*Proc. First Tampere Sem. Linear Models* (1983)
pp. 9—36, © Dept. of Math. Sci., Univ. Tampere, 1985

# A Unified Approach to Inference from Linear Models*

by

**C. Radhakrishna Rao**

**University of Pittsburgh, U.S.A., and
the Indian Statistical Institute, New Delhi, India**

## CONTENTS

* Based on an invited lecture series presented at The First International Tampere Seminar on Linear Statistical Models and their Applications, University of Tampere, Tampere, Finland, August 30 - September 2, 1983.

## 1. GENERALIZED INVERSE OF MATRICES

Generalized inverse of matrices has a wide range of applications in statistics; especially its applications in linear models are well known. It is therefore quite appropriate that g-inverses should be covered in a seminar or linear models. In this chapter we will recall what is known about g-inverses, and discuss some recent work on the subject, which provides a generalization and unification of the earlier results.

### 1.1 Background

The first major contribution in this area was made by E.H. Moore in 1920. He considered a linear transformation **A** which maps points in a vector space $\mathscr{V}$ into a vector space $\mathscr{W}$. These spaces may be of different dimensions. If transformation **A** is bijective, then there exists a unique inverse for it. But in cases where **A** is not bijective, Moore gave the following definition: **G** is "the general reciprocal" of **A** if it satisfies the conditions

$$\mathbf{AG} = \mathbf{P}_{\mathscr{A}}, \tag{1.1}$$

$$\mathbf{GA} = \mathbf{P}_{\mathscr{G}}, \tag{1.2}$$

where $\mathscr{A}$ is $\mathscr{R}(\mathbf{A})$, the range of **A**, $\mathscr{G}$ is the range of **G**, and $\mathbf{P}_{\mathscr{A}}$ and $\mathbf{P}_{\mathscr{G}}$ denote the orthogonal projection operators on $\mathscr{A}$ and $\mathscr{G}$ respectively.

This approach is basically geometric. However, Moore considered only vector spaces endowed with an inner product, and the concept of norm was involved in his definition.

In 1955 R. Penrose defined the generalized inverse of a matrix transformation. Let **A** be an $m \times n$ matrix with complex elements. Then the generalized inverse of **A**, denoted by **G**, is such that it satisfies the four conditions

$$\mathbf{AGA} = \mathbf{A}, \tag{1.3}$$

$$\mathbf{GAG} = \mathbf{G}, \tag{1.4}$$

$$\mathbf{AG} \text{ is hermitian}, \tag{1.5}$$

$$\mathbf{GA} \text{ is hermitian}. \tag{1.6}$$

Penrose's approach was purely algebraic. In the literature on g-inverse this kind of a definition is very often used.

In a way, Moore's and Penrose's definitions are quite similar. If we consider the vector spaces as Euclidean spaces $E^n$ and $E^m$ and we have the usual inner product $<\chi,\xi> = \mathbf{x}'\mathbf{y}$, then these two definitions are equivalent. On the whole Moore's definition is somewhat more general because the concept of a projection operator envisages a more general inner product, namely $<\mathbf{x},\mathbf{y}> = \mathbf{x}'\mathbf{U}\mathbf{y}$, where $\mathbf{U}$ is a positive definite matrix.

In 1955, C.R. Rao constructed an inverse of a singular matrix for use in computing least squares estimates of parameters in the Gauss-Markoff model and their variances and covariances. This inverse was different from the Moore-Penrose inverse.

C.R. Rao (1962) introduced a general definition of a g-inverse in the form $\mathbf{AGA} = \mathbf{A}$ and in 1967 provided a classification of generalized inverses as shown in Table 1.1, where $\mathbf{A}$ is an $m \times n$ matrix and $\mathscr{V} = E^n$ and $\mathscr{W} = E^m$ are Euclidean spaces furnished with inner products. All g-inverses in Table 1.1 can be obtained in seeking particular solutions to a nonhomogeneous equation $\mathbf{Ax} = \mathbf{y}$. If the equation is consistent, but the matrix $\mathbf{A}$ is singular or rectangular, then a solution to the equation can be obtained by writing $\mathbf{x} = \mathbf{A}^{-}\mathbf{y}$. Here $\mathbf{G} = \mathbf{A}^{-}$ is a matrix that satisfies the condition $\mathbf{AGA} = \mathbf{A}$, and it is simply called a g-inverse of $\mathbf{A}$ and denoted by $\mathbf{A}^{-}$ (Rao 1962).

Of course, there may be more than one solution to the equation $\mathbf{Ax} = \mathbf{y}$. The solution with the smallest norm can be found by using a different kind of g-inverse, namely the minimum norm g-inverse $\mathbf{A}_m^{-}$. Now $\mathbf{A}_m^{-}$ is a matrix $\mathbf{G}$ that satisfies $\mathbf{GA} = \mathbf{I} - \mathbf{P}_{\mathscr{K}}$, where $\mathscr{K}$ is the kernel space of $\mathbf{A}$ and $\mathbf{P}_{\mathscr{K}}$ is the orthogonal projector on $\mathscr{K}$ using any inner product associated with the norm (to be minimized).

If the equation $\mathbf{Ax} = \mathbf{y}$ is not consistent, then it can be solved by means of least squares, by minimizing the norm of $\mathbf{y} - \mathbf{Ax}$. Again, there is a g-inverse, the least squares g-inverse $\mathbf{A}_l^{-}$, that gives a solution to this optimization problem. This time $\mathbf{A}_l^{-}$ is a $\mathbf{G}$ that satisfies $\mathbf{AG} = \mathbf{P}_{\mathscr{A}}$, where $\mathbf{P}_{\mathscr{A}}$ is the orthogonal projector on $\mathscr{A}$ using the inner product associated with the norm.

Finally, if we have an inconsistent equation $\mathbf{Ax} = \mathbf{y}$ and we want to find a minimum norm least squares solution to it, this can be obtained by using

**TABLE 1.1**

*Special types of g-inverses*

| Notation | Name | Condition |
|---|---|---|
| $\mathbf{A}^{-}$ | g-inverse | $\mathbf{AGA} = \mathbf{A}$ |
| $\mathbf{A}_m^{-}$ | minimum norm g-inverse | $\mathbf{GA} = \mathbf{I} - \mathbf{P}_{\mathscr{K}}$<br>[$\mathscr{K}$ is the kernel space of $\mathbf{A}$] |
| $\mathbf{A}_l^{-}$ | least squares g-inverse | $\mathbf{AG} = \mathbf{P}_{\mathscr{A}}$ |
| $\mathbf{A}^{+}$ | minimum norm least squares g-inverse | $\mathbf{GA} = \mathbf{P}_{\mathscr{G}}$, $\mathbf{AG} = \mathbf{P}_{\mathscr{A}}$ |

another generalized inverse satisfying the conditions $\mathbf{GA} = \mathbf{P}_{\mathscr{G}}$ and $\mathbf{AG} = \mathbf{P}_{\mathscr{A}}$, which is the Moore-Penrose inverse.

Thus we have seen that study of the equation $\mathbf{Ax} = \mathbf{y}$ furnishes motivation for these kinds of definitions of g-inverses. The approach outlined in Rao (1967a) is partly algebraic and partly geometric.

For a detailed discussion of the different types of g-inverses listed in Table 1.1, their applications and generalizations, reference may be made to the book by Rao and Mitra (1971).

An entirely different approach is to define g-inverses through certain optimization problems (Rao 1980, p. 18). Let us suppose that $\mathbf{A}$ is a given matrix and $\mathbf{G}$ is a corresponding g-inverse. If the true inverse exists, then $\mathbf{AG}$ is the identity matrix. Otherwise, let us find a $\mathbf{G}$ that minimizes the difference between $\mathbf{I}$ and $\mathbf{AG}$, i.e.,

$$\min_{\mathbf{G}} \|\mathbf{I} - \mathbf{AG}\| = \|\mathbf{I} - \mathbf{AX}\|. \tag{1.7}$$

The minimizer $\mathbf{X}$ turns out to be the least squares g-inverse $\mathbf{A}_l^-$. If instead of (1.7) we minimize the difference between $\mathbf{I}$ and $\mathbf{GA}$, then the matrix $\mathbf{Y}$ that satisfies

$$\min_{\mathbf{G}} \|\mathbf{I} - \mathbf{GA}\| = \|\mathbf{I} - \mathbf{YA}\| \tag{1.8}$$

is the minimum norm g-inverse.

Finally, let us first find the class of $\mathbf{X}$'s that minimizes (1.7), and then use these $\mathbf{X}$'s in order to minimize $\|\mathbf{I} - \mathbf{XB}\|$ with respect to $\mathbf{B}$. Then the matrix $\mathbf{X}$ that satifies the conditions

$$\min_{\mathbf{G}} \|\mathbf{I} - \mathbf{AG}\| = \|\mathbf{I} - \mathbf{AX}\|, \tag{1.9}$$

$$\min_{\mathbf{B}} \|\mathbf{I} - \mathbf{XB}\| = \|\mathbf{I} - \mathbf{XA}\|, \tag{1.10}$$

simultaneously is the Moore-Penrose inverse $\mathbf{A}^+$.

It may look as if the solutions to these optimization problems depend upon the type of norm we use. However, it is shown in Rao (1980) that a wide class of norms, which von Neumann (1937) called unitarily invariant norms, give the same solutions.

## 1.2 $\mathscr{L}\mathscr{M}$ N-inverse

Let us consider a linear transformation

$$\mathbf{A} : \mathscr{V} \to \mathscr{W}.$$

In general **A** may not be a mapping onto the space $\mathscr{W}$ but only into it. In the following we shall find an inverse transformation of **A** which would be defined everywhere on the vector space $\mathscr{W}$. Let us suppose that the range space of **A**, denoted by $\mathscr{A}$, is a proper subspace in $\mathscr{W}$, and represent by $\mathscr{L}$ a direct complement of $\mathscr{A}$ in $\mathscr{W}$, i.e.,

$$\mathscr{L} \oplus \mathscr{A} = \mathscr{W}. \tag{1.11}$$

Further, let $\mathscr{K}$ be the kernel space of **A** (i.e., the set of all vectors **x** such that $\mathbf{Ax} = \mathbf{O}$), and let $\mathscr{M}$ be a direct complement of $\mathscr{K}$ in $\mathscr{V}$, i.e., we have the decomposition

$$\mathscr{M} \oplus \mathscr{K} = \mathscr{V}. \tag{1.12}$$

If $\mathbf{A} : \mathscr{V} \rightarrow \mathscr{W}$ is bijective, then there exists a unique inverse for it. Otherwise an inverse may be defined only in a special sense and for a specific purpose.

The complements $\mathscr{L}$ and $\mathscr{M}$ can be chosen, for instance, in the following way. Let **G** be a linear transformation that satisfies $\mathbf{AGA} = \mathbf{A}$. Then it is easy to show that

$$\mathscr{R}(\mathbf{GA}) \oplus \mathscr{K} = \mathscr{V} \tag{1.13}$$

and hence we can choose $\mathscr{R}(\mathbf{GA})$ for $\mathscr{M}$. Further, the mapping **A** restricted to $\mathscr{M}$,

$$\mathbf{A} \mid \mathscr{M} : \mathscr{M} \rightarrow \mathscr{A},$$

is bijective and it has a unique inverse which is

$$\mathbf{G} \mid \mathscr{A} : \mathscr{A} \rightarrow \mathscr{M}.$$

Let $\mathscr{L} = \mathscr{R}(\mathbf{I} - \mathbf{AG})$ so that $\mathscr{L}$ is a direct complement of $\mathscr{A}$. Then **G** maps the points in $\mathscr{L}$ to $\mathscr{K}$, i.e.,

$$\mathbf{G} \mid \mathscr{L} = (\mathbf{G} - \mathbf{GAG}) \mid \mathscr{L} = \mathbf{N} \mid \mathscr{L} : \mathscr{L} \rightarrow \mathscr{K}, \tag{1.14}$$

where $\mathbf{N} = \mathbf{G} - \mathbf{GAG}$. Thus we have found an inverse transformation of **A** defined in the entire vector space $\mathscr{W}$, which maps the points in the range space of **A** onto $\mathscr{M}$ and the points in $\mathscr{L}$ into the kernel space of **A**.

