least-squares estimate $\hat{\beta}_1$ of the subset of parameters β_1 (say) of $\beta' = [\beta_1':\beta_2']$ can be determined by

$$AP_{(2)}\left(\begin{bmatrix}\mathbf{I}\\ \mathbf{0}\end{bmatrix}\right) = \frac{|[\mathbf{X}_1:\mathbf{y}]'[\mathbf{X}_1:\mathbf{y}]|}{|[\mathbf{X}_1:\mathbf{A}:\mathbf{y}]'[\mathbf{X}_1:\mathbf{A}:\mathbf{y}]|} \tag{3.110}$$

where $\mathbf{X} = [\mathbf{X}_1:\mathbf{X}_2]$ is partitioned conformably with $\beta' = [\beta_1':\beta_2']$. This follows directly from the fact that each parameter is estimated independently from the others due to the orthogonality of the columns of $\mathbf{X}$.

Similarly, with the columns of $\mathbf{X}$ being orthogonal, for any subset $\mathbf{L}$ of linear functions $\ell'\beta$ of β

$$AP_{(2)}(\mathbf{L}) = \frac{|\,[\mathbf{XL}:\mathbf{y}]'[\mathbf{XL}:\mathbf{y}]\,|}{|\,[\mathbf{XL}:\mathbf{A}:\mathbf{y}]'[\mathbf{XL}:\mathbf{A}:\mathbf{y}]\,|} \tag{3.111}$$

which definition is consistent with (3.108) since $\mathbf{X}'\mathbf{X} = (\mathbf{X}'\mathbf{X})^{-1} = \mathbf{I}_p$.

Now, for any $\mathbf{X}$ of full rank we observe that $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1/2}$ is a matrix whose columns are orthogonal, and thus in the reparametrized model $\mathbf{y} = \mathbf{X}^*\beta^* + \mathbf{e}$ where $\mathbf{X}^* = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1/2}$, $\beta^* = (\mathbf{X}'\mathbf{X})^{-1/2}\beta$ the statistic $AP_{(2)}(\mathbf{L}^*)$ can be defined and motivated as in (3.111) for any set of linear functions $\mathbf{L}^{*\prime}\beta^*$ of β^*. Rephrasing the problem in the original coordinates yields the definition of $AP_{(2)}(\mathbf{L})$ as in (3.108).

The generalization of $AP_{(2)}(\mathbf{L})$ for arbitrary variance-covariance structure $\mathbf{V}$ is straightforward and we define

$$AP_{(2)}(\mathbf{L}) = \frac{|\,\mathbf{L}_2'\mathbf{V}^*\mathbf{L}_2\,|}{|\,\mathbf{L}_1'\mathbf{V}^*\mathbf{L}_1\,|} \tag{3.112}$$

where $\mathbf{L}_1 = [\mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^{-1}\mathbf{L}:\mathbf{y}]$ and $\mathbf{L}_2 = [\mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^{-1}\mathbf{L}:\mathbf{A}:\mathbf{y}]$, with $\mathbf{A}$ respectively set equal to $\mathbf{A} = \mathbf{V}_2$, $\mathbf{A} = [\mathbf{0}:\mathbf{I}]'$ and $\mathbf{A} = \mathbf{P}_2$ corresponding to three types of outlier.

$AP_{(2)}(\mathbf{L})$ factorizes similarly to $AP_{(2)}$ in (3.105), and any rank and column space conditions on $\mathbf{X}$, $\mathbf{A}$ and $\mathbf{L}$ can be dropped. When $\mathbf{L}'\beta$ is a subset of estimable linear functions $\ell'\beta$ of β, then $AP_{(2)}(\mathbf{L})$ is defined as the ratio of the product of the nonzero eigenroots of $\mathbf{L}_2'\mathbf{V}^*\mathbf{L}_2$ and $\mathbf{L}_1'\mathbf{V}^*\mathbf{L}_1$ respectively. In this case, $(\mathbf{X}'\mathbf{V}^*\mathbf{X})^{-1}$ is replaced by $(\mathbf{X}'\mathbf{V}^*\mathbf{X})^-$.

If $AP_{(2)}(\mathbf{L})$ is factorized similar to the factorization of $AP_{(2)}$ in (3.105), then the second term $|\,\mathbf{A}'\mathbf{M}_{(\mathbf{L})}\mathbf{A}\,|$ is of special interest, giving the potential influence of $\mathbf{y}_2$ on the estimate $\mathbf{L}'\hat{\beta}$ of $\mathbf{L}'\beta$. $\mathbf{M}_{(\mathbf{L})}$ denotes here the $\mathbf{M}$-matrix computed with respect to the design matrix $\mathbf{X}_{(\mathbf{L})} = \mathbf{X}(\mathbf{X}'\mathbf{V}^*\mathbf{X})^-\mathbf{L}$. When in the partitioned linear model

$$\mathbf{y} = [\mathbf{X}_1:\mathbf{X}_2]\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix} + \mathbf{e}\ ;\quad cov(\mathbf{e}) = \sigma^2\mathbf{V} \tag{2.131}$$

tests show that the variables $\mathbf{X}_2$ are insignificant, then the model (2.131) is usually reduced to the model

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{e}\ ;\quad cov(\mathbf{e}) = \sigma^2\mathbf{V}\ . \tag{3.113}$$

The set of influential observations under the models (2.131) and (3.113) may differ considerably. When e.g. an observation y_j is not influential under (2.131) but influential under (3.113), then the question might be of interest whether y_j was influential at least on $\mathbf{X}_1\hat{\beta}_1$ under (2.131). In other words, the question might be of interest whether y_j became influential because of the reduction of the model, or whether y_j was already influential on the estimate of the parameters β_1 corresponding to the significant variables $\mathbf{X}_1$ in the full model, which influence was otherwise obscured by a relatively low influence on the estimate $\hat{\beta}_2$ corresponding to the insignificant or "unimportant" variables in the model.

3.3 OUTLIERS AND INFLUENCE IN THE VARIANCE COMPONENTS MODEL

We consider the variance components model $(\mathbf{y}, \mathbf{X}\beta, \Sigma\sigma_i^2\mathbf{V}_i)$ with m variance components $(\sigma_1^2,\ldots, \sigma_m^2)$ and arbitrary known $\mathbf{V}_1, \ldots, \mathbf{V}_m$. This model can be written alternatively in the form

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e} \; ; \quad cov(\mathbf{e}) = \sigma_1^2\mathbf{V}_1 + \cdots + \sigma_m^2\mathbf{V}_m \; . \tag{3.114}$$

Inherent in this model is the assumption that the error term $\mathbf{e}$ can be decomposed into m uncorrelated components $\mathbf{e}_i$ such that

$$\mathbf{e} = \mathbf{e}_1 + \cdots + \mathbf{e}_m \; , \quad \text{and} \tag{3.115}$$

$$cov(\mathbf{e}_i) = \sigma_i^2\mathbf{V}_i \; , \quad i = 1, \ldots, m \; . \tag{3.116}$$

With the exception of special cases, as given in Section 1.4, the BLUE $\mathbf{X}\hat{\beta}$ for $\mathbf{X}\beta$ is not known when the variance components $(\sigma_1^2, \ldots, \sigma_m^2)$ are unknown. In this case the variance components $(\sigma_1^2, \ldots, \sigma_m^2)$ must be estimated before $\mathbf{X}\beta$ is estimated, e.g. by MINQUE as discussed by Rao (1973 pp. 302–305), or $\mathbf{X}\beta$ and $(\sigma_1^2, \ldots, \sigma_m^2)$ are estimated simultaneously, for example by maximum likelihood (ML).

In the following we assume that $\mathbf{X}\beta$ and $(\sigma_1^2, \ldots, \sigma_m^2)$ are estimated by maximum likelihood, where $\mathbf{y}$ and $\mathbf{e}$ follow a normal distribution.

As a direct generalization to the problem of outliers in a $LM(\mathbf{y}, \mathbf{X}\beta, \sigma^2\mathbf{V})$ with the single variance component σ^2, we propose some approaches to adjust for possible outliers in the data and indicate the appropriate likelihood-ratio-tests (LRT's).

Generally, when the model (3.114) is partitioned as

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \beta + \mathbf{e} \; ; \quad cov(\mathbf{e}) = \sum_{i=1}^{m}\sigma_i^2\mathbf{V}_i = \sum_{i=1}^{m}\sigma_i^2 \begin{bmatrix} \mathbf{V}_{11}^{(i)} & \mathbf{V}_{12}^{(i)} \\ \mathbf{V}_{21}^{(i)} & \mathbf{V}_{22}^{(i)} \end{bmatrix} \tag{3.117}$$

the set of "suspicious" observations is $\mathbf{y}_2$, in the manner of the previous sections of this chapter.

Additive outliers

To adjust for an additive shift in the mean of $\mathbf{y}_2$ we fit the adjusted model

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 : \mathbf{0} \\ \mathbf{X}_2 : \mathbf{I} \end{bmatrix} \begin{bmatrix} \beta \\ \lambda \end{bmatrix} + \mathbf{e} \; ; \quad cov(\mathbf{e}) = \sum_{i=1}^{m}\sigma_i^2\mathbf{V}_i \tag{3.118}$$

From the discussion of the reduced data model corresponding to the adjusted model (3.44) it is clear that the ML-estimates $\mathbf{X}_1\tilde{\beta}$ for $\mathbf{X}\beta$ and $(\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_m^2)$ for $(\sigma_1^2, \ldots, \sigma_m^2)$ under model (3.118) are equivalently obtained in the reduced data model

$$\mathbf{y}_1 = \mathbf{X}_1\beta + \mathbf{e}_1 \; ; \quad cov(\mathbf{e}_1) = \sum_{i=1}^{m}\sigma_i^2\mathbf{V}_{11}^{(i)} \; . \tag{3.119}$$

Similarily, along the lines of the development in Section 3.2, the ML-estimate $\tilde{\mathbf{y}}_2$ for $\mathbf{y}_2$ is given by

$$\tilde{\mathbf{y}}_2 = \mathbf{X}_2\tilde{\beta} + \tilde{\lambda} = \mathbf{y}_2 - \tilde{\mathbf{V}}_{21}\tilde{\mathbf{V}}_{11}^{-}\tilde{\mathbf{e}}_1 \ , \tag{3.120}$$

where $\tilde{\mathbf{e}}_1 = \mathbf{y}_1 - \mathbf{X}_1\tilde{\beta}$ and $\tilde{\mathbf{V}} = \sum_{i=1}^{m}\tilde{\sigma}_i^2\mathbf{V}_i$. If $R(\mathbf{X}_2) \subset R(\mathbf{X}_1)$, the ML-estimate $\tilde{\lambda}$ of λ can be obtained from (3.120) as

$$\tilde{\lambda} = \mathbf{y}_2 - \mathbf{X}_2\tilde{\beta} - \tilde{\mathbf{V}}_{21}\tilde{\mathbf{V}}_{11}^{-}\tilde{\mathbf{e}}_1 \ . \tag{3.121}$$

The LRT for the hypothesis H_0: $\lambda = \mathbf{0}$ under model (3.118), which constitutes the LRT for the hypothesis that $\mathbf{y}_2$ is not an A-outlier against the alternative that $\mathbf{y}_2$ is an A-outlier, is obtained by comparing the maximum likelihood under the models (3.117) and (3.118) respectively.

Distributional outliers

Let $J \neq \emptyset$ be a subset of $\{1, \ldots, m\}$. Then, to adjust for distributional outliers in the error components $\mathbf{e}_2^{(j)}$, $j \in J$, we fit the adjusted model

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 : \sum_j \sigma_j^2\mathbf{V}_{12}^{(j)} \\ \mathbf{X}_2 : \sum_j \sigma_j^2\mathbf{V}_{22}^{(j)} \end{bmatrix} \begin{bmatrix} \beta \\ \lambda \end{bmatrix} + \mathbf{e} \ ; \quad cov(\mathbf{e}) = \sum_{i=1}^{m}\sigma_i^2\mathbf{V}_i \ . \tag{3.122}$$

Which components we would include in the set J of suspicious error components would, in a practical situation, depend on the data and on the suspicions we would have about the error components involved in the outlier $\mathbf{y}_2$.

If J contains only one component j (say), then (3.122) reduces to

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 : \mathbf{V}_{12}^{(j)} \\ \mathbf{X}_2 : \mathbf{V}_{22}^{(j)} \end{bmatrix} \begin{bmatrix} \beta \\ \lambda \end{bmatrix} + \mathbf{e} \ ; \quad cov(\mathbf{e}) = \sum_{i=1}^{m}\sigma_i^2\mathbf{V}_i \ . \tag{3.123}$$

and ML-estimates for $\mathbf{X}\beta$, λ and $(\sigma_1^2, \ldots, \sigma_m^2)$ can be computed using standard methods, as well as the LRT-statistic for the hypothesis H_0: $\lambda = \mathbf{0}$.

However, if $card(J) > 1$, then the variance components σ_j^2, $j \in J$ appear in the mean of $\mathbf{y}$ and the ML-estimation of the parameters in the model could be complicated.

If $J = \{1, \ldots, m\}$, i.e. we fit all components, then it is clear that

$$\tilde{\mathbf{y}}_2 = \mathbf{X}_2\tilde{\beta} + (\sum_{i=1}^{m}\tilde{\sigma}_i^2\mathbf{V}_{22}^{(j)})\tilde{\lambda} \ , \tag{3.124}$$

i.e. the fitted value for $\mathbf{y}_2$ in (3.122) with $J = \{1, \ldots, m\}$ is $\mathbf{y}_2$ itself.

Transformational outliers

In a manner similar to the treatment of transformational outliers in the general linear model we fit the adjusted model

$$\text{(3.125)} \qquad \mathbf{y} = [\mathbf{X} : \textstyle\sum_J \sigma_j^2 \mathbf{P}_2^{(j)}] \begin{bmatrix} \beta \\ \lambda \end{bmatrix} + \mathbf{e} \; ; \quad cov(\mathbf{e}) = \sum_{i=1}^{m} \sigma_i^2 \mathbf{V}_i$$

Here J denotes the set of error components $\mathbf{e}_j$ of whose respective principal axes a subset $\mathbf{P}_2^{(j)}$ is fitted in (3.125), where

$$\text{(3.126)} \qquad \mathbf{V}_i = [\mathbf{P}_1^{(i)} : \mathbf{P}_2^{(i)}] \begin{bmatrix} \Delta_1^{(i)} & \\ & \Delta_2^{(i)} \end{bmatrix} \begin{bmatrix} \mathbf{P}_1^{(i)\prime} \\ \mathbf{P}_2^{(i)\prime} \end{bmatrix}, \quad i = 1, \ldots, m$$

is the SVD of $\mathbf{V}_i$.

If J contains only one component j (say), then σ_j^2 can be eliminated from the mean of $\mathbf{y}$ in (3.125), as is the case in (3.123), and standard methods can be used to compute the ML-estimates for the parameters in the model and the LRT-statistic for the hypothesis $H_0: \lambda = \mathbf{0}$.

If J contains precisely two components, without loss of generality the components 1 and 2, then we can proceed as follows: we compute a nonsingular matrix $\mathbf{T}$ such that

$$\text{(3.127)} \qquad \mathbf{T}(\sigma_1^2 \mathbf{V}_1 + \sigma_2^2 \mathbf{V}_2)\mathbf{T}' = \sigma_1^2 \Delta_1 + \sigma_2^2 \Delta_2 = \sigma_2^2 \begin{bmatrix} \Delta_{11} & \\ & \Delta_{12} \end{bmatrix} + \sigma_2^2 \begin{bmatrix} \Delta_{21} & \\ & \Delta_{22} \end{bmatrix}$$

where Δ_1 and Δ_2 are diagonal matrices. Such a $\mathbf{T}$ does always exist and essentially its computation involves the computation of the eigenvectors of $\mathbf{V}_1^g \mathbf{V}_2$, where $\mathbf{V}_1^g$ denotes the Moore-Penrose inverse of $\mathbf{V}_1$ (Flury, 1983).

Now transform the model (3.125), where $J = \{1, 2\}$, by $\mathbf{T}$ to obtain

$$\text{(3.128)} \qquad \begin{bmatrix} \mathbf{y}_1^* \\ \mathbf{y}_2^* \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^* : & \mathbf{0} \\ \mathbf{X}_2^* : \sigma_1^2 \Delta_{12} + \sigma_1^2 \Delta_{22} & \end{bmatrix} \begin{bmatrix} \beta \\ \lambda \end{bmatrix} + \mathbf{e}^* \; ;$$

$$cov(\mathbf{e}^*) = \sigma_1^2 \Delta_1 + \sigma_1^2 \Delta_2 + \sum_{i=3}^{m} \sigma_i^2 \mathbf{V}_i^* \, .$$

In model (3.128), fitting $[\mathbf{0} : \sigma_1^2 \Delta_{12} + \sigma_2^2 \Delta_{22}]'$ is equivalent to fitting $[\mathbf{0} : \mathbf{I}]'$, and we can proceed as if $\mathbf{y}_2^*$ was an $\mathbf{A}$-outlier in (3.128).

Essentially the same simplification is possible when J contains an arbitrary number of components, but the matrices $\mathbf{V}_j$, $j \in J$, are simultaneously diagonalizable. That is, there is a nonsingular matrix $\mathbf{T}$ such that

$$\text{(3.129)} \qquad \mathbf{T}\mathbf{V}_j\mathbf{T}' = \Delta_j = \begin{bmatrix} \Delta_{j1} & \\ & \Delta_{j2} \end{bmatrix}, \quad \text{for all } j \in J \; .$$

Then the model (3.125) can be transformed by $\mathbf{T}$ to obtain

$$\text{(3.130)} \qquad \begin{bmatrix} \mathbf{y}_1^* \\ \mathbf{y}_2^* \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^* : & \mathbf{0} \\ \mathbf{X}_2^* : \sum_J \sigma_j^2 \Delta_{j2} & \end{bmatrix} \begin{bmatrix} \beta \\ \lambda \end{bmatrix} + \mathbf{e}^* \; ; \quad cov(\mathbf{e}^*) = \sum_J \sigma_j^2 \Delta_j + \sum_J \sigma_j^2 \mathbf{V}_j^* \, .$$

As above, the model (3.130) can be treated as if $\mathbf{y}_2^*$ was an A-outlier, and ML-estimates for the parameters in the model, as well as the LRT-statistic for the hypothesis $H_0: \lambda = \mathbf{0}$ can be computed using standard methods and equations (3.120) and (3.121).

Influential observations

To assess the influence of a set of data points $\mathbf{y}_2$ (say) in the variance components model (3.117), we propose to use Cook's distance $C_{(2)}$ and the Andrews-Pregibon statistic $AP_{(2)}$, as generalized in the previous section for the case where $\mathbf{V}$ is arbitrary and known, replacing the known matrix $\mathbf{V}$ by its ML-estimate $\hat{\mathbf{V}} = \sum_{i=1}^{m} \hat{\sigma}_i^2 \mathbf{V}_i$ under the model (3.117).

A different course of action would be to replace $\mathbf{V}$ by $\tilde{\mathbf{V}}$, where $\tilde{\mathbf{V}} = \sum_{i=1}^{m} \tilde{\sigma}_i^2 \mathbf{V}_i$ is the ML-estimate for $\mathbf{V}$ in the respective adjusted models (3.118), (3.122) and (3.125). An advantage would be that the estimate $\tilde{\mathbf{V}}$ for $\mathbf{V}$ is not contaminated by the possible outlier $\mathbf{y}_2$. However, the use of $\hat{\mathbf{V}}$ instead of $\tilde{\mathbf{V}}$ involves far less computation ($\tilde{\mathbf{V}}$ would have to be computed anew for all subsets of observations under investigation), and using the same matrix $\hat{\mathbf{V}}$ for a variety of subsets of observations facilitates the comparison of the corresponding values of the influence measure in question.

CHAPTER 4

Outliers and Influence in Multivariate Models

In this chapter we consider the problem of outliers and influential observations in normal multivariate models. We distinguish three types of outlier, similar to those applied in the general linear model, namely *distributional, additive* and *transformational (outliers in principal components)*. In the manner of the previous chapter we adjust the model for possible outliers, and while doing so we fit, in the case of D- and T-outliers, a new type of dummy variable. Since the variance-covariance structure of an observational vector in a multivariate sample is unknown in general, the extra variable fitted to adjust respectively for D- and T-outliers in the data is unknown and has to be estimated from the data, in addition to the unknown parameters associated with the new variable.

When in the multivariate regression model the set of data points suspected to be outlying can be arranged in a way that these data points occur in a nested pattern, then the maximum likelihood estimation of the unknown parameters in the models respectively adjusted for D-, A- and T-outliers can be performed in closed form, and thus likelihood ratio test statistics for the hypotheses that no outliers are present are obtained in closed form. For an arbitrary outlier pattern, and for the problem of outliers in the generalized multivariate regression model (growth curve model), the EM-algorithm can be applied to obtain maximum likelihood estimates iteratively.

In the generalized multivariate regression model the influence of a subset of the data points can be assessed using the methods applied in the general linear model, replacing the unknown variance-covariance matrix of the observational vectors by its maximum likelihood estimate.

Test-statistics for the detection of outliers in multivariate samples have been proposed by Siotani (1959) and Wilks (1963). Hawkins (1980) devotes a chapter to multivariate outliers, but emphasis in the literature has been on the case where one or more observational vectors are assumed to be outliers. We emphasize here the typification of outliers when only a subset of the components of one or more observational vectors are suspected to be outlying, thus extending the typification of outliers in the general linear model to multivariate models. When all components of the observational vectors in question are assumed to be outlying, the different types of outlier can not be distinguished and the likelihood ratio test statistics corresponding to each type of outlier are identical, and equivalent to the Wilks (1963) statistic.

4.1 THE MULTIVARIATE REGRESSION MODEL

We consider the multivariate linear regression model

$$\underset{(m\times n)}{\mathbf{Y}} = \underset{(m\times p)}{\mathbf{X}} \cdot \underset{(p\times n)}{\mathbf{B}} + \underset{(m\times n)}{\mathbf{E}} \tag{4.1}$$

where $\mathbf{Y}$ is the matrix of observations, $\mathbf{X}$ is the known design matrix, $\mathbf{B}$ is the matrix of regression parameters and $\mathbf{E}$ is an unobservable matrix of error components.

The rows of $\mathbf{E}$ are assumed to be stochastically independent and identically distributed as $N_n(\mathbf{0},\mathbf{V})$, where $\mathbf{V}$ is an arbitrary but positive definite and symmetric matrix which is generally unknown. The rows of $\mathbf{Y} = [\mathbf{y}_1 : \cdots : \mathbf{y}_m]'$ are the observations, that is, we have a sample of m independently distributed observational vectors $\mathbf{y}_1, \ldots, \mathbf{y}_m$ whose respective means are specified by the corresponding row of the matrix $\mathbf{XB}$.

If $\mathbf{a}_1, \ldots, \mathbf{a}_k$ are the column vectors of an arbitrary $l\times k$-matrix $\mathbf{A} = [\mathbf{a}_1 : \cdots : \mathbf{a}_k]$, then

$$vec(\mathbf{A}) = [\mathbf{a}_1' : \cdots : \mathbf{a}_k']' \tag{4.2}$$

denotes the $l\cdot k\times 1$-vector of column vectors of $\mathbf{A}$. With this notation we can write the model (4.1) in the alternative form

$$vec(\mathbf{Y}') = (\mathbf{X}\otimes\mathbf{I}_n)\, vec(\mathbf{B}') + vec(\mathbf{E}') \tag{4.1a}$$

$$\Leftrightarrow \quad \mathbf{y} = (\mathbf{X}\otimes\mathbf{I}_n)\beta + \mathbf{e};\quad cov(\mathbf{e}) = \Sigma = (\mathbf{I}_m\otimes\mathbf{V})\,.$$

where $\mathbf{y} = [\mathbf{y}_1' : \ldots : \mathbf{y}_m']' = vec(\mathbf{Y}')$, $\beta = [\beta_1' : \ldots : \beta_p']' = vec(\mathbf{B}')$, $\mathbf{e} = [\mathbf{e}_1' : \ldots : \mathbf{e}_m']' = vec(\mathbf{E}')$ and $\otimes$ denotes the Kronecker product of matrices.

Obviously the model (4.1a) and thus (4.1) is a special case of the general linear model (2.1), with a design matrix of the form $\Xi = (\mathbf{X}\otimes\mathbf{I}_n)$ and a variance-covariance matrix $\Sigma = \mathbf{I}_m\otimes\mathbf{V}$ in block diagonal form. As pointed out above, $\mathbf{V}$ is in general assumed to be arbitrary but positive definite and symmetric. As opposed to the usual assumption in the general linear model, $\mathbf{V}$ is unknown, and thus Σ is only known to have some structure, namely the block diagonal form $\mathbf{I}_m\otimes\mathbf{V}$.

It is well-known that the maximum likelihood estimate (MLE) $\mathbf{X}\hat{\mathbf{B}}$ for $\mathbf{XB}$ under model (4.1) is

$$\mathbf{X}\hat{\mathbf{B}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y} \tag{4.3}$$

and with

$$\hat{\mathbf{E}} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}')\mathbf{Y} = (\mathbf{I}-\mathbf{M})\mathbf{Y} = \mathbf{N}\mathbf{Y} \tag{4.4}$$

the MLE $\hat{\mathbf{V}}$ for $\mathbf{V}$ is given by

$$m\cdot\hat{\mathbf{V}} = \hat{\mathbf{E}}'\hat{\mathbf{E}} = \mathbf{Y}'\mathbf{N}\mathbf{Y}\,. \tag{4.5}$$

The matrix $\hat{\mathbf{V}}$ is positive definite w.p.1 if $n \le m - rank(\mathbf{X})$. In the following we assume that this condition is satisfied.

4.2 OUTLIERS IN THE MULTIVARIATE REGRESSION MODEL

Noting that the multivariate regression model (4.1) is a special case of the general linear model (2.1), no conceptual difficulties arise in extending the methods to adjust the general linear model for different types of outlier to the multivariate regression model.

If J, an arbitrary subset of data points y_{ij} from the data matrix $\mathbf{Y} = (y_{ij})$ is the set of observations suspected to be outlying, then the model (4.1) adjusted for D- and A-outliers respectively is

$$\mathbf{Y} = \mathbf{XB} + \Theta\mathbf{V} + \mathbf{E} \tag{4.6}$$

if J is assumed to be a set of D-outliers, and

$$\begin{aligned} \mathbf{Y} &= \mathbf{XB} + \Theta\mathbf{I}_n + \mathbf{E} \\ &= \mathbf{XB} + \Theta + \mathbf{E} \end{aligned} \tag{4.7}$$

if J is assumed to be a set of A-outliers.

The parameter matrix $\Theta = (\theta_{ij})$ is of the order $m \times n$, and θ_{ij} is *a-priori* specified to be zero if and only if $y_{ij} \notin J$.

The adjusted model (4.7) causes no problems since the extra variables fitted are dummy variables of the form $\mathbf{u}_j$, where $\mathbf{u}_j$ is the j-th unit vector. However, since the matrix $\mathbf{V}$ is unknown, a new type of dummy variable is fitted in the model (4.6) when adjusting for D-outliers, and a similar problem arises when the model (4.1) is adjusted for T-outliers:

Let $\mathbf{V} = \mathbf{P}\triangle\mathbf{P}'$ be the SVD of $\mathbf{V}$, then the (uncentered) PC's $\mathbf{Y}^*$ of $\mathbf{Y}$ are given by

$$\mathbf{Y}^* = \mathbf{YP} \ . \tag{4.8}$$

If J^*, an arbitrary subset of PC's y^*_{ij} of $\mathbf{Y}^* = (y^*_{ij})$ is the set of PC's suspected to outlying, then the model (4.1) adjusted for the T-outliers J^* is

$$\mathbf{Y} = \mathbf{XB} + \Theta\mathbf{P}' + \mathbf{E} \ . \tag{4.9}$$

As above, the parameter matrix $\Theta = (\theta_{ij})$ is of the order $m \times n$, and θ_{ij} is *a-priori* specified to be zero if and only if $y^*_{ij} \notin J^*$.

Of course, if $\mathbf{V}$ is unknown $\mathbf{P}$ is unknown. As in the model (4.6) the dummy variables fitted in the model (4.9) are also unknown.

In the development that follows we will concentrate on the special case when the data matrix $\mathbf{Y}$ and the matrix of PC's $\mathbf{Y}^*$ can be arranged in such a way that the sets J and J^* respectively form a rectangular submatrix of $\mathbf{Y}$ and $\mathbf{Y}^*$. In this case, maximum likelihood estimation of the unknown parameters in the adjusted models (4.6), (4.7) and (4.9), can be performed in closed form, thus showing that fitting the 'unknown' dummy variables in the models (4.6) and (4.9) is a valid approach. More generally, maximum likelihood

estimation can be performed in closed form when the outliers occur in a nested pattern. For an arbitrary outlier pattern we present the corresponding versions of the EM-algorithm with respect to three types of outlier, to obtain maximum likelihood estimates iteratively.

4.2.1 Distributional outliers

We consider the adjusted model (4.6) and assume that possibly after some rearrangement of the rows and columns of $\mathbf{Y}$ (and the corresponding rearrangement of the rows of $\mathbf{X}$) the set J of data points suspected to be outlying forms a submatrix of the data matrix $\mathbf{Y}$. That is, we assume that $\mathbf{Y}$ can be partitioned as

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_{11} : \mathbf{Y}_{12} \\ \mathbf{Y}_{21} : \mathbf{Y}_{22} \end{bmatrix} = [\mathbf{Y}_1 : \mathbf{Y}_2] \tag{4.10}$$

where the $k \times l$-submatrix $\mathbf{Y}_{22}$ consists of the data points which are possibly outlying. In this case, the adjusted model (4.6) can be written as

$$\begin{bmatrix} \mathbf{Y}_{11} : \mathbf{Y}_{12} \\ \mathbf{Y}_{21} : \mathbf{Y}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} [\mathbf{B}_1 : \mathbf{B}_2] + \begin{bmatrix} \mathbf{0} \\ \Theta \end{bmatrix} [\mathbf{V}_{21} : \mathbf{V}_{22}] + \begin{bmatrix} \mathbf{E}_{11} : \mathbf{E}_{12} \\ \mathbf{E}_{21} : \mathbf{E}_{22} \end{bmatrix} \tag{4.11}$$

where $\mathbf{X} = [\mathbf{X}_1' : \mathbf{X}_2']'$ and $\mathbf{B} = [\mathbf{B}_1 : \mathbf{B}_2]$ are conformably partitioned, i.e. $\mathbf{X}_2$ is $k \times p$ and $\mathbf{B}_2$ is $p \times l$. The parameter matrix Θ is here $k \times l$, and $\mathbf{V}_{22}$ is a $l \times l$-submatrix of

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_{11} : \mathbf{V}_{12} \\ \mathbf{V}_{21} : \mathbf{V}_{22} \end{bmatrix} . \tag{4.12}$$

For $\mathbf{X}_2\mathbf{B}_2$ to be estimable under the adjusted model (4.11) we require, as in the univariate case, that $R(\mathbf{X}_2) \subset R(\mathbf{X}_1)$.

