The Gauss-Markov Theorem in Multivariate Analysis

by

Morris L. Eaton*
Department of Theoretical Statistics
University of Minnesota

Technical Report No. 422

*This research was supported in part by the National Science Foundation Grant MCS 81 00762.

Abstract

The Gauss-Markov (G.M.) Theorem is formulated in a vector space setting general enough to cover the linear models of multivariate analysis. The connection between the G.M. Theorem and linear unbiased prediction is explored. The results are applied to the problem of assessing the bias of certain types of "plug in" estimators of variance in linear model problems where covariances are estimated.

Key Words: Linear model theory, Gauss-Markov Theorem, prediction, weighted least squares, iteratively reweighted least squares, bias in variance estimates.

AMS 1980 subject classifications: Primary 62J99, Secondary, 62F11, 62H99.

§1 Introduction

The work in this paper arose, in part, from an attempt to understand a phenomenon observed in Freedman and Peters (1982, 1983). Here is the problem. In a linear regression model $y = X\beta + \varepsilon$ with $y \in R^n$, $\beta \in R^k$ and X a known $n \times k$ matrix of rank k, suppose one is interested in estimating a linear combination of the β's, say $c'\beta$ with $c \in R^k$ and obtaining some idea of the variance of this estimate. Assuming $E\varepsilon = 0$ and $\mathrm{Cov}(\varepsilon) = \Sigma$, the variance of the Gauss-Markov estimator of $c'\beta$ is $\alpha^2 = c'(X'\Sigma^{-1}X)^{-1}c$ - assuming Σ and X have full rank. In the models of interest to Freedman and Peters (econometric type models), Σ is not known, but is often assumed to be known up to a few unknown parameters. These parameters are estimated from the data (perhaps by an iterative procedure) to yield $\tilde{\Sigma}$ as an estimate of Σ. Then, pretending that $\Sigma = \tilde{\Sigma}$, one calculates the "pretend" Gauss-Markov estimator, say $\tilde{\beta}$, of β. This gives $c'\tilde{\beta}$ as an estimate of $c'\beta$ and $\tilde{\alpha}^2 = c'(X'\tilde{\Sigma}^{-1}X)^{-1}c$ as an estimate of $\alpha_0^2 = \mathrm{var}(c'\tilde{\beta})$. In many situations, there is an asymptotic justification for this procedure (see Arnold (1981), Chapter 10 and Williams (1975)). However, there is some evidence to suggest that $\tilde{\alpha}^2$ underestimates α_0^2 by a lot, even in situations where one believes the asymptotics should be valid (Freedman and Peters (1983)).

In an attempt to understand this "underestimation problem" in the generality of econometric models, there are some immediate technical problems. The design matrix X is often singular and a number of linear side conditions on β are ordinarily imposed. Thus, the description "$y = X\beta + \varepsilon$" isn't quite right. But the side conditions can be added to the model description, and generalized inverses can be used to give formulas for Gauss-Markov and "pretend" Gauss-Markov estimators. A more serious technical problem is that $\tilde{\Sigma}$ can depend on the data in a very complicated way. Of course, $\tilde{\Sigma}$ appears in

all of the formulas (including generalized inverses) for the "pretend" Gauss-Markov estimators. All in all, working directly with the explicit matrix formulas seemed quite hopeless.

An alternate approach which sometimes yields useful information concerning linear model problems is the so called coordinate-free approach described in Kruskal (1961, 1968). The setting for this approach is a finite dimensional inner product space. However, the "proper" inner product depends on Σ. When Σ is unknown and has to be estimated, the inner product appears to be more of a hindrance than a help. Extending the Kruskal work, it was shown (Eaton (1972, 1978)) that linear model theory could be done in vector spaces without inner products. Essentially, the present paper modifies and extends results in Eaton (1978) and gives an application to the "underestimation problem". This vector space approach led to an understanding of the linear model problems described above which I was unable to get from either matrix methods or inner product space methods. This is not to say that the latter methods would not yield solutions to these problems, but rather that the former appears to shed some new light on certain linear model problems.

In what follows, it is assumed that the reader is familiar with finite dimensional vector space theory as can be found in Halmos (1958). In Section 2, we set notation and briefly describe means and covariances for random vectors. What we mean by a linear model is set down in Section 3.