Now let us choose for $\mathscr{M}$ and $\mathscr{L}$ *any direct* complements of $\mathscr{K}$ and $\mathscr{A}$ respectively, and for **N** any specified linear transformation that takes points of $\mathscr{L}$ into $\mathscr{K}$ and the points of $\mathscr{A}$ into the null vector, and ask whether there is a **G** associated with them.

We show that such a **G** exists, is unique and $\mathscr{R}(\mathbf{GA}) = \mathscr{M}$, $\mathscr{R}(\mathbf{I} - \mathbf{GA}) = \mathscr{L}$ and $\mathbf{N} = \mathbf{G} - \mathbf{GAG}$. Figure 1.1 illustrates the situation.

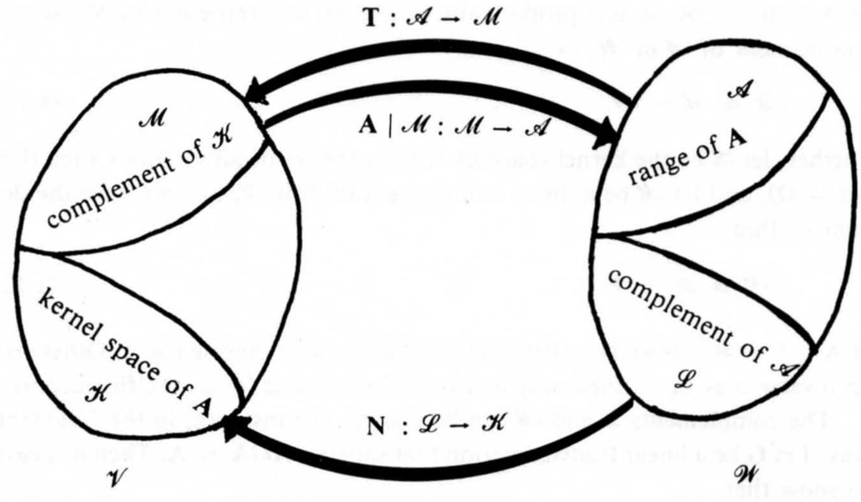
$\mathbf{T} : \mathscr{A} \to \mathscr{M}$
$\mathbf{A} \mid \mathscr{M} : \mathscr{M} \to \mathscr{A}$
$\mathscr{M}$
complement of $\mathscr{K}$
kernel space of $\mathbf{A}$
$\mathscr{K}$
$\mathscr{V}$
$\mathscr{A}$
range of $\mathbf{A}$
complement of $\mathscr{A}$
$\mathscr{L}$
$\mathbf{N} : \mathscr{L} \to \mathscr{K}$
$\mathscr{W}$

**Figure 1.1.** *Illustration of the $\mathscr{L}\,\mathscr{M}$ $\mathbf{N}$-inverse.*

Let us denote by $\mathbf{P}_{\mathscr{A}\cdot\mathscr{L}}$ the projection operator on $\mathscr{A}$ along a complementary subspace $\mathscr{L}$ and by $\mathbf{P}_{\mathscr{M}\cdot\mathscr{K}}$ the projection operator on $\mathscr{M}$ along $\mathscr{K}$. Notice that these operators need not be orthogonal projectors; the only condition is that the subspaces along which and on which we project are complementary and span the whole space. Now the following properties hold:

$$\mathbf{P}_{\mathscr{A}\cdot\mathscr{L}} + \mathbf{P}_{\mathscr{L}\cdot\mathscr{A}} = \mathbf{I}, \tag{1.15}$$

$$\mathbf{P}_{\mathscr{M}\cdot\mathscr{K}} + \mathbf{P}_{\mathscr{K}\cdot\mathscr{M}} = \mathbf{I}, \tag{1.16}$$

$$\mathbf{AP}_{\mathscr{K}\cdot\mathscr{M}} = \mathbf{O} \text{ and } \mathbf{AP}_{\mathscr{M}\cdot\mathscr{K}} = \mathbf{A}. \tag{1.17}$$

Using these projection operators we can define a g-inverse as follows (Rao and Yanai 1984). First we give a definition of an $\mathscr{L}\,\mathscr{M}$ $\mathbf{N}$-inverse specifying a restriction on $\mathbf{N}$.

DEFINITION 1.1. *Let $\mathscr{M}$ and $\mathscr{L}$ be any chosen complements of $\mathscr{K}$ and $\mathscr{A}$ in their respective spaces. The mapping $\mathbf{A} \mid \mathscr{M} : \mathscr{M} \to \mathscr{A}$ is bijective and has a unique inverse $\mathbf{T} : \mathscr{A} \to \mathscr{M}$. Further, let $\mathbf{N} : \mathscr{W} \to \mathscr{V}$ be a specified linear transformation such that* $\mathbf{AN} = \mathbf{O}$ and $\mathbf{NA} = \mathbf{O}$. *Then a linear transformation $\mathbf{G} : \mathscr{W} \to \mathscr{V}$ is said to be an $\mathscr{L}\,\mathscr{M}$ $\mathbf{N}$-inverse of $\mathbf{A}$ if*

$$\mathbf{G} \mid \mathscr{A} = \mathbf{T} \textit{ and } \mathbf{G} \mid \mathscr{L} = \mathbf{N} \mid \mathscr{L}, \tag{1.18}$$

*where the second condition can also be written as* $\mathbf{GP}_{\mathscr{L}\cdot\mathscr{A}} = \mathbf{N}$.

For any choice of $\mathscr{L}$, $\mathscr{M}$ and $\mathbf{N}$, there always exists an $\mathscr{L}\mathscr{M}\mathbf{N}$-inverse and it is unique. This uniqueness is easily shown, since if $\mathbf{G}_1$ and $\mathbf{G}_2$ are two $\mathscr{L}\mathscr{M}$ $\mathbf{N}$-inverses then we have

$$(\mathbf{G}_1 - \mathbf{G}_2) \mid \mathscr{A} = \mathbf{O} \text{ and } (\mathbf{G}_1 - \mathbf{G}_2) \mid \mathscr{L} = \mathbf{O} \tag{1.19}$$

which imply that $\mathbf{G}_1 = \mathbf{G}_2$. Its existence is verified by noticing that the matrix $\mathbf{G}$,

$$\mathbf{G} = \mathbf{G}_o\mathbf{A}\mathbf{G}_o + \mathbf{N}, \tag{1.20}$$

where

$$\mathbf{G}_o = \mathbf{P}_{\mathscr{M}\cdot\mathscr{K}}\mathbf{A}^{-}\mathbf{P}_{\mathscr{A}\cdot\mathscr{L}}, \tag{1.21}$$

satisfies the conditions for the $\mathscr{L}\mathscr{M}\mathbf{N}$-inverse.

Now let us examine how Definition 1.1 is related to the Moore-Penrose type of definition. It can be shown (Rao and Yanai 1984) that the following statements are all equivalent:

$$\mathbf{G} \text{ is an } \mathscr{L}\mathscr{M}\mathbf{N}\text{-inverse;} \tag{1.22}$$

$$\mathbf{GA} = \mathbf{P}_{\mathscr{M}\cdot\mathscr{K}} \text{ and } \mathbf{GP}_{\mathscr{L}\cdot\mathscr{A}} = \mathbf{N}; \tag{1.23}$$

$$\mathbf{GA} = \mathbf{P}_{\mathscr{M}\cdot\mathscr{K}}, \quad \mathbf{AG} = \mathbf{P}_{\mathscr{A}\cdot\mathscr{L}} \text{ and } \mathbf{G} - \mathbf{GAG} = \mathbf{N}; \tag{1.24}$$

$$\mathbf{AGA} = \mathbf{A},\ \mathscr{R}(\mathbf{G} \mid \mathscr{A}) = \mathscr{M} \text{ and } \mathbf{GP}_{\mathscr{L}\cdot\mathscr{A}} = \mathbf{N}. \tag{1.25}$$

From the condition (1.24) by taking $\mathbf{N} = \mathbf{O}$ we get

$$\mathbf{G} = \mathbf{GAG},\ \mathscr{M} = \mathscr{G}, \tag{1.26}$$

in which case

$$\mathbf{GA} = \mathbf{P}_{\mathscr{G}\cdot\mathscr{K}} \text{ and } \mathbf{AG} = \mathbf{P}_{\mathscr{A}\cdot\mathscr{L}}. \tag{1.27}$$

Thus the $\mathscr{L}\mathscr{M}\mathbf{N}$-inverse is a generalization of the definition given by Moore, since $\mathbf{G} - \mathbf{GAG}$ does not have to be $\mathbf{O}$ but can be chosen as any transformation $\mathbf{N}$ that takes the points in $\mathscr{L}$ into $\mathscr{K}$ and the projection operators need not be orthogonal. We represent an $\mathscr{L}\mathscr{M}\mathbf{N}$-inverse by $\mathbf{A}^{-}_{lmn}$ when $\mathbf{N} \neq \mathbf{O}$ and by $\mathbf{A}^{+}_{lm}$ when $\mathbf{N} = \mathbf{O}$.

Note that (1.27) by itself does not have a unique solution for $\mathbf{G}$. But uniqueness can be achieved by demanding conditions such as orthogonality of the projection operators as done by Moore.

If we specify only one or two of the arbitrary elements $\mathscr{L}$, $\mathscr{M}$ and $\mathbf{N}$, we get different types of g-inverses. For instance, if we specify $\mathscr{M}$ but neither $\mathscr{L}$ nor **N** we get an $\mathscr{M}$-inverse. Then the following statements are equivalent:

$$\mathbf{G} \text{ is an } \mathscr{M}\text{-inverse;} \tag{1.28}$$

$$\mathbf{GA} = \mathbf{P}_{\mathscr{M}\cdot\mathscr{K}}; \tag{1.29}$$

$$\mathbf{AGA} = \mathbf{A} \text{ and } \mathscr{R}(\mathbf{G} \mid \mathscr{A}) = \mathscr{M}. \tag{1.30}$$

We denote an $\mathscr{M}$-inverse by $\mathbf{A}_m^-$. Now again this definition does not involve the concept of inner product. But if we choose $\mathscr{V}$ as an inner product space and $\mathscr{M}$ as the orthogonal complement of $\mathscr{K}$, then we have

$$\min_{\mathbf{Ax}\,=\,\mathbf{y}} \|\mathbf{x}\| = \|\mathbf{A}_m^- \mathbf{y}\|, \tag{1.31}$$

where $\mathbf{Ax} = \mathbf{y}$ is a consistent equation. In other words, under these circumstances the $\mathscr{M}$-inverse is the minimum norm g-inverse of **A**. If $\mathbf{A}^-$ is a g-inverse satisfying $\mathbf{AA}^-\mathbf{A} = \mathbf{A}$, then an $\mathscr{M}$-inverse can be found by writing

$$\mathbf{G} = \mathbf{P}_{\mathscr{M}\cdot\mathscr{K}}\mathbf{A}^-. \tag{1.32}$$

Similarly, if we fix only $\mathscr{L}$, then we get the following equivalent statements:

$$\mathbf{G} \text{ is an } \mathscr{L}\text{-inverse;} \tag{1.33}$$

$$\mathbf{AG} = \mathbf{P}_{\mathscr{A}\cdot\mathscr{L}}; \tag{1.34}$$

$$\mathbf{AGA} = \mathbf{A} \text{ and } \mathbf{AGP}_{\mathscr{L}\cdot\mathscr{A}} = \mathbf{O}. \tag{1.35}$$

An $\mathscr{L}$-inverse is represented by $\mathbf{A}_l^-$. Now if $\mathscr{W}$ is an inner product space and $\mathscr{L}$ is the orthogonal complement of $\mathscr{A}$, then the $\mathscr{L}$-inverse equals the least squares inverse, i.e.,

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{Ax}\| = \|\mathbf{y} - \mathbf{AA}_l^- \mathbf{y}\|. \tag{1.36}$$

One solution to (1.34) can be obtained by writing

$$\mathbf{G} = \mathbf{A}^-\mathbf{P}_{\mathscr{A}\cdot\mathscr{L}}, \tag{1.37}$$

where $\mathbf{AA}^-\mathbf{A} = \mathbf{A}$.