The ML-estimates $\mathbf{X}\hat{\mathbf{B}}$, $\hat{\mathbf{E}}$ and $\hat{\mathbf{V}}$ of the unknowns $\mathbf{XB}$, $\mathbf{E}$ and $\mathbf{V}$ in model (4.1) are given by equations (4.3) through (4.5). Let now $\mathbf{X}\tilde{\mathbf{B}}$, $\tilde{\mathbf{E}}$, $\tilde{\Theta}$ and $\tilde{\mathbf{V}}$ denote the ML-estimates of $\mathbf{XB}$, $\mathbf{E}$, Θ and $\mathbf{V}$ in the adjusted model (4.11). If $l = n$, then the submatrices $\mathbf{Y}_{11}$, $\mathbf{Y}_{21}$, $\mathbf{E}_{11}$, $\mathbf{E}_{21}$ and $\mathbf{B}_1$ vanish and the model (4.11) is obtained as

$$\begin{bmatrix} \mathbf{Y}_{12} \\ \mathbf{Y}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \mathbf{B}_2 + \begin{bmatrix} \mathbf{0} \\ \Theta \end{bmatrix} \mathbf{V} + \begin{bmatrix} \mathbf{E}_{12} \\ \mathbf{E}_{22} \end{bmatrix} \tag{4.13}$$

Clearly, the ML-estimates $\mathbf{X}\tilde{\mathbf{B}}_2 = \mathbf{X}\tilde{\mathbf{B}}$, $\tilde{\mathbf{E}}$ and $\tilde{\mathbf{V}}$ for $\mathbf{XB}$, $\mathbf{E}$ and $\mathbf{V}$ under the model (4.13) are obtained in the reduced model

$$\mathbf{Y}_{12} = \mathbf{X}_1\mathbf{B}_2 + \mathbf{E}_{12} \tag{4.14}$$

using the formulae (4.3) through (4.5), replacing $\mathbf{Y}$ by $\mathbf{Y}_{12}$ and $\mathbf{X}$ by $\mathbf{X}_1$. Then Θ can be estimated by $\tilde{\Theta} = (\mathbf{Y}_{22} - \mathbf{X}_2\tilde{\mathbf{B}}_2)\tilde{\mathbf{V}}^{-1}$.

In the following we assume that $l < n$. Transforming the model (4.11) from the right by the transformation

(4.15) $$\mathbf{T} = \begin{bmatrix} \mathbf{I} & : \mathbf{0} \\ -\mathbf{V}_{22}^{-1}\mathbf{V}_{21} & : \mathbf{I} \end{bmatrix}$$

we observe that the MLE $\mathbf{X}_1\tilde{\mathbf{B}}_2$ for $\mathbf{X}_1\mathbf{B}_2$ is given by

(4.16) $$\mathbf{X}_1\tilde{\mathbf{B}}_2 = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^-\mathbf{X}_1'\mathbf{Y}_{12} \quad ,$$

i.e. $\mathbf{X}_1\mathbf{B}_2$ is estimated from uncontaminated data alone, and similarly

(4.17) $$\tilde{\mathbf{E}}_{22} = \mathbf{0} \quad .$$

But clearly, using (4.17) we have

(4.18) $$\mathbf{Y}_{22} - \mathbf{X}_2\tilde{\mathbf{B}}_2 = \tilde{\Theta}\tilde{\mathbf{V}}_{22}$$

which implies

(4.19) $$\tilde{\Theta} = (\mathbf{Y}_{22} - \mathbf{X}_2\tilde{\mathbf{B}}_2)\tilde{\mathbf{V}}_{22}^{-1}$$

where the MLE $\tilde{\mathbf{V}}_{22}$ for $\mathbf{V}_{22}$ is given by

(4.20)
$$\begin{aligned} m \cdot \tilde{\mathbf{V}}_{22} &= [\tilde{\mathbf{E}}_{12}' : \tilde{\mathbf{E}}_{22}'] \begin{bmatrix} \tilde{\mathbf{E}}_{12} \\ \tilde{\mathbf{E}}_{22} \end{bmatrix} \\ &= \tilde{\mathbf{E}}_{12}'\tilde{\mathbf{E}}_{12} \ , \quad \text{from (4.17)} \\ &= \mathbf{Y}_{12}'(\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^-\mathbf{X}_1')\mathbf{Y}_{12} \ , \quad \text{from (4.16)} \ , \end{aligned}$$

which is also a function of uncontaminated data alone.

Similarly to (4.20), the MLE $\tilde{\mathbf{V}}_{21}$ for $\mathbf{V}_{21}$ is obtained by

(4.21)
$$\begin{aligned} m \cdot \tilde{\mathbf{V}}_{21} &= [\tilde{\mathbf{E}}_{12}' : \tilde{\mathbf{E}}_{22}'] \begin{bmatrix} \tilde{\mathbf{E}}_{11} \\ \tilde{\mathbf{E}}_{21} \end{bmatrix} \\ &= \tilde{\mathbf{E}}_{12}'\tilde{\mathbf{E}}_{11} \ , \quad \text{from (4.17)} \\ &= (\mathbf{Y}_{12} - \mathbf{X}_1\tilde{\mathbf{B}}_2)'(\mathbf{Y}_{11} - \mathbf{X}_1\tilde{\mathbf{B}}_1) \\ &= \mathbf{Y}_{12}'(\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^-\mathbf{X}_1')(\mathbf{Y}_{11} - \mathbf{X}_1\tilde{\mathbf{B}}_1) \ , \quad \text{from (4.16)} \\ &= \mathbf{Y}_{12}'(\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^-\mathbf{X}_1')\mathbf{Y}_{11} \ , \quad \text{since } (\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^-\mathbf{X}_1')\mathbf{X}_1 = \mathbf{0} \\ &= \tilde{\mathbf{E}}_{12}'\mathbf{Y}_{11} \quad . \end{aligned}$$

We note that as a consequence of (4.21) we have that

(4.22) $$\tilde{\mathbf{E}}_{12}'\mathbf{Y}_{11} = \tilde{\mathbf{E}}_{12}'\tilde{\mathbf{E}}_{11} = \tilde{\mathbf{E}}_{12}'\hat{\mathbf{E}}_{11} \quad .$$

Now, knowing $\tilde{\Theta}$, $\tilde{\mathbf{V}}_{22}$ and $\tilde{\mathbf{V}}_{21}$ from (4.19), (4.20) and (4.21) $\mathbf{X}\tilde{\mathbf{B}}_1$ can be computed as

(4.23)
$$\begin{aligned} \mathbf{X}\tilde{\mathbf{B}}_1 &= \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}' \begin{bmatrix} \mathbf{Y}_{11} \\ \mathbf{Y}_{21} - \tilde{\Theta}\tilde{\mathbf{V}}_{21} \end{bmatrix} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}' \begin{bmatrix} \mathbf{Y}_{11} \\ \mathbf{Y}_{21} - (\mathbf{Y}_{22} - \mathbf{X}_2\tilde{\mathbf{B}}_2)\tilde{\mathbf{V}}_{22}^{-1}\tilde{\mathbf{E}}_{12}'\mathbf{Y}_{11} \div m \end{bmatrix} \end{aligned}$$

and with

(4.24) $$\tilde{\mathbf{E}}_{11} = \mathbf{Y}_{11} - \mathbf{X}_1\tilde{\mathbf{B}}_1 \quad , \text{ and}$$

(4.25) $$\tilde{\mathbf{E}}_{21} = \mathbf{Y}_{21} - \mathbf{X}_2\tilde{\mathbf{B}}_1 - \tilde{\Theta}\tilde{\mathbf{V}}_{21}$$

the MLE $\tilde{\mathbf{V}}_{11}$ for $\mathbf{V}_{11}$ is obtained by

(4.26) $$m \cdot \tilde{\mathbf{V}}_{11} = [\tilde{\mathbf{E}}_{11}' : \tilde{\mathbf{E}}_{21}'] \begin{bmatrix} \tilde{\mathbf{E}}_{11} \\ \tilde{\mathbf{E}}_{21} \end{bmatrix} .$$

Thus the ML-estimates $\mathbf{X}\tilde{\mathbf{B}}$, $\tilde{\mathbf{E}}$ and $\tilde{\Theta}$ and $\tilde{\mathbf{V}}$ for $\mathbf{XB}$, $\mathbf{E}$, Θ and $\mathbf{V}$ in the adjusted model (4.11) can be written in closed form. As in the univariate case we have that $\tilde{\mathbf{E}}_{22} = \mathbf{0}$ in the adjusted model (4.11), that is, the fitted value for the mean of $\mathbf{Y}_{22}$ in model (4.11) is the observed $\mathbf{Y}_{22}$ itself.

Alternatively to the representation above in equations (4.16) through (4.26), the ML-estimates $\mathbf{X}\tilde{\mathbf{B}}$, $\tilde{\mathbf{E}}$, $\tilde{\Theta}$ and $\tilde{\mathbf{V}}$ can be given in terms of the ML-estimates $\mathbf{X}\hat{\mathbf{B}}$, $\hat{\mathbf{E}}$ and $\hat{\mathbf{V}}$ under the original model (4.1). This alternative representation facilitates the computation of the adjusted estimates under model (4.11) by using quantities which are known from the initial analysis of the original model (4.1).

Since $\tilde{\mathbf{E}}_{22} = \mathbf{0}$ from (4.17) we can write

(4.27)
$$\begin{aligned} \mathbf{0} &= \tilde{\mathbf{E}}_{22} \\ &= [\mathbf{0} : \mathbf{I}] \begin{bmatrix} \tilde{\mathbf{E}}_{12} \\ \tilde{\mathbf{E}}_{22} \end{bmatrix} \\ &= [\mathbf{0} : \mathbf{I}] \, (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}') \begin{bmatrix} \mathbf{Y}_{12} \\ \mathbf{Y}_{22} - \tilde{\Theta}\tilde{\mathbf{V}}_{22} \end{bmatrix} \\ &= [\mathbf{0} : \mathbf{I}] \begin{bmatrix} \mathbf{N}_{11} : \mathbf{N}_{12} \\ \mathbf{N}_{21} : \mathbf{N}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{Y}_{12} \\ \mathbf{Y}_{22} - \tilde{\Theta}\tilde{\mathbf{V}}_{22} \end{bmatrix} \\ &= \hat{\mathbf{E}}_{22} - \mathbf{N}_{22}\tilde{\Theta}\tilde{\mathbf{V}}_{22} \quad . \end{aligned}$$

This implies

(4.28) $$\tilde{\Theta}\tilde{\mathbf{V}}_{22} = \mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_{22}$$

where the nonsingularity of $\mathbf{N}_{22}$ follows from $R(\mathbf{X}_2) \subset R(\mathbf{X}_1)$. Consequently $\mathbf{X}\tilde{\mathbf{B}}_2$ can be written as

(4.29)
$$\begin{aligned} \mathbf{X}\tilde{\mathbf{B}}_2 &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}' \begin{bmatrix} \mathbf{Y}_{12} \\ \mathbf{Y}_{22} - \mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_{22} \end{bmatrix} \\ &= \mathbf{X}\hat{\mathbf{B}}_2 - \mathbf{M}_2\mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_{22} \quad . \end{aligned}$$

Further, using (4.29)

(4.30) $m \cdot \tilde{\mathbf{V}}_{22}$

$$= \tilde{\mathbf{E}}_2'\tilde{\mathbf{E}}_2$$

$$= \left(\begin{bmatrix}\mathbf{Y}_{12}\\ \mathbf{Y}_{22}\end{bmatrix} - \begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{bmatrix}\tilde{\mathbf{B}}_2 - \begin{bmatrix}\mathbf{0}\\ \mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_{22}\end{bmatrix}\right)' \left(\begin{bmatrix}\mathbf{Y}_{12}\\ \mathbf{Y}_{22}\end{bmatrix} - \begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{bmatrix}\tilde{\mathbf{B}} - \begin{bmatrix}\mathbf{0}\\ \mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_{22}\end{bmatrix}\right)$$

$$= (\hat{\mathbf{E}}_2 - \mathbf{N}_2\mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_{22})'(\hat{\mathbf{E}}_2 - \mathbf{N}_2\mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_{22})$$

$$= \hat{\mathbf{E}}_2'\hat{\mathbf{E}}_2 - \hat{\mathbf{E}}_{22}'\mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_{22}$$

$$= m \cdot \hat{\mathbf{V}}_{22} - \hat{\mathbf{E}}_{22}'\mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_{22} \quad .$$

From (4.21)

(4.31) $m \cdot \tilde{\mathbf{V}}_{21} = \tilde{\mathbf{E}}_2'\tilde{\mathbf{E}}_1$

$$= [\tilde{\mathbf{E}}_{12}' : \tilde{\mathbf{E}}_{22}'] \begin{bmatrix}\mathbf{Y}_{11} - \mathbf{X}_1\tilde{\mathbf{B}}_1\\ \mathbf{0}\end{bmatrix}$$

$$= [\tilde{\mathbf{E}}_{12}' : \tilde{\mathbf{E}}_{22}'] \begin{bmatrix}\mathbf{Y}_{11} - \mathbf{X}_1\hat{\mathbf{B}}_1\\ \hat{\mathbf{E}}_{21}\end{bmatrix}, \quad \text{since } \tilde{\mathbf{E}}_{22} = \mathbf{0} \text{ and } \tilde{\mathbf{E}}_{12}'\mathbf{X}_1 = \mathbf{0}$$

$$= \left(\begin{bmatrix}\hat{\mathbf{E}}_{12}\\ \hat{\mathbf{E}}_{22}\end{bmatrix} - \begin{bmatrix}\mathbf{N}_{12}\\ \mathbf{N}_{22}\end{bmatrix}\mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_{22}\right)' \begin{bmatrix}\hat{\mathbf{E}}_{11}\\ \hat{\mathbf{E}}_{21}\end{bmatrix}$$

$$= \hat{\mathbf{E}}_2'\hat{\mathbf{E}}_1 - \hat{\mathbf{E}}_{22}'\mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_{21}, \quad \text{since } [\mathbf{N}_{21} : \mathbf{N}_{22}] \begin{bmatrix}\hat{\mathbf{E}}_{11}\\ \hat{\mathbf{E}}_{21}\end{bmatrix} = \hat{\mathbf{E}}_{21}$$

$$= m \cdot \hat{\mathbf{V}}_{21} - \hat{\mathbf{E}}_{22}'\mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_{21} \quad .$$

Finally, the MLE $\tilde{\mathbf{V}}_{11}$ for $\mathbf{V}_{11}$ is given by

(4.32)

$$m \cdot \tilde{\mathbf{V}}_{11}$$

$$= [\tilde{\mathbf{E}}_{11}' : \tilde{\mathbf{E}}_{21}'] \begin{bmatrix}\tilde{\mathbf{E}}_{11}\\ \tilde{\mathbf{E}}_{21}\end{bmatrix}$$

$$= \left(\begin{bmatrix}\mathbf{Y}_{11}\\ \mathbf{Y}_{21}\end{bmatrix} - \begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{bmatrix}\tilde{\mathbf{B}}_1 - \begin{bmatrix}\mathbf{0}\\ \tilde{\Theta}\tilde{\mathbf{V}}_{21}\end{bmatrix}\right)' \left(\begin{bmatrix}\mathbf{Y}_{11}\\ \mathbf{Y}_{21}\end{bmatrix} - \begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{bmatrix}\tilde{\mathbf{B}}_1 - \begin{bmatrix}\mathbf{0}\\ \tilde{\Theta}\tilde{\mathbf{V}}_{21}\end{bmatrix}\right)$$

$$= \left(\begin{bmatrix}\mathbf{Y}_{11}\\ \mathbf{Y}_{21}\end{bmatrix} - \mathbf{M}\begin{bmatrix}\mathbf{Y}_{11}\\ \mathbf{Y}_{21} - \tilde{\Theta}\tilde{\mathbf{V}}_{21}\end{bmatrix} - \begin{bmatrix}\mathbf{0}\\ \tilde{\Theta}\tilde{\mathbf{V}}_{21}\end{bmatrix}\right)' \left(\begin{bmatrix}\mathbf{Y}_{11}\\ \mathbf{Y}_{21}\end{bmatrix} - \mathbf{M}\begin{bmatrix}\mathbf{Y}_{11}\\ \mathbf{Y}_{21} - \tilde{\Theta}\tilde{\mathbf{V}}_{21}\end{bmatrix} - \begin{bmatrix}\mathbf{0}\\ \tilde{\Theta}\tilde{\mathbf{V}}_{21}\end{bmatrix}\right)$$

$$= (\mathbf{N}\mathbf{Y}_1 - \mathbf{N}_2\tilde{\Theta}\tilde{\mathbf{V}}_{21})'(\mathbf{N}\mathbf{Y}_1 - \mathbf{N}_2\tilde{\Theta}\tilde{\mathbf{V}}_{21})$$

$$= (\hat{\mathbf{E}}_1 - \mathbf{N}_2\tilde{\Theta}\tilde{\mathbf{V}}_{21})'(\hat{\mathbf{E}}_1 - \mathbf{N}_2\tilde{\Theta}\tilde{\mathbf{V}}_{21})$$

$$= \hat{\mathbf{E}}_1'\hat{\mathbf{E}}_1 - \hat{\mathbf{E}}_1'\mathbf{N}_2\tilde{\Theta}\tilde{\mathbf{V}}_{21} - \tilde{\mathbf{V}}_{12}\tilde{\Theta}'\mathbf{N}_2'\hat{\mathbf{E}}_1 + \tilde{\mathbf{V}}_{12}\tilde{\Theta}'\mathbf{N}_2'\mathbf{N}_2\tilde{\Theta}\tilde{\mathbf{V}}_{21}$$

$$= m \cdot \hat{\mathbf{V}}_{11} - \hat{\mathbf{E}}_{21}'\tilde{\Theta}\tilde{\mathbf{V}}_{21} - \tilde{\mathbf{V}}_{12}\tilde{\Theta}'\hat{\mathbf{E}}_{21} + \tilde{\mathbf{V}}_{12}\tilde{\Theta}'\mathbf{N}_{22}\tilde{\Theta}\tilde{\mathbf{V}}_{21}$$

$$= m \cdot \hat{\mathbf{V}}_{11} - \hat{\mathbf{E}}_{21}'\mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_{22}(\hat{\mathbf{E}}_2'\hat{\mathbf{E}}_2 - \hat{\mathbf{E}}_{22}'\mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_{22})^{-1}(\hat{\mathbf{E}}_2'\hat{\mathbf{E}}_1 - \hat{\mathbf{E}}_{22}'\mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_{21})$$
$$- (\hat{\mathbf{E}}_1'\hat{\mathbf{E}}_2 - \hat{\mathbf{E}}_{21}'\mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_{22})(\hat{\mathbf{E}}_2'\hat{\mathbf{E}}_2 - \hat{\mathbf{E}}_{22}'\mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_{22})^{-1}\hat{\mathbf{E}}_{22}'\mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_{21}$$
$$+ (\hat{\mathbf{E}}_1'\hat{\mathbf{E}}_2 - \hat{\mathbf{E}}_{21}'\mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_{22})(\hat{\mathbf{E}}_2'\hat{\mathbf{E}}_2 - \hat{\mathbf{E}}_{22}'\mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_{22})^{-1}\hat{\mathbf{E}}_{22}'\mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_{22} \cdot$$
$$\cdot (\hat{\mathbf{E}}_2'\hat{\mathbf{E}}_2 - \hat{\mathbf{E}}_{22}'\mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_{22})^{-1}(\hat{\mathbf{E}}_2'\hat{\mathbf{E}}_1 - \hat{\mathbf{E}}_{22}'\mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_{22}) .$$

These results are summarized in the following

Theorem 4.1 (Schall)

The ML-estimates $\mathbf{X}\tilde{\mathbf{B}}$, $\tilde{\Theta}$ and $\tilde{\mathbf{V}}$ for $\mathbf{XB}$, Θ and $\mathbf{V}$ in the adjusted model (4.11) are given by

(i) $$\mathbf{X}[\tilde{\mathbf{B}}_1 : \tilde{\mathbf{B}}_2]$$
$$= \mathbf{X}[\hat{\mathbf{B}}_1 : \hat{\mathbf{B}}_2] - \mathbf{M}_2\tilde{\Theta}[\tilde{\mathbf{V}}_{21} : \tilde{\mathbf{V}}_{22}]$$
$$= \mathbf{X}[\hat{\mathbf{B}}_1 : \hat{\mathbf{B}}_2] - \mathbf{M}_2\mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_{22}[(\hat{\mathbf{E}}_2'\hat{\mathbf{E}}_2 - \hat{\mathbf{E}}_{22}'\mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_{22})^{-1}(\hat{\mathbf{E}}_2'\hat{\mathbf{E}}_1 - \hat{\mathbf{E}}_{22}'\mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_{21}) : \mathbf{I}]$$

(ii) $$\tilde{\Theta} = \mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_{22}\tilde{\mathbf{V}}_{22}^{-1}$$
$$= m \cdot \mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_{22}(\hat{\mathbf{E}}_2'\hat{\mathbf{E}}_2 - \hat{\mathbf{E}}_{22}'\mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_{22})^{-1}$$

(iii) $$m \cdot \tilde{\mathbf{V}} = \hat{\mathbf{E}}'\hat{\mathbf{E}} - \begin{bmatrix} \hat{\mathbf{E}}_{21}'\tilde{\Theta}\tilde{\mathbf{V}}_{21} + \tilde{\mathbf{V}}_{12}\tilde{\Theta}'\hat{\mathbf{E}}_{21} - \tilde{\mathbf{V}}_{12}\tilde{\Theta}'\mathbf{N}_{22}\tilde{\Theta}\tilde{\mathbf{V}}_{21} : \hat{\mathbf{E}}_{22}'\mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_{21} \\ \hat{\mathbf{E}}_{21}'\mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_{22} \qquad\qquad\qquad\qquad\qquad\qquad : \hat{\mathbf{E}}_{22}'\mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_{22} \end{bmatrix}$$

where $\mathbf{X}\hat{\mathbf{B}}$, $\hat{\mathbf{E}}$ and $\hat{\mathbf{V}}$ are the ML-estimates for $\mathbf{XB}$, $\mathbf{E}$ and $\mathbf{V}$ under the original model (4.1).

□

We note that because of the removal of the error term $\mathbf{E}_{22}$ from the estimation of $\mathbf{V}$ in the adjusted model (4.11) the estimate $\tilde{\mathbf{V}}$ for $\mathbf{V}$ as given by Theorem 4.1 is biased. An estimate $\tilde{\tilde{\mathbf{V}}}$ for $\mathbf{V}$ in the adjusted model (4.11) corrected for the bias is obtained by

(4.33) $$\tilde{\tilde{\mathbf{V}}} = \begin{bmatrix} \tilde{\tilde{\mathbf{V}}}_{11} : \tilde{\tilde{\mathbf{V}}}_{12} \\ \tilde{\tilde{\mathbf{V}}}_{21} : \tilde{\tilde{\mathbf{V}}}_{22} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{V}}_{11} + \frac{k}{m-k} \cdot \tilde{\mathbf{V}}_{12}\tilde{\mathbf{V}}_{22}^{-1}\tilde{\mathbf{V}}_{21} : \frac{m}{m-k} \cdot \tilde{\mathbf{V}}_{12} \\ \frac{m}{m-k} \cdot \tilde{\mathbf{V}}_{21} \qquad\qquad\qquad : \frac{m}{m-k} \cdot \tilde{\mathbf{V}}_{22} \end{bmatrix} .$$

The LRT-statistic for testing the hypothesis $H_0 : \Theta = \mathbf{0}$ in model (4.11), which is the hypothesis that $\mathbf{Y}_{22}$ is not a D-outlier, is obtained by comparing the maximum likelihood under model (4.1) with the maximum likelihood under model (4.11) :

Corollary 4.1.1

The LRT-statistic for the hypothesis $H_0 : \Theta = \mathbf{0}$ in the adjusted model (4.11) is given by

(4.34) $$\chi^2 = m \cdot \ln \frac{|\hat{\mathbf{V}}|}{|\tilde{\mathbf{V}}|}$$

where $\hat{\mathbf{V}}$ and $\tilde{\mathbf{V}}$ are respectively the ML-estimates for $\mathbf{V}$ under the models (4.1) and (4.11). By the general theory of likelihood ratio tests, as presented by Rao (1973), the statistic χ^2 is asymptotically ($m \to \infty$) distributed chi-squared with $k \cdot l$ degrees of freedom. □

4.2.2 Outliers by additive shifts

As in the previous section we consider the adjusted model (4.7) assuming that possibly after some rearrangement of the rows and columns of the data matrix $\mathbf{Y}$ the set J of possible A-outliers forms a $k \times l$-submatrix $\mathbf{Y}_{22}$ of $\mathbf{Y}$ as given in (4.10).

The adjusted model (4.7) can then be written as

$$\text{(4.35)} \qquad \begin{bmatrix} \mathbf{Y}_{11} : \mathbf{Y}_{12} \\ \mathbf{Y}_{21} : \mathbf{Y}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} [\mathbf{B}_1 : \mathbf{B}_2] + \begin{bmatrix} \mathbf{0} \\ \Theta \end{bmatrix} [\mathbf{0} : \mathbf{I}] + \begin{bmatrix} \mathbf{E}_{11} : \mathbf{E}_{12} \\ \mathbf{E}_{21} : \mathbf{E}_{22} \end{bmatrix}$$

where the parameter matrix Θ is $k \times l$. Again we require $R(\mathbf{X}_2) \subset R(\mathbf{X}_1)$ for $\mathbf{X}_2\mathbf{B}_2$ to be estimable in the model (4.35).

If $l = n$, the model (4.35) is equivalent to the model (4.13) since the parameter matrix Θ in (4.35) is a reparametrization of the parameter matrix Θ in (4.11), and *vice versa*, with $\Theta_{(4.35)} = \Theta_{(4.11)} \cdot \mathbf{V}$. Thus, if a set of complete observational vectors is specified as outlying, A- and D-outliers can not be distinguished. We proceed assuming that $l < n$.

The ML-estimate $\mathbf{X}\tilde{\mathbf{B}}$ for $\mathbf{XB}$ under the adjusted model (4.35) is obtained by

$$\text{(4.36)} \qquad \begin{aligned} \mathbf{X}\tilde{\mathbf{B}} &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}' \left(\begin{bmatrix} \mathbf{Y}_{11} : \mathbf{Y}_{12} \\ \mathbf{Y}_{21} : \mathbf{Y}_{22} \end{bmatrix} - \begin{bmatrix} \mathbf{0} \\ \tilde{\Theta} \end{bmatrix} [\mathbf{0} : \mathbf{I}] \right) \\ &= \mathbf{X}[\hat{\mathbf{B}}_1 : \hat{\mathbf{B}}_2 - \mathbf{M}_2\tilde{\Theta}] \end{aligned}$$

where $\tilde{\Theta}$ is the MLE for Θ under model (4.35). Thus

$$\text{(4.37)} \qquad \mathbf{X}\tilde{\mathbf{B}}_1 = \mathbf{X}\hat{\mathbf{B}}_1 \ ,$$

$$\text{(4.38)} \qquad \tilde{\mathbf{E}}_1 = \hat{\mathbf{E}}_1 \ , \quad \text{and}$$

$$\text{(4.39)} \qquad m \cdot \tilde{\mathbf{V}}_{11} = \tilde{\mathbf{E}}_1'\tilde{\mathbf{E}}_1 = \hat{\mathbf{E}}_1'\hat{\mathbf{E}}_1 = m \cdot \hat{\mathbf{V}}_{11} \ .$$

To compute $\mathbf{X}\tilde{\mathbf{B}}_2$ we note that similar to the relationship (3.53) in the general linear model we obtain

$$\text{(4.40)} \qquad \begin{aligned} &\mathbf{Y}_{22} - \mathbf{X}_2\tilde{\mathbf{B}}_2 - \tilde{\Theta} \\ &= \tilde{\mathbf{E}}_{22} \\ &= \tilde{\mathbf{E}}_{21}\tilde{\mathbf{V}}_{11}^{-1}\tilde{\mathbf{V}}_{12} \\ &= \hat{\mathbf{E}}_{21}\hat{\mathbf{V}}_{11}^{-1}\hat{\mathbf{E}}_1'\tilde{\mathbf{E}}_2 \div m \ , \quad \text{from (4.38), (4.39)} \\ &= \hat{\mathbf{E}}_{21}\hat{\mathbf{V}}_{11}^{-1}[\hat{\mathbf{E}}_{11}' : \hat{\mathbf{E}}_{21}'] \left(\begin{bmatrix} \mathbf{Y}_{12} \\ \mathbf{Y}_{22} \end{bmatrix} - \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \tilde{\mathbf{B}}_2 - \begin{bmatrix} \mathbf{0} \\ \tilde{\Theta} \end{bmatrix} \right) \div m \ . \end{aligned}$$

But $\mathbf{X}\tilde{\mathbf{B}}_2 = \mathbf{X}\hat{\mathbf{B}}_2 - \mathbf{M}_2\tilde{\Theta}$ from (4.36), which with (4.40) yields

(4.41) $$\mathbf{Y}_{22} - \mathbf{X}_2\hat{\mathbf{B}}_2 + \mathbf{M}_{22}\tilde{\Theta} - \tilde{\Theta}$$
$$= \hat{\mathbf{E}}_{21}\hat{\mathbf{V}}_{11}^{-1}[\hat{\mathbf{E}}_{11}' : \hat{\mathbf{E}}_{21}'](\begin{bmatrix}\hat{\mathbf{E}}_{12}\\ \hat{\mathbf{E}}_{22}\end{bmatrix} + \mathbf{M}_2\tilde{\Theta} - \begin{bmatrix}\mathbf{0}\\ \tilde{\Theta}\end{bmatrix}) \div m \quad .$$

This equation is equivalent to

(4.42) $$\hat{\mathbf{E}}_{22} - \mathbf{N}_{22}\tilde{\Theta} = \hat{\mathbf{E}}_{21}\hat{\mathbf{V}}_{11}^{-1}(\hat{\mathbf{V}}_{12} - \hat{\mathbf{E}}_{21}'\tilde{\Theta} \div m)$$

since $\hat{\mathbf{E}}_1'\mathbf{N}_2 = \hat{\mathbf{E}}_{21}'$, so that

(4.43) $$\tilde{\Theta} = (\mathbf{N}_{22} - \hat{\mathbf{E}}_{21}\hat{\mathbf{V}}_{11}^{-1}\hat{\mathbf{E}}_{21}' \div m)^{-1}(\hat{\mathbf{E}}_{22} - \hat{\mathbf{E}}_{21}\hat{\mathbf{V}}_{11}^{-1}\hat{\mathbf{V}}_{12}) \quad .$$

Now the MLE $\mathbf{X}\tilde{\mathbf{B}}_2$ for $\mathbf{X}\mathbf{B}_2$ can be computed from (4.36), and with $\tilde{\mathbf{E}}_{12} = \mathbf{Y}_{12} - \mathbf{X}_1\tilde{\mathbf{B}}_2$ and $\tilde{\mathbf{E}}_{22} = \mathbf{Y}_{22} - \mathbf{X}_2\tilde{\mathbf{B}}_2 - \tilde{\Theta}$, $\tilde{\mathbf{V}}_{11}$ and $\tilde{\mathbf{V}}_{12}$ are obtained.