A formulation and proof of our version of the G.M. Theorem is given in Section 4. This version shows quite clearly that: (i) the geometry of linear models is determined by the underlying covariance and not by some "external" coordinate system or inner product and (ii) best linear unbiased estimators do not depend on the quadratic measure of loss one uses. These claims are made precise in Theorem 4.1.

The application of the theory to the "underestimation problem" begins in Section 6. Here, a wide class of estimators of Σ is defined. If $\tilde{\mu}$ denotes a "pretend" Gauss-Markov estimator of μ based on one of these estimators, then conditions under which $\tilde{\mu}$ is an unbiased estimate of μ are given in Proposition 6.1. In Section 7, conditions are given so that $\alpha_0^2 = \alpha^2 + \tau$ where $\tau > 0$ (in the notation of the first paragraph above). This result is also extended to the prediction problem.

Finally, in Section 8 the "underestimation problem" is treated explicitly. Two sources of bias are pinpointed, even if $\tilde{\Sigma}$ happens to be a good estimate of Σ. In particular, it is shown that if $\tilde{\Sigma}$ is an unbiased estimator of Σ, then $E\tilde{\alpha}^2 \leqslant \alpha^2$, with a strict inequality in most cases. Thus, when $\tilde{\alpha}^2$ is used to estimate α_0^2, one has that $E\tilde{\alpha}^2 \leqslant \alpha^2 = \alpha_0^2 - \tau$ where $\tau > 0$. In effect, the positive quantity τ is being estimated to be zero.

§2 Preliminaries

Vector spaces are denoted by V, W,... and their dual spaces by V′,W′,... . All vector spaces are finite dimensional and the canonical identification of V with V′′ is always made. For $\xi \in V'$ and $x \in V$, the value of ξ at x is denoted by $[\xi,x]$ - see Halmos (1958, p. 21) for a discussion. If M is a subspace of V, $M^0 \subseteq V'$ is the annihilator of M. The vector space of linear transformations from V to W is $L(V,W)$ and for $A \in L(V,W)$, $R(A)$ is the range of A and $N(A)$ is the null space of A. Also, A′ denotes the adjoint A so A′ is the unique linear transformation in $L(W',V')$ which satisfies

$$[\xi,Ax] = [A'\xi,x] \tag{2.1}$$

for all $\xi \in W'$ and $x \in V$. The basic relations

$$\begin{aligned} (R(A))^0 &= N(A') \\ (N(A))^0 &= R(A') \end{aligned} \tag{2.2}$$

will be used without mention (see Halmos (1958, p. 88)).

If V is the direct sum of subspaces M and N, we write $V = M \oplus N$. Each direct sum decomposition of V yields a projection P on M along N and a projection I - P on N along M. Such projections satisfy $P^2 = P$, $R(P) = M$ and $N(P) = N$. Conversely, if $A \in L(V,V)$ satisfies $A^2 = A$, then A is the projection on $R(A)$ along $N(A)$ - see Halmos (1958, p. 73).

For $A \in L(V,W)$, the function $H(\xi,x) = [\xi,Ax]$ is bilinear on $W' \times V$. Conversely, if H_1 is bilinear on $W' \times V$, then there is a unique $A_1 \in L(V,W)$ such that $H_1(\xi,x) = [\xi,A_1x]$. In the particular case that $V = W'$, a bilinear function H is called symmetric if $H(\xi,\eta) = H(\xi,\eta)$, $\xi,\eta \in W'$. For symmetric H's, if $H(\xi,\xi) \geqslant 0$ for all ξ, then H is non-negative definite (written $H \geqslant 0$) and H is positive definite (written $H > 0$) if $H(\xi,\xi) > 0$ for all $\xi \neq 0$. If

$A \in L(W', W)$ corresponds to a symmetric H, then A is called symmetric and we write $A \geq 0$ $(A > 0)$ whenever $H \geq 0$ $(H > 0)$.

Suppose H is non-negative definite on $W' \times W'$ and H corresponds to $A \in L(W', W)$ via the equation $H(\xi, \eta) = [\xi, A\eta]$. Here are two useful facts which are easy to prove.