Finally, if we specify both $\mathscr{L}$ and $\mathscr{M}$ then the following statements are equivalent:

$$\mathbf{G} \text{ is an } \mathscr{L}\,\mathscr{M}\text{-inverse;} \tag{1.38}$$

$$\mathbf{GA} = \mathbf{P}_{\mathscr{M}\cdot\mathscr{K}} \text{ and } \mathbf{AG} = \mathbf{P}_{\mathscr{A}\cdot\mathscr{L}}; \tag{1.39}$$

$$\mathbf{AGA} = \mathbf{A},\ \mathscr{R}(\mathbf{G} \mid \mathscr{A}) = \mathscr{M} \text{ and } \mathbf{AGP}_{\mathscr{L}\cdot\mathscr{A}} = \mathbf{O}. \tag{1.40}$$

An $\mathscr{L}\,\mathscr{M}$-inverse is denoted by $\mathbf{A}_{lm}^{-}$. A solution satisfying any of the conditions (1.38)—(1.40) can be obtained through

$$\mathbf{G} = \mathbf{P}_{\mathscr{M}\cdot\mathscr{K}}\mathbf{A}^{-}\mathbf{P}_{\mathscr{A}\cdot\mathscr{L}}, \tag{1.41}$$

where $\mathbf{A}^{-}$ is defined as before.

So we have seen that once we have an $\mathbf{A}^{-}$ satisfying $\mathbf{A}\mathbf{A}^{-}\mathbf{A} = \mathbf{A}$ we can derive all kinds of g-inverses by using projection operators. Rao and Yanai (1984) have given explicit expressions for $\mathbf{A}_{m}^{-}$, $\mathbf{A}_{l}^{-}$ and $\mathbf{A}_{lm}^{+}$.

## 1.3 Applications

As already mentioned, the g-inverses have a great variety of applications in statistics. Now let us review some useful results where the concept of a g-inverse is applied. One application is the explicit representation of projection operators. For instance the orthogonal projection operator on the range space of $\mathbf{A}$ can be expressed as

$$\mathrm{P}_{\mathscr{A}} = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-}\mathbf{A}', \tag{1.42}$$

when the inner product is defined as $<\mathbf{x},\mathbf{y}> = \mathbf{x}'\mathbf{y}$, and

$$\mathbf{P}_{\mathscr{A}} = \mathbf{A}(\mathbf{A}'\mathbf{U}\mathbf{A})^{-}\mathbf{A}'\mathbf{U}, \tag{1.43}$$

when the inner product is a more general one $<\mathbf{x},\mathbf{y}> = \mathbf{x}'\mathbf{U}\mathbf{y}$, where $\mathbf{U}$ is a positive definite matrix. [See Rao (1967a), where the expressions (1.42) and (1.43) were first reported.] By means of projection operators we get a result which is often used in the theory of linear models, namely

$$\mathbf{A} = \mathbf{P}_{\mathscr{A}}\mathbf{A} = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-}\mathbf{A}'\mathbf{A}. \tag{1.44}$$

Another field of application is the theory of multivariate normal distributions. The variance-covariance matrix of multivariate normal variable is in general assumed to be nonsingular, so that in the very definition of the density function there occurs the inverse of the variance-covariance matrix. However, the density function can be given, even if the variance-covariance matrix is singular, by using a generalized inverse (Rao 1973a, pp. 527—528).

It is often useful to study the matrix product $\mathbf{A}\mathbf{A}^{-}$. Although $\mathbf{A}^{-}$ is usually not unique, some elements of $\mathbf{A}\mathbf{A}^{-}$ may be. For instance, the *ith* column of $\mathbf{A}\mathbf{A}^{-}$ is the unit vector $\mathbf{e}_i$ (the *ith* column of I) if and only if the *ith* row vector of $\mathbf{A}$ is independent of the other row vectors of $\mathbf{A}$ (Rao 1981, p. 3). The conditions under which certain elements of $\mathbf{A}\mathbf{A}^{-}$ are unique are very important in multivariate analysis, e.g. in defining multiple correlation and partial correlations when the variance-covariance matrix is singular.

## 2. UNIFIED THEORY OF LINEAR ESTIMATION

### 2.1 The model

Let us consider the general Gauss-Markoff model

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon,\ \mathscr{E}(\epsilon) = \mathbf{O},\ \mathscr{D}(\epsilon) = \sigma^2\mathbf{V}, \tag{2.1}$$

where **X** and **V** are known matrices of order $n \times m$ and $n \times n$, respectively, and $\beta$ and $\sigma^2$ are unknown parameters. A heuristic principle for finding an estimate for $\beta$ is to look for a point in the expectation space, i.e., in the range space of **X**, that would be closest to the observed point **Y**. This leads us to minimize some chosen norm

$$\|\mathbf{Y} - \mathbf{X}\beta\| \tag{2.2}$$

with respect to $\beta$. Naturally, the expression (2.2) depends upon the type of norm we use. If we have the usual inner product, $<\mathbf{a},\mathbf{b}> = \mathbf{a}'\mathbf{b}$, then

$$\|\mathbf{Y} - \mathbf{X}\beta\|^2 = (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta), \tag{2.3}$$

which gives us the ordinary least squares solution. If instead of (2.3) we minimize

$$\|\mathbf{Y} - \mathbf{X}\beta\|^2 = (\mathbf{Y} - \mathbf{X}\beta)'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\beta), \tag{2.4}$$

then we have implicitly used an inner product of the form $<\mathbf{a},\mathbf{b}> = \mathbf{a}'\mathbf{V}^{-1}\mathbf{b}$. This approach leads to the Aitken least squares theory. We note, however, that this kind of definition requires **V** to be nonsingular. In the unified theory of linear estimation we determine the appropriate norm to be minimized in the general case, when we impose no restrictions on the rank of matrix **V**.

It is well known that in the ordinary and in Aitken's least squares theory the optimal point for (2.2) is obtained using a projection operator. If **P** is the orthogonal projection operator on the range space of **X**, then the point in $\mathscr{R}(\mathbf{X})$ that minimizes (2.2) is **PY**, which is the coordinate free estimator of $\mathscr{E}(\mathbf{Y})$. However, if an estimator of $\beta$ itself is required, then it is obtained from

$$\mathbf{PY} = \mathbf{X}\hat{\beta}. \tag{2.5}$$

Let us recall some facts about the model (2.1) in the general case. Denote by **Z** a matrix of maximum rank such that

$$\mathbf{X}'\mathbf{Z} = \mathbf{O}. \tag{2.6}$$

Then the subspaces $\mathscr{R}(\mathbf{X})$ and $\mathscr{R}(\mathbf{VZ})$ are disjoint and

$$\varrho(\mathbf{V} : \mathbf{X}) = \varrho(\mathbf{X} : \mathbf{VZ}),\ \varrho(\mathbf{VZ}) = \varrho(\mathbf{V} : \mathbf{X}) - \varrho(\mathbf{X}), \tag{2.7}$$

where $\varrho(\cdot)$ denotes the rank. Further, the observation vector $\mathbf{Y}$ has the property

$$\mathbf{Y} \in \mathscr{R}(\mathbf{V} : \mathbf{X}) = \mathscr{R}(\mathbf{X} : \mathbf{VZ}) \text{ with probability 1.} \tag{2.8}$$

The concept of *identifiability* is a central one in the discussion of linear models. A linear parametric function $\mathbf{p}'\beta$ is said to be identifiable if

$$\mathbf{p}'\beta_1 \neq \mathbf{p}'\beta_2 \Rightarrow \mathbf{X}\beta_1 \neq \mathbf{X}\beta_2 \tag{2.9}$$

for which a necessary and sufficient condition is that

$$\mathbf{p} \in \mathscr{R}(\mathbf{X}'). \tag{2.10}$$

Condition (2.10) is also the condition for unbiased estimability of $\mathbf{p}'\beta$ by a linear function of the observations. In general, identifiability is a much more important concept than unbiasedness; it does not make sense to estimate nonidentifiable parametric functions.

## 2.2 Generalized projection

We will first give the definition of a *generalized projection operator* (cf. Rao 1974, Rao and Yanai 1979). Let $\mathscr{A} = \mathscr{R}(\mathbf{A})$ and $\mathscr{B} = \mathscr{R}(\mathbf{B})$ be subspaces of $E^n$ such that

$$\mathscr{A} \cap \mathscr{B} = \{\mathbf{O}\} \tag{2.11}$$

and

$$\mathscr{A} + \mathscr{B} = \mathscr{R}(\mathbf{A} : \mathbf{B}) = \mathscr{S} \subset E^n, \tag{2.12}$$

where $\mathscr{S}$ need not be the whole space $E^n$. Now, every vector $\mathbf{u} \in \mathscr{S}$ has a unique decomposition

$$\mathbf{u} = \mathbf{u}_A + \mathbf{u}_B,\ \mathbf{u}_A \in \mathscr{A} \text{ and } \mathbf{u}_B \in \mathscr{B}. \tag{2.13}$$

We say that $\mathbf{P}_{\mathbf{A}\mid\mathbf{B}}$ is the generalized projection operator on $\mathscr{A}$ along $\mathscr{B}$ if

$$\mathbf{P}_{\mathbf{A}\mid\mathbf{B}}\mathbf{u} = \mathbf{u}_A \text{ for all } \mathbf{u} \in \mathscr{S}. \tag{2.14}$$

The projector $\mathbf{P}_{\mathbf{A}\mid\mathbf{B}}$ so defined need only satisfy the conditions

$$\mathbf{P}_{\mathbf{A}\mid\mathbf{B}}\mathbf{A} = \mathbf{A},\ \mathbf{P}_{\mathbf{A}\mid\mathbf{B}}\mathbf{B} = \mathbf{O}. \tag{2.15}$$

Note that $\mathbf{P}_{\mathbf{A}\mid\mathbf{B}}$ is not necessarily unique and it is not necessarily idempotent.

Let us now consider the model (2.1) where $\mathbf{X}$ and $\mathbf{V}$ may be deficient in rank. We want to find the linear function of $\mathbf{Y}$ that would estimate $\mathbf{X}\beta$ unbiasedly and have the smallest dispersion error. Let $\mathbf{PY}$ and $\mathbf{LY}$ be unbiased estimators of $\mathbf{X}\beta$, i.e.,

$$\mathscr{E}(\mathbf{PY}) = \mathbf{PX}\beta = \mathbf{X}\beta \tag{2.16}$$

and

$$\mathscr{E}(\mathbf{LY}) = \mathbf{LX}\beta = \mathbf{X}\beta. \tag{2.17}$$

Now (2.16) and (2.17) hold if

$$\mathbf{PX} = \mathbf{X} \text{ and } \mathbf{LX} = \mathbf{X}. \tag{2.18}$$

Note that in the case of singular **V** the condition $\mathbf{PX} = \mathbf{X}$ is actually only sufficient for the unbiasedness of **PY**. This is so because of the restrictions due to the singularity of **V** which become known when the observations are available; see e.g. Rao (1973b). However, if there are other unbiased estimators not satisfying the condition (2.18), then they are equivalent, with probability one, to those satisfying the condition (2.18). So there is no loss of generality in using the condition (2.18) as necessary and sufficient for unbiasedness.