These results are summarized and an explicit formulation is given by the following theorem, which is essentially due to Anderson (1957). However, we represent here the adjusted estimates $\mathbf{X}\tilde{\mathbf{B}}$, $\tilde{\mathbf{V}}$ and $\tilde{\Theta}$ in terms of the estimates $\mathbf{X}\hat{\mathbf{B}}$, $\hat{\mathbf{V}}$ and $\hat{\mathbf{E}}$ under the original model (4.1).

Theorem 4.2 (Anderson, 1957)

The ML-estimates $\tilde{\Theta}$, $\mathbf{X}\tilde{\mathbf{B}}$ and $\tilde{\mathbf{V}}$ for Θ, $\mathbf{X}\mathbf{B}$ and $\mathbf{V}$ in the adjusted model (4.35) are obtained from

(i) $$\tilde{\Theta} = (\mathbf{N}_{22} - \hat{\mathbf{E}}_{21}\hat{\mathbf{V}}_{11}^{-1}\hat{\mathbf{E}}_{21}' \div m)^{-1}(\hat{\mathbf{E}}_{22} - \hat{\mathbf{E}}_{21}\hat{\mathbf{V}}_{11}^{-1}\hat{\mathbf{V}}_{12}) \quad .$$

(ii) $$\mathbf{X}\tilde{\mathbf{B}} = \mathbf{X}[\tilde{\mathbf{B}}_1 : \tilde{\mathbf{B}}_2]$$
$$= \mathbf{X}[\hat{\mathbf{B}}_1 : \tilde{\mathbf{B}}_2 - \mathbf{M}_2\tilde{\Theta}]$$
$$= \mathbf{X}[\hat{\mathbf{B}}_1 : \hat{\mathbf{B}}_2 - (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}_2'(\mathbf{N}_{22} - \hat{\mathbf{E}}_{21}\hat{\mathbf{V}}_{11}^{-1}\hat{\mathbf{E}}_{21}' \div m)^{-1}(\hat{\mathbf{E}}_{22} - \hat{\mathbf{E}}_{21}\hat{\mathbf{V}}_{11}^{-1}\hat{\mathbf{V}}_{12})] \quad .$$

(iii) $$m \cdot \tilde{\mathbf{V}} = (\mathbf{Y} - \mathbf{X}\tilde{\mathbf{B}} - \begin{bmatrix}\mathbf{0}\\ \tilde{\Theta}\end{bmatrix}[\mathbf{0} : \mathbf{I}])'(\mathbf{Y} - \mathbf{X}\tilde{\mathbf{B}} - \begin{bmatrix}\mathbf{0}\\ \tilde{\Theta}\end{bmatrix}[\mathbf{0} : \mathbf{I}])$$
$$= ([\hat{\mathbf{E}}_1 : \hat{\mathbf{E}}_2] - [\mathbf{0} : \mathbf{N}_2\tilde{\Theta}])'([\hat{\mathbf{E}}_1 : \hat{\mathbf{E}}_2] - [\mathbf{0} : \mathbf{N}_2\tilde{\Theta}])$$
$$= \begin{bmatrix}\hat{\mathbf{E}}_1'\hat{\mathbf{E}}_1 & : \hat{\mathbf{E}}_1'\hat{\mathbf{E}}_2 - \hat{\mathbf{E}}_{21}'\tilde{\Theta}\\ \hat{\mathbf{E}}_2'\hat{\mathbf{E}}_1 - \tilde{\Theta}'\hat{\mathbf{E}}_{21} & : \hat{\mathbf{E}}_2'\hat{\mathbf{E}}_2 - \tilde{\Theta}'\hat{\mathbf{E}}_{22} - \hat{\mathbf{E}}_{22}'\tilde{\Theta} + \tilde{\Theta}'\mathbf{N}_{22}\tilde{\Theta}\end{bmatrix}$$
$$= \hat{\mathbf{E}}'\hat{\mathbf{E}} - \begin{bmatrix}\mathbf{0} & : \hat{\mathbf{E}}_{21}'\tilde{\Theta}\\ \tilde{\Theta}'\hat{\mathbf{E}}_{21} & : \hat{\mathbf{E}}_{22}'\tilde{\Theta} + \tilde{\Theta}'\hat{\mathbf{E}}_{22} - \tilde{\Theta}'\hat{\mathbf{N}}_{22}\tilde{\Theta}\end{bmatrix}$$
$$= m \cdot \hat{\mathbf{V}} - \begin{bmatrix}\mathbf{0} & : \hat{\mathbf{E}}_{21}'\tilde{\Theta}\\ \tilde{\Theta}'\hat{\mathbf{E}}_{21} & : \hat{\mathbf{E}}_{22}'\tilde{\Theta} + \tilde{\Theta}'\hat{\mathbf{E}}_{22} - \tilde{\Theta}'\hat{\mathbf{N}}_{22}\tilde{\Theta}\end{bmatrix} \quad .$$

□

Similarly to the concluding remarks of the previous section we note that the estimate $\tilde{\mathbf{V}}$ for $\mathbf{V}$ in the adjusted model (4.35) as given by Theorem 4.2 is biased, due to the missing data $\mathbf{Y}_{22}$. An estimate $\tilde{\tilde{\mathbf{V}}}$ for $\mathbf{V}$ corrected for the bias is obtained by

$$\tilde{\tilde{\mathbf{V}}} = \begin{bmatrix} \tilde{\tilde{\mathbf{V}}}_{11} : \tilde{\tilde{\mathbf{V}}}_{12} \\ \tilde{\tilde{\mathbf{V}}}_{21} : \tilde{\tilde{\mathbf{V}}}_{22} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{V}}_{11} : & \tilde{\mathbf{V}}_{12} \\ \tilde{\mathbf{V}}_{21} : \frac{m}{m-k} \cdot \tilde{\mathbf{V}}_{22} - \frac{k}{m-k} \cdot \tilde{\mathbf{V}}_{21}\tilde{\mathbf{V}}_{11}^{-1}\tilde{\mathbf{V}}_{12} \end{bmatrix} . \tag{4.44}$$

This result is due to Orchard and Woodbury (1972).

Corollary 4.2.1

The LRT-statistic for the hypothesis $H_0 : \Theta = \mathbf{0}$ in the adjusted model (4.35), which is the hypothesis that $\mathbf{Y}_{22}$ is not an A-outlier, is given by $\chi^2 = m \cdot \ln(|\hat{\mathbf{V}}|/|\tilde{\mathbf{V}}|)$, which is asymptotically ($m \to \infty$) distributed chi-squared with $k \cdot l$ degrees of freedom.

□

The results of this section are also applicable when $\mathbf{Y}_{22}$ is not a set of possible outliers but a set of missing data points. That is, the data matrix $\mathbf{Y}$ is incomplete in such a way that possibly after some rearrangement of the rows and columns of $\mathbf{Y}$ the missing data points form a submatrix $\mathbf{Y}_{22}$ of $\mathbf{Y}$. In this case, the missing data points can be replaced by arbitrary values (e.g. $\mathbf{Y}_{22} = \mathbf{0}$) and the resulting ML-estimate $\tilde{\mathbf{Y}}_{22} = \mathbf{X}_2\tilde{\mathbf{B}}_2$ under model (4.35) is the missing data estimate for $\mathbf{Y}_{22}$.

4.2.3 Mean shifts and outliers in principal components

As the third type of outlier in the multivariate regression model (4.1) we treat in this section transformational (T-) outliers or outliers in principal components. Gnanadesikan and Kettenring (1972) and Hawkins (1974, 1980) used principal components to detect outliers in multivariate data, but the development here is different in presenting outliers in principal components as a proper type of outlier.

When a subset of l (say) principal components corresponding to the principal axes $\mathbf{P}_2$ (say) is specified as outlying for all observations belonging to a subset $\mathbf{Y}_2$ (say) of the observations $\mathbf{Y}$ in the model (4.1), then, possibly after some rearrangement of the rows and columns of $\mathbf{Y}$, the adjusted model (4.9) can be written as

$$\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \mathbf{B} + \begin{bmatrix} \mathbf{0} \\ \Theta \end{bmatrix} \mathbf{P}_2' + \begin{bmatrix} \mathbf{E}_1 \\ \mathbf{E}_2 \end{bmatrix} . \tag{4.45}$$

The parameter matrix Θ is here $k \times l$, that is, l principal components corresponding to the principal axes $\mathbf{P}_2$ are specified as outlying, for k observations $\mathbf{Y}_2$. The matrix $\mathbf{Y}_2$ is thus $k \times n$ and $\mathbf{X}_2$ is $k \times p$. Note that $\mathbf{Y}$ and $\mathbf{E}$ are partitioned differently from the partitioning in the previous sections, and thus $\mathbf{Y}_1$, $\mathbf{Y}_2$ and $\mathbf{E}_1$, $\mathbf{E}_2$ now denote different submatrices of $\mathbf{Y}$ and $\mathbf{E}$ respectively. The matrix $\mathbf{P}_2$ is a $n \times l$-submatrix of $\mathbf{P} = [\mathbf{P}_1 : \mathbf{P}_2]$, where

(3.4)
$$\mathbf{V} = \mathbf{P}\triangle\mathbf{P}' = [\mathbf{P}_1 : \mathbf{P}_2] \begin{bmatrix} \triangle_1 & \\ & \triangle_2 \end{bmatrix} \begin{bmatrix} \mathbf{P}_1' \\ \mathbf{P}_2' \end{bmatrix}$$

is the SVD of $\mathbf{V}$.

When $l=n$, the model (4.45) is a reparametrization of the model (4.13), and a reparametrization of the model (4.35) where $l=n$. Thus, if complete observational vectors are specified as outliers, D-, A- and T-outliers can not be distinguished. In the following we assume that $l<n$.

A model which is more general than (4.45) is the model

(4.46)
$$\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \mathbf{B} + \begin{bmatrix} \mathbf{0} \\ \mathbf{A}\Theta \end{bmatrix} \mathbf{P}_2' + \begin{bmatrix} \mathbf{E}_1 \\ \mathbf{E}_2 \end{bmatrix}$$

where $\mathbf{A}$ is an arbitrary known $k\times a$-matrix, so that Θ is here $a\times l$. The model (4.46) allows for an arbitrary mean shift in a subset of l principal components of $\mathbf{Y}_2$. Taking $\mathbf{A} = \mathbf{I}_k$ we obtain the model (4.45) as a special case. Without loss of generality we take $a \leq k$.

If the adjusted model (4.46) is transformed from the right by $\mathbf{P} = [\mathbf{P}_1 : \mathbf{P}_2]$ we obtain

(4.47)
$$\begin{bmatrix} \mathbf{Y}_1\mathbf{P}_1 : \mathbf{Y}_1\mathbf{P}_2 \\ \mathbf{Y}_2\mathbf{P}_1 : \mathbf{Y}_2\mathbf{P}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \mathbf{B}\mathbf{P} + \begin{bmatrix} \mathbf{0} \\ \mathbf{A}\Theta \end{bmatrix} [\mathbf{0} : \mathbf{I}] + \begin{bmatrix} \mathbf{E}_1\mathbf{P}_1 : \mathbf{E}_1\mathbf{P}_2 \\ \mathbf{E}_2\mathbf{P}_1 : \mathbf{E}_2\mathbf{P}_2 \end{bmatrix}$$

and thus with $\mathbf{Y}_2\mathbf{P}_2 = \mathbf{X}_2\mathbf{B}\mathbf{P}_2 + \mathbf{A}\Theta + \mathbf{E}_2\mathbf{P}_2$ we have the mean shift $\mathbf{A}\Theta$ in the PC's $\mathbf{Y}_2\mathbf{P}_2$.

The maximum likelihood estimation of the unknowns in the adjusted model (4.46) can be performed in closed form, as will be shown below. The ML-estimates $\mathbf{X}\tilde{\mathbf{B}}$, $\tilde{\Theta}$ and $\tilde{\mathbf{V}}$ for $\mathbf{X}\mathbf{B}$, Θ and $\mathbf{V}$ in the model (4.45) are then obtained by taking $\mathbf{A} = \mathbf{I}_k$ in the corresponding formulae for the ML-estimates $\mathbf{X}\tilde{\mathbf{B}}$, $\tilde{\Theta}$ and $\tilde{\mathbf{V}}$ in the model (4.46).

The MLE $\tilde{\mathbf{V}}$ for $\mathbf{V}$ in model (4.46) is given by

(4.48)
$$\begin{aligned} m\cdot\tilde{\mathbf{V}} &= \left(\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix} - \begin{bmatrix} \mathbf{0} \\ \mathbf{A}\tilde{\Theta} \end{bmatrix} \tilde{\mathbf{P}}_2'\right)'\mathbf{N}\left(\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix} - \begin{bmatrix} \mathbf{0} \\ \mathbf{A}\tilde{\Theta} \end{bmatrix} \tilde{\mathbf{P}}_2'\right) \\ &= \hat{\mathbf{E}}'\hat{\mathbf{E}} + \tilde{\mathbf{P}}_2\tilde{\Theta}'\mathbf{A}'\mathbf{N}_{22}\mathbf{A}\tilde{\Theta}\tilde{\mathbf{P}}_2' \\ &\quad - \tilde{\mathbf{P}}_2\tilde{\Theta}'\mathbf{A}'\hat{\mathbf{E}}_2 \\ &\quad - \hat{\mathbf{E}}_2'\mathbf{A}\tilde{\Theta}\tilde{\mathbf{P}}_2' \quad . \end{aligned}$$

Using the identities

(4.49)
$$\begin{aligned} \mathbf{A}\tilde{\Theta} &= \mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'(\mathbf{Y}_2 - \mathbf{X}_2\tilde{\mathbf{B}})\tilde{\mathbf{P}}_2 \\ &= \mathbf{M}_\mathbf{A}(\mathbf{Y}_2 - \mathbf{X}_2\tilde{\mathbf{B}})\tilde{\mathbf{P}}_2\,, \quad \mathbf{M}_\mathbf{A} = \mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'\,, \quad \text{and} \end{aligned}$$

(4.50)
$$\begin{aligned} \mathbf{X}_2\tilde{\mathbf{B}} &= \mathbf{X}_2(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\left(\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix} - \begin{bmatrix} \mathbf{0} \\ \mathbf{A}\Theta \end{bmatrix} \tilde{\mathbf{P}}_2'\right) \\ &= \mathbf{X}_2\hat{\mathbf{B}} - \mathbf{X}_2(\mathbf{X}'\mathbf{X})^-\mathbf{X}_2'\mathbf{A}\tilde{\Theta}\tilde{\mathbf{P}}_2' \\ &= \mathbf{X}_2\hat{\mathbf{B}} - \mathbf{M}_{22}\mathbf{A}\tilde{\Theta}\tilde{\mathbf{P}}_2' \end{aligned}$$

we conclude that

$$\text{(4.51)} \qquad \begin{aligned} \mathbf{A}\tilde{\Theta} &= \mathbf{M}_A(\mathbf{Y}_2 - \mathbf{X}_2\hat{\mathbf{B}} + \mathbf{M}_{22}\mathbf{A}\tilde{\Theta}\tilde{\mathbf{P}}_2')\tilde{\mathbf{P}}_2 \\ &= \mathbf{M}_A(\hat{\mathbf{E}}_2\tilde{\mathbf{P}}_2 + \mathbf{M}_{22}\mathbf{A}\tilde{\Theta}) \\ &= \mathbf{M}_A(\hat{\mathbf{E}}_2\tilde{\mathbf{P}}_2 + \mathbf{M}_{22}\mathbf{M}_A\mathbf{A}\tilde{\Theta}) \end{aligned}$$

which yields

$$\text{(4.52)} \qquad \mathbf{A}\tilde{\Theta} = (\mathbf{I} - \mathbf{M}_A\mathbf{M}_{22}\mathbf{M}_A)^{-1}\mathbf{M}_A\hat{\mathbf{E}}_2\tilde{\mathbf{P}}_2 \quad .$$

Now we proceed from (4.48):

$$\text{(4.53)} \qquad \begin{aligned} m \cdot \tilde{\mathbf{V}} &= \hat{\mathbf{E}}'\hat{\mathbf{E}} + \tilde{\mathbf{P}}_2\tilde{\Theta}'\mathbf{A}'\mathbf{N}_{22}\mathbf{A}\tilde{\Theta}\tilde{\mathbf{P}}_2' \\ &\quad - \tilde{\mathbf{P}}_2\tilde{\Theta}'\mathbf{A}'\hat{\mathbf{E}}_2 \\ &\quad - \hat{\mathbf{E}}_2'\mathbf{A}\tilde{\Theta}\tilde{\mathbf{P}}_2' \end{aligned}$$

$$\begin{aligned} &= \hat{\mathbf{E}}'\hat{\mathbf{E}} + \tilde{\mathbf{P}}_2\tilde{\Theta}'\mathbf{A}'\mathbf{N}_{22}\mathbf{A}\tilde{\Theta}\tilde{\mathbf{P}}_2' \\ &\quad - \tilde{\mathbf{P}}_2\tilde{\mathbf{P}}_2'\hat{\mathbf{E}}_2'\mathbf{M}_A(\mathbf{I} - \mathbf{M}_A\mathbf{M}_{22}\mathbf{M}_A)^{-1}\mathbf{M}_A\hat{\mathbf{E}}_2 \\ &\quad - \hat{\mathbf{E}}_2'\mathbf{M}_A(\mathbf{I} - \mathbf{M}_A\mathbf{M}_{22}\mathbf{M}_A)^{-1}\mathbf{M}_A\hat{\mathbf{E}}_2\tilde{\mathbf{P}}_2\tilde{\mathbf{P}}_2' \quad , \quad \text{from (4.52)} \end{aligned}$$

$$\begin{aligned} &= \hat{\mathbf{E}}'\hat{\mathbf{E}} + \tilde{\mathbf{P}}_2\tilde{\Theta}'\mathbf{A}'\mathbf{N}_{22}\mathbf{A}\tilde{\Theta}\tilde{\mathbf{P}}_2' \\ &\quad - 2 \cdot \tilde{\mathbf{P}}_2\tilde{\mathbf{P}}_2'\hat{\mathbf{E}}_2'\mathbf{M}_A(\mathbf{I} - \mathbf{M}_A\mathbf{M}_{22}\mathbf{M}_A)^{-1}\mathbf{M}_A\hat{\mathbf{E}}_2\tilde{\mathbf{P}}_2\tilde{\mathbf{P}}_2' \\ &\quad - \tilde{\mathbf{P}}_2\tilde{\mathbf{P}}_2'\hat{\mathbf{E}}_2'\mathbf{M}_A(\mathbf{I} - \mathbf{M}_A\mathbf{M}_{22}\mathbf{M}_A)^{-1}\mathbf{M}_A\hat{\mathbf{E}}_2\tilde{\mathbf{P}}_1\tilde{\mathbf{P}}_1' \\ &\quad - \tilde{\mathbf{P}}_1\tilde{\mathbf{P}}_1'\hat{\mathbf{E}}_2'\mathbf{M}_A(\mathbf{I} - \mathbf{M}_A\mathbf{M}_{22}\mathbf{M}_A)^{-1}\mathbf{M}_A\hat{\mathbf{E}}_2\tilde{\mathbf{P}}_2\tilde{\mathbf{P}}_2' \quad , \\ &\qquad\qquad \text{using } \tilde{\mathbf{P}}_1\tilde{\mathbf{P}}_1' + \tilde{\mathbf{P}}_2\tilde{\mathbf{P}}_2' = \mathbf{I} \end{aligned}$$

$$\begin{aligned} &= \hat{\mathbf{E}}'\hat{\mathbf{E}} + \tilde{\mathbf{P}}_2\tilde{\Theta}'\mathbf{A}'\mathbf{N}_{22}\mathbf{A}\tilde{\Theta}\tilde{\mathbf{P}}_2' \\ &\quad - 2 \cdot \tilde{\mathbf{P}}_2\tilde{\mathbf{P}}_2'\hat{\mathbf{E}}_2'\mathbf{M}_A(\mathbf{I} - \mathbf{M}_A\mathbf{M}_{22}\mathbf{M}_A)^{-1}\mathbf{M}_A\hat{\mathbf{E}}_2\tilde{\mathbf{P}}_2\tilde{\mathbf{P}}_2' \\ &\quad - \tilde{\mathbf{P}}_2\tilde{\mathbf{P}}_2'\hat{\mathbf{E}}_2'\mathbf{M}_A(\mathbf{I} - \mathbf{M}_A\mathbf{M}_{22}\mathbf{M}_A)^{-1}\mathbf{M}_A\hat{\mathbf{E}}_2\tilde{\mathbf{P}}_1\tilde{\mathbf{P}}_1' \\ &\quad - (\mathbf{I} - \tilde{\mathbf{P}}_2\tilde{\mathbf{P}}_2')\hat{\mathbf{E}}_2'\mathbf{M}_A(\mathbf{I} - \mathbf{M}_A\mathbf{M}_{22}\mathbf{M}_A)^{-1}\mathbf{M}_A\hat{\mathbf{E}}_2\tilde{\mathbf{P}}_2\tilde{\mathbf{P}}_2' \end{aligned}$$

$$\begin{aligned} &= \hat{\mathbf{E}}'\hat{\mathbf{E}} + \tilde{\mathbf{P}}_2\tilde{\Theta}'\mathbf{A}'\mathbf{N}_{22}\mathbf{A}\tilde{\Theta}\tilde{\mathbf{P}}_2' \\ &\quad - \tilde{\mathbf{P}}_2\tilde{\mathbf{P}}_2'\hat{\mathbf{E}}_2'\mathbf{M}_A(\mathbf{I} - \mathbf{M}_A\mathbf{M}_{22}\mathbf{M}_A)^{-1}\mathbf{M}_A\hat{\mathbf{E}}_2\tilde{\mathbf{P}}_2\tilde{\mathbf{P}}_2' \\ &\quad - \tilde{\mathbf{P}}_2\tilde{\mathbf{P}}_2'\hat{\mathbf{E}}_2'\mathbf{M}_A(\mathbf{I} - \mathbf{M}_A\mathbf{M}_{22}\mathbf{M}_A)^{-1}\mathbf{M}_A\hat{\mathbf{E}}_2\tilde{\mathbf{P}}_1\tilde{\mathbf{P}}_1' \\ &\quad - \hat{\mathbf{E}}_2'\mathbf{M}_A(\mathbf{I} - \mathbf{M}_A\mathbf{M}_{22}\mathbf{M}_A)^{-1}\mathbf{M}_A\hat{\mathbf{E}}_2(\mathbf{I} - \tilde{\mathbf{P}}_1\tilde{\mathbf{P}}_1') \end{aligned}$$

$$= \hat{\mathbf{E}}'\hat{\mathbf{E}} + \tilde{\mathbf{P}}_2\tilde{\mathbf{P}}_2'\hat{\mathbf{E}}_2'\mathbf{M}_A(\mathbf{I} - \mathbf{M}_A\mathbf{M}_{22}\mathbf{M}_A)^{-1}\mathbf{N}_{22}(\mathbf{I} - \mathbf{M}_A\mathbf{M}_{22}\mathbf{M}_A)^{-1}\mathbf{M}_A\hat{\mathbf{E}}_2\tilde{\mathbf{P}}_2\tilde{\mathbf{P}}_2'$$
$$- \hat{\mathbf{E}}_2'\mathbf{M}_A(\mathbf{I} - \mathbf{M}_A\mathbf{M}_{22}\mathbf{M}_A)^{-1}\mathbf{M}_A\hat{\mathbf{E}}_2$$
$$+ \tilde{\mathbf{P}}_1\tilde{\mathbf{P}}_1'\hat{\mathbf{E}}_2'\mathbf{M}_A(\mathbf{I} - \mathbf{M}_A\mathbf{M}_{22}\mathbf{M}_A)^{-1}\hat{\mathbf{E}}_2\tilde{\mathbf{P}}_1\tilde{\mathbf{P}}_1'$$
$$- \tilde{\mathbf{P}}_2\tilde{\mathbf{P}}_2'\hat{\mathbf{E}}_2'\mathbf{M}_A(\mathbf{I} - \mathbf{M}_A\mathbf{M}_{22}\mathbf{M}_A)^{-1}\hat{\mathbf{E}}_2\tilde{\mathbf{P}}_2\tilde{\mathbf{P}}_2'$$

$$= \hat{\mathbf{E}}'\hat{\mathbf{E}} - \hat{\mathbf{E}}_2'\mathbf{M}_A(\mathbf{I} \quad \mathbf{M}_A\mathbf{M}_{22}\mathbf{M}_A)^{-1}\mathbf{M}_A\hat{\mathbf{E}}_2$$
$$+ \tilde{\mathbf{P}}_1\tilde{\mathbf{P}}_1'\hat{\mathbf{E}}_2'\mathbf{M}_A(\mathbf{I} - \mathbf{M}_A\mathbf{M}_{22}\mathbf{M}_A)^{-1}\mathbf{M}_A\hat{\mathbf{E}}_2\tilde{\mathbf{P}}_1\tilde{\mathbf{P}}_1'$$
$$- \tilde{\mathbf{P}}_2\tilde{\mathbf{P}}_2'\hat{\mathbf{E}}_2'\mathbf{M}_A(\mathbf{I} - \mathbf{M}_A\mathbf{M}_{22}\mathbf{M}_A)^{-1}(\mathbf{M}_{22} - \mathbf{M}_A\mathbf{M}_{22}\mathbf{M}_A) \cdot$$
$$\cdot (\mathbf{I} - \mathbf{M}_A\mathbf{M}_{22}\mathbf{M}_A)^{-1}\mathbf{M}_A\hat{\mathbf{E}}_2\tilde{\mathbf{P}}_2\tilde{\mathbf{P}}_2' \ .$$

Now let

(4.54) $$[\mathbf{P}_1^* : \mathbf{P}_2^*] \begin{bmatrix} \Delta_1^* & \\ & \Delta_2^* \end{bmatrix} \begin{bmatrix} \mathbf{P}_1^{*\prime} \\ \mathbf{P}_2^{*\prime} \end{bmatrix}$$

be the SVD of $\hat{\mathbf{E}}'\hat{\mathbf{E}} - \hat{\mathbf{E}}_2'\mathbf{M}_A(\mathbf{I} - \mathbf{M}_A\mathbf{M}_{22}\mathbf{M}_A)^{-1}\mathbf{M}_A\hat{\mathbf{E}}_2$, and further let

(4.55) $$\mathbf{Q}_1\tilde{\Delta}_1\mathbf{Q}_1' \ , \quad \text{and}$$

(4.56) $$\mathbf{Q}_2\tilde{\Delta}_2\mathbf{Q}_2'$$

be respectively the SVD's of $\Delta_1^* + \mathbf{P}_1^{*\prime}\hat{\mathbf{E}}_2'\mathbf{M}_A(\mathbf{I} - \mathbf{M}_A\mathbf{M}_{22}\mathbf{M}_A)^{-1}\mathbf{M}_A\hat{\mathbf{E}}_2\mathbf{P}_1^*$ and $\Delta_2^* - \mathbf{P}_2^{*\prime}\hat{\mathbf{E}}_2'\mathbf{M}_A\ (\mathbf{I} - \mathbf{M}_A\mathbf{M}_{22}\mathbf{M}_A)^{-1}\ (\mathbf{M}_{22} - \mathbf{M}_A\mathbf{M}_{22}\mathbf{M}_A)\ (\mathbf{I} - \mathbf{M}_A\mathbf{M}_{22}\mathbf{M}_A)^{-1}\ \mathbf{M}_A\hat{\mathbf{E}}_2\mathbf{P}_2^*$. Then $\tilde{\mathbf{V}}$ must be taken as

(4.57) $$m \cdot \tilde{\mathbf{V}} = [\mathbf{P}_1^*\mathbf{Q}_1 : \mathbf{P}_2^*\mathbf{Q}_2] \begin{bmatrix} \tilde{\Delta}_1 & \\ & \tilde{\Delta}_2 \end{bmatrix} \begin{bmatrix} \mathbf{Q}_1'\mathbf{P}_1^{*\prime} \\ \mathbf{Q}_2'\mathbf{P}_2^{*\prime} \end{bmatrix}$$
$$= [\tilde{\mathbf{P}}_1 : \tilde{\mathbf{P}}_2] \begin{bmatrix} \tilde{\Delta}_1 & \\ & \tilde{\Delta}_2 \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{P}}_1' \\ \tilde{\mathbf{P}}_2' \end{bmatrix}$$

where $\tilde{\mathbf{P}}_1$ and $\tilde{\mathbf{P}}_2$ are respectively taken as $\tilde{\mathbf{P}}_1 = \mathbf{P}_1^*\mathbf{Q}_1$ and $\tilde{\mathbf{P}}_2 = \mathbf{P}_2^*\mathbf{Q}_2$ as indicated by (4.57). That $\tilde{\mathbf{V}}$ as given by (4.57) in fact satisfies the relationship (4.53) is shown by evaluating $\mathbf{P}^{*\prime}\tilde{\mathbf{V}}\mathbf{P}^*$:

(4.58) $$m \cdot \mathbf{P}^{*\prime}\tilde{\mathbf{V}}\mathbf{P}^*$$
$$= \begin{bmatrix} \mathbf{P}_1^{*\prime} \\ \mathbf{P}_2^{*\prime} \end{bmatrix} [\mathbf{P}_1^*\mathbf{Q}_1 : \mathbf{P}_2^*\mathbf{Q}_2] \begin{bmatrix} \tilde{\Delta}_1 & \\ & \tilde{\Delta}_2 \end{bmatrix} \begin{bmatrix} \mathbf{Q}_1'\mathbf{P}_1^{*\prime} \\ \mathbf{Q}_2'\mathbf{P}_2^{*\prime} \end{bmatrix} [\mathbf{P}_1^* : \mathbf{P}_2^*]$$
$$= \begin{bmatrix} \mathbf{Q}_1 : \mathbf{0} \\ \mathbf{0} : \mathbf{Q}_2 \end{bmatrix} \begin{bmatrix} \tilde{\Delta}_1 & \\ & \tilde{\Delta}_2 \end{bmatrix} \begin{bmatrix} \mathbf{Q}_1' : \mathbf{0} \\ \mathbf{0} : \mathbf{Q}_2' \end{bmatrix}$$
$$= \begin{bmatrix} \mathbf{Q}_1\tilde{\Delta}_1\mathbf{Q}_1' : & \mathbf{0} \\ \mathbf{0} & : \mathbf{Q}_2\tilde{\Delta}_2\mathbf{Q}_2' \end{bmatrix} \ .$$

This implies

(4.59) $m \cdot \tilde{\mathbf{V}}$

$$= m \cdot \mathbf{P}^*\mathbf{P}^{*\prime}\tilde{\mathbf{V}}\mathbf{P}^*\mathbf{P}^{*\prime}$$

$$= \mathbf{P}_1^*\mathbf{Q}_1\tilde{\Delta}_1\mathbf{Q}_1'\mathbf{P}_1^{*\prime} + \mathbf{P}_2^*\mathbf{Q}_2\tilde{\Delta}_2\mathbf{Q}_2'\mathbf{P}_2^{*\prime}$$

$$= \mathbf{P}_1^*\Delta_1^*\mathbf{P}_1^{*\prime} + \mathbf{P}_2^*\Delta_2^*\mathbf{P}_2^{*\prime}$$

$$+ \mathbf{P}_1^*\mathbf{P}_1^{*\prime}\hat{\mathbf{E}}_2'\mathbf{M}_\mathbf{A}(\mathbf{I} - \mathbf{M}_\mathbf{A}\mathbf{M}_{22}\mathbf{M}_\mathbf{A})^{-1}\mathbf{M}_\mathbf{A}\hat{\mathbf{E}}_2\mathbf{P}_1^*\mathbf{P}_1^{*\prime}$$

$$- \mathbf{P}_2^*\mathbf{P}_2^{*\prime}\hat{\mathbf{E}}_2'\mathbf{M}_\mathbf{A}(\mathbf{I} - \mathbf{M}_\mathbf{A}\mathbf{M}_{22}\mathbf{M}_\mathbf{A})^{-1}(\mathbf{M}_{22} - \mathbf{M}_\mathbf{A}\mathbf{M}_{22}\mathbf{M}_\mathbf{A})\,(\mathbf{I} - \mathbf{M}_\mathbf{A}\mathbf{M}_{22}\mathbf{M}_\mathbf{A})^{-1}\mathbf{M}_\mathbf{A}\hat{\mathbf{E}}_2\mathbf{P}_2^*\mathbf{P}_2^{*\prime}$$

which is precisely the last term in (4.53) since $\mathbf{P}_1^*\Delta_1^*\mathbf{P}_1^{*\prime} + \mathbf{P}_2^*\Delta_2^*\mathbf{P}_2^{*\prime}$ is the SVD of $\hat{\mathbf{E}}'\hat{\mathbf{E}} - \hat{\mathbf{E}}_2'\mathbf{M}_\mathbf{A}(\mathbf{I} - \mathbf{M}_\mathbf{A}\mathbf{M}_{22}\mathbf{M}_\mathbf{A})^{-1}\mathbf{M}_\mathbf{A}\hat{\mathbf{E}}_2$ from (4.54), and $\mathbf{P}_i^*\mathbf{P}_i^{*\prime} = \tilde{\mathbf{P}}_i\tilde{\mathbf{P}}_i'$, $i = 1,2$.