Lemma 2.1: Let H and A be non-negative definite as above. Then

(i) $N(A) = \{\xi \mid [\xi, A\xi] = 0\}$

(ii) For any subspace $M \subseteq W$, $M \cap A(M^{o}) = \{0\}$.

Now, suppose Y is a random vector taking values in a vector space V. The sigma algebra of V is generated by the obvious family of open sets on V. If $E\left|[\xi, Y]\right| < +\infty$ for all $\xi \in V'$, then the function $\xi \to E[\xi, Y]$ is a well defined linear function on V'. Hence there exists a unique $\mu \in V$ such that

$$E[\xi, Y] = [\xi, \mu], \quad \xi \in V' \tag{2.3}$$

The vector μ is the mean vector of Y and is denoted by $EY = \mu$. If $A \in L(V, W)$ and $w_0 \in W$, then $E(AY + w_0) = A\mu + w_0$.

To define the covariance of a random vector $Y \in V$, first assume that $\mathrm{var}([\xi, Y]) < +\infty$ for all $\xi \in V'$ where var denotes variance. Thus, the function

$$H(\xi_1, \xi_2) = \mathrm{cov}([\xi_1, Y], [\xi_2, Y]) \tag{2.4}$$

is well defined on $V' \times V'$, is bilinear, symmetric and non-negative definite. Thus there is a unique symmetric $\Sigma \in L(V', V)$ with $\Sigma \geq 0$ which satsifies

$$\mathrm{cov}([\xi_1, Y], [\xi_2, Y]) = [\xi_1, \Sigma\xi_2] \tag{2.5}$$

for all $\xi_1, \xi_2 \in V'$. The linear transformation Σ is called the covariance of Y and we write $\Sigma = \mathrm{Cov}(Y)$. It is easy to show that

$$\mathrm{Cov}(AY + w_0) = A\,\mathrm{Cov}(Y)\,A' \tag{2.6}$$

for $A \in L(V,W)$ and $w_0 \in W$. A bit more generally, consider $Y_i \in V_i$ such that $\text{var}([\xi_i, Y_i]) < +\infty$ for $\xi_i \in V_i'$, $i = 1,2$. Then the bilinear function

$$H(\xi_1, \xi_2) \equiv \text{cov}([\xi_1, Y_1], [\xi_2, Y_2]) \tag{2.7}$$

is well defined and is given by a linear transformation $\Sigma_{12} \in L(V_2', V_1)$ satisfying

$$\text{cov}([\xi_1, Y_1], [\xi_2, Y_2]) = [\xi_1, \Sigma_{12}\xi_2] \tag{2.8}$$

for all $\xi_i \in V_i'$, $i = 1,2$. Also, $\Sigma_{12}' \in L(V_1', V_2)$ is denoted by Σ_{21} and satisfies

$$\text{cov}([\xi_1, Y_1], [\xi_2, Y_2]) = [\xi_2, \Sigma_{21}\xi_1] \tag{2.9}$$

for $\xi_i \in V_i'$, $i = 1,2$. The transformation Σ_{12} is sometimes called the cross covariance between Y_1 and Y_2 (in that order).

Definition 2.1: The random vectors Y_1 and Y_2 are uncorrelated if for all $\xi_i \in V_i$, $i = 1,2$,

$$\text{cov}([\xi_1, Y_1], [\xi_2, Y_2]) = 0. \tag{2.10}$$

Obviously, Y_1 and Y_2 are uncorrelated iff $\Sigma_{12} = 0$. The following result will be used in the sequel.

Proposition 2.1: Suppose $Y_1 \in V_1$ and $Y_2 \in V_2$ are uncorrelated and let H_1 be any bilinear function on $V_1 \times V_2$. Then,

$$EH_1(Y_1, Y_2) = H_1(EY_1, EY_2). \tag{2.11}$$

Proof: Let $\xi_1, \ldots, \xi_m$ be a basis for V_1' and $\eta_1, \ldots, \eta_n$ be a basis for V_2'. For $\xi \in V_1'$ and $\eta \in V_2'$, let $\xi \otimes \eta$ be the bilinear function on $V_1 \times V_2$ defined by $(\xi \otimes \eta)(x_1, x_2) = [\xi, x_1][\eta, x_2]$. Standard arguments now show that the collection