An unbiased estimator **PY** is said to have the smallest dispersion error if

$$\mathscr{E}(\mathbf{PY} - \mathbf{X}\beta)(\mathbf{PY} - \mathbf{X}\beta)' \leq \mathscr{E}(\mathbf{LY} - \mathbf{X}\beta)(\mathbf{LY} - \mathbf{X}\beta)' \tag{2.19}$$

for all unbiased estimators **LY**. Such an estimator **PY** is called the *minimum dispersion unbiased estimator* (MDUE) of $\mathbf{X}\beta$. Condition (2.19) means that the difference between the dispersion matrix of **LY** and that of **PY** is nonnegative definite (n.n.d.). This is a very strong requirement. For instance, it can be shown that if (2.19) holds, then

$$\mathscr{E}(\mathbf{PY} - \mathbf{X}\beta)'\mathbf{B}(\mathbf{PY} - \mathbf{X}\beta) \leq \mathscr{E}(\mathbf{LY} - \mathbf{X}\beta)'\mathbf{B}(\mathbf{LY} - \mathbf{X}\beta) \tag{2.20}$$

for all n.n.d. matrices **B**. Thus **PY** minimizes simultaneously every compound loss function of the form (2.20), e.g., the trace of the dispersion matrix. Similarly any monotone function of the eigenvalues of the dispersion matrix will be minimized.

Let **P** and **L** satisfy the condition (2.18). Since

$$(\mathbf{L} - \mathbf{P})\mathbf{X} = \mathbf{O}, \tag{2.21}$$

**L** can be written in the form

$$\mathbf{L} = \mathbf{P} + \mathbf{M}, \tag{2.22}$$

with

$$\mathbf{MX} = \mathbf{O}. \tag{2.23}$$

With this notation the condition (2.19) becomes

$$\begin{aligned}\mathscr{E}(\mathbf{PY} - \mathbf{X}\beta)(\mathbf{PY} - \mathbf{X}\beta)' &\leq \mathscr{E}[(\mathbf{P} + \mathbf{M})\mathbf{Y} - \mathbf{X}\beta][(\mathbf{P} + \mathbf{M})\mathbf{Y} - \mathbf{X}\beta]' \\ &= \mathscr{E}(\mathbf{PY} - \mathbf{X}\beta)(\mathbf{PY} - \mathbf{X}\beta)' + \mathscr{E}\mathbf{M}(\mathbf{Y} - \mathbf{X}\beta)(\mathbf{Y} - \mathbf{X}\beta)'\mathbf{M}' \\ &\quad + \mathscr{E}\mathbf{P}(\mathbf{Y} - \mathbf{X}\beta)(\mathbf{Y} - \mathbf{X}\beta)'\mathbf{M}' + \mathscr{E}\mathbf{M}(\mathbf{Y} - \mathbf{X}\beta)(\mathbf{Y} - \mathbf{X}\beta)'\mathbf{P}'.\end{aligned} \tag{2.24}$$

A necessary and sufficient condition for the inequality (2.24) to hold for all $\mathbf{M}$ that satisfy $\mathbf{MX} = \mathbf{O}$ (cf. Rao 1973a, p. 317), is that

$$\mathscr{E}\mathbf{P}(\mathbf{Y} - \mathbf{X}\beta)(\mathbf{Y} - \mathbf{X}\beta)'\mathbf{M}' = \mathbf{PVM}' = \mathbf{O}. \tag{2.25}$$

Therefore $\mathbf{PY}$ is the MDUE of $\mathbf{X}\beta$ if and only if

$$\begin{aligned} &\mathbf{PX} = \mathbf{X},\ \mathbf{PVM}' = \mathbf{O}\ \forall\ \mathbf{M} : \mathbf{MX} = \mathbf{O} \\ \Leftrightarrow\ &\mathbf{PX} = \mathbf{X},\ \mathscr{R}(\mathbf{VP}') \subset \mathscr{R}(\mathbf{X}) \\ \Leftrightarrow\ &\mathbf{PX} = \mathbf{X},\ \mathbf{PVZ} = \mathbf{O}, \end{aligned} \tag{2.26}$$

where $\mathbf{Z} = \mathbf{X}^{\perp}$ is a matrix as defined in (2.6). We recognize that, according to (2.15), the matrix $\mathbf{P}$ that satisfies conditions (2.26) is the generalized projection operator on $\mathscr{R}(\mathbf{X})$ along $\mathscr{R}(\mathbf{VZ})$, denoted by $\mathbf{P}_{\mathbf{X}\,|\,\mathbf{VZ}}$, i.e., the MDUE of $\mathbf{X}\beta$ is $\mathbf{P}_{\mathbf{X}\,|\,\mathbf{VZ}}\,\mathbf{Y}$ which is the projection of $\mathbf{Y}$ on the column space of $\mathbf{X}$.

Solving for $\mathbf{P}$ from the equations (2.26) we get the following expressions:

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}, \text{ when } \mathbf{V} \text{ is nonsingular}, \tag{2.27}$$

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}', \text{ when } \mathbf{V} = \mathbf{I}, \tag{2.28}$$

and in general, whatever may be the ranks of $\mathbf{X}$ and $\mathbf{V}$,

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{T}\mathbf{X})^{-}\mathbf{X}'\mathbf{T}, \tag{2.29}$$

where

$$\mathbf{T} = (\mathbf{V} + \mathbf{XUX}')^{-} \tag{2.30}$$

and $\mathbf{U}$ is any matrix that satisfies

$$\varrho(\mathbf{V} + \mathbf{XUX}') = \varrho(\mathbf{V} : \mathbf{X}). \tag{2.31}$$

There always exists a $\mathbf{U}$ that satisfies the condition (2.31). For instance, the choice $\mathbf{U} = \mathbf{I}$ has this property.

Next we consider the parametric function $\mathbf{p}'\beta$. We assume that $\mathbf{p} \in \mathscr{R}(\mathbf{X}')$ or in other words that $\mathbf{p}'\beta$ is identifiable. Let us denote $\mathbf{P} = \mathbf{P}_{\mathbf{X}\,|\,\mathbf{VZ}}$. Then the minimum variance unbiased estimator (MVUE) of $\mathbf{p}'\beta$ is $\lambda'\mathbf{PY}$ if

$$\mathbf{p} = \mathbf{X}'\lambda. \tag{2.32}$$

It is easy to check that this holds. Since

$$\mathscr{E}(\lambda'\mathbf{PY}) = \lambda'\mathbf{PX}\beta = \lambda'\mathbf{X}\beta = \mathbf{p}'\beta, \tag{2.33}$$

the expression $\lambda'\mathbf{PY}$ is unbiased for $\mathbf{p}'\beta$. If $\lambda_1$ and $\lambda_2$ are two vectors satisfying (2.32), then

$$(\lambda_1' - \lambda_2')\mathbf{PX} = \mathbf{O}' \text{ and } (\lambda_1' - \lambda_2')\mathbf{PVZ} = \mathbf{O}' \tag{2.34}$$

and therefore

$$(\lambda_1' - \lambda_2')\mathbf{PY} = \mathrm{O} \text{ with probability } 1. \tag{2.35}$$

Thus, although there may be several vectors $\lambda$ that satisfy (2.32), the estimator $\lambda'\mathbf{PY}$ is always unique. Finally, we will check the minimum variance property. Let $\mathbf{q}'\mathbf{Y}$ be another unbiased estimator of $\mathbf{p}'\beta$. The unbiasedness condition

$$\mathscr{E}(\mathbf{q}'\mathbf{Y}) = \mathbf{q}'\mathbf{X}\beta = \mathbf{p}'\beta \tag{2.36}$$

implies that

$$\mathbf{p} = \mathbf{X}'\mathbf{q}. \tag{2.37}$$

It is easily seen that $\mathbf{q}'\mathbf{PY}$ is another unbiased estimator of $\mathbf{p}'\beta$. Now, because $\mathbf{PY}$ is the MDUE of $\mathbf{X}\beta$, we have

$$\mathscr{E}(\mathbf{Y} - \mathbf{X}\beta)(\mathbf{Y} - \mathbf{X}\beta)' \geq \mathscr{E}\mathbf{P}(\mathbf{Y} - \mathbf{X}\beta)(\mathbf{Y} - \mathbf{X}\beta)'\mathbf{P}' \tag{2.38}$$

and therefore

$$\mathscr{E}\mathbf{q}'(\mathbf{Y} - \mathbf{X}\beta)(\mathbf{Y} - \mathbf{X}\beta)'\mathbf{q} \geq \mathscr{E}\mathbf{q}'\mathbf{P}(\mathbf{Y} - \mathbf{X}\beta)(\mathbf{Y} - \mathbf{X}\beta)'\mathbf{P}'\mathbf{q}, \tag{2.39}$$

or in other words, $\mathbf{q}'\mathbf{PY}$ has a smaller variance than $\mathbf{q}'\mathbf{Y}$. But since $\mathbf{q}'\mathbf{PY}$ equals $\lambda'\mathbf{PY}$ for all $\mathbf{q}$ satisfying (2.37), we note that $\lambda'\mathbf{PY}$ is the MVUE of $\mathbf{p}'\beta$.

Substituting for $\mathbf{P}$ the explicit expression given in (2.29) we get

$$\begin{aligned}\lambda'\mathbf{PY} &= \lambda'\mathbf{X}(\mathbf{X}'\mathbf{TX})^{-}\mathbf{X}'\mathbf{TY} \\ &= \mathbf{p}'(\mathbf{X}'\mathbf{TX})^{-}\mathbf{X}'\mathbf{TY} = \mathbf{p}'\hat{\beta},\end{aligned} \tag{2.40}$$

defining

$$\hat{\beta} = (\mathbf{X}'\mathbf{TX})^{-}\mathbf{X}'\mathbf{TY}. \tag{2.41}$$

We note that the $\hat{\beta}$ as defined in (2.41) minimizes

$$(\mathbf{Y} - \mathbf{X}\beta)'\mathbf{T}(\mathbf{Y} - \mathbf{X}\beta) = (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{V} + \mathbf{XUX}')^{-}(\mathbf{Y} - \mathbf{X}\beta). \tag{2.42}$$

Thus we have found a generalization of the Gauss—Markoff—Aitken theory of least squares, applicable to all situations.

The estimator $\hat{\beta}$ in (2.41) can always be used whether $\mathbf{V}$ is singular or not. If the subspaces $\mathscr{R}(\mathbf{X})$ and $\mathscr{R}(\mathbf{VZ})$ do not cover the whole space $E^n$, then $\hat{\beta}$ may not be unique; it depends e.g. upon the choice of generalized inverse in (2.41). Nevertheless, the estimator $\mathbf{p}'\hat{\beta}$ will always be unique.

Thus far we have only studied the estimation of parameter $\beta$. An estimator of the other unknown parameter $\sigma^2$ can be obtained in a way similar to that in the nonsingular case. An unbiased estimator of $\sigma^2$ is

$$\begin{aligned}\hat{\sigma}^2 &= f^{-1}(\mathbf{Y} - \mathbf{X}\hat{\beta})'\mathbf{T}(\mathbf{Y} - \mathbf{X}\hat{\beta}) \\ &= f^{-1}\mathbf{Y}'\mathbf{T}(\mathbf{I} - \mathbf{P})\mathbf{Y},\end{aligned} \tag{2.43}$$

where $\mathbf{P}$ is given in (2.29) and

$$f = \varrho(\mathbf{V} : \mathbf{X}) - \varrho(\mathbf{X}) \tag{2.44}$$

denotes the degrees of freedom.

The variances and covariances of the estimators are of the form

$$\mathscr{Var}(\mathbf{p}'\hat{\beta}) = \sigma^2\mathbf{p}'[(\mathbf{X}'\mathbf{T}\mathbf{X})^- - \mathbf{U}]\mathbf{p} \tag{2.45}$$

and

$$\mathscr{Cov}(\mathbf{p}'\hat{\beta},\mathbf{q}'\hat{\beta}) = \sigma^2\mathbf{p}'[(\mathbf{X}'\mathbf{T}\mathbf{X})^- - \mathbf{U}]\mathbf{q}. \tag{2.46}$$

It is worth noting how matrix $\mathbf{U}$ appears in these representations. Some papers have been published giving wrong results because the role of the matrix $\mathbf{U}$ has been ignored. In Section 2.3 it will be shown how the problem of estimation can be treated without introducing the matrix $\mathbf{U}$.