Thus a PCA of the model (4.46) is performed by computing three SVD's as given above in (4.54) through (4.56). We formulate this procedure as an algorithm.

Algorithm 4.3 (Schall)

PCA of the adjusted model (4.46)

Step 1: Compute the SVD $\mathbf{P}^*\Delta^*\mathbf{P}^{*\prime}$ of $\hat{\mathbf{E}}'\hat{\mathbf{E}} - \hat{\mathbf{E}}_2'\mathbf{M}_\mathbf{A}(\mathbf{I} - \mathbf{M}_\mathbf{A}\mathbf{M}_{22}\mathbf{M}_\mathbf{A})^{-1}\mathbf{M}_\mathbf{A}\hat{\mathbf{E}}_2$.

Step 2: Choose the l principal axes $\mathbf{P}_2^*$ whose corresponding principal components $\mathbf{Y}_2\mathbf{P}_2^*$ are shifted by $\mathbf{A}\Theta$, to obtain a partitioning of $\mathbf{P}^*\Delta^*\mathbf{P}^{*\prime}$ as

(4.54)
$$\mathbf{P}^*\Delta^*\mathbf{P}^{*\prime} = [\mathbf{P}_1^* : \mathbf{P}_2^*] \begin{bmatrix} \Delta_1^* & \\ & \Delta_2^* \end{bmatrix} \begin{bmatrix} \mathbf{P}_1^{*\prime} \\ \mathbf{P}_2^{*\prime} \end{bmatrix} .$$

Step 3: Compute the SVD's $\mathbf{Q}_1\tilde{\Delta}_1\mathbf{Q}_1'$ and $\mathbf{Q}_2\tilde{\Delta}_2\mathbf{Q}_2'$ of
$\Delta_1^* + \mathbf{P}_1^{*\prime}\hat{\mathbf{E}}_2'\mathbf{M}_\mathbf{A}(\mathbf{I} - \mathbf{M}_\mathbf{A}\mathbf{M}_{22}\mathbf{M}_\mathbf{A})^{-1}\mathbf{M}_\mathbf{A}\hat{\mathbf{E}}_2\mathbf{P}_1^*$ and
$\Delta_2^* - \mathbf{P}_2^{*\prime}\hat{\mathbf{E}}_2'\mathbf{M}_\mathbf{A}(\mathbf{I} - \mathbf{M}_\mathbf{A}\mathbf{M}_{22}\mathbf{M}_\mathbf{A})^{-1}\,(\mathbf{M}_{22} - \mathbf{M}_\mathbf{A}\mathbf{M}_{22}\mathbf{M}_\mathbf{A})\,(\mathbf{I} - \mathbf{M}_\mathbf{A}\mathbf{M}_{22}\mathbf{M}_\mathbf{A})^{-1}\mathbf{M}_\mathbf{A}\hat{\mathbf{E}}_2\mathbf{P}_2^*$.

Step 4: (i) The MLE $\tilde{\mathbf{V}}$ of $\mathbf{V}$ under model (4.45) is given by

(4.57)
$$m \cdot \tilde{\mathbf{V}} = [\mathbf{P}_1^*\mathbf{Q}_1 : \mathbf{P}_2^*\mathbf{Q}_2] \begin{bmatrix} \tilde{\Delta}_1 & \\ & \tilde{\Delta}_2 \end{bmatrix} \begin{bmatrix} \mathbf{Q}_1'\mathbf{P}_1^{*\prime} \\ \mathbf{Q}_2'\mathbf{P}_2^{*\prime} \end{bmatrix} = m \cdot \tilde{\mathbf{P}}\tilde{\Delta}\tilde{\mathbf{P}}'$$

and thus the estimated principal axes of $\mathbf{Y}'$ are $[\mathbf{P}_1^*\mathbf{Q}_1 : \mathbf{P}_2^*\mathbf{Q}_2]$ with corresponding estimated variances of the principal components of $\mathbf{Y}$ given by $\tilde{\Delta} \div m$.

(ii) The MLE $\mathbf{A}\tilde{\Theta}$ for $\mathbf{A}\Theta$ is given by

(4.52)
$$\mathbf{A}\tilde{\Theta} = (\mathbf{I} - \mathbf{M}_\mathbf{A}\mathbf{M}_{22}\mathbf{M}_\mathbf{A})^{-1}\mathbf{M}_\mathbf{A}\hat{\mathbf{E}}_2\tilde{\mathbf{P}}_2 .$$

(iii) The MLE $\mathbf{X}\tilde{\mathbf{B}}$ for $\mathbf{X}\mathbf{B}$ is given by

(4.50)
$$\mathbf{X}\tilde{\mathbf{B}} = \mathbf{X}\hat{\mathbf{B}} - \mathbf{M}_2\mathbf{A}\tilde{\Theta}\tilde{\mathbf{P}}_2' .$$

□

Note that Algorithm 4.3 can be seen as a two-stage maximum likelihood method. In step 2 of the algorithm a set of l principal axes $\mathbf{P}_2^*$ is chosen, and based on that choice the ML-estimates of $\mathbf{XB}$, $\mathbf{A}\Theta$ and $\mathbf{V}$ are computed in step 4. If ML-estimates based on all possible subsets of l principal axes are computed, and the subset yielding the highest likelihood is selected, then the procedure is a two stage maximum likelihood method.

We also note that once the set $\mathbf{P}_2^*$ of l principal axes is chosen, and thus a partitioning of the space $\mathbb{R}^n$ into the subspaces spanned by the columns of $\mathbf{P}_1^*$ and $\mathbf{P}_2^*$ respectively, this partitioning is preserved by the principal axes obtained by maximum likelihood. That is, $C(\mathbf{P}_1^*) = C(\tilde{\mathbf{P}}_1) = C(\mathbf{P}_1^*\mathbf{Q}_1)$ and $C(\mathbf{P}_2^*) = C(\tilde{\mathbf{P}}_2) = C(\mathbf{P}_2^*\mathbf{Q}_2)$.

The likelihood ratio test statistic for the hypothesis $H_0 : \mathbf{A}\Theta = \mathbf{0}$ which is the hypothesis that no mean shift is present, is obtained by comparing the maximum likelihood under the original model (4.1) and the adjusted model (4.46) respectively.

Corollary 4.3.1

The LRT-statistic for the hypothesis $H_0 : \mathbf{A}\Theta = \mathbf{0}$ under the adjusted model (4.46) is given by

(4.60)
$$\begin{aligned} \chi^2 &= m \cdot \ln \frac{|\tilde{\mathbf{V}}|}{|\hat{\mathbf{V}}|} \\ &= m \cdot \ln \frac{|\tilde{\Delta}|}{|\hat{\mathbf{V}}|} \end{aligned}$$

which is asymptotically ($m \to \infty$) distributed chi-squared with $rank(\mathbf{A}) \cdot l$ degrees of freedom.

□

As noted above, taking $\mathbf{A} = \mathbf{I}_k$ we obtain the model (4.45) which is the model (4.1) adjusted for outliers in principal components or T-outliers. The computations in Algorithm 4.3 are considerably simplified by the fact that $\mathbf{A} = \mathbf{I}$, $\mathbf{M}_\mathbf{A} = \mathbf{I}$ and $\mathbf{M}_{22} - \mathbf{M}_\mathbf{A}\mathbf{M}_{22}\mathbf{M}_\mathbf{A} = \mathbf{0}$. The corresponding version of Algorithm 4.3 for the outlier problem is

Algorithm 4.3.2

PCA of the adjusted model (4.45)

Step 1: Compute the SVD $\mathbf{P}^*\triangle^*\mathbf{P}^{*\prime}$ of $\hat{\mathbf{E}}'\hat{\mathbf{E}} - \hat{\mathbf{E}}_2'\mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_2$.

Step 2: Choose the l principal axes $\mathbf{P}_2^*$ whose corresponding principal components $\mathbf{Y}_2\mathbf{P}_2^*$ are outlying to obtain a partitioning of $\mathbf{P}^*\triangle^*\mathbf{P}^{*\prime}$ as

(4.57a)
$$\mathbf{P}^*\triangle^*\mathbf{P}^{*\prime} = [\mathbf{P}_1^* : \mathbf{P}_2^*] \begin{bmatrix} \triangle_1^* & \\ & \triangle_2^* \end{bmatrix} \begin{bmatrix} \mathbf{P}_1^{*\prime} \\ \mathbf{P}_2^{*\prime} \end{bmatrix} .$$

Step 3: Compute the SVD $\mathbf{Q}_1\tilde{\Delta}_1\mathbf{Q}_1'$ of $\Delta_1^* + \mathbf{P}_1^{*\prime}\hat{\mathbf{E}}_2'\mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_2\mathbf{P}_1^*$.

Step 4: (i) The MLE $\tilde{\mathbf{V}}$ of $\mathbf{V}$ under model (4.43) is given by

(4.59a)
$$m \cdot \tilde{\mathbf{V}} = [\mathbf{P}_1^*\mathbf{Q}_1 : \mathbf{P}_2^*] \begin{bmatrix} \tilde{\Delta}_1 & \\ & \Delta_2^* \end{bmatrix} \begin{bmatrix} \mathbf{Q}_1'\mathbf{P}_1^{*\prime} \\ \mathbf{P}_2^{*\prime} \end{bmatrix}$$

and thus the estimated principal axes of $\mathbf{Y}'$ are $[\mathbf{P}_1^*\mathbf{Q}_1 : \mathbf{P}_2^*]$ with corresponding estimated variances of the principal components given by $\tilde{\Delta}_1 \div m$ and $\Delta_2^* \div m$.

(ii) The MLE $\mathbf{A}\tilde{\Theta}$ for $\mathbf{A}\Theta$ is given by

(4.52a)
$$\tilde{\Theta} = \mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_2\mathbf{P}_2^*$$

(iii) The MLE $\mathbf{X}\tilde{\mathbf{B}}$ for $\mathbf{XB}$ is given by

(4.50a)
$$\begin{aligned} \mathbf{X}\tilde{\mathbf{B}} &= \mathbf{X}\hat{\mathbf{B}} - \mathbf{M}_2\tilde{\Theta}\mathbf{P}_2^{*\prime} \\ &= \mathbf{X}\hat{\mathbf{B}} - \mathbf{M}_2\mathbf{N}_{22}^{-1}\hat{\mathbf{E}}_2\mathbf{P}_2^*\mathbf{P}_2^{*\prime} \ . \end{aligned}$$

□

As before in Sections 4.2.1 and 4.2.2 we note that the estimate $\tilde{\mathbf{V}}$ for $\mathbf{V}$ as given by (4.59a) is biased, and an estimate $\tilde{\tilde{\mathbf{V}}}$ for $\mathbf{V}$ corrected for the bias is obtained by

(4.59b)
$$\tilde{\tilde{\mathbf{V}}} = [\mathbf{P}_1^*\mathbf{Q}_1 : \mathbf{P}_2^*] \begin{bmatrix} \frac{1}{m} \cdot \tilde{\Delta}_1 : & \mathbf{0} \\ \mathbf{0} & : \frac{1}{m-k} \cdot \Delta_2^* \end{bmatrix} \begin{bmatrix} \mathbf{Q}_1'\mathbf{P}_1^{*\prime} \\ \mathbf{P}_2^{*\prime} \end{bmatrix} .$$

The LRT-statistic for testing the hypothesis $H_0 : \Theta = \mathbf{0}$ in the adjusted model (4.45) is obtained in a manner similar to Corollary 4.3.1.

Flury (1984, 1985) treated the model of common principal components. He assumed that the n-variate random vectors $\mathbf{y}_i$ $(i = 1, \ldots, k)$ are independently distributed as $N_n(\mu_i, \mathbf{V}_i)$, where $\mu_i \in \mathbb{R}^n$ and $\mathbf{V}_i$ are positive definite and symmetric. The hypothesis of common principal components is that the variance-covariance matrices $\mathbf{V}_i$ are simultaneously diagonalizable, that is, there exists an orthogonal matrix $\mathbf{P}$ such that

(4.61)
$$\mathbf{P}'\mathbf{V}_i\mathbf{P} = \Delta_i \ , \quad i = 1, \ldots, k \ .$$

The method amounts to assuming common principal axes for the random vectors $\mathbf{y}_1, \ldots, \mathbf{y}_k$, but allowing for different variances of the corresponding principal components.

In the second paper, Flury (1985) generalized the model of common principal components in a way that only a subset of the principal components of the random vectors $\mathbf{y}_i$, $i = 1, \ldots, k$ are common to all random vectors. During the development of this section, in contrast, we assumed common principal components with common variances for all observational vectors, but for a subset of observations and a subset of principal components we allowed for a mean shift, as opposed to a shift in the variance which was treated by Flury (1985).

4.2.4 Nested outlier patterns

More general than the outlier pattern we have treated in the previous sections, when the data points suspected to be outlying form a submatrix $\mathbf{Y}_{22}$ of the data matrix $\mathbf{Y}$, is the nested outlier pattern. In this case we assume that the data matrix $\mathbf{Y}$ of the multivariate regression model (4.1) can be partitioned as

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_{11} & \cdots & \mathbf{Y}_{1q} \\ \vdots & & \vdots \\ \mathbf{Y}_{q1} & \cdots & \mathbf{Y}_{qq} \end{bmatrix} \tag{4.62}$$

such that the rectangular submatrices

$$\{\mathbf{Y}_{ij} \mid j > q-i+1\} \tag{4.63}$$

contain the data points suspected to be outlying. The submatrices $\mathbf{Y}_{ij}$ of $\mathbf{Y}$ given by (4.63) are the submatrices below the contragredient diagonal of $\mathbf{Y}$. Naturally, the partitioning of $\mathbf{Y}$ as in (4.62), (4.63) may have been obtained after a suitable rearrangement of the rows and columns of $\mathbf{Y}$, and a corresponding rearrangement of the rows of $\mathbf{X}$ in model (4.1).

Conformably with the partitioning of $\mathbf{Y}$ we partition $\mathbf{X}$, $\mathbf{B}$ and $\mathbf{V}$ as

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_q \end{bmatrix}, \tag{4.64}$$

$$\mathbf{B} = [\mathbf{B}_1 \cdots \mathbf{B}_q], \quad \text{and}$$

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_{11} & \cdots & \mathbf{V}_{1q} \\ \vdots & & \vdots \\ \mathbf{V}_{q1} & \cdots & \mathbf{V}_{qq} \end{bmatrix}.$$

The principal submatrices $\mathbf{Y}_{ii}$ of $\mathbf{Y}$ are of order $k_i \times l_i$ $(i = 1, \ldots, q)$, which implies that the submatrices $\mathbf{X}_i$ of $\mathbf{X}$ are of order $k_i \times p$, the submatrices $\mathbf{B}_i$ of $\mathbf{B}$ are of order $p \times l_i$ and the principal submatrices $\mathbf{V}_{ii}$ of $\mathbf{V}$ are of order $l_i \times l_i$ $(i = 1, \ldots, q)$. Clearly, to conform with the notation before, we have that $\sum_{i=1}^{q} k_i = m$ and $\sum_{i=1}^{q} l_i = n$.

When the suspected outlier pattern is nested in the form (4.62), (4.63), then the ML-estimation of the unknowns in the corresponding models adjusted for outliers can still be performed in closed form, extending the methods presented in Sections 4.2.1 through 4.2.3.

The outlier pattern described by (4.62) and (4.63) we call nested of order $q-1$. Clearly, the pattern treated in Sections 4.2.1 through 4.2.3 and corresponding models (4.11), (4.35) and (4.45) is nested of order $q-1 = 1$, and is thus a special case of the pattern (4.62), (4.63).

Distributional outliers

The model (4.1), adjusted for D-outliers in a nested pattern, is

$$(4.65)\qquad \begin{bmatrix} \mathbf{Y}_{11} \cdots \mathbf{Y}_{1q} \\ \vdots \quad\quad \vdots \\ \mathbf{Y}_{q1} \cdots \mathbf{Y}_{qq} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_q \end{bmatrix} [\mathbf{B}_1 \cdots \mathbf{B}_q] + \begin{bmatrix} \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & & \Theta_{2q} \\ & & \vdots \\ \mathbf{0} & \Theta_{q2} \cdots & \Theta_{qq} \end{bmatrix} \begin{bmatrix} \mathbf{V}_{11} \cdots \mathbf{V}_{1q} \\ \vdots \quad\quad \vdots \\ \mathbf{V}_{q1} \cdots \mathbf{V}_{qq} \end{bmatrix} + \begin{bmatrix} \mathbf{E}_{11} \cdots \mathbf{E}_{1q} \\ \vdots \quad\quad \vdots \\ \mathbf{E}_{q1} \cdots \mathbf{E}_{qq} \end{bmatrix}$$

In (4.65) we require $q \geq 2$. For $q = 2$ the model (4.11) is obtained as a special case. The model (4.65) is in turn a special case of (4.6).

Maximum-likelihood estimation of the unknowns in (4.65) can be performed in closed form by the following algorithm. To simplify the notation, let

$$(4.66)\qquad \mathbf{X}_a = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_{a+1} \end{bmatrix}, \quad \mathbf{M}_a = \mathbf{X}_a(\mathbf{X}_a'\mathbf{X}_a)^{-}\mathbf{X}_a', \quad \mathbf{N}_a = \mathbf{I} - \mathbf{M}_a$$

$$(4.67)\qquad \mathbf{V}_a = \begin{bmatrix} \mathbf{V}_{q-a,q-a} & \cdots & \mathbf{V}_{q-a,q} \\ \vdots & & \vdots \\ \mathbf{V}_{q,q-a} & \cdots & \mathbf{V}_{qq} \end{bmatrix},$$

$$(4.68)\qquad \mathbf{Y}_{a,i} = \begin{bmatrix} \mathbf{Y}_{1,i} \\ \vdots \\ \mathbf{Y}_{a+1,i} \end{bmatrix}, \quad a \in \{0, \ldots, q-1\}$$

$$(4.69)\qquad \mathbf{E}_a = [\mathbf{E}_{a+2,q-a} \cdots \mathbf{E}_{a+2,q}], \quad \text{and}$$

$$(4.70)\qquad \Theta_a = [\Theta_{a+2,q-a} \cdots \Theta_{a+2,q}], \quad a \in \{0, \ldots, q-2\}.$$

Algorithm 4.5 (Schall)

ML-estimation in the adjusted model (4.65)

Step 0: Set $a = 0$

Step 1: (i) Estimate $\mathbf{X}_a\mathbf{B}_{q-a}$ by

$$(4.71)\qquad \mathbf{X}_a\tilde{\mathbf{B}}_{q-a} = \mathbf{M}_a\mathbf{Y}_{a,q-a}$$

(ii) Estimate $\mathbf{V}_{q-a,i}$ $(i = 1, \ldots, q-a)$ by

$$(4.72)\qquad m \cdot \tilde{\mathbf{V}}_{q-a,i} = \mathbf{Y}_{a,q-a}'\mathbf{N}_a\mathbf{Y}_{a,i}$$

(iii) Estimate Θ_a by

$$(4.73)\qquad \tilde{\Theta}_a = \tilde{\mathbf{E}}_a\tilde{\mathbf{V}}_a^{-1}.$$

Step 2: Set

(4.74)
$$\mathbf{Y}_{a+2,i} = \mathbf{Y}_{a+2,i} - \tilde{\Theta}_a \begin{bmatrix} \tilde{\mathbf{V}}_{q-a,i} \\ \vdots \\ \tilde{\mathbf{V}}_{qi} \end{bmatrix}, \quad i = 1, \ldots, q-a-1$$

Step 3: Set $a = a+1$ and proceed with step 1.

□

The algorithm stops in step 1 (ii) when $a = q-1$.

Using Algorithm 4.5 we obtain a biased estimate $\tilde{\mathbf{V}}$ for $\mathbf{V}$, and if $q > 2$ the estimates $\mathbf{X}\tilde{\mathbf{B}}_j$ for $\mathbf{X}\mathbf{B}_j$ $(j = 1, \ldots, q-2)$ are biased. If we wish to correct for the bias, then the number m in step 1 (ii) must be replaced by

(4.75)
$$c_a = \sum_{i=1}^{a+1} k_i \quad .$$

Further, for $a \geq 1$, $\mathbf{V}_{q-a,j}$ $(j = 1, \ldots, q-a)$ is estimated by

(4.76)
$$c_a \cdot \tilde{\mathbf{V}}_{q-a,j} = \mathbf{Y}'_{a,q-a}\mathbf{N}_a\mathbf{Y}_{a,j} + \sum_{i=1}^{a} k_{i+1} [\tilde{\mathbf{V}}_{q-a,q-a+i} \cdots \tilde{\mathbf{V}}_{q-a,q}]\tilde{\mathbf{V}}_{a-i}^{-1} \begin{bmatrix} \tilde{\mathbf{V}}_{q-a+i,j} \\ \vdots \\ \tilde{\mathbf{V}}_{q,j} \end{bmatrix}$$

in step 1 (ii).

Each time the algorithm enters step 1, a row of blocks or submatrices of the parameter matrix Θ is estimated, that is, its nonzero submatrices $\Theta_a = [\Theta_{a+2,q-a} : \cdots : \Theta_{a+2,q}]$. Thereafter, the corresponding row of blocks of the data matrix $\mathbf{Y}$ is adjusted in step 2, and the algorithm proceeds as if the adjusted $\mathbf{Y}_{a+2,1}, \ldots, \mathbf{Y}_{a+2,q-a-1}$ had been observed. In this manner, the parameter matrix Θ is estimated row by row, starting from the top. For $q = 2$ we obtain the procedure outlined in Section 4.2.1. The adjustment for bias in (4.75) and (4.76) generalizes the adjustment (for $q = 2$) in (4.33).

Outliers by additive shifts

The model (4.1) adjusted for A-outliers occurring in a nested pattern is given by

(4.77)
$$\begin{bmatrix} \mathbf{Y}_{11} \cdots \mathbf{Y}_{1q} \\ \vdots \quad\quad \vdots \\ \mathbf{Y}_{q1} \cdots \mathbf{Y}_{qq} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_q \end{bmatrix} [\mathbf{B}_1 : \cdots : \mathbf{B}_q] + \begin{bmatrix} \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & & \Theta_{2q} \\ & & \vdots \\ \mathbf{0}\ \Theta_{q2} & \cdots & \Theta_{qq} \end{bmatrix} + \begin{bmatrix} \mathbf{E}_{11} \cdots \mathbf{E}_{1q} \\ \vdots \quad\quad \vdots \\ \mathbf{E}_{q1} \cdots \mathbf{E}_{qq} \end{bmatrix} .$$

Anderson (1957) and Rubin (1974) treated a model of this type in the context of missing data estimation where the set of submatrices (4.63) of the data matrix $\mathbf{Y}$ represents missing data points rather than A-outliers. But the problem is equivalent to the A-outlier problem, as far as ML-estimation of the unknowns in model (4.77) is concerned, as we have noted previously.

Let

(4.78) $$\mathbf{Y}_i = \begin{bmatrix} \mathbf{Y}_{1i} \\ \vdots \\ \mathbf{Y}_{qi} \end{bmatrix}, \ \mathbf{E}_i = \begin{bmatrix} \mathbf{E}_{1i} \\ \vdots \\ \mathbf{E}_{qi} \end{bmatrix}, \ i \in \{1, \ldots, q\}$$

(4.79) $$\mathbf{V}_a = \begin{bmatrix} \mathbf{V}_{11} & \cdots & \mathbf{V}_{1a} \\ \vdots & & \vdots \\ \mathbf{V}_{a1} & \cdots & \mathbf{V}_{aa} \end{bmatrix},$$

(4.80) $$\mathbf{N}_a = \mathbf{I} - \begin{bmatrix} \mathbf{X}_{q-a+1} \\ \vdots \\ \mathbf{X}_q \end{bmatrix} (\mathbf{X}'\mathbf{X})^{-}[\mathbf{X}_{q-a+1} : \cdots : \mathbf{X}_q], \quad \text{and}$$

(4.81) $$\Theta_a = \begin{bmatrix} \Theta_{q-a+1,a+1} \\ \vdots \\ \Theta_{q,a+1} \end{bmatrix}, \quad a \in \{1, \ldots, q-1\}$$

Then ML-estimation in model (4.77) is performed by

Algorithm 4.6 (Anderson, 1957)

ML-estimation in the adjusted model (4.77)

Step 0: Set $a = 1$

Step 1: (i) Estimate $\mathbf{X}\mathbf{B}_a$ by

(4.82) $$\mathbf{X}\tilde{\mathbf{B}}_a = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}_a$$

(ii) Estimate $\mathbf{V}_{i,a}$ $(i = 1, \ldots, a)$ by

(4.83) $$m \cdot \tilde{\mathbf{V}}_{i,a} = (\mathbf{Y}_i - \mathbf{X}\tilde{\mathbf{B}}_i)'(\mathbf{Y}_a - \mathbf{X}\tilde{\mathbf{B}}_a)$$

(iii) Estimate Θ_a by

(4.84) $$\tilde{\Theta}_a = (\mathbf{N}_a - \tilde{\mathbf{A}}\tilde{\mathbf{V}}_a^{-1}\tilde{\mathbf{A}}')^{-1}(\hat{\mathbf{C}} - \tilde{\mathbf{A}}\tilde{\mathbf{V}}_a^{-1}[\tilde{\mathbf{E}}_1 : \cdots : \tilde{\mathbf{E}}_a]'\hat{\mathbf{E}}_{a+1} \div m), \text{ where}$$

$$\tilde{\mathbf{A}} = \begin{bmatrix} \tilde{\mathbf{E}}_{q-a+1,1} & \cdots & \tilde{\mathbf{E}}_{q-a+1,a} \\ \vdots & & \vdots \\ \tilde{\mathbf{E}}_{q1} & \cdots & \tilde{\mathbf{E}}_{q,a} \end{bmatrix},$$

$$\hat{\mathbf{C}} = \begin{bmatrix} \hat{\mathbf{E}}_{q-a+1,a+1} \\ \vdots \\ \hat{\mathbf{E}}_{q,a+1} \end{bmatrix} ,$$

and $\hat{\mathbf{E}}$ are the estimated residuals from model (4.1).

Step 2: Set

(4.85) $$\mathbf{Y}_{i,a+1} = \mathbf{Y}_{i,a+1} - \tilde{\Theta}_{i,a+1} , \quad i = q-a+1, \ldots, q .$$

Step 3: Set $a = a+1$ and proceed with step 1.

□

The algorithm stops in step 1 (ii) when $a = q$. Each time the algorithm enters step 1, a column of blocks of the parameter matrix Θ is estimated, that is, its nonzero submatrices $\Theta_a = [\Theta'_{q-a+1,a+1} : \cdots : \Theta'_{q,a+1}]'$. Thereafter, the corresponding column of blocks of the data matrix $\mathbf{Y}$ is adjusted in step 2, and the algorithm proceeds as if the adjusted $\mathbf{Y}_{q-a+1,a+1}, \ldots, \mathbf{Y}_{q,a+1}$ had been observed. In this manner, Θ is estimated column by column, starting from the left.