## 2.3 Least squares with restrictions

The first ones to consider the least squares estimation with a singular $\mathbf{V}$ were A. J. Goldman and M. Zelen in 1964. They reduced the problem to one of least squares theory with restrictions on the parameters. Certain refinements to this method have been introduced by Rao and Mitra (1971).

The singularity of $\mathbf{V}$ means that there exists a matrix $\mathbf{N} \neq \mathbf{O}$ of rank $n - \varrho(\mathbf{V})$ such that

$$\mathbf{N}'\mathbf{V} = \mathbf{O}, \tag{2.47}$$

which implies that

$$\mathbf{N}'\mathbf{Y} = \mathbf{N}'\mathbf{X}\beta \text{ with probability 1.} \tag{2.48}$$

Let $\mathbf{V}^-$ be any g-inverse of $\mathbf{V}$ and let $\hat{\beta}$ be such that

$$\min_{\mathbf{N}'\mathbf{X}\beta = \mathbf{N}'\mathbf{Y}} (\mathbf{Y}-\mathbf{X}\beta)'\mathrm{V}^-(\mathbf{Y}-\mathbf{X}\beta) = (\mathbf{Y}-\mathbf{X}\hat{\beta})'\mathbf{V}^-(\mathbf{Y}-\mathbf{X}\hat{\beta}) = R_o^2. \tag{2.49}$$

The value of $\hat{\beta}$ depends on the g-inverse we use in (2.49). But despite this, the expression $\mathbf{p}'\hat{\beta}$ is unique and it turns out to be the MVUE of the parametric function $\mathbf{p}'\beta$. Similarly, the expression $f^{-1}R_o^2$, where $f = \varrho(\mathbf{V} : \mathbf{X}) - \varrho(\mathbf{X})$, does not depend on the choice of $\mathbf{V}^-$ in (2.49); $f^{-1}R_o^2$ is an unbiased estimator of $\sigma^2$. If the observation vector $\mathbf{Y}$ is normally distributed, i.e.,

$$\mathbf{Y} \sim N_n(\mathbf{X}\beta,\sigma^2\mathbf{V}), \tag{2.50}$$

then the distribution of $R_o^2$ is

$$R_o^2 \sim \sigma^2\chi^2(f). \tag{2.51}$$

In order to test a consistent hypothesis

$$\mathbf{K}'\beta = \mathbf{w} \tag{2.52}$$

we impose another set of restrictions on (2.49) and solve

$$R_1{}^2 = \min_{\substack{\mathbf{N}'\mathbf{X}\beta = \mathbf{N}'\mathbf{Y} \\ \mathbf{K}'\beta = \mathbf{w}}} (\mathbf{Y} - \mathbf{X}\beta)'\mathbf{V}^-(\mathbf{Y} - \mathbf{X}\beta). \tag{2.53}$$

Now,

$$R_1{}^2 - R_o{}^2 \sim \sigma^2\chi^2(h),\ h = \varrho[\mathscr{D}(\mathbf{K}'\hat{\beta})] \tag{2.54}$$

and is independent of $R_o{}^2$, and therefore

$$\frac{R_l{}^2 - R_o{}^2}{h} \div \frac{R_o{}^2}{f} \sim F(h,f). \tag{2.55}$$

The expression in (2.55) has central $F$ distribution with $h$ and $f$ degrees of freedom when the hypothesis is true and otherwise noncentral.

## 2.4 The Inverse Partitioned Matrix (IPM) approach

Another unified approach to linear estimation is the IPM method developed by C. R. Rao (1971). Again we consider the model (2.1), where **X** and **V** may be deficient in rank. Let

$$\begin{pmatrix} \mathbf{V} & \mathbf{X} \\ \mathbf{X}' & \mathbf{O} \end{pmatrix}^- = \begin{pmatrix} \mathbf{C}_1 & \mathbf{C}_2' \\ \mathbf{C}_3 & -\mathbf{C}_4 \end{pmatrix} \tag{2.56}$$

for any g-inverse. Now it can be shown e.g. that a parametric function $\mathbf{p}'\beta$ is estimable only if

$$\mathbf{p}'\mathbf{C}_2\mathbf{X} = \mathbf{p}' \text{ or } \mathbf{p}'\mathbf{C}_3\mathbf{X} = \mathbf{p}'. \tag{2.57}$$

The MVUE of an identifiable function $\mathbf{p}'\beta$ is $\mathbf{p}'\hat{\beta}$, where

$$\hat{\beta} = \mathbf{C}_2\mathbf{Y} \text{ or } \mathbf{C}_3\mathbf{Y} \text{ (which may not be the same).} \tag{2.58}$$

The dispersion matrix of $\hat{\beta}$ is $\sigma^2\mathbf{C}_4$ in the sense that

$$\begin{cases} \mathscr{V}\!ar(\mathbf{p}'\hat{\beta}) = \sigma^2\mathbf{p}'\mathbf{C}_4\mathbf{p}, \\ \mathscr{C}\!ov(\mathbf{p}'\hat{\beta},\mathbf{q}'\hat{\beta}) = \sigma^2\mathbf{p}'\mathbf{C}_4\mathbf{q} = \sigma^2\mathbf{q}'\mathbf{C}_4\mathbf{p}. \end{cases} \tag{2.59}$$

For $\sigma^2$ we obtain an unbiased estimator as

$$\hat{\sigma}^2 = f^{-1}\mathbf{Y}'\mathbf{C}_1\mathbf{Y}, \text{ where } f = \varrho(\mathbf{V} : \mathbf{X}) - \varrho(\mathbf{X}). \tag{2.60}$$

Hence it can be seen that through an inverse partitioned matrix of the form (2.56) we get all the necessary ingredients for the estimation of parameters. In a similar way, the IPM approach can be used in inference problems.

Let $\mathbf{S}'\hat{\beta}$ be the vector of the MVUEs of $k$ estimable parametric functions $\mathbf{S}'\beta$ and let $R_o^2 = \mathbf{Y}'\mathbf{C}_1\mathbf{Y}$. If $\mathbf{Y} \sim N_n(\mathbf{X}\beta,\sigma^2\mathbf{V})$, then $\mathbf{S}'\hat{\beta}$ and $R_o^2$ are independently distributed,

$$\mathbf{S}'\hat{\beta} \sim N_k(\mathbf{S}'\beta,\sigma^2\mathbf{S}'\mathbf{C}_4\mathbf{S}),\ R_o^2 \sim \sigma^2\chi^2(f), \tag{2.61}$$

where $f$ is as defined in (2.60). Further, let

$$\mathbf{S}'\beta = \mathbf{w} \tag{2.62}$$

be a null hypothesis. The hypothesis is consistent if and only if

$$\mathbf{G}\mathbf{G}^{-}\mathbf{u} = \mathbf{u},\ \mathbf{u} = \mathbf{S}'\hat{\beta} - \mathbf{w}, \tag{2.63}$$

where

$$\mathbf{G} = \mathbf{S}'\mathbf{C}_4\mathbf{S}. \tag{2.64}$$

The test statistic for the consistent hypothesis (2.62) is

$$\mathbf{F} = \frac{\mathbf{u}'\mathbf{G}^{-}\mathbf{u}}{h} \div \frac{R_o^2}{f},\ h = \varrho(\mathbf{G}), \tag{2.65}$$

and it has central $F$ distribution with $h$ and $f$ degrees of freedom when the hypothesis is true and otherwise noncentral.

### 2.5 Errors in observations

In the linear model an elementary assumption is that the observation vector $\mathbf{Y}$ belongs to the subspace generated by the columns of $\mathbf{X}$ and $\mathbf{V}$. However, in a practical estimation situation there may be errors in observations, e.g. rounding-off error, so that $\mathbf{Y}$ may not actually be confined to this particular subspace. Therefore it appears to be a good procedure to project the observed vector $\mathbf{Y}$ onto $\mathscr{R}(\mathbf{V} : \mathbf{X})$ and then use this projection of $\mathbf{Y}$ in the computations (Rao 1978). As a matter of fact, such correction of observations can be carried out by making certain modifications in the generalized inverse we use in estimation.

## 3. ALTERNATIVE CRITERIA OF ESTIMATION

The estimation criteria used in Chapter 2, such as unbiasedness and minimum variance, are sometimes criticized. In this chapter we introduce an alternative set of criteria where we do not use any of these customary concepts.

Let us consider the linear model (2.1), where for simplicity we suppose that $\mathbf{X}$ and $\mathbf{V}$ are of full rank. Now the singular value decomposition of $\mathbf{V}^{-1/2}\mathbf{X}$ can be written as

$$\mathbf{V}^{-1/2}\mathbf{X} = \mathbf{P}\Delta\mathbf{Q}', \tag{3.1}$$

where $\mathbf{P}_{n \times m}$ has orthogonal columns, $\Delta_{m \times m}$ is a diagonal matrix with positive elements and $\mathbf{Q}_{m \times m}$ is an orthogonal matrix. Let $\mathbf{R}_{n \times (n - m)}$ be such a matrix that

$$\mathbf{T} = (\mathbf{P} : \mathbf{R}) \tag{3.2}$$

is orthogonal. Now the model can be transformed to

$$\begin{pmatrix}\mathbf{Y}_1\\ \mathbf{Y}_2\end{pmatrix} = \mathbf{T}'\mathbf{V}^{-1/2}\mathbf{Y}, \tag{3.3}$$

where

$$\mathbf{Y}_1 = \mathbf{P}'\mathbf{V}^{-1/2}\mathbf{Y} = \Delta\mathbf{Q}'\beta + \mathbf{P}'\mathbf{V}^{-1/2}\epsilon \tag{3.4}$$

and

$$\mathbf{Y}_2 = \mathbf{R}'\mathbf{V}^{-1/2}\mathbf{Y} = \mathbf{R}'\mathbf{V}^{-1/2}\epsilon. \tag{3.5}$$

Writing $\theta = \Delta\mathbf{Q}'\beta$, the model (2.1) is equivalent to the two uncorrelated models

$$\mathbf{Y}_1 = \theta + \epsilon_1 \text{ and } \mathbf{Y}_2 = \epsilon_2, \tag{3.6}$$

where the variance-covariance matrices of $\epsilon_1$ and $\epsilon_2$ are

$$\mathscr{D}(\epsilon_1) = \sigma^2 I_m,\ \mathscr{D}(\epsilon_2) = \sigma^2 I_{n-m}. \tag{3.7}$$

We can hence replace the original model by a new one where $m$ observations ($\mathbf{Y}_1$) depend upon the parameter $\theta$ and the other $n - m$ observations ($\mathbf{Y}_2$) are uncorrelated with $\mathbf{Y}_1$ and do not depend on $\theta$, so that $\mathbf{Y}_2$ has no information on $\theta$. Now we lay down the following criteria for the estimation of an identifiable parametric function $\mathbf{p}'\beta = \mathbf{q}'\theta$:

(i) The estimator of $\mathbf{q}'\theta$ is function of $\mathbf{Y}_1$ only, say $f(\mathbf{Y}_1)$.

(ii) The estimator is *method-consistent* (MC), i.e., if there is no error in $\mathbf{Y}_1$ then the estimator $f(\mathbf{Y}_1)$ coincides with the true value $\mathbf{q}'\theta$.

Condition (ii) implies that

$$f(\theta) = \mathbf{q}'\theta \text{ for all } \theta \in E^m, \tag{3.8}$$

i.e.,

$$f(\mathbf{Y}_1) = \mathbf{q}'\mathbf{Y}_1, \tag{3.9}$$

which is the least squares estimator of $\mathbf{q}'\theta = \mathbf{p}'\beta$. Thus, we have arrived at the least squares estimator without introducing concepts such as linearity of estimator or unbiasedness or minimum variance.

## 4. GENERALIZED RIDGE REGRESSION

Let us assume that in the model (2.1) the parameter vector $\beta$ itself is a random variable with

$$\mathscr{E}(\beta) = (1, \ldots, 1)'g = \mathbf{1}g \tag{4.1}$$

and

$$\mathscr{D}(\beta) = \sigma^2 k^{-1}\mathbf{I}_m. \tag{4.2}$$

Then the observation vector $\mathbf{Y}$ has the expectation

$$\mathscr{E}(\mathbf{Y}) = \mathbf{X}\mathbf{1}g, \tag{4.3}$$

and the variance-covariance matrix of $\beta$ and $\mathbf{Y}$ is (cf. Rao 1973a, p. 234)

$$\mathscr{D}\begin{pmatrix}\beta \\ \mathbf{Y}\end{pmatrix} = \begin{pmatrix}\sigma^2 k^{-1}\mathbf{I} & \sigma^2 k^{-1}\mathbf{X}' \\ \sigma^2 k^{-1}\mathbf{X} & \sigma^2 k^{-1}\mathbf{X}\mathbf{X}' + \sigma^2\mathbf{V}\end{pmatrix} \tag{4.4}$$

In this case an estimator of $\beta$ can be obtained by writing down the regression of $\beta$ on $\mathbf{Y}$, i.e.,

$$\beta^{(b)} = \mathbf{1}g + \mathbf{X}'(\mathbf{X}\mathbf{X}' + k\mathbf{V})^{-1}(\mathbf{Y} - \mathbf{X}\mathbf{1}g). \tag{4.5}$$

This expression gives the Bayes estimator of $\beta$, where the terms $g$ and $k$ may be regarded as prior parameters.

Next we consider some special cases. If $g = 0$ and $\mathbf{V} = \mathbf{I}$, then we have

$$\begin{aligned}\beta^{(b)} &= \mathbf{X}'(\mathbf{X}\mathbf{X}' + k\mathbf{I})^{-1}\mathbf{Y} \\ &= (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y},\end{aligned} \tag{4.6}$$

which is the ordinary ridge regression estimator. Hence the ridge estimator is found to be a Bayes estimator of $\beta$ under the assumption that the regression coefficients are independent and have a priori values equal to zero.

If $g = 0$ and $\mathbf{V}$ is nonsingular, then the estimator $\beta^{(b)}$ is of the form

$$\begin{aligned}\beta^{(b)} &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} \\ &= [\mathbf{I} - \mathbf{U}(\mathbf{U} + k^{-1}\mathbf{I})^{-1}]\beta^{(l)},\end{aligned} \tag{4.7}$$

where

$$\mathbf{U} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \tag{4.8}$$

and $\beta^{(l)}$ denotes the (generalized) least squares estimator of $\beta$:

$$\beta^{(l)} = (\mathbf{X}'\mathbf{V}^{-1}\,\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}.$$

When $g = 0$ but $\mathbf{V}$ may be singular we get the expression

$$\beta^{(b)} = \mathbf{X}'(\mathbf{X}\mathbf{X}' + k\mathbf{V})^{-}\mathbf{Y}. \tag{4.9}$$

Finally, if $g \neq 0$ and $\mathbf{V}$ is nonsingular, the estimator (4.5) can be written as

$$\begin{aligned}\beta^{(b)} &= \mathbf{1}g + (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\mathbf{1}g)\\ &= \beta^{(l)} - \mathbf{U}(\mathbf{U} + k^{-1}\mathbf{I})^{-1}(\beta^{(l)} - \mathbf{1}g).\end{aligned} \tag{4.10}$$

In general we do not know a priori what the values of $g$ and $k$ are. Fortunately, however, these parameters can be estimated from the observations as follows:

$$\hat{g} = \mathbf{1}'\mathbf{U}^{-1}\beta^{(l)} \div \mathbf{1}'\mathbf{U}^{-1}\mathbf{1} \tag{4.11}$$

and

$$\hat{k} = a \div b, \tag{4.12}$$

where $a$ and $b$ are determined from

$$(n - m + 2)a = (\mathbf{Y} - \mathbf{X}\beta^{(l)})'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\beta^{(l)}), \tag{4.13}$$

$$\begin{aligned}(m - 3)[\frac{b}{m-1}(\mathrm{tr}\,\mathbf{U}^{-1} - \frac{\mathbf{1}'\mathbf{U}^{-2}\mathbf{1}}{\mathbf{1}'\mathbf{U}^{-1}\mathbf{1}}) + a]&\\ = (\beta^{(l)} - \mathbf{1}\hat{g})'\mathbf{U}^{-1}(\beta^{(l)} - \mathbf{1}\hat{g}).&\end{aligned} \tag{4.14}$$

Substituting $\hat{g}$ and $\hat{k}$ in (4.10) we get the estimator

$$\beta^{(e)} = \mathbf{1}\hat{g} + (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} + \hat{k}\mathbf{I})^{-1}\mathbf{X}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\mathbf{1}\hat{g}). \tag{4.15}$$

This estimator may not be better than $\beta^{(l)}$ in terms of a compound loss function, but it may be a good competitor to the usual ridge estimator.

## 5. SIMULTANEOUS ESTIMATION AND PREDICTION

### 5.1 Simultaneous estimation

In some applications of linear estimation there may be several models to be estimated simultaneously. The method of simultaneous estimation was originally suggested by R. A. Fisher and was further developed by Fairfield Smith (1936), C. R. Henderson (1950), V. G. Panse (1946) and C. R. Rao (1952, 1953). The subject has been revided with the remarkable contribution by Stein (1955). We shall illustrate the Stein phenomenon for vector parameters. [See Efron and Morris (1972) for a general discussion.]

In this section we will study estimation from $k$ linear models

$$\mathbf{Y}_i = \mathbf{X}\beta_i + \epsilon_i \, , \; i = 1 \, , \ldots , \mathrm{k}. \tag{5.1}$$

In these models $(\beta_i, \epsilon_i)$ are assumed to be i.i.d. random variables with

$$\mathscr{E}\begin{pmatrix}\beta_i \\ \epsilon_i\end{pmatrix} = \begin{pmatrix}\beta \\ \mathbf{O}\end{pmatrix} , \; \mathscr{D}\begin{pmatrix}\beta_i \\ \epsilon_i\end{pmatrix} = \begin{pmatrix}\mathbf{F} & \mathbf{O} \\ \mathbf{O} & \sigma^2\mathbf{V}\end{pmatrix} , \tag{5.2}$$

where $\beta$, $\mathbf{F}$ and $\sigma^2$ are unknown.

Let $\beta_i^{(l)}$ denote the least squares estimator of $\beta_i$ from the $ith$ model and let $\mathbf{U}$ be defined as in (4.8). Then the regression of $\beta_i$ on $\mathbf{Y}_i$ can be written as

$$\beta_i^{(b)} = \beta_i^{(l)} - \sigma^2\mathbf{U}(\mathbf{F} + \sigma^2\mathbf{U})^{-1}(\beta_i^{(l)} - \beta). \tag{5.3}$$

The estimator $\beta_i^{(b)}$ is the Bayes estimator of $\beta_i$ under the assumptions (5.2). It can be shown to have the following property:

$$\begin{aligned} \mathscr{E}_{\beta_i}\mathscr{E}_{\mathbf{Y}_i} (\beta_i^{(b)} - \beta_i)(\beta_i^{(b)} - \beta_i)' &= \sigma^2\mathbf{U} - \sigma^4\mathbf{U}(\mathbf{F} + \sigma^2\mathbf{U})^{-1}\mathbf{U} \\ \leq \sigma^2\mathbf{U} &= \mathscr{E}(\beta_i^{(l)} - \beta_i)(\beta_i^{(l)} - \beta_i)' \, , \; i = 1 \, , \ldots , k, \end{aligned} \tag{5.4}$$

where $\mathscr{E}_{\beta_i}$ and $\mathscr{E}_{\mathbf{Y}_i}$ represent the expectations taken over the distribution of $\mathbf{Y}_i$ for a fixed $\beta_i$ and over the prior distribution of the vectors $\beta_i$, respectively. Inequality (5.4) tells us that the Bayes estimator of $\beta_i$ has a smaller *mean dispersion error* (MDE) than the least squares estimator. Consequently, the MDE connected with all the estimators $\beta_i^{(b)}$, $i = 1 \, , \ldots , k$, is

$$\begin{aligned} \mathbf{Q}^{(b)} &= \frac{1}{k} \sum_1^k \mathscr{E}(\beta_i^{(b)} - \beta_i)(\beta_i^{(b)} - \beta_i)' \\ &= \sigma^2\mathbf{U} - \sigma^4\mathbf{U}(\mathbf{F} + \sigma^2\mathbf{U})^{-1}\mathbf{U}. \end{aligned} \tag{5.5}$$

The unknown parameters $\beta$, $\mathbf{F}$ and $\sigma^2$ may be estimated from the observations in the following way:

$$k\beta_{\bullet} = \sum_1^k \beta_i^{(l)}, \tag{5.6}$$

$$k(n - m)\sigma^2_{\bullet} = \sum_1^k (\mathbf{Y}_i{}'\mathbf{V}^{-1}\mathbf{Y}_i - \mathbf{Y}_i{}'\mathbf{V}^{-1}\mathrm{X}\beta_i^{(l)}) = W, \tag{5.7}$$

$$(k - 1)(\mathbf{F}_{\bullet} + \sigma^2_{\bullet}\mathbf{U}) = \sum_1^k (\beta_i^{(l)} - \beta_{\bullet})(\beta_i^{(l)} - \beta_{\bullet})' = \mathbf{B}. \tag{5.8}$$

Substituting $\beta_{\bullet}$, $\mathbf{F}_{\bullet}$ and $\sigma^2_{\bullet}$ in the Bayes estimator (5.3) we obtain the empirical Bayes estimator

$$\beta_i^{(e)} = \beta_i^{(l)} - cW\mathbf{U}\mathbf{B}^{-1}(\beta_i^{(l)} - \beta_{\bullet}) \, , \; i = 1 \, , \ldots , k. \tag{5.9}$$

The constant $c$ is determined by minimizing

$$\mathscr{E} \sum_1^k (\beta_i^{(e)} - \beta_i)(\beta_i^{(e)} - \beta_i)', \tag{5.10}$$

under the assumption that $\epsilon_i$ and $\beta_i$ have independent multivariate normal distributions. The minimizer $c$ turns out to be

$$c = (k - m - 2)/(kn - km + 2), \tag{5.11}$$

and with this choice of $c$ the MDE for the estimators $\beta_i^{(e)}$ is

$$\begin{aligned}\mathbf{Q}^{(e)} &= \frac{1}{k} \sum_1^k \mathscr{E}(\beta_i^{(e)} - \beta_i)(\beta_i^{(e)} - \beta_i)' \\ &= \sigma^2\mathbf{U} - \frac{\sigma^4(n - m)(k - m - 2)}{k(n - m) + 2} \mathbf{U}(\mathbf{F} + \sigma^2\mathbf{U})^{-1}\mathbf{U},\end{aligned} \tag{5.12}$$

provided $k \geq m + 2$ (Rao 1975).

It can be seen from the expressions (5.5) and (5.12) that the following inequality holds:

$$\mathbf{Q}^{(l)} > \mathbf{Q}^{(e)} > \mathbf{Q}^{(b)}, \tag{5.13}$$

where $\mathbf{Q}^{(l)}$ is the MDE connected with the least squares estimators $\beta_i^{(l)}$. Although the empirical Bayes estimator is not as good as the Bayes estimator, i.e., when true $\beta$, $\mathbf{F}$ and $\sigma^2$ are known, it is better than the least squares estimator for all $k \geq m + 2$. The larger $k$ becomes, the better will the empirical Bayes estimator perform. Stein made the remarkable observation that when $m = 1$, $(Q^{(e)} \mid \beta_1, \ldots, \beta_k) < (Q^{(l)} \mid \beta_1, \ldots, \beta_k)$, i.e., when the expectations in (5.12) are taken for fixed $\beta_1, \ldots, \beta_k$. The same results hold when $m > 1$.