Transformational Outliers

Finally, the adjusted model for T-outliers or outliers in PC's occurring in a nested pattern is

(4.86) $$\begin{bmatrix} \mathbf{Y}_{11} \cdots \mathbf{Y}_{1q} \\ \vdots \quad\quad \vdots \\ \mathbf{Y}_{q1} \cdots \mathbf{Y}_{qq} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_q \end{bmatrix} [\mathbf{B}_1 \cdots \mathbf{B}_q] + \begin{bmatrix} \mathbf{0} & \cdots \mathbf{0} \\ \vdots & \Theta_{2q} \\ & \vdots \\ \mathbf{0}\ \Theta_{q2} & \cdots \Theta_{qq} \end{bmatrix} \begin{bmatrix} \mathbf{P}'_1 \\ \vdots \\ \mathbf{P}'_q \end{bmatrix} + \begin{bmatrix} \mathbf{E}_{11} \cdots \mathbf{E}_{1q} \\ \vdots \quad\quad \vdots \\ \mathbf{E}_{q1} \cdots \mathbf{E}_{qq} \end{bmatrix}$$

where

(4.87) $$\mathbf{V} = [\mathbf{P}_1 : \cdots : \mathbf{P}_q] \begin{bmatrix} \Delta_1 & & \\ & \ddots & \\ & & \Delta_q \end{bmatrix} \begin{bmatrix} \mathbf{P}'_1 \\ \vdots \\ \mathbf{P}'_q \end{bmatrix}$$

is the SVD of $\mathbf{V}$.

Let

(4.88) $$\mathbf{S}_a = \hat{\mathbf{E}}'_a \hat{\mathbf{E}}_a , \quad a \in \{1, \ldots, q\}$$

where $\hat{\mathbf{E}}_a$ is the matrix of estimated residuals in the model

(4.89) $$\begin{bmatrix} \mathbf{Y}_{11} \cdots \mathbf{Y}_{1q} \\ \vdots \quad\quad \vdots \\ \mathbf{Y}_{a1} \cdots \mathbf{Y}_{aq} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_a \end{bmatrix} [\mathbf{B}_1 \cdots \mathbf{B}_q] + \begin{bmatrix} \mathbf{E}_{11} \cdots \mathbf{E}_{1q} \\ \vdots \quad\quad \vdots \\ \mathbf{E}_{a1} \cdots \mathbf{E}_{aq} \end{bmatrix}$$

The ML-estimate $\tilde{\mathbf{V}}$ of $\mathbf{V}$ under model (4.86) is computed by

Algorithm 4.7 (Schall)

PCA of the adjusted model (4.86)

Step 0: (i) Set $a = 1$

(ii) Compute the SVD $\mathbf{S}_1 = \mathbf{R}_1\mathbf{D}_1\mathbf{R}_1' = \mathbf{V}_1$ of $\mathbf{S}_1$.

Step 1: Choose the l_{q-a+1} outlying principal axes $\tilde{\mathbf{P}}_{q-a+1}$ from $\mathbf{R}_a$ to obtain a partitioning of $\mathbf{R}_a$ and $\mathbf{D}_a$ as

(4.90) $$\mathbf{R}_a = [\mathbf{R}_{a+1} : \tilde{\mathbf{P}}_{q-a+1}] \text{ and } \mathbf{D}_a = \begin{bmatrix} \mathbf{D}_{a+1} & \\ & \tilde{\Delta}_{q-a+1} \end{bmatrix}$$

respectively, resulting in a partitioning of the SVD of $\mathbf{V}_a$ as

(4.91) $$\mathbf{V}_a = [\mathbf{R}_{a+1} : \tilde{\mathbf{P}}_{q-a+1} : \cdots : \tilde{\mathbf{P}}_q] \begin{bmatrix} \mathbf{D}_{a+1} & & & \\ & \tilde{\Delta}_{q-a+1} & & \\ & & \ddots & \\ & & & \tilde{\Delta}_q \end{bmatrix} \begin{bmatrix} \mathbf{R}_{a+1}' \\ \tilde{\mathbf{P}}_{q-a+1}' \\ \vdots \\ \tilde{\mathbf{P}}_q' \end{bmatrix}$$

Step 2: Compute the SVD

(4.92) $$\mathbf{D}_{a+1} + \mathbf{R}_{a+1}'(\mathbf{S}_{a+1} - \mathbf{S}_a)\mathbf{R}_{a+1} = \mathbf{Q}_a\mathbf{D}_a^*\mathbf{Q}_a'$$

of $\mathbf{D}_{a+1} + \mathbf{R}_{a+1}'(\mathbf{S}_{a+1} - \mathbf{S}_a)\mathbf{R}_{a+1}$.

Step 3: Set

(4.93) $$\mathbf{R}_{a+1} = \mathbf{R}_{a+1}\mathbf{Q}_a \text{ and } \mathbf{D}_{a+1} = \mathbf{D}_a^* \quad .$$

Step 4: Set $a = a+1$ and proceed with step 1.

□

The algorithm stops in step 1 when $a = q$, thus $\mathbf{R}_a = \mathbf{P}_1$ and $\mathbf{D}_a = \tilde{\Delta}_1$ and $\mathbf{R}_{a+1}$ and $\mathbf{D}_{a+1}$ vanish. The ML-estimate $\tilde{\mathbf{V}}$ of $\mathbf{V}$ under model (4.86) is then given by

(4.94) $$m \cdot \tilde{\mathbf{V}} = \mathbf{V}_q = [\tilde{\mathbf{P}}_1 : \cdots : \tilde{\mathbf{P}}_q] \begin{bmatrix} \tilde{\Delta}_1 & & \\ & \ddots & \\ & & \tilde{\Delta}_q \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{P}}_1' \\ \vdots \\ \tilde{\mathbf{P}}_q' \end{bmatrix}$$

where $\mathbf{V}_q$ is obtained from (4.91). This estimate for $\mathbf{V}$ is biased, and an estimate $\tilde{\tilde{\mathbf{V}}}$ for $\mathbf{V}$ which is corrected for the bias is obtained by

(4.95) $$\tilde{\tilde{\mathbf{V}}} = [\tilde{\mathbf{P}}_1 : \cdots : \tilde{\mathbf{P}}_q] \begin{bmatrix} \tilde{\tilde{\Delta}}_1 & & \\ & \ddots & \\ & & \tilde{\tilde{\Delta}}_q \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{P}}_1' \\ \vdots \\ \tilde{\mathbf{P}}_q' \end{bmatrix}$$

where $\tilde{\tilde{\Delta}}_i = (m \div \sum_{j=1}^{q-i+1} k_j) \cdot \tilde{\Delta}_i$. Now the other unknowns in model (4.86) are estimated by

(4.96) $$\begin{bmatrix} \tilde{\Theta}_{q-i+1,i+1} \\ \vdots \\ \tilde{\Theta}_{q,i+1} \end{bmatrix} = \mathbf{N}_i^{-1} \begin{bmatrix} \hat{\mathbf{E}}_{q-i+1,1} & \vdots \cdots \vdots & \hat{\mathbf{E}}_{q-i+1,q} \\ \vdots & & \vdots \\ \hat{\mathbf{E}}_{q,1} & \vdots \cdots \vdots & \hat{\mathbf{E}}_{qq} \end{bmatrix} \tilde{\mathbf{P}}_{i+1} \ , \ i = 1, \ldots, q-1 \ .$$

where the matrix $\mathbf{N}_i$ is the trailing principal $(\sum_{j=q-i+1}^{q} k_j) \times (\sum_{j=q-i+1}^{q} k_j)$-submatrix of $\mathbf{N} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$, and $\mathbf{XB}$ is estimated by

(4.97) $$\mathbf{X}\tilde{\mathbf{B}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'(\mathbf{Y} - \begin{bmatrix} \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & & \tilde{\Theta}_{2q} \\ & & \vdots \\ \mathbf{0} & \tilde{\Theta}_{q2} \cdots & \tilde{\Theta}_{qq} \end{bmatrix} \tilde{\mathbf{v}}') \ .$$

4.2.5 Maximum-likelihood estimation using the EM-algorithm

In the general case, when outliers occur in an arbitrary pattern, ML-estimation of the parameters in the adjusted models (4.6), (4.7) and (4.9) is not possible in closed form.

For the case of A-outliers, which we noted is equivalent to classical missing data estimation, there exist several well-known techniques to estimate the parameters in model (4.7), by iterative algorithms such as the EM-algorithm, Fisher-scoring and the Newton-Raphson method. Dempster, Laird and Rubin (1977) give an extensive treatment of the EM-algorithm, and the respective merits of the EM-, scoring-, and Newton-Raphson methods are discussed. Wu (1983) corrects an error in the EM theory and obtains several convergence results for the EM-algorithm.

In the following we present the corresponding versions of the EM-algorithm to perform the ML-estimation of the parameters in the adjusted models (4.6), (4.7) and (4.9).

Similar to the notation of Orchard and Woodbury (1972), let each observational vector $\mathbf{y}_i$, that is, each row vector of $\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_m \end{bmatrix}$ be written as

(4.98) $$\mathbf{y}_i = \mathbf{y}_{i,c} + \mathbf{y}_{i,o} \ , \quad i = 1, \ldots, m$$

where $\mathbf{y}_{i,c}$ is the portion of clean data points of $\mathbf{y}_i$, with zero in each position corresponding to an outlying data point, and $\mathbf{y}_{i,o}$ is the outlying portion, with zero in the positions corresponding to the clean data points of $\mathbf{y}_i$. These partitionings of the observational vectors may vary from observation to observation, depending on which components, if any, are outlying in $\mathbf{y}_i$.

Conformably with (4.98), the means of $\mathbf{y}_{i,c}$ and $\mathbf{y}_{i,o}$ are respectively denoted by $\mathbf{x}_i\mathbf{B}_c$ and $\mathbf{x}_i\mathbf{B}_o$, where $\mathbf{x}_i$ is the i-th row of $\mathbf{X}$. Finally, the variance-covariance matrix $\mathbf{V}$ is partitioned for each i as

(4.99) $$\mathbf{V} = \mathbf{V}^{(i)}_{c,c} + \mathbf{V}^{(i)}_{o,c} + \mathbf{V}^{(i)}_{c,o} + \mathbf{V}^{(i)}_{o,o} \ , \quad i = 1, \ldots, m$$

corresponding to the partitioning of $\mathbf{y}_i$ as in (4.98), that is, $\mathbf{V}^{(i)}_{c,c}$ contains the variances and covariances of the clean data points of $\mathbf{y}_i$, with zero in the positions corresponding to outlying data points, and similarly $\mathbf{V}^{(i)}_{o,c}$, $\mathbf{V}^{(i)}_{c,o}$ and $\mathbf{V}^{(i)}_{o,o}$. The inverse $(\mathbf{V}^{(i)}_{c,c} + \mathbf{V}^{(i)}_{o,o})^{-1}$ is partitioned in the same manner as

(4.100) $$(\mathbf{V}^{(i)}_{c,c} + \mathbf{V}^{(i)}_{o,o})^{-1} = (\mathbf{V}^{(i)}_{c,c})^{-1} + (\mathbf{V}^{(i)}_{o,o})^{-1} \ .$$

For the model (4.7), that is, for the model adjusted for A-outliers occurring in an arbitrary pattern, ML-estimation of the parameters is performed by

Algorithm 4.8 (Orchard and Woodbury, 1972)

ML-estimation in the adjusted model (4.7)

Step 0: Set $\tilde{\Theta} = \mathbf{0}$, and
$\tilde{\mathbf{V}} = \mathbf{I}$ (say)

Step 1: (i) Estimate $\mathbf{XB}$ by

(4.101) $$\mathbf{X}\tilde{\mathbf{B}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}' \ (\mathbf{Y} - \tilde{\Theta})$$

(ii) Estimate $\mathbf{V}$ by

(4.102) $$\tilde{\mathbf{V}} = [(\mathbf{Y} - \mathbf{X}\tilde{\mathbf{B}} - \tilde{\Theta})'(\mathbf{Y} - \mathbf{X}\tilde{\mathbf{B}} - \tilde{\Theta}) + \sum_{i=1}^{m} (\tilde{\mathbf{V}}^{(i)}_{o,o} - \tilde{\mathbf{V}}^{(i)}_{o,c}(\tilde{\mathbf{V}}^{(i)}_{c,c})^{-1}\tilde{\mathbf{V}}^{(i)}_{c,o})] \div m$$

(iii) Estimate $\Theta = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_m \end{bmatrix}$ by

(4.103) $$\tilde{\theta}_i = \mathbf{y}_{i,o} - \mathbf{x}_i\tilde{\mathbf{B}}^{(i)}_o - (\mathbf{y}_{i,c} - \mathbf{x}_i\tilde{\mathbf{B}}^{(i)}_c) \ (\tilde{\mathbf{V}}^{(i)}_{c,c})^{-1} \ \tilde{\mathbf{V}}^{(i)}_{c,o} \ , \quad i = 1, \ldots, m \ .$$

Step 2: Proceed with step 1 until the parameter estimates stabilize.

□

Note that by adding the term $\sum_{i=1}^{m} (\tilde{\mathbf{V}}^{(i)}_{o,o} - \tilde{\mathbf{V}}^{(i)}_{o,c} \ (\tilde{\mathbf{V}}^{(i)}_{c,c})^{-1} \ \tilde{\mathbf{V}}^{(i)}_{c,o}) \div m$ in step 1 (ii) while estimating $\mathbf{V}$ the estimate $\tilde{\mathbf{V}}$ is adjusted for bias.

In the model (4.6), which is the model adjusted for D-outliers, the appropriate version of the EM-algorithm is

Algorithm 4.9 (Schall)

ML-estimation in the adjusted model (4.6)

Step 0: Set $\tilde{\Theta} = \mathbf{0}$, and
$\tilde{\mathbf{V}} = \mathbf{I}$ (say)

Step 1: (i) Estimate $\mathbf{XB}$ by

(4.104) $$\mathbf{X}\tilde{\mathbf{B}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\ (\mathbf{Y} - \tilde{\Theta}\tilde{\mathbf{V}})$$

(ii) Estimate $\mathbf{V}$ by

(4.105) $$\tilde{\mathbf{V}} = [(\mathbf{Y} - \mathbf{X}\tilde{\mathbf{B}} - \tilde{\Theta}\tilde{\mathbf{V}})'(\mathbf{Y} - \mathbf{X}\tilde{\mathbf{B}} - \tilde{\Theta}\tilde{\mathbf{V}}) + \sum_{i=1}^{m} (\tilde{\mathbf{V}}^{(i)}_{o,o} + \tilde{\mathbf{V}}^{(i)}_{o,c} + \tilde{\mathbf{V}}^{(i)}_{c,o} + \tilde{\mathbf{V}}^{(i)}_{c,o}(\tilde{\mathbf{V}}^{(i)}_{o,o})^{-1}\tilde{\mathbf{V}}^{(i)}_{o,c})] \div m$$

(iii) Estimate $\Theta = \begin{bmatrix} \theta_1 \\ : \\ \theta_m \end{bmatrix}$ by

(4.106) $$\tilde{\theta}_i = (\mathbf{y}_{i,o} - \mathbf{x}_i\tilde{\mathbf{B}}^{(i)}_o)\ (\tilde{\mathbf{V}}^{(i)}_{o,o})^{-1} \quad , \quad i = 1, \ldots, m\ .$$

Step 2: Proceed with step 1 until the parameter estimates stabilize.

□

Again, the term $\sum_{i=1}^{m} (\tilde{\mathbf{V}}^{(i)}_{o,o} + \tilde{\mathbf{V}}^{(i)}_{o,c} + \tilde{\mathbf{V}}^{(i)}_{c,o} + \tilde{\mathbf{V}}^{(i)}_{c,o}\ (\tilde{\mathbf{V}}^{(i)}_{o,o})^{-1}\ \tilde{\mathbf{V}}^{(i)}_{o,c}) \div m$ is added in step 1 (ii) to correct for bias.

Finally, we give the EM-algorithm for the model (4.9), the model adjusted for T-outliers. The application of the EM-algorithm will usually be preceeded by a PCA of the data under the original model (4.1), to obtain preliminary estimates $\hat{\mathbf{Y}}^* = \mathbf{Y}\hat{\mathbf{P}}$ of the PC's $\mathbf{Y}^*$ of $\mathbf{Y}$, and to specify the outlying PC's after an inspection of $\hat{\mathbf{Y}}^*$. Each row vector $\mathbf{y}^*_i$ of $\mathbf{Y}^*$ will then be partitioned similarly to (4.98) into an outlying and a clean portion. Accordingly, the SVD $\mathbf{V} = \mathbf{P}\triangle\mathbf{P}'$ of $\mathbf{V}$ is partitioned as

(4.107) $$\mathbf{V} = \mathbf{P}^{(i)}_c\triangle^{(i)}_c\mathbf{P}^{(i)\prime}_c + \mathbf{P}^{(i)}_o\triangle^{(i)}_o\mathbf{P}^{(i)\prime}_o, \quad i = 1, .., m$$

conformably with the partitioning of $\mathbf{y}^*_i = \mathbf{y}^*_{i,c} + \mathbf{y}^*_{i,o}$.

Algorithm 4.10 (Schall)

ML-estimation in the adjusted model (4.9)

Step 0: Set $\tilde{\Theta} = \mathbf{0}$, and
$\tilde{\mathbf{V}} = \tilde{\mathbf{P}}\tilde{\triangle}\tilde{\mathbf{P}}' = \mathbf{I} \cdot \mathbf{I} \cdot \mathbf{I}$.

Step 1: (i) Estimate $\mathbf{XB}$ by

(4.108) $$\mathbf{X}\tilde{\mathbf{B}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\ (\mathbf{Y} - \tilde{\Theta}\tilde{\mathbf{P}}')$$

(ii) Estimate $\mathbf{V}$ by

(4.109) $$\tilde{\mathbf{V}} = [(\mathbf{Y} - \mathbf{X}\tilde{\mathbf{B}} - \tilde{\Theta}\tilde{\mathbf{P}})'(\mathbf{Y} - \mathbf{X}\tilde{\mathbf{B}} - \tilde{\Theta}\tilde{\mathbf{P}}') + \sum_{i=1}^{m}\tilde{\mathbf{P}}_o^{(i)}\tilde{\Delta}_o^{(i)}\tilde{\mathbf{P}}_o^{(i)\prime}] \div m$$

(iii) Compute the SVD $\tilde{\mathbf{V}} = \tilde{\mathbf{P}}\ \tilde{\Delta}\ \tilde{\mathbf{P}}'$ of $\tilde{\mathbf{V}}$.

(iv) Estimate $\Theta = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_m \end{bmatrix}$ by

(4.110) $$\tilde{\theta}_i = (\mathbf{y}_i - \mathbf{x}_i\tilde{\mathbf{B}})\ \tilde{\mathbf{P}}_o^{(i)}, \quad i = 1, \ldots, m\ .$$

Step 2: Proceed with step 1 until the parameter estimates stabilize.

□

Clearly, once the ML-estimate $\tilde{\mathbf{V}}$ for $\mathbf{V}$ in the corresponding adjusted model (4.6), (4.7) or (4.9) is obtained, the LRT-statistic for testing the hypothesis $H_0 : \Theta = \mathbf{0}$ is given by $\chi^2 = m \cdot \ln(|\hat{\mathbf{V}}|/|\tilde{\mathbf{V}}|)$, which is asymptotically distributed chi-squared with degrees of freedom equal to the number of outliers specified.

4.3 OUTLIERS AND INFLUENCE IN THE GROWTH CURVE MODEL

4.3.1 The growth curve model

The generalized multivariate regression model (growth curve model), which was first introduced by Potthoff and Roy (1964), can be written in the form

(4.111) $$\underset{(m\times n)}{\mathbf{Y}} = \underset{(m\times p)}{\mathbf{X}} \cdot \underset{(p\times a)}{\mathbf{B}} \cdot \underset{(a\times n)}{\mathbf{A}} + \underset{(m\times n)}{\mathbf{E}}\ ,$$

where $\mathbf{Y}$ is the matrix of observations, $\mathbf{X}$ and $\mathbf{A}$ are known matrices, $\mathbf{B}$ is the matrix of parameters and $\mathbf{E}$ is an unobservable matrix of error components. As in the multivariate regression model (4.1), the rows of $\mathbf{E}$ are assumed to be independent and identically distributed as $N_n(\mathbf{0},\mathbf{V})$, where $\mathbf{V}$ is an arbitrary but positive definite and symmetric matrix.

For $\mathbf{A} = \mathbf{I}_n$ we obtain the model (4.1), and thus the multivariate regression model is a special case of the growth curve model (4.111). The model (4.111) can alternatively be written in the form

(4.111a) $$\mathrm{vec}(\mathbf{Y}') = (\mathbf{X} \otimes \mathbf{A}')\mathrm{vec}(\mathbf{B}') + \mathrm{vec}(\mathbf{E}')$$

$$\Leftrightarrow\ \mathbf{y} = (\mathbf{X} \otimes \mathbf{A}')\beta + \mathbf{e}\ ;\ \ \mathrm{cov}(\mathbf{e}) = \Sigma = \mathbf{I}_m \otimes \mathbf{V}\ .$$

The model (4.111a), and thus the model (4.111), is a special case of the general linear model (2.1), with a design matrix of the form $\Xi = (\mathbf{X} \otimes \mathbf{A}')$ and a variance-covariance matrix $\Sigma = \mathbf{I}_m \otimes \mathbf{V}$ in block diagonal form. However, $\mathbf{V}$ is generally assumed to be unknown.

Khatri (1966) showed that the MLE $\mathbf{X}\hat{\mathbf{B}}\mathbf{A}$ for $\mathbf{XBA}$ under model (4.111) is given by

$$\mathbf{X}\hat{\mathbf{B}}\mathbf{A} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}\mathbf{S}^{-1}\mathbf{A}'(\mathbf{A}\mathbf{S}^{-1}\mathbf{A}')^{-}\mathbf{A} \tag{4.112}$$

where $\mathbf{S} = \mathbf{Y}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}')\mathbf{Y}$, and with

$$\hat{\mathbf{E}} = \mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}\mathbf{A} \tag{4.113}$$

the MLE $\hat{\mathbf{V}}$ for $\mathbf{V}$ is given by

$$\begin{aligned} m \cdot \hat{\mathbf{V}} &= \hat{\mathbf{E}}'\hat{\mathbf{E}} \\ &= \mathbf{A}'(\mathbf{A}\mathbf{S}^{-1}\mathbf{A}')^{-}\mathbf{A} \quad . \end{aligned} \tag{4.114}$$

Testing problems in the growth curve model, including some missing data problems, are treated by Kariya (1985).

4.3.2 Outliers in the growth curve model

Since the model (4.111) is also a special case of the general linear model (2.1), we can directly extend the methods to adjust the general linear model for different types of outlier to the multivariate general linear model (4.111), in the manner of Section 4.2 .

The model (4.111) respectively adjusted for D-, A- and T-outliers is given by

$$\mathbf{Y} = \mathbf{XBA} + \Theta\mathbf{V} + \mathbf{E} \ , \tag{4.115}$$

$$\mathbf{Y} = \mathbf{XBA} + \Theta + \mathbf{E} \ , \quad \text{and} \tag{4.116}$$

$$\mathbf{Y} = \mathbf{XBA} + \Theta\mathbf{P}' + \mathbf{E} \ , \tag{4.117}$$

where the parameter matrix $\Theta = (\theta_{ij})$ is of the order $m \times n$, and θ_{ij} is *a-priori* specified to be zero if the observation y_{ij} or the PC y^*_{ij} respectively are not in the subset of data points or PC's suspected to be outlying.

Unlike the situation in the multivariate regression model (4.1), the parameters in the adjusted models (4.115) through (4.117) can not be estimated in closed form even for simple outlier patterns. Thus iterative methods like the EM-algorithm must be used to perform maximum likelihood estimation. Algorithms 4.8 through 4.10 are easily generalized for the generalized multivariate regression model which is adjusted for outliers.

Algorithm 4.9.1

ML-estimation in the adjusted model (4.115)

Step 0: Set $\tilde{\Theta} = \mathbf{0}$, and

$$\tilde{\mathbf{V}} = \mathbf{I} \quad \text{(say)}$$

Step 1: (i) Estimate $\mathbf{XBA}$ by

(4.118) $$\mathbf{X}\tilde{\mathbf{B}}\mathbf{A} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'(\mathbf{Y} - \tilde{\Theta}\tilde{\mathbf{V}})\tilde{\mathbf{V}}^{-1}\mathbf{A}'(\mathbf{A}\tilde{\mathbf{V}}^{-1}\mathbf{A}')^{-}\mathbf{A}$$

(ii) Estimate $\mathbf{V}$ by

(4.119) $$\tilde{\mathbf{V}} = [(\mathbf{Y} - \mathbf{X}\tilde{\mathbf{B}}\mathbf{A} - \tilde{\Theta}\tilde{\mathbf{V}})'(\mathbf{Y} - \mathbf{X}\tilde{\mathbf{B}}\mathbf{A} - \tilde{\Theta}\tilde{\mathbf{V}}) + \sum_{i=1}^{m} (\tilde{\mathbf{V}}_{o,o}^{(i)} + \tilde{\mathbf{V}}_{o,c}^{(i)} + \tilde{\mathbf{V}}_{c,o}^{(i)} + \tilde{\mathbf{V}}_{c,o}^{(i)}(\tilde{\mathbf{V}}_{o,o}^{(i)})^{-1}\tilde{\mathbf{V}}_{o,c}^{(i)})] \div m$$

(iii) Estimate $\Theta = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_m \end{bmatrix}$ by

(4.120) $$\tilde{\theta}_i = (\mathbf{y}_{i,o} - \mathbf{x}_i\tilde{\mathbf{B}}\mathbf{A}_o^{(i)})\,(\tilde{\mathbf{V}}_{o,o}^{(i)})^{-1} \quad , \quad i = 1, \ldots, m \, .$$

Step 2: Proceed with step 1 until the parameter estimates stabilize.

□

Algorithm 4.8.1

ML-estimation in the adjusted model (4.116)

Step 0: Set $\tilde{\Theta} = \mathbf{0}$, and

$$\tilde{\mathbf{V}} = \mathbf{I} \quad \text{(say)}$$

Step 1: (i) Estimate $\mathbf{XBA}$ by

(4.121) $$\mathbf{X}\tilde{\mathbf{B}}\mathbf{A} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'(\mathbf{Y} - \tilde{\Theta})\tilde{\mathbf{V}}^{-1}\mathbf{A}'(\mathbf{A}\tilde{\mathbf{V}}^{-1}\mathbf{A}')^{-}\mathbf{A}$$

(ii) Estimate $\tilde{\mathbf{V}}$ by

(4.122) $$\tilde{\mathbf{V}} = [(\mathbf{Y} - \mathbf{X}\tilde{\mathbf{B}}\mathbf{A} - \tilde{\Theta})'(\mathbf{Y} - \mathbf{X}\tilde{\mathbf{B}}\mathbf{A} - \tilde{\Theta}) + \sum_{i=1}^{m} (\tilde{\mathbf{V}}_{o,o}^{(i)} - \tilde{\mathbf{V}}_{o,c}^{(i)}\,(\tilde{\mathbf{V}}_{c,c}^{(i)})^{-1}\tilde{\mathbf{V}}_{c,o}^{(i)})] \div m$$

(iii) Estimate $\Theta = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_m \end{bmatrix}$ by

(4.123) $$\tilde{\theta}_i = \mathbf{y}_{i,o} - \mathbf{x}_i\tilde{\mathbf{B}}\mathbf{A}_o^{(i)} - (\mathbf{y}_{i,c} - \mathbf{x}_i\tilde{\mathbf{B}}\mathbf{A}_c^{(i)})(\tilde{\mathbf{V}}_{c,c}^{(i)})^{-1}\tilde{\mathbf{V}}_{c,o}^{(i)} \quad , \quad i = 1, \ldots, m \, .$$

Step 2: Proceed with step 1 until the parameter estimates stabilize.

□

Algorithm 4.10.1

ML-estimation in the adjusted model (4.117)

Step 0: Set $\tilde{\Theta} = \mathbf{0}$, and

$$\tilde{\mathbf{V}} = \tilde{\mathbf{P}}\tilde{\Delta}\tilde{\mathbf{P}}' = \mathbf{I} \cdot \mathbf{I} \cdot \mathbf{I} \quad \text{(say)}$$

Step 1: (i) Estimate **XBA** by

(4.124) $$\mathbf{X}\tilde{\mathbf{B}}\mathbf{A} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'(\mathbf{Y} - \tilde{\Theta}\tilde{\mathbf{P}}')\tilde{\mathbf{V}}^{-1}\mathbf{A}'(\mathbf{A}\tilde{\mathbf{V}}^{-1}\mathbf{A}')^{-}\mathbf{A}$$

(ii) Estimate **V** by

(4.125) $$\tilde{\mathbf{V}} = [(\mathbf{Y} - \mathbf{X}\tilde{\mathbf{B}}\mathbf{A} - \tilde{\Theta}\tilde{\mathbf{P}}')'(\mathbf{Y} - \mathbf{X}\tilde{\mathbf{B}}\mathbf{A} - \tilde{\Theta}\tilde{\mathbf{P}}') + \sum_{i=1}^{m} \tilde{\mathbf{P}}_o^{(i)} \tilde{\Delta}_o^{(i)} \tilde{\mathbf{P}}_o^{(i)\prime}] \div m$$

(iii) Compute the SVD $\tilde{\mathbf{V}} = \tilde{\mathbf{P}}\tilde{\Delta}\tilde{\mathbf{P}}'$ of $\tilde{\mathbf{V}}$.

(iv) Estimate $\Theta = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_m \end{bmatrix}$ by

(4.126) $$\theta_i = (\mathbf{y}_i - \mathbf{x}_i\tilde{\mathbf{B}}\mathbf{A})\tilde{\mathbf{P}}_o^{(i)} , \quad i = 1, \ldots, m .$$

Step 2: Proceed with step 1 until the parameter estimates stabilize.

□

In Algorithm 4.8.1 through 4.10.1 we use the notation of the previous section, where $\mathbf{y}_i$, **V**, **P** and Δ are decomposed into components corresponding to outlying and clean data points, as given in equations (4.98) through (4.100) and (4.107). Here the matrix **A** is similarly decomposed as

(4.127) $$\mathbf{A} = \mathbf{A}_c^{(i)} + \mathbf{A}_o^{(i)} , \quad i = 1, \ldots, m$$

corresponding to the decomposition of $\mathbf{y}_i = \mathbf{y}_{i,c} + \mathbf{y}_{i,o}$, that is, $\mathbf{x}_i\mathbf{B}\mathbf{A}_c^{(i)}$ is the mean of $\mathbf{y}_{i,c}$ and $\mathbf{x}_i\mathbf{B}\mathbf{A}_o^{(i)}$ is the mean of $\mathbf{y}_{i,o}$.

4.3.3 Influence in the growth curve model

The variance-covariance matrix Σ of the growth curve model (4.111) is known to have the block diagonal form $\Sigma = (\mathbf{I}_m \otimes \mathbf{V})$, but **V** is generally assumed to be unknown. To assess the influence of a set of data points J (say) in the growth curve model (4.111), we propose to use the statistics Cook's distance $C_{(J)}$ and the Andrews-Pregibon statistic $AP_{(J)}$, as

generalized in the previous chapter for the case where the variance-covariance structure of a linear model is arbitrary but known, replacing the unknown matrix $\mathbf{V}$ by its ML-estimate $\hat{\mathbf{V}}$ under the model (4.111). With $\hat{\Sigma} = (\mathbf{I}_m \otimes \hat{\mathbf{V}})$ the statistics $C_{(J)}$ and $AP_{(J)}$ can be computed for any subset J of the data points (y_{ij}), in three different versions each depending on whether we see J as a set of possible D-, A- or T-outliers.