### 5.2 Simultaneous prediction

We will next study the prediction of future observations in the setup of $k$ linear models. Let us consider the $ith$ model $\mathbf{Y}_i = \mathbf{X}\beta_i + \epsilon_i$, and a future observation $y_i$ with the structure

$$\begin{aligned}y_i &= \mathbf{x}'\beta_i + \eta_i, \\ \mathscr{D}\begin{pmatrix}\epsilon_i \\ \eta_i\end{pmatrix} &= \sigma^2\begin{pmatrix}\mathbf{V} & \mathbf{a} \\ \mathbf{a}' & b\end{pmatrix}, \quad i = 1, \ldots, k.\end{aligned} \tag{5.14}$$

Given $\mathbf{Y}_i$ we want to predict the additional observation $y_i$ ; let the predictor of $y_i$ be of the form $\mathbf{p}'\mathbf{Y}_i$. We choose as our loss function

$$\mathscr{E}\sum_1^k (\mathbf{p}'\mathbf{Y}_i - y_i)^2, \tag{5.15}$$

in which case we try to find an estimator that minimizes the average mean square error over all our future predictions.

If $\beta_i^{(l)}$ and $\beta_i^{(e)}$ are as defined in (5.9), then

$$\begin{aligned}&\mathscr{E}\sum_1^k \{y_i - [\mathbf{x}'\beta_i^{(e)} + \mathbf{a}'\mathbf{V}^{-1}(\mathbf{Y}_i - \mathbf{X}\beta_i^{(e)})]\}^2\\ &\leq \mathscr{E}\sum_1^k \{y_i - [\mathbf{x}'\beta_i^{(l)} + \mathbf{a}'\mathbf{V}^{-1}(\mathbf{Y}_i - \mathbf{X}\beta_i^{(l)})]\}^2. \qquad (5.16)\end{aligned}$$

The result (5.16) shows that

$$\mathbf{x}'\beta_i^{(e)} + \mathbf{a}'\mathbf{V}^{-1}(\mathbf{Y}_i - \mathbf{X}\beta_i^{(e)}) \qquad (5.17)$$

is a better predictor of $y_i$ than the least squares predictor

$$\mathbf{x}'\beta_i^{(l)} + \mathbf{a}'\mathbf{V}^{-1}(\mathbf{Y}_i - \mathbf{X}\beta_i^{(l)}) \qquad (5.18)$$

under the compound loss function (5.15).

## 6. DIRECT OR INVERSE REGRESSION

In this chapter we will present two examples where both direct and inverse regression are applied. In our first example we assume that we have to predict a person's head breadth, knowing his head length. We assume further that there are available some previous measurements of head lengths and head breadths of individuals. Preferably these individuals should belong to the same racial, age etc. group as the person we are considering.

Let *HL* and *HB* denote the head length and head breadth respectively of an individual. We suppose that *HB* depends on *HL* linearly and write down the regression of *HB* on *HL*

$$\widehat{HB} = \alpha + \beta\, HL. \qquad (6.1)$$

Coefficients $\alpha$ and $\beta$ can be estimated from the previous measurements (in the usual manner). If the person has $HL = a$ then a predictor of his *HB* is obtained as follows

$$\widehat{HB} = \hat{\alpha} + \hat{\beta}\, a, \qquad (6.2)$$

where $\hat{\alpha}$ and $\hat{\beta}$ are estimates of the coefficients in (6.1).

The procedure we have used here is direct regression. An alternative method would have been to compute the regression of *HL* on *HB*

$$\widehat{HL} = \alpha_1 + \beta_1\, HB, \qquad (6.3)$$

estimate $\alpha_1$ and $\beta_1$ from this model, substitute $a$ for $HL$ and then solve for $HB$:

$$HB = (a - \hat{\alpha}_1)/\hat{\beta}_1. \tag{6.4}$$

This method of estimation is called inverse regression.

In this simple example it is not possible to say which one of the two approaches would be definitely superior to the other. If only one of the variables is stochastic, then it is reasonable to take regression of the stochastic variable on the nonstochastic one. But here variables $HL$ and $HB$ are quite similar in character, so that no distinction can be made on this basis.

In our second example there may be seen some advantages in one approach over the other. We suppose that we have two repeated measurements, say $m_1$ and $m_2$, on the blood pressure of an individual. If $t$ is the true value of the blood pressure of that individual, then

$$\begin{cases} m_1 = t + \epsilon_1 \\ m_2 = t + \epsilon_2, \end{cases} \tag{6.5}$$

where

$$\begin{cases} \mathscr{E}(\epsilon_i) = 0, \ \mathscr{Var}(\epsilon_i) = \sigma_\epsilon^2 \ , i = 1,2; \\ \mathscr{Cov}(\epsilon_1,\epsilon_2) = 0. \end{cases} \tag{6.6}$$

If we regard (6.5) as a simple linear model with one unknown parameter, then the least squares estimator of $t$ is

$$\bar{m} = (m_1 + m_2)/2. \tag{6.7}$$

In fact, this is a result given by inverse regression.

On the other hand, we can consider $(t, m_1, m_2)$ as three measurements on an individual, one of which is unobservable. Then we may predict $t$ by the regression of $t$ on $m_1$ and $m_2$, i.e., by

$$\hat{t} = \alpha + \beta_1 m_1 + \beta_2 m_2 \tag{6.8}$$

computed from the distribution of $(t, m_1, m_2)$ in the population. From the structure of observations (6.6) we derive the dispersion matrix of $(t, m_1, m_2)$ for the population of individuals:

$$\mathscr{D}\begin{pmatrix} t \\ m_1 \\ m_2 \end{pmatrix} = \begin{pmatrix} \sigma_t^2 & \sigma_t^2 & \sigma_t^2 \\ \sigma_t^2 & \sigma_t^2 + \sigma_\epsilon^2 & \sigma_t^2 \\ \sigma_t^2 & \sigma_t^2 & \sigma_t^2 + \sigma_\epsilon^2 \end{pmatrix}, \tag{6.9}$$

where $\sigma_t^2$ is the variance of true blood pressure over the individuals. Of course, all three variables have the same expectation; let us denote it by $\tau$.

With this notation we write down the regression of $t$ on $m_1$ and $m_2$:

$$\hat{t} = \tau + \frac{2\sigma_t^2}{2\sigma_t^2 + \sigma_\epsilon^2}(\bar{m} - \tau). \tag{6.10}$$

This is the estimator obtained through direct regression.

The errors of prediction connected with estimators (6.7) and (6.10) are

$$\mathscr{E}(\bar{m} - t)^2 = \frac{1}{2}\sigma_\epsilon^2 \tag{6.11}$$

and

$$\mathscr{E}(\hat{t} - t)^2 = \frac{1}{2}\sigma_\epsilon^2 \frac{2\sigma_t^2}{2\sigma_t^2 + \sigma_\epsilon^2}. \tag{6.12}$$

It can be seen from these expressions that $\hat{t}$ has a smaller prediction error than $\bar{m}$. In other words, the direct regression estimator is better than the inverse regression (least squares) estimator when we consider the overall mean square errors averaged over all future predictions.

Estimator (6.10) is not usually applicable as such because of the unknown parameters in it. The unknowns can, however, be estimated provided we have previous data available, consisting of repeated measurements of blood pressure on $n$ individuals. On these data we perform an analysis of variance and compute the mean squares between individuals, denoted by $A$, and within individuals, denoted by $B$. The expectations of these two mean squares are

$$\mathscr{E}(A) = 2\sigma_t^2 + \sigma_\epsilon^2 \tag{6.13}$$

and

$$\mathscr{E}(B) = \sigma_\epsilon^2, \tag{6.14}$$

and therefore they can be used for the estimation of variances $\sigma_t^2$ and $\sigma_\epsilon^2$. The total mean of all our previous measurements forms an unbiased estimator of the average blood pressure $\tau$. With these expressions we are able to write down an empirical Bayes estimator of $t$ as

$$\hat{d} = \hat{\tau} + \frac{2\hat{\sigma}_t^2}{2\hat{\sigma}_t^2 + \hat{\sigma}_\epsilon^2}(\bar{m} - \hat{\tau}), \tag{6.15}$$

which, after rearrangement of terms, becomes

$$\begin{aligned} \hat{d} &= \bar{m} - \frac{\hat{\sigma}_\epsilon^2}{2\hat{\sigma}_t^2 + \hat{\sigma}_\epsilon^2}(\bar{m} - \hat{\tau}), \\ &= \bar{m} - \frac{B}{A}(\bar{m} - \hat{\tau}). \end{aligned} \tag{6.16}$$

The terms $A$, $B$ and $\hat{\tau}$ may be updated every time new data become available, and thus the estimator can be improved.

In can be shown that this empirical estimator has a smaller prediction error than the least squares estimator when $A$ and $B$ are obtained from a large sample, so that

$$\mathscr{E}(\bar{m}-t)^2 \geq \mathscr{E}(\hat{d}-t)^2 \geq \mathscr{E}(\hat{t}-t)^2. \tag{6.17}$$

Should we therefore use predictor (6.16) for the estimation of blood pressure instead of $\bar{m}$? When we examine the estimator (6.16) we notice that if $\bar{m}$ is large compared with the average $\hat{\tau}$, then $\hat{d}$ tends to give a smaller value than $\bar{m}$. On the other hand, if $\bar{m}$ is small compared with $\hat{\tau}$, then $\hat{d}$ becomes large. Thus the estimator $\hat{d}$ shrinks estimates towards the average, so that high blood pressures are underestimated and low ones overestimated. If we are trying to estimate blood pressure for diagnostic purposes this may be an undesirable feature. Therefore, in this case it may be more appropriate to apply inverse instead of direct regression. However, in other situations where the quadratic loss function is meaningful, estimator $\hat{d}$ may be better than $\bar{m}$.

## 7. MODEL SELECTION FOR PREDICTION

In the following we give an example concerning model selection for prediction purposes. The example is based on a study where a dentist wanted to predict the growth of a tooth.

The tooth had been measured weekly for 6 weeks, and the size of the tooth in the *7th* week was to be predicted from these measurements. If the growth is to be estimated by a polynomial, then what order polynomial should we select? It seems plausible that a maximum use of data is achieved by fitting a fifth degree polynomial to it. But, generally this is not so as demonstrated below. We had measurements for 7 weeks on 13 boys. For each boy, a polynomial of the k*th* degree was fitted to the first 6 measurements and the seventh value was predicted and compared with the observed value. The mean square errors between the observed and predicted values for the 13 boys are as follows for different values of $k$:

| $k$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| M.S.E. | 1.02 | 0.80 | 0.91 | 1.03 | 5.32 |

From the above it is seen that the second degree polynomial provided the best prediction. On the other hand, the fifth degree polynomial produced very bad predictions.

When model selection is under consideration, such criteria as Mallow's $C_p$, Akiake's information criterion and Fisher's forward and backward selection of variables are usually cited. But in this example none of these could discover the superiority of the second degree polynomial. It is therefore suggested that the procedure we are going to adopt for future observations should be tested whenever previous information is available. The failure of the fifth degree polynomial was due to the fact that we have used too many variables (5) with too few observations (6). Although a high degree equation is more precise mathematically and theoretically, paucity of observations makes estimation imprecise.

Reference may be made to Rao (1984) and Rao and Boudreau (1984) for a discussion of prediction procedures in growth models.

## ACKNOWLEDGEMENTS

The author wishes to thank Pirkko Welin for preparing this text with great care and diligence based on my lectures at the seminar.

## BIBLIOGRAPHY

Efron, B. and Morris, C. (1972). Empirical Bayes on vector observations. *Biometrika,* 55, 335—347.

Goldman, A. J. & Zelen, M. (1964). Weak generalized inverse and minimum variance unbiased estimation. *J. Res. Nat. Bur. Standards, Sect. B,* 68, 151—172.

Henderson, C. R. (1950). Estimation of genetic parameters. (Abstract) *Ann. Math. Statist.,* 21, 309—310.

Moore, E. H. (1920). On the reciprocal of the general algebraic matrix. (Abstract) *Bull. Amer. Math. Soc.,* 26, 394—395.

Panse, V. G. (1964). An application of discriminant fuction for selection in poultry. *J. Genetics,* (London) 47, 242—253.

Penrose, R. (1955). A generalized inverse for matrices. *Proc. Cambridge Philos. Soc.,* 51, 406—413.

Rao, C. R. (1952). *Advanced Statistical Methods in Biometric Research.* John Wiley, 1952 and Heffner, 1971.

Rao, C. R. (1953). Discriminant function for genetic selection. *Sankhyā,* 12, 229—246.

Rao, C. R. (1955). Analysis of dispersion for multiply classified data with unequal numbers in cells. *Sankhyā,* 15, 253—280.

Rao, C. R. (1962). A note on a generalized inverse of matrix with applications to problems in mathematical statistics. *J. Roy. Statist. Soc., Ser.B,* 24, 152—342.

Rao, C. R. (1967a). Calculus of generalized inverse of matrices, Part I: General theory, *Sankhyā, Ser. A,* 29, 317—342.

Rao, C. R. (1967b). Least squares theory using an estimated dispersion matrix and its application to measurement of signals. *Proc. Fifth Berkeley Symposium on Math. Statist. and Prob.,* 1, L. LeCam and J. Neyman, *eds.* Univ. of California Press, Berkeley, 355—372.

Rao, C. R. (1971). Unified theory of linear estimation. *Sankhyā, Ser. A,* 33, 371—394, and correction *Sankhyā, Ser. A,* 34, 477.

Rao, C. R. (1973a). *Linear Statistical Inference and Its Applications.* Second Edition, Wiley, New York.

Rao, C. R. (1973b). Representations of best linear unbiased estimators in the Gauss-Markoff model with a singular dispersion matrix. *J. Multivariate Anal.,* 3, 276—292.

Rao, C. R. (1974). Projectors, generalized inverses and the BLUE's. *J. Roy. Statist. Soc., Ser. B,* 36, 442—448.

Rao, C. R. (1975). Simultaneous estimation of parameters in different linear models and applications to biometric problems. *Biometrics,* 31, 545—554.

Rao, C. R. (1978). Choice of best linear estimators in the Gauss-Markoff model with a singular dispersion matrix. *Comm. Statist. A — Theory Methods,* 7, 1199—1208.

Rao, C. R. (1980). Matrix approximations and reduction of dimensionality in multivariate statistical analysis. *Multivariate Analysis — V,* P. R. Krishnaiah, *ed.,* North-Holland, 3—22.

Rao, C. R. (1981). A lemma on g-inverse of a matrix and computation of correlation coefficients in the singular case. *Comm. Statist. A — Theory Methods,* 10, 1—10.

Rao, C. R. (1983). Inference from linear models with fixed effects: recent results and some problems. *Proceedings of the 50-th anniversary of the Stat. Lab. of Iowa State,* in press.

Rao, C. R. (1984). Prediction of future observations in polynomial growth curve models. *Sankhyā,* in press.

Rao, C. R. and Boudreau, R. (1984). Prediction of future observations in a factor analytic type growth model. *Proceedings of the Sixth Symposium on Multivariate Analysis,* P. R. Krishnaiah, *ed.,* in press.

Rao, C. R. & Mitra, S. K. (1971). *Generalized Inverse of Matrices and its Applications.* Wiley, New York.

Rao, C. R. & Yanai, H. (1979). General definition and decomposition of projectors and some applications to statistical problems. *J. Statist. Plann. Inference,* 3, 1—17.

Rao, C. R. & Yanai, H. (1984). Generalized inverse of linear transformations: a geometric approach. *Linear Algebra Appl.,* in press.

Smith, H. Fairfield (1936). A discriminant function for plant selection. *Ann. Eugenics,* (London) 7, 240—260.

Stein, C. (1955). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symposium on Math. Statist. and Prob.,* 1.

von Neumann, J. (1937). Some matrix inequalities and metrization of metric spaces. *Tomsk Univ. Rev.,* 1, 286—299.

Zyskind, G. (1967). On canonical forms, nonnegative covariance matrices and best and simple least squares linear estimators in linear models. *Ann. Math. Statist.,* 38, 1092—1109.

*May 1984*

*Department of Mathematics and Statistics*
*University of Pittsburgh*
*Pittsburgh, PA 15260*
*U.S.A.*

Pukkila
Puntanen
Editors

Proceedings of the First International Tampere Seminar on
Linear Statistical Models and their Applications

# *PROCEEDINGS*

of the

# First International Tampere Seminar on Linear Statistical Models and their Applications

EDITED BY Tarmo Pukkila and Simo Puntanen
DEPARTMENT OF MATHEMATICAL SCIENCES
UNIVERSITY OF TAMPERE, FINLAND

# PROCEEDINGS

of the

# First International Tampere Seminar on Linear Statistical Models and their Applications

*University of Tampere, Tampere, Finland*
*August 30th to September 2nd, 1983*

Edited by

Tarmo Pukkila and Simo Puntanen

Published by

Department of Mathematical Sciences/Statistics, University of Tampere
P.O. BOX 607, SF-33101 Tampere, Finland

1985

Report No. A 138

ISBN 951-44-1693-7
ISSN 0356-4231

Printed by the University of Tampere Printing Service
Composition by Vammalan Kirjapaino Oy, Vammala

# Preface

The First International Tampere Seminar on Linear Statistical Models and their Applications was held at the University of Tampere, Tampere, Finland, during the period August 30—September 2, 1983. The seminar brought together, from at least nine different countries, more than 100 researchers in linear statistical models and related areas.

The main speakers in the seminar were Professor C. Radhakrishna Rao (University of Pittsburgh, U.S.A., and the Indian Statistical Institute, New Delhi, India) and George P. H. Styan (McGill University, Montréal, Canada). There is no doubt that the whole seminar audience greatly enjoyed the excellent lecture series given by these two outstanding statisticians. Professor Rao's topic was »A Unified Approach to Inference from Linear Models» and Professor Styan's was »Schur Complements and Linear Statistical Models». The editors of these *Proceedings* are extremely grateful to Professors Rao and Styan for the effort they put into their fine articles for this volume; we are very proud to publish them.

This was the first time Professor Rao had visited Finland and certainly his visit was a great honour for the whole statistical community of Finland. It is also a great honour to the University of Tampere that Professor Rao, one of the fathers of modern statistics, has consented to accept an Honorary Doctorate in connection with the University's 60th anniversary in May 1985.

Both Professor Rao's and Professor Styan's stimulating personalities and their vital participation in all seminar activities made a great impression on the organizers. The role of Professor Styan in planning the seminar was indeed invaluable. Also our sincere thanks go to him for his very useful advice and cooperation in preparing the *Proceedings* for publication.

The invited talks at the seminar were given by Professors R. W. Farebrother (University of Manchester, U.K.), Johan Fellman (Swedish School of Economics, Helsinki), Hannu Niemi (University of Helsinki) and Bimal Kumar Sinha (University of Pittsburgh, U.S.A.). My special thanks are due to these speakers, as well as to the speakers in the contributed paper sessions, for an excellent series of talks and for their kind cooperation in preparing the papers for publication. I also wish to thank Professors Jerzy K. Baksalary (Academy of Agriculture, Poznań, Poland), J. A. Melamed (Tbilisi, U.S.S.R.) and Der-Shin Chang (Hsinchu, Taiwan), who, unfortunately, for unavoidable reasons, were not able to participate in the seminar. The last two, however, do have their contributions in the *Proceedings.*

All papers in this volume have been refereed and I therefore wish to thank the anonymous referees for their efforts. The deadline for the first versions of the papers was September 2, 1983. On the front page of each paper is the affiliation of the author as of August 1983 and at the end of the paper the current affiliation and the date when the final version was received.

Financially the seminar was supported by the University of Tampere, the Academy of Finland, the City of Tampere, and various associations and companies whose names are listed on page viii. All deserve our sincerest thanks for their invaluable support. The City of Tampere also very kindly invited all participants to a Civic Reception in the Town Hall.

The seminar was organized by a local committee within the Department of Mathematical Sciences/Statistics, consisting of Paula Hietala, Pentti Huuhtanen, Päivi Laurila, Erkki Liski, Simo Puntanen and myself. Also Raija Leppälä, Olavi Stenman and Pirkko Welin very kindly gave their help in various arrangements. I am deeply grateful to each and every one of my colleagues for this tremendous cooperation.

Jointly with Simo Puntanen, as the editors of these *Proceedings,* I wish to give special thanks to Pirkko Welin for her kind assistance in preparing this volume.

I would also like to thank Professor and Mrs. Eino Haikala for their warm hospitality in connection with the seminar. The organizers of the seminar are also grateful to the University Rector J. K. Visakorpi for his address during the inauguration of the seminar.

Our sincere thanks also go to all of the participants in the seminar, whether reading papers or not, for it was their participation which made the seminar a success. The organizers were also glad to see the active interest shown in the social programme of the seminar. It is my impression that everyone will have warm memories of the Seminar Dinner and Civic Reception — even hot ones of the Sauna Party.

The success of the seminar is due also to Professors Gustav Elfving (University of Helsinki) and Timo Mäkeläinen (University of Helsinki). Both acted as chairmen during several sessions in the seminar. Besides this, Professor Elfving gave an inspiring opening address entitled »Finnish Mathematical Statistics in the Past». I am deeply grateful for their help and for their cooperation in carrying out the seminar.

While preparing this volume for publication, we were deeply saddened that Professor Gustav Elfving passed away on March 25, 1984. He was instrumental in the promotion of statistics in Finland as well as of our seminar. As the seminar organizers we were very impressed not only by his active participation in mathematical and statistical discussions but also by his great and warm sociability: he was one of the real old-time gentlemen. It is to Professor Elfving that we dedicate this volume.

*Tampere, February 1985*

*Tarmo Pukkila*
*Seminar Director*

THE FINANCIAL SUPPORT
OF THE FOLLOWING INSTITUTIONS
IS GRATEFULLY ACKNOWLEDGED:

*THE ACADEMY OF FINLAND*
*THE TAMPERE UNIVERSITY FOUNDATION*
*THE TAMPERE CITY SCIENCE FOUNDATION*

*KANSALLIS-OSAKE-PANKKI*

*FINNAIR OY*

*OY INTERNATIONAL BUSINESS MACHINES AB*
*THE FINNISH STATISTICAL SOCIETY*
*THE RESEARCH INSTITUTE OF THE FINNISH ECONOMY*

# Contents

## Opening Address

## Invited Lecture Series

## Invited Talks

## Contributed Papers

*Dedicated to the memory of*

*Gustav Elfving*

# IN MEMORIAM

## Gustav Elfving
## 1908—1984

While these *Proceedings* were in print we learned that the prominent contributor to the Seminar, Professor Gustav Elfving had died on the 25th of March 1984 at his home in Helsinki.

Erik Gustav Elfving was born in 1908 in Helsinki. He graduated in 1930 in Mathematics and took his doctoral degree in 1934, both at the University of Helsinki. He was appointed Professor of Mathematics at the same university in 1948 and retired in 1975.

Elfving wrote his doctoral thesis under the guidance of Rolf Nevanlinna on the latter's value distribution theory. Elfving's first paper in probability in 1937 marks the beginning of a long and distinguished career in probability and statistics. His writings published for the international professional audience contain papers on Markov chains, point processes, sufficiency and complete classes, design of linear experiments and sample designs, test item selection, order statistics, exact distributions and expansions of distributions, quality control, nonparametric statistics, and Bayesian statistics. Characteristic to them are many original ideas, clear formulation, and elegant presentation stressing central ideas.

Perhaps the single most influential piece of Elfving's writings — and one in a subject very close to this Seminar — is the paper »Optimum allocation in linear regression theory» published in 1952 in the *Annals of Mathematical Statistics*. The problem of optimal design of linear experiments is here brought, for the first time in some generality, before a wide statistical audience. Today an extensive field of study, optimal design of experiments has since undergone profound technical development. Yet, Elfving's work remains part of its foundations. (Compare with Professor Fellman's lecture in this volume.)

During his retirement Elving took the role of a historian of mathematics. His lecture in this volume partly draws on a major work »*The History of Mathematics in Finland 1828—1918*» (Societas Scientiarum Fennica, Helsinki 1981, 195 pp.).

September 1984

Timo Mäkeläinen
Department of Mathematics
University of Helsinki
Helsinki, Finland