Alternatively, $\mathbf{V}$ can be replaced by its ML-estimate $\tilde{\mathbf{V}}$ under the corresponding adjusted model (4.115), (4.116) or (4.117). An advantage would be that the estimate $\tilde{\mathbf{V}}$ is not contaminated by the possible outliers J. On the other hand, using $\hat{\mathbf{V}}$ involves far less computation, since $\tilde{\mathbf{V}}$ would have to be computed anew for all subsets of observations under investigation. Further, using the matrix $\hat{\mathbf{V}}$ facilitates the comparison of the corresponding values of $C_{(J)}$ or $AP_{(J)}$, when these statistics are computed for a variety of subsets J.

Once the unknown matrix $\mathbf{V}$ is replaced by its estimate $\hat{\mathbf{V}}$ (or alternatively $\tilde{\mathbf{V}}$), and henceforth treated as if it was known, the methods of the previous chapter can directly be applied, and essentially nothing new is involved when $C_{(J)}$ or $AP_{(J)}$ are computed for a given subset J of data points. However, if the rows and columns of the data matrix $\mathbf{Y}$ can be arranged in a way that the set J of data points forms a submatrix $\mathbf{Y}_{22}$ of $\mathbf{Y}$, then the statistics $C_{(J)} = C_{(22)}$ and $AP_{(J)} = AP_{(22)}$ can written in a particularly neat form.

Thus we assume that possibly after some rearrangement of the rows and columns of $\mathbf{Y}$, and a corresponding rearrangement of the rows of $\mathbf{X}$ and the columns of $\mathbf{A}$, the set J of data points whose influence should be assessed forms a submatrix $\mathbf{Y}_{22}$ of the data matrix $\mathbf{Y}$. Then the adjusted models (4.115) and (4.116) can respectively be written as

$$(4.128)\qquad \begin{bmatrix} \mathbf{Y}_{11} : \mathbf{Y}_{12} \\ \mathbf{Y}_{21} : \mathbf{Y}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \mathbf{B}[\mathbf{A}_1 : \mathbf{A}_2] + \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_k \end{bmatrix} \Theta[\mathbf{V}_{21} : \mathbf{V}_{22}] + \begin{bmatrix} \mathbf{E}_{11} : \mathbf{E}_{12} \\ \mathbf{E}_{21} : \mathbf{E}_{22} \end{bmatrix},$$

and

$$(4.129)\qquad \begin{bmatrix} \mathbf{Y}_{11} : \mathbf{Y}_{12} \\ \mathbf{Y}_{21} : \mathbf{Y}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \mathbf{B}[\mathbf{A}_1 : \mathbf{A}_2] + \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_k \end{bmatrix} \Theta[\mathbf{0} : \mathbf{I}_l] + \begin{bmatrix} \mathbf{E}_{11} : \mathbf{E}_{12} \\ \mathbf{E}_{21} : \mathbf{E}_{22} \end{bmatrix}.$$

All partitions are conformable, that is, with $\mathbf{Y}_{22}$ being of order $k \times l$ we have that $\mathbf{X}_2$ is of order $k \times p$, $\mathbf{A}_2$ is of order $a \times l$ and $\mathbf{V}_{22}$ is of order $l \times l$.

Alternatively, the models (4.128) and (4.129) can be written in the vectorized form

$$(4.128\text{a})\qquad \mathbf{y} = \left(\begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \otimes \begin{bmatrix} \mathbf{A}_1' \\ \mathbf{A}_2' \end{bmatrix} : \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_k \end{bmatrix} \otimes \begin{bmatrix} \mathbf{V}_{12} \\ \mathbf{V}_{22} \end{bmatrix} \right) \begin{bmatrix} \beta \\ \theta \end{bmatrix} + \mathbf{e}, \quad cov(\mathbf{e}) = (\mathbf{I}_m \otimes \mathbf{V}),$$

and

$$(4.129\text{a})\qquad \mathbf{y} = \left(\begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \otimes \begin{bmatrix} \mathbf{A}_1' \\ \mathbf{A}_2' \end{bmatrix} : \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_k \end{bmatrix} \otimes \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_l \end{bmatrix} \right) \begin{bmatrix} \beta \\ \theta \end{bmatrix} + \mathbf{e}, \quad cov(\mathbf{e}) = (\mathbf{I}_m \otimes \mathbf{V}).$$

Let

$$\text{(4.130)} \quad \mathbf{N} = \mathbf{I}_m \otimes \mathbf{V} - \mathbf{X}(\mathbf{X}'\mathbf{X})\mathbf{X}' \otimes \mathbf{A}'(\mathbf{A}\mathbf{V}^{-1}\mathbf{A}')^{-}\mathbf{A}$$

$$\mathbf{N}_{22} = \mathbf{I}_k \otimes \mathbf{V}_{22} - \mathbf{X}_2(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}_2' \otimes \mathbf{A}_2'(\mathbf{A}\mathbf{V}^{-1}\mathbf{A}')^{-}\mathbf{A}_2 , \quad \text{and}$$

$$\text{(4.131)} \quad \mathbf{M} = \mathbf{I}_m \otimes \mathbf{V}^{-1} - \mathbf{X}(\mathbf{X}'\mathbf{X})\mathbf{X}' \otimes \mathbf{V}^{-1}\mathbf{A}'(\mathbf{A}\mathbf{V}^{-1}\mathbf{A}')^{-}\mathbf{A}\mathbf{V}^{-1} ,$$

$$\mathbf{M}_2 = ([\mathbf{0} : \mathbf{I}_k] \otimes [\mathbf{0} : \mathbf{I}_l]\, \mathbf{M} \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{I}_k \end{bmatrix} \otimes \mathbf{I}_n \right) , \quad \text{and}$$

$$\mathbf{M}_{22} = \mathbf{M}_2 \left(\mathbf{I}_k \otimes \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_l \end{bmatrix} \right) .$$

Further, for an arbitrary matrix $\mathbf{C}$, let $vec'(\mathbf{C})$ denote the vector $(vec(\mathbf{C}))'$.

In the manner of the previous chapter, Q_{kl}^D and Q_{kl}^A denote respectively the outlier sum of squares due to the $k \cdot l$ D- or A-outliers $\mathbf{Y}_{22}$, and $C_{(22)}^D$, $C_{(22)}^A$, $AP_{(22)}^D$ and $AP_{(22)}^A$ are the corresponding versions of Cook's distance and the Andrews-Pregibon statistic. Then we have, using Theorem 3.1 and Theorem 3.2

$$\text{(4.132)} \quad Q_{kl}^D = vec'(\hat{\mathbf{E}}_{22}')\mathbf{N}_{22}^{-}\, vec(\hat{\mathbf{E}}_{22}') , \quad \text{and}$$

$$\text{(4.133)} \quad Q_{kl}^A = vec'(\hat{\mathbf{E}}_2')\mathbf{M}_2'\mathbf{M}_{22}^{-}\mathbf{M}_2\, vec(\hat{\mathbf{E}}_2')$$

under models (4.128) and (4.129) respectively, where $\mathbf{E}_2 = [\mathbf{E}_{21} : \mathbf{E}_{22}]$. Thus, using (3.86) through (3.88), we obtain $C_{(22)}^D$ and $C_{(22)}^A$ as

$$\text{(4.134)} \quad C_{(22)}^D = \frac{Q_{kl}^D}{vec'(\hat{\mathbf{E}}')\Sigma^{-1}vec(\hat{\mathbf{E}}')} \cdot \left(\frac{vec'(\hat{\mathbf{E}}_{22}')\mathbf{N}_{22}^{-}\Sigma_{22}\mathbf{N}_{22}^{-}\, vec(\hat{\mathbf{E}}_{22})}{Q_{kl}^D} - 1 \right) ,$$

and

$$\text{(4.135)} \quad C_{(22)}^A = \frac{Q_{kl}^A}{vec'(\hat{\mathbf{E}}')\Sigma^{-1}vec(\hat{\mathbf{E}}')} \cdot \left(\frac{vec'(\hat{\mathbf{E}}_2')\mathbf{M}_2'\mathbf{M}_{22}^{-}\Sigma^{22}\mathbf{M}_{22}^{-}\mathbf{M}_2\, vec(\hat{\mathbf{E}}_2')}{Q_{kl}^A} - 1 \right) ,$$

where $\Sigma_{22} = (\mathbf{I}_k \otimes \mathbf{V}_{22})$ and $\Sigma^{22} = (\Sigma^{-1})_{22}$. Similarly, using (3.106) and (3.107), $AP_{(22)}^D$ and $AP_{(22)}^A$ are given by

$$\text{(4.136)} \quad AP_{(22)}^D = \frac{vec'(\hat{\mathbf{E}}')\Sigma^{-1}vec(\hat{\mathbf{E}}') - Q_{kl}^D}{vec'(\hat{\mathbf{E}}')\Sigma^{-1}vec(\hat{\mathbf{E}}')} \cdot |\mathbf{N}_{22}| , \quad \text{and}$$

$$\text{(4.137)} \quad AP_{(22)}^A = \frac{vec'(\hat{\mathbf{E}}')\Sigma^{-1}vec(\hat{\mathbf{E}}') - Q_{kl}^A}{vec'(\hat{\mathbf{E}}')\Sigma^{-1}vec(\hat{\mathbf{E}})} \cdot |\mathbf{M}_{22}| .$$

Now, replacing $\Sigma = (\mathbf{I}_m \otimes \mathbf{V})$ by $\hat{\Sigma} = (\mathbf{I}_m \otimes \hat{\mathbf{V}})$ we obtain the sample versions $\hat{C}_{(22)}$ and $\hat{AP}_{(22)}$ of $C_{(22)}$ and $AP_{(22)}$, using the sample outlier sum of squares $\hat{Q}_{kl}^D$ and $\hat{Q}_{kl}^A$.

Similar to the development above we can obtain formulae for Q^T_{kl}, $C^T_{(22)}$ and $AP^T_{(22)}$. But as pointed out in Sections 3.2.1 and 3.2.2, those statistics are directly obtained in the model (4.117) which is transformed from the right by the transformation $\mathbf{T} = \mathbf{P}$, using either Q^D_{kl}, $C^D_{(22)}$ and $AP^D_{(22)}$ or Q^A_{kl}, $C^A_{(22)}$ and $AP^A_{(22)}$ in the transformed model, which has a diagonal variance-covariance structure. Similarly, the sample versions $\hat{Q}^T_{kl}$, $\hat{C}^T_{(22)}$ and $\hat{AP}^T_{(22)}$ of Q^T_{kl}, $C^T_{(22)}$ and $AP^T_{(22)}$ are obtained in the model (4.117) transformed by $\mathbf{T} = \hat{\mathbf{P}}$.

4.4 AN EXAMPLE: FISHER's IRIS DATA

Anderson (1935) published a data set on three species of iris: *versicolor, virginica* and *setosa*. The four variables are (1) sepal length, (2) sepal width, (3) petal length and (4) petal width, and 50 observations were taken on each species. Fisher (1936) used the data for an example of discriminant analysis, and since then the data have been commonly used to demonstrate multivariate techniques.

A mean is fitted to each of the variables in a given sample, and thus the model for each of the three samples is

$$\underset{(50\times 4)}{\mathbf{Y}} = \underset{(50\times 1)}{\mathbf{1}_{50}} \cdot \underset{(1\times 4)}{\mathbf{B}} + \underset{(50\times 4)}{\mathbf{E}}, \quad \Sigma = \underset{(4\times 4)}{(\mathbf{I}_{50} \otimes \mathbf{V})} . \tag{4.138}$$

The model (4.138) is a multivariate location model, a special case of the multivariate regression model (4.1). The symbol $\mathbf{1}_m$ denotes the vector $(1, \ldots, 1)'$ of dimension m.

The ML-estimates $\hat{\mathbf{B}}$ and $\hat{\mathbf{V}}$ for $\mathbf{B}$ and $\mathbf{V}$ are given in Table 4.11, as well as the estimated principal axes $\hat{\mathbf{P}}$ and the estimated variances $\hat{\Delta}$ of the corresponding principal components.

The EM-algorithms 4.8 through 4.10 have been programmed, to perform ML-estimation in the adjusted models (4.6), (4.7) and (4.9). A program listing is included as an appendix. The convergence of the EM-algorithm is known to be linear, and can be slow, as pointed out by Dempster, Laird and Rubin (1977). For this data, however, convergence was generally rather rapid, particularly for simple outlier patterns and few outliers. But even when a relatively large number of outliers (>10) was fitted, generally less than 15 iterations were required to obtain a precision in the tenth decimal.

To locate D- and A-outliers in the data, each data point y_{ij} was specified as a single outlier, and the corresponding χ^2-value from the LRT was computed. This value follows approximately a χ^2_1-distribution, and thus those data points y_{ij} were singled out for further investigation which yielded a χ^2-value exceeding, respectively, the $(1-0.05)$-, the $(1-0.025)$-, and the $(1-0.005)$-fractile of the χ^2_1-distribution. The individual χ^2-values for the observations in the three samples are given in Tables 4.13, 4.18 and 4.23, and the values exceeding the $\chi^2_1(0.95)$-fractile are printed in bold type, the largest being marked by an asterisk. We note that the negative χ^2-values result from the bias adjustment made by the EM-procedure.

For the samples of *versicolor* and *virginica* there is little agreement between the location and the number of those data points which yield high χ^2-values when specified respectively as D- and A-outlier. The sample of *setosa*, however, is exceptional in showing high agreement between A- and D-outliers, which could be explained by the fact that the sample variance-covariance matrix $\hat{\mathbf{V}}$ for *setosa* is closest to being diagonal of all three sample covariance matrices.

The computation of 1200 individual χ^2-values to locate A- and D-outliers in the three samples using an iterative algorithm like the EM-algorithm takes a large amount of computing time. Alternatively, the χ^2-values could be computed in closed form using the formulae given in Sections 4.2.1 and 4.2.2. A third possibility, which involves the least number of computations, is to compute for each data point y_{ij} the statistic $\hat{Q}_{ij}$, that is, the sample outlier sum of squares corresponding to the observation y_{ij}. In the multivariate location model, using the statistic $\hat{Q}_{ij}^D$ is equivalent to using the scaled residuals $\hat{e}_{ij}/\sqrt{\hat{V}_{jj}}$, and similarly using $\hat{Q}_{ij}^T$ is equivalent to using the scaled PC-residuals $\hat{e}_{ij}^*/\sqrt{\hat{\Delta}_j}$. Finally, using $\hat{Q}_{ij}^A$ is equivalent to using the scaled residuals in the model transformed by $\hat{\mathbf{V}}^{-1}$. Thus we have

(4.139)
$$\hat{Q}_{ij}^D = \frac{\hat{e}_{ij}^2}{\hat{V}_{jj}} \cdot \frac{m}{m-1} ,$$

$$\hat{Q}_{ij}^A = \frac{\hat{\varepsilon}_{ij}^2}{(\hat{\mathbf{V}}^{-1})_{jj}} \cdot \frac{m}{m-1} , \quad \text{where } \hat{\mathbf{E}}\hat{\mathbf{V}}^{-1} = (\hat{\varepsilon}_{ij}) , \quad \text{and}$$

$$\hat{Q}_{ij}^T = \frac{(\hat{e}_{ij}^*)^2}{\hat{\Delta}_j} \cdot \frac{m}{m-1} , \quad \text{where } \hat{\mathbf{E}}\hat{\mathbf{P}} = \hat{\mathbf{E}}^* = (\hat{e}_{ij}^*) .$$

To locate T-outliers in the data, the scaled PC-residuals $\hat{e}_{ij}^*/\sqrt{\hat{\Delta}_j}$ were computed, and they are given in Tables 4.12, 4.17 and 4.22. The scaled PC-residuals are approximately distributed as $N(0,1)$, and similarly to the procedure above, those PC's y_{ij}^* whose corresponding scaled PC-residual yielded a tail probability smaller than 0.025, 0.0125 or 0.0025 were singled out for further investigation. The residuals with a tail probability smaller than 0.025 appear in bold type in Tables 4.12, 4.17 and 4.22, and the absolutely largest is marked by an asterisk.

Thus for each of the three samples, and with respect to D-, A- and T-outliers, three sets of suspicious data points were obtained. Tables 4.13 through 4.15, 4.18 through 4.20 and 4.23 through 4.25 present the results obtained when the respective sets of data points were specified as outliers. The subscripts of the data points specified as outliers, the number of outliers and the resulting χ^2-statistic are listed. In addition, the adjusted estimates $\tilde{\mathbf{B}}$ and $\tilde{\Delta}$ for $\mathbf{B}$ and Δ are given. The χ^2-statistic can be compared with a χ_k^2-distribution, where k is the number of outlying data points in question. Unfortunately it is difficult to determine a level of significance at which to reject the null-hypothesis. Ideally the chi-square value for k outliers would have to be compared with the $(1-\alpha)$-fractile of the distribution of the maximum of $c_k = \binom{n}{k}$ possible χ^2-statistics from the data. Since this distribution is generally intractable, the $(1-\alpha/c_k)$-fractile of the χ_k^2-distribution is used as an approxima-

tion, resulting from the first Bonferroni inequality. Hawkins (1980) notes that this approximation is conservative but generally very good for $k = 1$. For $k>1$, however, the approximation can be extremely conservative. We will keep this in mind while screening the data outliers.

The $(1-0.05/c_k)$-fractiles of the χ^2_k-distribution for $n \cdot m = 200$ and $k = 1, \ldots, 7$ were computed using the routine MDCH of the IMSLIB (1985)-Library. They are

(4.140)

k	1	2	3	4	5	6	7
$\chi^2_k(1-0.05/c_k)$	13.5	25.8	37.4	48.4	59.0	69.2	78.8

We note that the χ^2_7-value appears to be unreliable, and so is possibly the χ^2_6-value. The computation of $\chi^2_8(1-0.05/c_8)$ failed since the CDC CYBER 170 treated $(1-0.05/c_8)$ as 1.0 in single precision. In the following analysis we only use the values for $k = 1, \ldots, 4$.

The $(1-0.025/c_1) = (1-0.025/200)$-fractile of $N(0,1)$ is 3.66, and for the distribution of the maximum of 200 independent $N(0,1)$-variates we obtain a $(1-0.025)$-fractile of 3.02. Those two fractiles can be used as values against which the scaled PC-residuals may be compared.

Versicolor

The largest χ^2_1-value for D-outliers is 6.44, much smaller than 13.5, and thus the hypothesis that D-outliers are in the sample is rejected. Similarly we reject the hypothesis of T-outliers, with the absolute value of the largest scaled PC-residual being 2.51 compared with 3.66 or even 3.02. The largest χ^2_1-value for A-outliers, however, is 13.06, with the second largest being 8.10. This warrants some further investigation in view of the conservative nature of the Bonferroni-approximation. The χ^2_2-value for observations (19,2) and (49,3) is 21.19 from Table 4.15, compared with $\chi^2_2(1-0.05/c_2) = 25.8$. It appears that (19,2) and (49,3) are A-outliers.

Virginica

The largest χ^2_1-value for D-outliers is 6.73, the largest χ^2_1-value for A-outliers is 8.42 and the largest absolute value of the PC-residuals is 2.48, all well below Bonferroni significance. The hypothesis that outliers are present is rejected.

Setosa

The sample of *setosa* is interesting since it shows great agreement between D-, A- and T-outliers. The largest χ^2_1-values for D- and A-outliers are respectively 11.95 and 12.96, both for the data point (44,4). The largest absolute PC-residual is 3.28, also appearing in the observational vector $\mathbf{y}_{44}$. All those values are close to significance, but the hypothesis

of A-outliers in the data appears best supported by the data. The χ^2_2-value for observations (42,2) and (44,4) specified as D- and A-outliers respectively is 19.21 and 22.77, compared with $\chi^2_2(1-0.05/c_2) = 25.8$, and similarly the χ^2_3-values are respectively 24.29, 32.41 and 26.26 for D-, A- and T-outliers, compared with $\chi^2_3(1-0.05/c_3) = 37.3$. Thus it appears that at most three A-outliers are present, data points (25,3), (42,2) and (44,4). The χ^2_5-value in Table 4.25 is nonsignificant.

In summary, it appears to be worthwhile to consider the three different types of outlier, distributional, additive, and transformational, since there may generally be little agreement between the sets of supicious data points corresponding to the respective types of outlier. There is no evidence in this data that complete observational vectors are outlying. It is generally a single component which signals a certain observational vector to be possibly outlying. Thus it is instructive to search the individual data points for outliers, and not only complete observational vectors.

Table 4.11 Initial estimates for Fisher's iris data

	Versicolor	*Virginica*	*Setosa*
$\hat{\mathbf{B}}$	[5.936 2.770 4.260 1.326]	[6.588 2.974 5.552 2.026]	[5.006 3.428 1.462 .246]
$\hat{\mathbf{V}}$	$\begin{bmatrix} .261 & & & \\ .083 & .096 & & \\ .179 & .081 & .216 & \\ .055 & .040 & .072 & .038 \end{bmatrix}$	$\begin{bmatrix} .396 & & & \\ .092 & .102 & & \\ .297 & .067 & .298 & \\ .048 & .047 & .048 & .074 \end{bmatrix}$	$\begin{bmatrix} .122 & & & \\ .097 & .141 & & \\ .016 & .011 & .030 & \\ .010 & .009 & .006 & .011 \end{bmatrix}$
$\hat{\mathbf{P}}$	$\begin{bmatrix} -.687 & .669 & .265 & .102 \\ -.305 & -.567 & .730 & -.229 \\ -.624 & -.343 & -.627 & -.316 \\ -.215 & -.335 & -.063 & .915 \end{bmatrix}$	$\begin{bmatrix} -.741 & -.165 & -.534 & .371 \\ -.203 & .749 & -.325 & -.541 \\ -.628 & -.169 & .652 & -.391 \\ -.124 & .619 & .429 & .646 \end{bmatrix}$	$\begin{bmatrix} -.669 & .593 & .440 & -.036 \\ -.734 & -.621 & -.274 & -.020 \\ -.097 & .490 & -.832 & -.240 \\ -.064 & .131 & -.195 & .970 \end{bmatrix}$
$\hat{\Delta}$	[.478 .071 .054 .010]	[.681 .104 .051 .034]	[.232 .036 .026 .009]

Table 4.12 *Versicolor:* residuals and scaled PC-residuals

residuals $\hat{e}_{ij}$				scaled PC-residuals $\hat{e}^*_{ij}/\sqrt{\hat{\Delta}_j}$			
1.06	.43	.44	.07	−1.67	1.10	1.36	−.62
.46	.43	.24	.17	−.92	−.28	1.19	.33
.96	.33	.64	.17	−1.73	.67	.36	−.20
−.44	−.47	−.26	−.03	.88	.27	−1.27	1.24
.56	.03	.34	.17	−.93	.70	−.23	1.05
−.24	.03	.24	−.03	.01	−.93	−.82	−1.33
.36	.53	.44	.27	−1.08	−1.13	.82	.28
−1.04	−.37	−.96	−.33	**2.16**	−.17	.34	−.17
.66	.13	.34	−.03	−1.02	.99	.26	−.95
−.74	−.07	−.36	.07	1.06	−1.33	−.11	1.25
−.94	−.77	−.76	−.33	**2.06**	.68	−1.35	.23
−.04	.23	−.06	.17	−.07	−.72	.80	1.24
.06	−.57	−.26	−.33	.52	**2.12**	−.93	−.81
.16	.13	.44	.07	−.64	−.53	−.61	−.86
−.34	.13	−.66	−.03	.88	−.24	1.82	1.23
.76	.33	.14	.07	−1.05	.94	1.51	.27
−.34	.23	.24	.17	−.04	−1.86	−.36	−.04
−.14	.07	−.16	−.33	.41	.42	.15	**−2.51***
.26	−.57	.24	.17	−.28	1.35	**−2.19**	**2.46**
−.34	−.27	−.36	−.23	.85	.48	−.20	−.67
−.04	.43	.54	.47	−.79	**−2.30**	−.28	1.64
.16	.03	−.26	−.03	.07	.72	.99	.70
.36	−.27	.64	.17	−.87	.45	**−2.21**	.57
.16	.03	.44	−.13	−.53	−.06	−.87	**−2.50**
.46	.13	.04	−.03	−.55	.87	.84	−.19
.66	.23	.14	.07	−.91	.90	1.08	.40
.86	.03	.54	.07	−1.38	1.32	−.40	−.22
.76	.23	.74	.37	−1.64	.00	−.51	1.37
.06	.13	.24	.17	−.39	−.64	−.21	.61
−.24	−.17	−.76	−.33	1.10	1.16	1.34	−.44
−.44	−.37	−.46	−.23	1.08	.57	−.36	−.22
−.44	−.37	−.56	−.33	1.20	.83	−.06	−.83
−.14	−.07	−.36	−.13	.53	.43	.63	.01
.06	−.07	.84	.27	−.88	−1.12	**−2.50**	.08
−.54	.23	.24	.17	.16	**−2.36**	−.59	−.25
.06	.63	.24	.27	−.64	−1.84	1.33	.38
.76	.33	.44	.17	−1.36	.43	.67	.23
.36	−.47	.14	−.03	−.27	1.77	−1.44	.78
−.34	.23	−.16	−.03	.38	−1.10	.78	−.62
−.44	−.27	−.26	−.03	.79	−.15	−.64	.77
−.44	−.17	.14	−.13	.42	−.75	−1.38	−1.69
.16	.23	.34	.07	−.59	−.61	−.03	−.77
−.14	−.17	−.26	−.13	.48	.51	.05	−.08
−.94	−.47	−.96	−.33	**2.10**	.30	.14	.17
−.34	−.07	−.06	−.03	.43	−.58	−.44	−.24
−.24	.23	−.06	−.13	.23	−.85	.65	−1.77
−.24	.13	−.06	−.03	.24	−.76	.31	−.60
−.26	.13	.04	−.03	−.35	.37	.61	−.40
−.84	−.27	−1.26	−.23	**2.16**	.38	1.67	1.71
−.24	.03	−.16	−.03	.37	−.42	.26	−.04

Table 4.13 *Versicolor:* individual bias adjusted χ^2-statistics

χ^2 for *D*-outlier y_{ij}				χ^2 for *A*-outlier y_{ij}			
3.42	.80	−.19	−.88	2.28	−.03	−.86	−.36
−.19	.96	−.75	−.23	−.60	−.18	−.50	−.87
2.75	.09	.89	−.25	.46	−.96	−1.00	−.94
−.31	1.29	−.71	−.99	−1.00	2.29	−.85	.56
.21	−1.00	−.48	−.23	.34	−.06	−.59	.02
−.81	−1.00	−.75	−.99	1.41	−.66	2.11	.47
−.52	2.03	−.13	.98	−.98	.22	−.98	−.80
3.30	.35	3.48	1.77	−.37	−.97	−.76	−.95
.69	−.84	−.49	−.99	−.15	−.91	−.79	.16
1.04	−.96	−.44	−.88	.17	−.97	−.52	1.09
2.28	**5.62**	1.55	1.66	−.68	1.89	−1.01	−1.00
−1.01	−.47	−.99	−.23	−.99	−.94	.40	.80
−1.00	2.31	−.73	1.68	.00	.90	−.80	.32
−.91	−.84	−.11	−.87	−.44	−.90	.66	−.43
−.61	−.85	.92	−.99	−.58	−.51	3.78	.54
1.22	.08	−.93	−.88	2.28	−.63	.48	−.99
−.60	−.48	−.76	−.25	1.85	−.54	−.34	−.96
−.95	−.97	−.91	1.65	−.62	.57	1.60	**6.17**
−.81	1.80	−.81	−.40	.93	**13.06***	−.19	**5.95**
−.58	−.24	−.41	.37	−.98	−1.00	−.92	−.42
−1.01	.69	.17	**4.92**	.62	−.99	−.96	3.58
−.91	−1.00	−.70	−.99	.39	−1.01	.80	−.67
−.55	−.32	.78	−.29	−1.01	3.36	.08	−.59
−.92	−1.00	−.21	−.65	.26	−.31	**5.67**	**5.55**
−.18	−.84	−1.00	−.99	.30	−.90	−.74	−.91
.69	−.48	−.92	−.87	1.54	−.96	.09	−.93
1.93	−1.00	.31	−.88	.78	−.54	−.91	−.89
1.19	−.50	1.51	2.76	−.29	−.08	−.87	1.23
−.99	−.83	−.74	−.20	−.94	−.99	−1.01	−.48
−.81	−.73	1.67	1.79	−.05	−.71	.20	−.51
−.28	.45	−.02	.36	−1.00	−.73	−1.01	−.90
−.29	.41	.45	1.90	−1.01	−.99	−.98	.00
−.94	−.96	−.40	−.59	−.83	−.97	−.62	−1.00
−1.00	−.97	2.05	.75	1.72	.66	3.41	−.88
.01	−.51	−.77	−.29	**4.59**	−.23	.57	−1.00
−1.00	3.21	−.77	.86	−.51	2.12	−.81	−.59
1.31	.12	−.12	−.23	.29	−.92	−.82	−.97
−.55	1.19	−.93	−.99	.77	**4.45**	−.98	−.61
−.58	−.47	−.89	−.99	−.10	.91	−.96	−.74
−.28	−.25	−.70	−.99	−.88	−.24	−.92	−.38
−.34	−.74	−.93	−.63	2.64	−.95	**4.14**	1.65
−.91	−.46	−.47	−.87	−.67	−.45	−.23	−.55
−.94	−.70	−.69	−.59	−.95	−.96	−.96	−.98
2.44	1.26	3.50	1.79	−.89	−.88	−.44	−1.01
−.57	−.96	−.99	−.99	−.27	−1.01	−.74	−.98
−.81	−.49	−.99	−.62	.35	2.34	.26	1.99
−.80	−.83	−.99	−.99	−.38	−.23	−.83	−.74
−.74	−.83	−1.00	−.99	−.75	−.75	−1.00	−.80
1.31	−.39	**6.44***	−.10	−.17	−.99	**8.10**	1.92
−.79	−1.00	−.89	−.99	−.87	−.87	−1.00	−1.01

Table 4.14 *Versicolor* (individually most likely *D*-outliers)

individual $\chi^2 \geq$	7.88 ≙ 0.5%	5.02 ≙ 2.5%	3.84 ≙ 5%
outliers	–	(11,2), (49,3)	(21,4)
number of outliers	0	2	3
χ^2-statistic	–	12.55	18.03
$\tilde{\mathbf{B}}$	–	[5.970 2.795 4.298 1.342]	[5.958 2.785 4.280 1.331]
$\tilde{\Delta}$	–	[.404 .071 .050 .010]	[.416 .065 .050 .009]

Table 4.15 *Versicolor* (individually most likely *A*-outliers)

individual $\chi^2 \geq$	7.88 ≙ 0.5%	5.02 ≙ 2.5%	3.84 ≙ 5%
outliers	(19,2), (49,3)	(18,4), (19,4), (24,3), (24,4)	(35,1), (38,2), (41,3)
number of outliers	2	6	9
χ^2-statistic	21.19	36.83	55.98
$\tilde{\mathbf{B}}$	[5.936 2.786 4.273 1.326]	[5.936 2.784 4.266 1.333]	[5.951 2.794 4.255 1.333]
$\tilde{\Delta}$	[.461 .070 .046 .008]	[.459 .069 .045 .006]	[.471 .059 .040 .005]

Table 4.16 *Versicolor* (largest scaled PC-residuals)

$\hat{e}^*_{ij}/\sqrt{\hat{\Delta}_j} \geq$	2.81 ≙ 0.5%	2.24 ≙ 2.5%	1.96 ≙ 5%
outliers	–	(18,4), (19,4), (21,2), (27,4) (34,3), (35,2)	(8,1), (11,1), (13,2), (19,3), (23,3), (44,1), (49,1)
number of outliers	0	6	13
χ^2-statistic	–	32.70	69.66
$\tilde{\mathbf{B}}$	–	[5.952 2.756 4.249 1.322]	[6.046 2.827 4.297 1.353]
$\tilde{\Delta}$	–	[.478 .059 .051 .006]	[.317 .053 .038 .007]

Table 4.17 *Virginica:* residuals and scaled PC-residuals

residuals $\hat{e}_{ij}$				scaled PC-residuals $\hat{e}^*_{ij}/\sqrt{\hat{\Delta}_j}$			
−.29	.33	.45	.47	−.23	1.58	**2.40**	−.83
−.79	−.27	−.45	−.13	1.14	−.24	.71	−.27
.51	.03	.35	.07	−.74	−.24	−.10	.48
−.29	−.07	.05	−.23	.27	−.48	.50	−1.26
−.09	.03	.25	.17	−.14	.31	1.21	−.17
1.01	.03	1.05	.07	−1.72	−.36	.73	.00
−1.69	−.47	−1.05	−.33	**2.48***	−.31	1.02	−.93
.71	−.07	.75	−.23	−1.16	−1.36	.15	−.73
.11	−.47	.25	−.23	−.14	−1.72	.70	.30
.61	.63	.55	.47	−1.19	1.76	.13	−.10
−.09	.23	−.45	−.03	.37	.76	−1.47	.03
−.19	−.27	−.25	−.13	.45	−.65	−.13	.52
.21	.03	−.05	.07	−.17	.12	−.55	.72
−.89	−.47	−.55	−.03	1.34	−.40	1.14	.68
−.79	−.17	−.45	.37	1.04	.95	1.52	1.20
−.19	.23	−.25	.27	.26	1.28	−.09	.45
−.09	.03	−.05	−.23	.15	−.30	−.41	−.94
1.11	.83	1.15	.17	**−2.10**	1.08	−.18	**−2.02**
1.11	−.37	1.35	.27	**−1.97**	−1.62	**2.31**	1.45
−.59	−.77	−.55	−.53	1.22	**−2.21**	−.08	.41
.31	.23	.15	.27	−.49	.81	−.12	.62
−.99	−.17	−.65	−.03	1.43	.39	.66	−.19
1.11	−.17	1.15	−.03	−1.82	−1.62	.88	.23
−.29	−.27	−.65	−.23	.86	−.58	−1.23	.82
.11	.33	.15	.07	−.30	.76	−.17	−.79
.61	.23	.45	−.23	−.91	−.46	−.91	−1.18
−.39	−.17	−.75	−.23	1.00	−.24	−1.43	.53
−.49	.03	−.65	−.23	.96	.22	−1.19	−.47
−.19	−.17	.05	.07	.16	−.19	.97	.29
.61	.03	.25	−.43	−.68	−1.20	−1.58	−.87
.81	−.17	.55	−.13	−1.08	−1.35	−.33	.55
1.31	.83	.85	−.03	**−2.02**	.75	−1.89	−1.68
−.19	−.17	.05	.17	.15	.00	1.16	.64
−.29	−.17	−.45	−.53	.72	−1.03	−1.37	−.96
−.49	−.37	.05	−.63	.59	−1.84	.64	**−2.19**
1.11	.03	.55	.27	−1.46	−.27	−.57	1.97
−.29	.43	.05	.37	.06	1.83	.91	−.62
−.19	.13	−.05	−.23	.21	−.02	−.32	−1.44
−.59	.03	−.75	−.23	1.13	.32	−1.24	−.46
.31	.13	−.15	.07	−.21	.35	−1.21	.85
.11	.13	.05	.37	−.22	.93	.40	1.07
.31	.13	−.45	.27	−.01	.89	−1.70	**2.19**
−.79	−.27	−.45	−.13	1.14	−.24	.71	−.27
.21	.23	.35	.27	−.55	.76	.70	−.01
.11	.33	.15	.47	−.36	1.53	.59	.62
.11	.03	−.35	.27	.12	.71	−.80	1.87
−.29	−.47	−.55	−.13	.81	−.90	−.47	1.55
−.09	.03	−.35	−.03	.34	.24	−.89	.40
−.39	.43	−.15	.27	.32	1.79	.39	−.75
−.69	.03	−.45	−.23	.99	.22	−.14	−1.30

Table 4.18 *Virginica:* individual bias adjusted χ^2-statistics

χ^2 for *D*-outlier y_{ij}				χ^2 for *A*-outlier y_{ij}			
−.83	−.11	−.44	1.77	**5.83**	−.45	**4.55**	.36
.60	−.27	−.33	−.80	−.05	−.94	−.73	−1.01
−.33	−1.00	−.60	−.94	−.62	−.84	−.98	−.94
−.80	−.96	−1.00	−.32	.35	−.79	.33	.07
−.99	−1.00	−.81	−.60	.01	−.99	.09	−.72
1.60	−1.00	2.84	−.94	−1.00	−.53	.19	−1.01
6.73*	.99	2.52	.27	3.40	−1.01	−.11	−.68
.24	−.96	.86	−.35	−1.01	−.95	.01	.22
−.98	1.22	−.81	−.34	−1.01	1.26	−.55	−.97
−.11	3.01	−.05	2.08	−1.00	.37	−.98	−.34
−.99	−.52	−.34	−1.00	−.23	−.11	.90	−.92
−.92	−.26	−.80	−.79	−.85	−.45	−.85	−1.01
−.90	−1.00	−1.00	−.94	−.23	−.94	−.31	−.84
.90	1.22	−.02	−1.00	−.47	.61	−.93	−.29
.47	−.74	−.38	.80	−.30	−.01	−.99	3.31
−.92	−.51	−.80	.02	−1.00	−.83	−.67	−.22
−.99	−1.00	−1.00	−.31	−.93	−.57	−.95	.14
1.80	**5.78**	3.13	−.67	−.37	**5.25**	1.10	.62
1.48	.03	**4.34**	−.25	−1.01	**8.42***	2.50	2.13
−.22	**5.27**	−.08	2.71	−.88	1.99	−.94	−.43
−.76	−.50	−.94	.03	−.78	−1.00	−.79	−.33
1.56	−.72	.43	−1.00	.15	−1.01	−.93	−.94
2.01	−.75	3.45	−1.00	−.93	.97	.55	−1.00
−.81	−.29	.41	−.34	.66	−.69	1.46	−.98
−.98	.06	−.94	−.94	−.84	.23	−.93	−.88
−.08	−.53	−.36	−.34	−.94	.03	−.96	1.36
−.64	−.73	.90	−.34	.35	−1.00	1.68	−.90
−.42	−1.00	.43	−.33	−.93	−.26	−.18	−.29
−.92	−.71	−1.00	−.94	−.71	−.53	−.66	−.69
−.12	−1.00	−.82	1.41	.06	−.67	−.81	2.77
.65	−.73	−.03	−.80	.26	.06	−.98	−.94
2.91	**5.53**	1.03	−1.00	−.25	**6.39**	−.96	2.86
−.92	−.72	−1.00	−.60	−.77	−.17	−.74	.04
−.81	−.73	−.36	2.91	−.87	−.66	−.72	2.55
−.50	.18	−1.00	**4.21**	2.44	−.91	3.32	**5.11**
2.04	−1.00	−.09	−.08	3.79	.97	.76	.82
−.82	.73	−1.00	.85	.99	.53	−.29	−.49
−.92	−.86	−1.00	−.32	−.46	.37	−.70	.74
−.15	−1.00	.91	−.34	−.94	−.14	.02	−.32
−.77	−.86	−.93	−.94	1.04	−1.01	1.17	−.91
−.98	−.86	−1.00	.95	−.92	−.85	−.85	1.13
−.80	−.88	−.42	−.13	**5.85**	−.72	**7.32**	1.35
.60	−.27	−.33	−.80	−.05	−.94	−.73	−1.01
−.90	−.51	−.60	.03	−.77	−.95	−.69	−.59
−.98	.02	−.94	2.18	−.97	−.95	−1.01	1.14
−.98	−1.00	−.62	−.04	1.74	−.49	2.59	1.29
−.81	1.20	−.03	−.81	.59	1.55	.82	−.59
−.99	−1.00	−.59	−1.00	−.41	−.99	.09	−1.01
−.65	.75	−.94	−.03	.34	1.20	−.82	−.89
.18	−1.00	−.34	−.33	.08	.21	−.83	.14

Table 4.19 *Virginica* (individually most likely *D*-outliers)

individual $\chi^2 \geq$	7.88 ≙ 0.5%	5.02 ≙ 2.5%	3.84 ≙ 5%
outliers	–	(7,1), (18,2), (20,2), (32,2)	(19,3), (35,4)
number of outliers	0	4	6
χ^2-statistic	–	24.17	37.94
$\tilde{\mathbf{B}}$	–	[6.614 2.959 5.573 2.017]	[6.587 2.963 5.544 2.028]
$\tilde{\Delta}$	–	[.529 .103 .048 .029]	[.462 .094 .047 .028]

Table 4.20 *Virginica* (individually most likely *A*-outliers)

individual $\chi^2 \geq$	7.88 ≙ 0.5%	5.02 ≙ 2.5%	3.84 ≙ 5%
outliers	(19,2)	(1,1), (18,2), (32,2) (35,4), (42,1), (42,3)	(1,3)
number of outliers	1	7	8
χ^2-statistic	8.42	38.71	37.69
$\tilde{\mathbf{B}}$	[6.588 2.990 5.552 2.026]	[6.600 2.959 5.565 2.036]	[6.598 2.959 5.564 2.036]
$\tilde{\Delta}$	[.693 .010 .048 .031]	[.694 .091 .040 .022]	[.702 .092 .040 .022]

Table 4.21 *Virginica* (largest scaled PC-residuals)

$\hat{e}^*_{ij} / \sqrt{\hat{\Delta}_j} \geq$	2.81 ≙ 0.5%	2.24 ≙ 2.5%	1.96 ≙ 5%
outliers	–	(1,3), (7,1), (19,3)	(18,1), (18,4), (19,1), (20,2), (32,1), (35,4), (42,4)
number of outliers	0	3	10
χ^2-statistic	–	16.22	49.86
$\tilde{\mathbf{B}}$	–	[6.632 2.987 5.562 2.023]	[6.550 2.984 5.494 2.021]
$\tilde{\Delta}$	–	[.608 .104 .041 .034]	[.461 .096 .041 .025]

Table 4.22 *Setosa:* residuals and scaled PC-residuals

residuals $\hat{e}_{ij}$				*scaled PC-residuals* $\hat{e}^*_{ij}/\sqrt{\hat{\Delta}_j}$			
.09	.07	−.06	−.05	−.22	−.13	.51	−.37
−.11	−.43	−.06	−.05	.82	.87	.81	−.19
−.31	−.23	−.16	−.05	.81	−.67	.44	.10
−.41	−.33	.04	−.05	1.06	−.14	−.69	−.35
−.01	.17	−.06	−.05	−.24	−.77	.07	−.35
.39	.47	.24	.15	−1.34	.42	−1.14	.73
−.41	−.03	−.06	.05	.61	−1.31	−.80	.88
−.01	−.03	.04	−.05	.05	.14	−.11	−.56
−.61	−.53	−.06	−.05	1.67	−.37	−.38	.03
−.11	−.33	.04	−.15	.66	.73	.25	−1.49
.39	.27	.04	−.05	−.96	.42	.47	−.78
−.21	−.03	.14	−.05	.31	−.23	−1.17	−.74
−.21	−.43	−.06	−.15	.97	.49	.66	−1.18
−.71	−.43	−.36	−.15	1.73	−1.86	.84	−.22
.79	.57	−.26	−.05	−1.92	−.08	**2.59**	−.23
.69	.97	.04	.15	**−2.47**	−.79	−.14	1.02
.39	.47	−.16	.15	−1.26	−.61	.92	1.75
.09	.07	−.06	.05	−.24	−.06	.39	.66
.69	.37	.24	.05	−1.59	1.62	−.03	−.39
.09	.37	.04	.05	−.71	−.78	−.64	.35
.39	−.03	.24	−.05	−.55	1.91	−.05	−1.23
.09	.27	.04	.15	−.57	−.39	−.59	1.40
−.41	.17	−.46	−.05	.40	**−3.06**	1.03	.82
.09	−.13	.24	.25	−.02	1.50	−1.06	**2.00**
−.21	−.03	.44	−.05	.25	.54	**−2.71**	−1.51
−.01	−.43	.14	−.05	.64	1.70	.06	−.73
−.01	−.03	.14	.15	.00	.53	−.86	1.24
.19	.07	.04	−.05	−.38	.44	.26	−.66
.19	−.03	−.06	−.05	−.21	.51	.95	−.38
−.31	−.23	.14	−.05	.75	.11	−1.10	−.66
−.21	−.33	.14	−.05	.76	.75	−.66	−.68
.39	−.03	.04	.15	−.53	1.53	.74	1.35
.19	.67	.04	−.15	−1.28	−1.59	−.63	−1.82
.49	.77	−.06	−.05	−1.85	−1.16	.41	−.67
−.11	−.33	.04	−.05	.65	.80	.13	−.46
−.01	−.23	−.26	−.05	.41	.02	1.77	.24
.49	.07	−.16	−.05	−.76	.87	**2.11**	−.27
−.11	.17	−.06	−.15	−.08	−1.15	−.09	−1.34
−.61	−.43	−.16	−.05	1.53	−.96	−.03	.26
.09	−.03	.04	−.05	−.09	.45	.16	−.60
−.01	.07	−.16	.05	−.08	−.63	.63	.96
−.51	−1.13	−.16	.05	**2.45**	1.71	1.30	1.40
−.61	−.23	−.16	−.05	1.23	−1.61	−.37	.22
−.01	.07	.14	.35	−.18	.35	−1.27	**3.28***
.09	.37	.44	.15	−.81	.32	**−2.81**	.36
−.21	−.43	−.06	.05	.94	.63	.42	.88
.09	.37	.14	−.05	−.72	−.59	−1.03	−.94
−.41	−.23	−.06	−.05	.93	−.72	−.34	−.11
.29	.27	.04	−.05	−.83	.10	.20	−.74
−.01	−.13	−.06	−.05	.22	.21	.57	−.29

Table 4.23 *Setosa:* individual bias adjusted χ^2-statistics

χ^2 for *D*-outlier y_{ij}				χ^2 for *A*-outlier y_{ij}			
−.94	−.97	−.88	−.81	−.90	−1.01	−.91	−.82
−.92	.31	−.88	−.82	−.19	.93	−.94	−.95
−.23	−.64	−.10	−.82	−.78	−1.01	−.49	−1.01
.38	−.24	−.96	−.82	−.29	−1.01	−.62	−.92
−1.01	−.80	−.88	−.81	−.84	−.49	−.96	−.84
.24	.55	.90	1.18	−1.01	−.59	−.22	−.12
.33	−1.00	−.89	−.75	2.07	.38	−.94	−.19
−1.01	−1.00	−.96	−.81	−1.01	−1.01	−.87	−.71
2.16	1.00	−.89	−.82	.13	−.98	−1.00	−1.01
−.92	−.27	−.96	.97	−.62	−.17	−.50	1.29
.29	−.49	−.96	−.82	−.04	−.99	−.99	−.36
−.67	−1.00	−.37	−.82	−.12	−.62	.37	−.60
−.67	.29	−.89	.98	−.68	.05	−1.00	.59
3.04	.17	3.43	.80	1.07	−.70	1.33	−.75
3.87	.99	.99	−.85	**4.74**	−.90	**3.88**	−.68
2.82	**6.15**	−.97	.99	−1.01	2.00	−.66	.21
.18	.48	−.19	1.08	−.85	−.72	2.12	1.89
−.94	−.97	−.88	−.74	−.96	−1.01	−.64	−.62
3.07	−.09	.84	−.76	2.09	−.48	−.09	−.93
−.94	−.01	−.96	−.74	−.41	.41	−1.00	−.86
.19	−1.01	.83	−.83	1.90	.81	1.01	.30
−.94	−.50	−.96	1.23	−.65	−.46	−.95	1.17
.10	−.84	**6.13**	−.86	2.88	3.41	**6.34**	−.69
−.95	−.91	.75	**5.21**	−.93	−.09	−.48	**4.52**
−.72	−1.01	**5.48**	−.85	.90	−.40	**9.57**	.65
−1.01	.26	−.39	−.83	.34	1.82	−.06	−.57
−1.01	−1.00	−.37	1.24	−.94	−.99	−.82	.97
−.70	−.97	−.96	−.81	−.56	−.92	−.95	−.57
−.70	−1.00	−.88	−.82	.13	−.47	−.84	−.79
−.24	−.65	−.37	−.82	−.36	−.99	.47	−.71
−.67	−.25	−.37	−.82	−1.01	−.64	.25	−.68
.20	−1.01	−.97	1.12	1.83	1.28	−.73	1.04
−.74	2.02	−.97	.77	.10	**4.32**	−.51	2.84
.93	3.39	−.89	−.83	−1.01	1.55	−.75	−.35
−.92	−.23	−.96	−.82	−.76	−.10	−.85	−.82
−1.01	−.66	1.36	−.82	−.02	−.05	1.56	−1.01
.90	−.98	−.19	−.83	**4.23**	.92	.72	−.79
−.92	−.81	−.89	.98	−.43	.25	−.99	1.07
2.14	.28	−.14	−.82	.57	−.97	−.71	−.98
−.94	−1.00	−.96	−.81	−.75	−.87	−.92	−.66
−1.01	−.97	−.11	−.74	−.99	−.95	.39	−.28
.68	**8.19**	−.32	−.81	.72	**9.80**	.37	1.31
2.08	−.66	−.16	−.83	2.44	.08	−.75	−.99
−1.01	−.98	−.51	**11.95***	−.20	−.98	−.98	**12.96***
−.95	−.16	**5.69**	.94	1.00	.82	**5.20**	−.39
−.67	.31	−.88	−.75	−.89	.26	−.77	−.23
−.94	−.04	−.38	−.82	−.40	.65	.06	−.25
-39	−.64	−.88	−.82	.06	−.85	−1.01	−.99
−.29	−.48	−.96	−.82	−.73	−.97	−.97	−.44
−1.01	−.89	−.88	−.81	−.79	−.77	−.93	−.89

Table 4.24 *Setosa* (individually most likely *D*-outliers)

individual $\chi^2 \geq$	7.88 ≙ 0.5%	5.02 ≙ 2.5%	3.84 ≙ 5%
outliers	(42,2), (44,4)	(16,2), (23,3), (24,4), (25,3), (45,3)	(15,1)
number of outliers	2	7	8
χ^2-statistic	19.21 (22.77)*	58.49 (56.41)*	61.82 (62.02)*
$\tilde{\mathbf{B}}$	[5.014 3.443 1.460 .241]	[4.986 3.414 1.449 0.232]	[4.965 3.399 1.446 .234]
$\tilde{\Delta}$	[.217 .035 .025 .007]	[.197 .031 .016 .006]	[.186 .031 .016 .006]

*χ^2-values if the same data points are specified as *A*-outliers.

Table 4.25 *Setosa* (individually most likely *A*-outliers)

individual $\chi^2 \geq$	7.88 ≙ 0.5%	5.02 ≙ 2.5%	3.84 ≙ 5%
outliers	(25,3), (42,2), (44,4)	(23,3), (45,3)	(15,1), (15,3), (24,4) (33,2), (37,1)
number of outliers	3	5	10
χ^2-statistic	32.41 (24.29)*	46.23 (41.98)*	78.69 (54.34)*
$\tilde{\mathbf{B}}$	[5.006 3.445 1.452 .240]	[5.006 3.445 1.452 .240]	[4.986 3.433 1.455 .234]
$\tilde{\Delta}$	[.212 .035 .021 .007]	[.211 .032 .017 .007]	[.193 .033 .011 .006]

*χ^2-values if the same data points are specified as *D*-outliers.

Table 4.26 *Setosa* (largest scaled PC-residuals)

$\hat{e}^*_{ij}/\sqrt{\hat{\Delta}_j} \geq$	2.81 ≙ 0.5%	2.24 ≙ 2.5%	1.96 ≙ 5%
outliers	(23,2), (44,4), (45,3)	(15,3), (16,1), (25,3), (42,1)	(24,4), (37,3)
number of outliers	3	7	9
χ^2-statistic	26.26	55.96	68.97
$\tilde{\mathbf{B}}$	[5.016 3.420 1.456 .239]	[5.016 3.419 1.456 .239]	[5.015 3.419 1.462 .234]
$\tilde{\Delta}$	[.232 .031 .023 .007]	[.184 .032 .016 .007]	[.184 .032 .013 .007]

4.5 APPENDIX

Program listing of program EMALG for the computation of MLE's and LRT-statistics in multivariate models adjusted for outliers using the EM-algorithm.

```
      PROGRAM EMALG(INPUT,OUTPUT,TAPE5=INPUT,TAPE6=OUTPUT)
C
C
      PARAMETER(N=4,M=50)
C
      REAL Y(M,N)
      REAL BHAT(N),VHAT(N,N),EHAT(M,N),ESTAR(M,N)
      INTEGER PA(M,N),O(M,N),C(M,N),NO(M)
C
      REAL PPRIME(N,N),DELTA(N)
C
      REAL THETAD(M,N),ETILDAD(M,N),VTILDAD(N,N), BTILDAD(N)
      REAL ETETD(M,N),VCORRD(N,N),VJD(N,N)VH1D(N,N)
      REAL VINVD(N,N),VH2D(N,N)
C
      REAL THETAA(M,N),ETILDAA(M,N), VTILDAA(N,N),BTILDAA(N)
      REAL ETETA(N,N),VCORRA(N,N),VJA(N,N)VH1A(N,N)
      REAL VINVA(N,N),VH2A(N,N)
C
      REAL THETAT(M,N),ETILDAT(M,N),VTILDAT(N,N),BTILDAT(N)
      REAL ETETT(N,N),VCORRT(N,N),VJT(N,N),VH1T(N,N)
      REAL PTILDAT(N,N),DELTAT(N)
C
      REAL WKAEREA(N)
C
C
C     READ DATA Y(I,J)
C
      DO 10 I=1,M
      READ (5,11)(Y(I,J),J=1,N)
10    CONTINUE
11    FORMAT(4F4.2)
C
C
C     COMPUTE INITIAL ESTIMATES BHAT,EHAT,VHAT
C
      DO 70 J=1,N
      BHAT(J)=0.0
      DO 71 I=1,M
      BHAT(J)=BHAT(J)+Y(I,J)
71    CONTINUE
      BHAT(J)=BHAT(J)/M
70    CONTINUE
C
      WRITE (6,77)(BHAT(I),I=1,N)
77    FORMAT(5H BHAT/,4F15.10)
C
      DO 72 I=1,M
      DO 73 J=1,N
```

```
      EHAT(I,J)=Y(I,J)-BHAT(J)
73    CONTINUE
72    CONTINUE
C
      WRITE(6,78)((EHAT(I,J),J=1,N),I=1,M)
78    FORMAT(5H EHAT/,50(4F8.2/))
      DO 74 I=1,N
      DO 75 J=1,N
      VHAT(I,J)=0.0
      DO 76 K=1,M
      VHAT(I,J)=VHAT(I,J)+EHAT(K,I)*EHAT(K,J)
76    CONTINUE
      VHAT(I,J)=VHAT(I,J)/M
75    CONTINUE
74    CONTINUE
C
      WRITE(6,79)((VHAT(I,J),J=1,N),I=1,N)
79    FORMAT(5H VHAT/,4(4F15.10/))
C
C     COMPUTE SVD OF VHAT
C
      DO 791 I=1,N
      DO 792 J=1,N
      PPRIME(I,J)=0.0
792   CONTINUE
791   CONTINUE
      DO 793 I=1,N
      PPRIME(I,I)=1.0
793   CONTINUE
C
      CALL LSVDF(VHAT,N,N,N,PPRIME,N,N,DELTA,IER)
C
      WRITE(6,794)((VHAT(I,J),J=1,N),I=1,N)
794   FORMAT(5H     P/,4(4F15.10/))
      WRITE(6,795)((PPRIME(I,J),J=1,N),I=1,N)
795   FORMAT(7H PPRIME/,4(4F15.10/))
      WRITE(6,796)(DELTA(I),I=1,N)
796   FORMAT(6H DELTA/,4F15.10/)
C
C     COMPUTE SCALED PC-RESIDUALS ESTAR(I,J)
C
      DO 230 I=1,M
      DO 231 J=1,N
      ESTAR(I,J)=0.0
      DO 232 K=1,N
      ESTAR (I,J)=ESTAR(I,J)+EHAT(I,K)*VHAT(K,J)
232   CONTINUE
      ESTAR (I,J)=ESTAR(I,J)/SQRT(DELTA(J))
231   CONTINUE
230   CONTINUE
C
      WRITE(6,233)((ESTAR(I,J),J=1,N),I=1,M)
233   FORMAT(/6H ESTAR/,50(4F8.2/))
C
      D=1.0
      DO 797 I=1,N
      D=D*DELTA(I)
```

```
797     CONTINUE
        WRITE(6,798)D
798     FORMAT(12H DETERMINANT,F25.20/)
C
C       READ OUTLIER-PATTERN NO(I), PA(I,J)
C
        DO 12 I=1,M
        READ(5,13)NO(I),(PA(I,J),J=1,N)
12      CONTINUE
13      FORMAT (2X,5I2)
C
C       DETERMINE COMPONENTS OF OUTLYING AND CLEAN DATA
C
        DO 200 J=1,M
        KO=1
        KC=1
        DO 201 I=1,N
        IF (PA(J,I).EQ.0) THEN
          C(J,KC)=I
          KC=KC+1
        ELSE
          O(J,KO)=I
          KO=KO+1
        ENDIF
201     CONTINUE
200     CONTINUE
C
C
C       STEP 0:
C
        DO 14 I=1,M
        DO 15 J=1,N
        THETAD(I,J)=0.0
        THETAA(I,J)=0.0
        THETAT(I,J)=0.0
15      CONTINUE
14      CONTINUE
C
        DO 16 I=1,N
        DO 17 J=1,N
        VTILDAD(I,J)=0.0
        VTILDAA(I,J)=0.0
        VTILDAT(I,J)=0.0
        PTILDAT(I,J)=0.0
17      CONTINUE
16      CONTINUE
        DO 18 I=1,N
        VTILDAD(I,I)=1.0
        VTILDAA(I,I)=1.0
        VTILDAT(I,I)=1.0
        PTILDAT(I,I)=1.0
        DELTAT(I)=1.0
18      CONTINUE
C
C
        DO 100 IT=1,20
C
C
```

```
C       STEP 1: (I) COMPUTE MEAN BTILDA
C
        DO 19 I=1,N
        BTILDAD(I)=0.0
        BTILDAA(I)=0.0
        BTILDAT(I)=0.0
        DO 20 J=1,M
        TH=0.0
        TT=0.0
        IF (NO(J).GT.0) THEN
          DO 21 K=1,NO(J)
          TH=TH+THETAD(J,O(J,K))*VTILDAD(O(J,K),I)
          TT=TT+THETAT(J,O(J,K))*PTILDAT(O(J,K),I)
21        CONTINUE
        ENDIF
        BTILDAD(I)=BTILDAD(I)+Y(J,I)-TH
        BTILDAA(I)=BTILDAA(I)+Y(J,I)-THETAA(J,I)
        BTILDAT(I)=BTILDAT(I)+Y(J,I)-TT
20      CONTINUE
        BTILDAD(I)=BTILDAD(I)/M
        BTILDAA(I)=BTILDAA(I)/M
        BTILDAT(I)=BTILDAT(I)/M
19      CONTINUE
C
        WRITE(6,28) (BTILDAD(I),I=1,N)
        WRITE(6,29)(BTILDAA(I),I=1,N)
        WRITE(6,281)(BTILDAT(I),I=1,N)
28      FORMAT(3H BD/,4F15.10)
29      FORMAT(3H BA/,4F15.10)
281     FORMAT(3H BT/,4F15.10)
C
C
C       STEP 1: (II) COMPUTE COVARIANCE MATRIX VTILDA
C
C       COMPUTE RESIDUALS ETILDA
C
        DO 22 I=1,M
        DO 23 J=1,N
        ETILDAD(I,J)=Y(I,J)-BTILDAD(J)
        ETILDAA(I,J)=Y(I,J)-BTILDAA(J)
        ETILDAT(I,J)=Y(I,J)-BTILDAT(J)
        IF (NO(I).GT.0) THEN
          TH=0.0
          TT=0.0
          DO 24 K=1,NO(I)
          TH=TH+THETAD(I,O(I,K))*VTILDAD(O(I,K),J)
          TT=TT+THETAT(I,O(I,K))*PTILDAT(O(I,K),J)
24        CONTINUE
          ETILDAD(I,J)=ETILDAD(I,J)-TH
          ETILDAA(I,J)=ETILDAA(I,J)-THETAA(I,J)
          ETILDAT(I,J)=ETILDAT(I,J)-TT
        ENDIF
23      CONTINUE
22      CONTINUE
C
C       COMPUTE ETILDA'ETILDA
C
        DO 25 I=1,N
```

```
      DO 26 J=1,N
      ETETD(I,J)=0.0
      ETETA(I,J)=0.0
      ETETT(I,J)=0.0
      DO 27 K=1,M
      ETETD(I,J)=ETETD(I,J)+ETILDAD(K,I)*ETILDAD(K,J)
      ETETA(I,J)=ETETA(I,J)+ETILDAA(K,I)*ETILDAA(K,J)
      ETETT(I,J)=ETETT(I,J)+ETILDAT(K,I)*ETILDAT(K,J)
27    CONTINUE
26    CONTINUE
25    CONTINUE
C
C     COMPUTE CORRECTION FOR BIAS VCORR
C
      DO 50 I=1,N
      DO 51 J=1,N
      VCORRD(I,J)=0.0
      VCORRA(I,J)=0.0
      VCORRT(I,J)=0.0
51    CONTINUE
50    CONTINUE
C
      DO 53 J=1,M
      IF (NO(J).GT.0) THEN
        DO 54 K1=1,N
        DO 55 K2=1,N
        VJD(K1,K2)=VTILDAD(K1,K2)
        VJA(K1,K2)=VTILDAA(K1,K2)
        VH1D(K1,K2)=VTILDAD(K1,K2)
        VH1A(K1,K2)=VTILDAA(K1,K2)
55      CONTINUE
54      CONTINUE
        IF (NO(J).LT.N) THEN
          DO 30 K1=1,(N-NO(J))
        DO 31 K2=1,(N-NO(J))
        VJD(C(J,K1),C(J,K2))=0.0
31      CONTINUE
30      CONTINUE
C
        DO 32 K1=1,(N-NO(J))
        DO 33 K2=1,NO(J))
        VH1D(C(J,K1),O(J,K2))=0.0
        VH1D(O(J,K2),C(J,K1))=0.0
        VH1A(C(J,K1),O(J,K2))=0.0
        VH1A(O(J,K2),C(J,K1))=0.0
33      CONTINUE
32      CONTINUE
C
        CALL LINV1F(VH1D,N,N,VINVD,5,WKAEREA,IER)
        CALL LINV1F(VH1A,N,N,VINVA,5,WKAEREA,IER)
C
        DO 34 K1=1,(N-NO(J))
        DO 35 K2=1,NO(J))
        VH2D(C(J,K1),O(J,K2))=0.0
        VH2A(O(J,K2),C(J,K1))=0.0
        DO 36 K=1,NO(J)
        VH2D(C(J,K1),O(J,K2))=VH2D(C(J,K1),O(J,K2))+
     *      VTILDAD(C(J,K1),O(J,K))*VINVD(O(J,K),O(J,K2))
        CONTINUE
        DO 361 K=1,(N-NO(J))
        VH2A(O(J,K2),C(J,K1))=VH2A(O(J,K2),C(J,K1))+
            VTILDAA(O(J,K2),C(J,K))*VINVA(C(J,K),C(J,K1))
```

```
36          CONTINUE
            CONTINUE
            CONTINUE
      *
361         DO 37 K1=1,(N-NO(J))
35          DO 38 K2=1,(N-NO(J))
34          VJD(C(J,K1),C(J,K2))=0.0
C           DO 39 K=1,NO(J)
            VJD(C(J,K1),C(J,K2))=VJD(C(J,K1),C(J,K2))+
               VH2D(C(J,K1),O(J,K))*VTILDAD(O(J,K),C(J,K2))
            CONTINUE
            CONTINUE
            CONTINUE
      *
39          DO 371 K1=1,NO(J)
38          DO 372 K2=1,NO(J)
37          VJA(O(J,K1),O(J,K2))=0.0
C           DO 391 K=1,(N-NO(J))
            VJA(O(J,K1),O(J,K2))=VJA(O(J,K1),O(J,K2))-
               VH2A(O(J,K1),C(J,K))*VTILDAA(C(J,K),O(J,K2))
            CONTINUE
            VJA(O(J,K1),O(J,K2))=VJA(O(J,K1),O(J,K2))+
               VTILDAA(O(J,K1),O(J,K2))
      *     CONTINUE
391         CONTINUE

      *   ENDIF
372
371         DO 375 K1=1,N
C           DO 376 K2=1,N
            VH1T(K1,K2)=PTILDAT(K2,K1)*DELTAT(K2)
C           CONTINUE
            CONTINUE

            DO 377 K1=1,N
376         DO 378 K2=1,N
375         VJT(K1,K2)=0.0
C           DO 379 K=1,NO(J)
            VJT(K1,K2)=VJT(K1,K2)+VH1T(K1,O(J,K))
               *PTILDAT(O(J,K),K2)
            CONTINUE
            CONTINUE

      *
379
378
377         CONTINUE
C
            DO 40 K1=1,N
            DO 41 K2=1,N
            VCORRD(K1,K2)=VCORRD(K1,K2)+VJD(K1,K2)
            VCORRT(K1,K2)=VCORRT(K1,K2)+VJT(K1,K2)
41          CONTINUE
40          CONTINUE
            DO 401 K1=1,NO(J)
            DO 411 K2=1,NO(J)
            VCORRA(O(J,K1),O(J,K2))=VCORRA(O(J,K1),O(J,K2))+
      *       VJA(O(J,K1),O(J,K2))
```

```
411       CONTINUE
401       CONTINUE
        ENDIF
53      CONTINUE
C
        DO 44 I=1,N
        DO 45 J=1,N
        VTILDAD(I,J)=(ETETD(I,J)+VCORRD(I,J))/M
        VTILDAA(I,J)=(ETETA(I,J)+VCORRA(I,J))/M
        VTILDAT(I,J)=(ETETT(I,J)+VCORRT(I,J))/M
45      CONTINUE
44      CONTINUE
C
C       COMPUTE SVD OF VTILDAT
C
        DO 430 I=1,N
        DO 431 J=1,N
        PTILDAT(I,J)=0.0
431     CONTINUE
        PTILDAT(I,I)=1.0
430     CONTINUE
C
        CALL LSVDF(VTILDAT,N,N,N,PTILDAT,N,N,DELTAT,IER)
C
C
C       STEP 1: (III) COMPUTE THETA
C
        DO 60 J=1,M
        IF (NO(J).GT.0) THEN
          DO 61 K1=1,N
          DO 62 K2=1,N
          VH1D(K1,K2)=VTILDAD(K1,K2)
          VH1A(K1,K2)=VTILDAA(K1,K2)
62        CONTINUE
61        CONTINUE
          IF (NO(J).LT.N) THEN
            DO 63 K1=1,(N-NO(J))
            DO 64 K2=1,NO(J))
            VH1D(C(J,K1),O(J,K2))=0.0
            VH1D(O(J,K2),C(J,K1))=0.0
            VH1A(C(J,K1),O(J,K2))=0.0
            VH1A(O(J,K2),C(J,K1))=0.0
64          CONTINUE
63          CONTINUE
C
            CALL LINV1F(VH1A,N,N,VINVA,5,WKAEREA,IER)
C
            DO 500 K1=1,(N-NO(J))
            DO 501 K2=1,NO(J))
            VH2A(C(J,K1),O(J,K2))=0.0
            DO 503 K=1,(N-NO(J))
            VH2A(C(J,K1),O(J,K2))=VH2A(C(J,K1),O(J,K2))+
     *      VINVA(C(J,K1),C(J,K))*VTILDAA(C(J,K),O(J,K2))
503         CONTINUE
501         CONTINUE
500         CONTINUE
            DO 504 K1=1,NO(J)
            TH=0.0
```

```
            DO 505 K=1,(N-NO(J))
            TH=TH+ETILDAA(J,C(J,K))*VH2A(C(J,K),O(J,K1))
505         CONTINUE
            THETAA(J,O(J,K1))=Y(J,O(J,K1))-BTILDAA(O(J,K1))-TH
504         CONTINUE
          ELSE
            DO 506 K=1,N
            THETAA(J,K)=Y(J,K)-BTILDAA(K)
506         CONTINUE
          ENDIF
C
          CALL LINV1F(VH1D,N,N,VINVD,5,WKAEREA,IER)
C
          DO 65 K1=1,NO(J)
          THETAD(J,O(J,K1))=0.0
          DO 66 K2=1,NO(J)
          THETAD(J,O(J,K1))=THETAD(J,O(J,K1))+
     *      (Y(J,O(J,K2))-BTILDAD(O(J,K2)))*
     *        VINVD(O(J,K2),O(J,K1))
66        CONTINUE
65        CONTINUE
C
          DO 67 K1=1,NO(J)
          THETAT(J,O(J,K1))=0.0
          DO 68 K2=1,N
          THETAT(J,O(J,K1))=THETAT(J,O(J,K1))+
     *      (Y(J,K2)-BTILDAT(K2))*PTILDAT(O(J,K1),K2)
68        CONTINUE
67        CONTINUE
C
        ENDIF
60      CONTINUE
C
100     CONTINUE
C
C       COMPUTE SVD OF VTILDAD
C
        DO 891 I=1,N
        DO 892 I=1,N
        PPRIME(I,J)=0.0
892     CONTINUE
891     CONTINUE
        DO 893 I=1,N
        PPRIME(I,I)=1.0
893     CONTINUE
C
        CALL LSVDF(VTILDAD,N,N,N,PPRIME,N,N,DELTA,IER)
C
        WRITE(6,894)((VTILDAD(I,J),J=1,N),I=1,N)
894     FORMAT(5H      PD/,4(4F15.10/))
        WRITE(6,895)((PPRIME(I,J),J=1,N),I=1,N)
895     FORMAT(7H PPRIMD/,4(4F15.10/))
        WRITE(6,896)(DELTA(I),I=1,N)
896     FORMAT(6H DELTD/,4(4F15.10/))
C
        DD=1.0
        DO 897 I=1,N
        DD=DD*DELTA(I)
```

```
897     CONTINUE
        WRITE(6,898)DD
898     FORMAT(12H DETERMINAND,F25.20/)
C
C       COMPUTE SVD OF VTILDAA
C
        DO 991 I=1,N
        DO 992 I=1,N
        PPRIME(I,J)=0.0
992     CONTINUE
991     CONTINUE
        DO 993 I=1,N
        PPRIME(I,I)=1.0
993     CONTINUE
C
        CALL LSVDF(VTILDAA,N,N,N,PPRIME,N,N,DELTA,IER)
C
        WRITE(6,994)((VTILDAA(I,J),J=1,N),I=1,N)
994     FORMAT(5H     PA/,4(4F15.10/))
        WRITE(6,995)(PPRIME(I,J),J=1,N),I=1,N)
995     FORMAT(7H   PPRIMA/,4(4F15.10/))
        WRITE(6,996)(DELTA(I),I=1,N)
996     FORMAT(6H   DELTA/,4(4F15.10/))
C
        DA=1.0
        DO 997 I=1,N
        DA=DA*DELTA(I)
997     CONTINUE
        WRITE(6,999)DA
999     FORMAT(12H   DETERMINANA,F25.20/)
C
        WRITE(6,522)((PTILDAT(I,J),J=1,N),I=1,N)
522     FORMAT(7H   PRIMET/,4(4F15.10/))
        WRITE(6,521)(DELTAT(I),I=1,N)
521     FORMAT(7H   DELTAT/,4F15.10)
        DT=1.0
        DO 69 I=1,N
        DT=DT*DELTAT(I)
69      CONTINUE
        WRITE(6,520)DT
520     FORMAT(12H   DETERMINATT,F25.20)
C
        CHID=M*ALOG(D/DD)
        CHIA=M*ALOG(D/DA)
        CHIT=M*ALOG(D/DT)
        WRITE(6.899)CHID,CHIA,CHIT
899     FORMAT(//5H   CHID,F15.10/5H   CHIA,F15.10/
     *             5H   CHIT,F15.10)
C
        END
```

As external routines were used the FORTRAN standard routines ALOG (natural logarithm) and SQRT (square root), as well as the routines LINV1F (computation of the inverse of a matrix) and LSVDF (singular value decomposition of a matrix) from the IMSL-Library (see IMSL, 1985).
The program is written in FORTRAN 5, and it was run on a CDC CYBER 170.

CHAPTER 5

A General Approach to Outliers

So far we have considered the problem of outliers in normal multivariate models, where three types of outlier were distinguished. We present now a general approach to outliers in multivariate models, distinguishing two types of outlier, *distributional* and *additive*. The concept of outliers in principal components, or *transformational* outliers, is closely linked to the assumption of multivariate normality, or at least approximative normality, and will thus not be considered in this general context.

It will turn out that the approach to D- and A-outliers in normal multivariate models, including the normal general linear model, is a special case of this general approach.

We consider the n-dimensional random vector $\mathbf{Y}$ with corresponding multivariate distribution function $F_{\mathbf{Y}}(\cdot)$. Let $\mathbf{y} \in \mathbb{R}^n$ be an observation of $\mathbf{Y}$, and let further $\mathbf{y}$ and $\mathbf{Y}$ be conformably partitioned as

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}, \quad \text{and } \mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix}, \tag{5.1}$$

where $\mathbf{y}_2$ and $\mathbf{Y}_2$ are of dimension k.

We denote the marginal distribution functions of $\mathbf{Y}_1$ and $\mathbf{Y}_2$ respectively by $F_{\mathbf{Y}_1}(\cdot)$ and $F_{\mathbf{Y}_2}(\cdot)$, and $F_{\mathbf{Y}_1}(\cdot \mid \mathbf{y}_2)$ denotes the distribution function of the conditional distribution of $\mathbf{Y}_1$ given $\mathbf{y}_2$, and similarly $F_{\mathbf{Y}_2}(\cdot \mid \mathbf{y}_1)$ denotes the distribution function of the conditional distribution of $\mathbf{Y}_2$ given $\mathbf{y}_1$.

The observations $\mathbf{y}_2$ are suspected to be outlying, and $\mathbf{y}_1$ is considered to be clean. Generally, we distinguish two types of outlier, distributional and additive.

Distributional outliers

We assume that the model for $\mathbf{y}$ is correct, that is, the distribution of $\mathbf{Y}$ is correctly specified by $F_{\mathbf{Y}}(\cdot)$, and $\mathbf{y}$ is a valid observation of $\mathbf{Y}$, but $\mathbf{y}_2$ is a rare observation from the tails of the distribution of $\mathbf{Y}_2$.

Then the test-statistic for the null hypothesis that $\mathbf{y}_2$ is not a D-outlier will be based on a comparison of $\mathbf{y}_2$ with $F_{\mathbf{Y}_2}(\cdot)$. If the null hypothesis is rejected, the reduced model for $\mathbf{y}_1$ is the conditional distribution of $\mathbf{Y}_1 \mid \mathbf{y}_2$, with distribution function $F_{\mathbf{Y}_1}(\cdot \mid \mathbf{y}_2)$. The adjusted data vector $\tilde{\mathbf{y}}$ is obtained as

$$\begin{bmatrix} \tilde{\mathbf{y}}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{y}_1 - [E(\mathbf{Y}_1 \mid \mathbf{y}_2) - E(\mathbf{Y}_1)] \\ E(\mathbf{Y}_2) \end{bmatrix}. \tag{5.2}$$

Additive outliers

Now we assume that $\mathbf{y}_2$ is not an observation from $\mathbf{Y}_2$, that is, $\mathbf{y}_2$ is arbitrary and totally unrelated to the underlying distribution of $\mathbf{Y}$. This situation can be modelled by assuming that an arbitrary and unknown vector λ has been added to the true observation $\mathbf{y}_2^*$, and only $\mathbf{y}_2 = \mathbf{y}_2^* + \lambda$ is available as data.

The test-statistic for the null hypothesis that $\mathbf{y}_2$ is not an A-outlier, or $\lambda = \mathbf{0}$, will be based on the comparison of $\mathbf{y}_2$ with $F_{\mathbf{Y}_2}(\cdot \mid \mathbf{y}_1)$. If the null hypothesis is rejected, the reduced model for $\mathbf{y}_1$ is the distribution of $\mathbf{Y}_1$ as before. The adjusted data $\tilde{\mathbf{y}}$ is given by

$$\begin{bmatrix} \tilde{\mathbf{y}}_1 \\ \tilde{\mathbf{y}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{y}_1 \\ E(\mathbf{Y}_2 \mid \mathbf{y}_1) \end{bmatrix} . \tag{5.3}$$

Comparison of D- and A-outliers

Table 5.1 presents a comparison of D- and A-outliers. We note the duality relationship between these two types of outlier, which was pointed out previously in Section 3.1.3 for the case of D- and A-outliers in the (normal) general linear model.

Table 5.1

Comparison of D- and A-outliers

	D-outlier	A-outlier
test-statistic	based on comparison of $\mathbf{y}_2$ with $F_{\mathbf{Y}_2}(\cdot)$	based on comparison of $\mathbf{y}_2$ with $F_{\mathbf{Y}_2}(\cdot \mid \mathbf{y}_1)$
reduced model for $\mathbf{y}_1$	distribution of $\mathbf{Y}_1 \mid \mathbf{y}_2$	distribution of $\mathbf{Y}_1$
adjusted data $\tilde{\mathbf{y}}$	$\tilde{\mathbf{y}} = \begin{bmatrix} \mathbf{y}_1 - [E(\mathbf{Y}_1 \mid \mathbf{y}_2) - E(\mathbf{Y}_1)] \\ E(\mathbf{Y}_2) \end{bmatrix}$	$\tilde{\mathbf{y}} = \begin{bmatrix} \mathbf{y}_1 \\ E(\mathbf{Y}_2 \mid \mathbf{y}_1) \end{bmatrix}$

□

Of course, the test-statistic for the null hypothesis that $\mathbf{y}_2$ is not an outlier can alternatively be based on the comparison of $T(\mathbf{y}_2)$ with $F_{T(\mathbf{Y}_2)}(\cdot)$ in the case of D-outliers, and with $F_{T(\mathbf{Y}_2)}(\cdot \mid \mathbf{y}_2)$ in the case of **A**-outliers, where $T(\cdot)$ is a statistic derived from $\mathbf{y}_2$.

Applications

When the distribution of $\mathbf{Y}$ is known, and thus the distribution function $F_{\mathbf{Y}}(\cdot)$, and the related marginal and conditional distributions, then tests for D- and A-outliers can directly be obtained as given above.

If the distribution of $\mathbf{Y}$ is not known but known to be a member of a family of distributions parametrized by $\theta \in \Theta$, i.e. the distribution function of $\mathbf{Y}$ is $F_{\mathbf{Y}}(\cdot \mid \theta)$, then we can estimate θ by $\tilde{\theta}$, and proceed as above with $F_{\mathbf{Y}}(\cdot \mid \tilde{\theta})$ and the related marginal and conditional distributions. The estimate $\tilde{\theta}$ is here obtained from $\mathbf{y}_1$ under the corresponding reduced model. Hence for D-outliers we estimate the parameter vector θ of $F_{\mathbf{Y}_1}(\cdot \mid \theta, \mathbf{y}_2)$ using the observation $\mathbf{y}_1$ of $\mathbf{Y}_1 \mid \mathbf{y}_2$, and for A-outliers we estimate the parameter vector θ of $F_{\mathbf{Y}_1}(\cdot \mid \theta)$, using the observation $\mathbf{y}_1$ of $\mathbf{Y}_1$.

We note that the approach to D- and A-outliers in the normal general linear model is a special case of this procedure, which is evident from the discussion in Chapter 3 of the reduced data models associated with the removal of D- and A-outliers respectively from the model.

However, in the general case, estimation in the reduced models may be difficult, or more complicated than in the complete models. But if the distribution function $F_{\mathbf{Y}}(\cdot \mid \theta)$ has a density $f_{\mathbf{Y}}(\cdot \mid \theta)$, then θ can be estimated in the full model using the EM-algorithm, i.e. by maximum likelihood.

The EM-algorithm proceeds as follows.

Algorithm 5.2 (Schall)

Step 0: Set $\tilde{\mathbf{y}} := \mathbf{y}$.

Step 1: (i) M-step: Estimate θ by its ML-estimate $\tilde{\theta}$ maximizing $L(\tilde{\mathbf{y}}, \theta) = f_{\mathbf{Y}}(\tilde{\mathbf{y}} \mid \theta)$.

(ii) E-step: Compute the adjusted data vector $\tilde{\mathbf{y}}$ as

$$\tilde{\mathbf{y}} = \begin{bmatrix} \mathbf{y}_1 - E(\mathbf{Y}_1 \mid \mathbf{y}_2, \tilde{\theta}) + E(\mathbf{Y}_1 \mid \tilde{\theta}) \\ E(\mathbf{Y}_2 \mid \tilde{\theta}) \end{bmatrix} ,$$

if $\mathbf{y}_2$ is suspected to be a D-outlier, or as

$$\tilde{\mathbf{y}} = \begin{bmatrix} \mathbf{y}_1 \\ E(\mathbf{Y}_2 \mid \mathbf{y}_1, \tilde{\theta}) \end{bmatrix} ,$$

if $\mathbf{y}_2$ is suspected to be an A-outlier.

Step 2: Proceed with step 1 until the parameter estimates stabilize.

□

If we can adjust for the outlier $\mathbf{y}_2$ by introducing a new parameter λ (say) for the distribution of $\mathbf{Y}$ (i.e. the density for the distribution of $\mathbf{Y}$ can be written as $f_{\mathbf{Y}}(\cdot \mid \theta, \lambda)$, and we obtain $f_{\mathbf{Y}}(\cdot \mid \theta)$ by setting $\lambda = \lambda_0$), as is the case with the normal distribution, then a LRT for the null hypothesis $H_0 : \lambda = \lambda_0$ can be obtained by comparing $\max_{\theta, \lambda} f_{\mathbf{Y}}(\mathbf{y} \mid \theta, \lambda)$ with $\max_{\theta} f_{\mathbf{Y}}(\mathbf{y} \mid \theta, \lambda_0)$.

Corresponding to two types of outlier we may distinguish two types of influence. If the distribution function of $\mathbf{Y}$ has a density $f_{\mathbf{Y}}(\cdot \mid \theta)$, $\theta \in \Theta$, we follow the approach by Cook and Weisberg (1982, pp. 182–186) and define a likelihood distance $LD_{(2)}$ with respect to the observations $\mathbf{y}_2$ as

$$LD_{(2)} = 2 \cdot [\ln f_{\mathbf{Y}}(\mathbf{y} \mid \hat{\theta}) - \ln f_{\mathbf{Y}}(\mathbf{y} \mid \tilde{\theta})] \ , \tag{5.4}$$

where $\hat{\theta}$ and $\tilde{\theta}$ are respectively the MLE's for θ under the observed data $\mathbf{y}$ and the adjusted data $\tilde{\mathbf{y}}$. If $\tilde{\mathbf{y}}$ is computed using (5.2), the influence of the possible D-outlier $\mathbf{y}_2$ is measured by (5.4), and similarly, if $\tilde{\mathbf{y}}$ is computed using (5.3), the influence of the possible A-outlier $\mathbf{y}_2$ is measured by (5.4). In a practical situation, $\tilde{\mathbf{y}}$ and $\tilde{\theta}$ may be computed using Algorithm 5.2 .

Cook and Weisberg (1982) suggest to calibrate $LD_{(2)}$ by comparison to the χ^2_p-distribution, where p is the dimension of the parameter vector θ. This comparison is motivated by the fact that an asymptotic $(1-\alpha) \cdot 100\%$ confidence region for θ is given by

$$C = \{\theta \mid 2 \cdot [\ln f_{\mathbf{Y}}(\mathbf{y} \mid \hat{\theta}) - \ln f_{\mathbf{Y}}(\mathbf{y} \mid \theta)] \leq \chi^2_p(1-\alpha)\} \ , \tag{5.5}$$

where $\chi^2_p(1-\alpha)$ is the $(1-\alpha)$-fractile of the χ^2_p-distribution.

BIBLIOGRAPHY

Aitken, A. C. (1935). On least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh*, **55**, 42–48.

Anderson, E. (1935). The irises of the Gaspe peninsula. *Bulletin of the American Iris Society*, **59**, 2–5.

Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, **52**, 200–203.

Andrews, D. F. (1972). Plots of high-dimensional data. *Biometrics*, **28**, 125–136.

Andrews, D. F. and Pregibon, D. (1978). Finding the outliers that matter. *Journal of the Royal Statistical Society, Ser. B*, **40**, 85–93.

Baksalary, J. K., Hanke, J. and Kala, R. (1980). Nonnegative definite solutions to some matrix equations occurring in distribution theory of quadratic forms. *Sankhyā, Ser. A*, **42**, 283–291.

Bhimasankaram, P. and Majumdar, D. (1980). Hermitian and nonnegative definite solutions of some matrix equations connected with distribution of quadratic forms. *Sankhyā, Ser. A*, **42**, 272–282.

Bose, R. C. (1944). The fundamental theorem of linear estimation (abstract). *Proceedings of 31st Indian Science Congress*, **3**, 5–6.

Businger, P. and Golub, G. H. (1965). Linear least squares solutions by Householder transformations. *Numerische Mathematik*, **7**, 269–276 [contribution I/8 in Wilkinson and Reinsch (1971)].

Chipman, J. S. (1964). On least squares with insufficient observations. *Journal of the American Statistical Association*, **59**, 1078–1111.

Cook, R. D. (1977). Detection of influential observations in regression. *Technometrics*, **19**, 15–18.

Cook, R. D. (1979). Influential observations in linear regression. *Journal of the American Statistical Association*, **74**, 169–174.

Cook, R. D. and Weisberg, S. (1979). Finding influential cases in linear regression: a review. Technical report No. 338, University of Minnesota, School of Statistics, St. Paul, Minnesota.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall, New York.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM-algorithm (with discussion). *Journal of the Royal Statistical Society, Ser. B*, **39**, 1–38.

Draper, N. R. (1961). Missing values in response surface designs. *Technometrics*, **3**, 389–398.

Draper, N. R. and John, J. A. (1981). Influential observations and outliers in regression. *Technometrics*, **23**, 21–26.

Dunne, T. T. (1982). Contributions to the theory of generalized inverses, the linear model and outliers. Unpublished PhD thesis, University of Cape Town, South Africa.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, 179–188.

Flury, B. N. (1983). Some relations between the comparisons of covariance matrices and principal component analysis. *Computational Statistics and Data Analysis*, **1**, 97–109.

Flury, B. N. (1984). Common principal components in k groups. *Journal of the American Statistical Association*, **79**, 892–898.

Flury, B. N. (1985). Two generalizations of the common principal components model. Technical report No. 11, Stanford University, Econometric Workshop, Stanford, California.

Gauß, C. F. (1821). *Theoria combinationis observationum erroribus minimus obnoxiae (pars prior)*. In: *Werke, IV*, (1973). Georg Olms Verlag, Hildesheim.

Gauß, C. F. (1823). *Theoria combinationis observationum erroribus minimus obnoxiae (pars posterior)*. In: *Werke, IV*, (1973). Georg Olms Verlag, Hildesheim.

Gentleman, J. F. and Wilk, M. B. (1975). Detecting outliers. II. Supplementing the direct analysis of residuals. *Biometrics*, **31**, 387–410.

Gnanadesikan, R. and Kettenring, J. R. (1972). Robust estimates, residuals and outlier detection with multiresponse data. *Biometrics*, **28**, 81–124.

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*, **68 B**, 151–172.

Golub, G. H. and Reinsch, C. (1970). Singular value decomposition and least squares solutions. *Numerische Mathematik*, **14**, 403–420 [contribution I/10 in Wilkinson and Reinsch (1971)].

Hawkins, D. M. (1974). The detection of errors in multivariate data using principal components. *Journal of the American Statistical Association*, **69**, 340–344.

Hawkins, D. M. (1980). *Identification of Outliers*. Chapman and Hall, London.

Hoaglin, D. C. and Welsh, R. (1978). The hat matrix in regression and ANOVA. *American Statistician*, **32**, 17–22.

IMSLIB (1985). International Mathematical and Statistical Library, Users Manual. IMSL Inc., Houston, Texas.

John, J. A. and Draper, N. R. (1978). On testing for two outliers or one outlier in two-way tables. *Technometrics*, **20**, 69–78.

Kariya, T. (1985). *Testing in the Multivariate General Linear Model*. Kinokuniya, Tokyo.

Kendall, M. G. and Stuart, A. (1973). *The Advanced Theory of Statistics*. Griffin, London.

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 a normal vector. *Journal of the Indian Statistical Association*, **1**, 61–70.

Khatri, C. G. (1966). A note on a MANOVA model applied to problems of growth curves. *Annals of the Institute of Statistical Mathematics*, **18**, 75–86.

Markoff, A. A. (1900). *Wahrscheinlichkeitsrechnung*. Teubner, Leipzig.

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.

Orchard, T. and Woodbury, M. A. (1972). A missing information principle: theory and applications. *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 697–715.

Potthoff, R. F. and Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve models. *Biometrika*, **51**, 313–326.

Pringle, R. M. and Rayner, A. A. (1971). *Generalized Inverse Matrices with Applications to Statistics*. Griffin, London.

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. (1972). Some recent results in linear estimation. *Sankhyā, Ser. A*, **34**, 369–378.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications*. Wiley, New York.

Rao, C. R. (1974). Projections, generalized inverses and BLUE's. *Journal of the Royal Statistical Society, Ser. A*, **36**, 442–448.

Rao, C. R. (1976). Estimation of parameters in a linear model. *Annals of Statistics*, **4**, 1023–1037.

Rubin, D. B. (1974). Characterizing the estimation of parameters in incomplete-data problems. *Journal of the American Statistical Association*, **69**, 467–474.

Schall, R. (1984). BLU-Schätzung im Linearen Modell bei unbekannten Varianzkomponenten. Unveröffentlichte Diplomarbeit, Universität Karlsruhe, Deutschland.

Schall, R. and Dunne, T. T. (1985). On outliers, recursive residuals, and missing data estimation for arbitrary known variance-covariance structure. Technical report No. 223, Stanford University, Department of Statistics, Stanford, California (submitted to *Annals of Statistics*).

Schall, R. and Dunne, T. T. (1986a). Transformation and reparametrization for best linear unbiased estimation of a linear model with arbitrary known variance. *Communications in Statistics*, to appear.

Schall, R. and Dunne, T. T. (1986b). A note on the chi-squaredness of quadratic forms. Technical report No. 225, Stanford University, Department of Statistics, Stanford, California (to appear in *Sankhyā, Ser. A*).

Schall, R. and Dunne, T. T. (1986c). A note on augmenting a linear model under arbitrary variance. Technical report No. 226, Stanford University, Department of Statistics, Stanford, California (submitted to *Communications in Statistics*).

Schall, R. and Dunne, T. T. (1986d). Additional variables and adjusted estimates with arbitrary known variance-covariance structure. Technical report No. 474, University of Minnesota, School of Statistics, St. Paul, Minnesota (submitted to *Communications in Statistics*).

Searle, S. R. (1971). *Linear Models*. Wiley, New York.

Siotani, M. (1959). The extreme value of the generalized distances of the individual points in the multivariate normal sample. *Annals of the Institute of Statistical Mathematics of Tokyo*, **10**, 183–206.

Wilkinson, J. H. and Reinsch, C. (1971). *Linear Algebra*. Springer-Verlag, Berlin.

Wilks, S. S. (1963). Multivariate statistical outliers. *Sankhyā, Ser. A*, **25**, 407–426.

Wu, C. F. J. (1983). On the convergence properties of the EM-algorithm. *Annals of Statistics*, **11**, 95–103.

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. 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.