$\{\xi_i \times \eta_j | i = 1,\ldots,m, j = 1,\ldots,m\}$ is a basis for the vector space of all bilinear functions on $V_1 \times V_2$. Thus,

$$H_1 = \sum_i \sum_j c_{ij}\, \xi_i \otimes \eta_j \tag{2.12}$$

where the c_{ij} are real numbers. Because both sides of (2.11) are linear in H_1, it suffices to verify (2.11) for $H_1 = \xi \otimes \eta$. Since Y_1 and Y_2 are uncorrelated, (2.11) holds for H_1's of the form $\xi \otimes \eta$. ■

Corollary 2.1: When Y_1 and Y_2 are uncorrelated, if $EY_1 = 0$ or $EY_2 = 0$, then $EH_1(Y_1, Y_2) = 0$.

The following result is used in the discussion of the prediction problem.

Proposition 2.2: Consider $Y_i \in V_i$, $i = 1,2$ with $\mathrm{Cov}(Y_i) = \Sigma_{ii}$, $i = 1,2$ and let Σ_{21} be the cross covariance between Y_2 and Y_1. Then

(i) the equation (in $C \in L(V_1, V_2)$) $C\Sigma_{11} = \Sigma_{21}$ has a solution C_0

(ii) for any solution C_0, the random vectors $Y_2 - C_0 Y_1$ and Y_1 are uncorrelated.

Proof: The proof is a minor variation of Proposition 3.33 in Eaton (1972). ■

For any random vector Y, $L(Y)$ denotes the distributional law of Y in V. A random vector $Y \in V$ is normal if $[\xi, Y]$ is univariate normal for all $\xi \in V'$. The existence and uniqueness (up to indexing by a mean vector and a covariance) of normal distributions on V can be demonstrated as in the inner product space case (for example, see Eaton (1983), Chapter 3).

§3 Linear Models

Suppose Y is a random vector in V which has a covariance. Essentially, a linear model for Y specifies two things:

(i) A linear subspace M of V (called the regression subspace) in which the mean vector of Y is assumed to lie.

(ii) A set γ which consists of the possible covariances for Y.

In these two specifications, an implicit assumption is that M and γ are not larger than they need be - consistent with a given experimental situation. At this point, the set γ is arbitrary, but further assumptions concerning γ will be made later.

It is customary in the literature to write a linear model as

$$Y = \mu + \varepsilon \tag{3.1}$$

where $\mu \in M$ is the mean vector of Y and ε is the error vector which satisfies $E\varepsilon = 0$ and $\mathrm{Cov}(\varepsilon) \in \gamma$. Thus, the regression subspace M and the covariance structure of ε (as given by the set γ) specify the first and second moment structure of Y. This is what we mean by a linear model for the observation vector Y.

Let Σ_0 be a fixed known covariance in $L(V',V)$. When

$$\gamma = \{c\Sigma_0 | c > 0\}, \tag{3.2}$$

the linear model for Y is often called a univariate linear model, no matter what the subspace M. This model is treated in detail in Scheffé (1957). To give one description of the classical multivariate linear model, take V to be the vector space $L_{p,n}$ of all $n \times p$ real matrices and let X: $n \times k$ be a fixed known matrix of rank k. Define M by

$$M = \{\mu | \mu \in L_{p,n},\ \mu = XB,\ B \in L_{p,k}\} \tag{3.3}$$

where B is a $k \times p$ matrix of unknown parameters. One common choice for γ is

$$\gamma = \{\Sigma | \Sigma = I_n \otimes \Gamma,\ \Gamma\colon\ p \times p \text{ is positive definite}\} \qquad (3.4)$$

Here, I_n is the $n \times n$ identity matrix, $\otimes$ denotes the Kronecker product and Γ is an unknown $p \times p$ positive definite matrix. A discussion of this model can be found in many multivariate texts - for example, see Anderson (1958), Rao (1973) or Eaton (1983). The notation used here is consistent with that in Eaton (1983).

§4 The Gauss-Markov Theorem

The Gauss-Markov (GM) Theorem has to do with the linear unbiased estimation of $\mu \in M \subseteq V$ when $Y = \mu + \varepsilon$ is a linear model for Y with regression subspace M. To be precise, let

$$A = \{A \mid A \in L(V,V),\ Ax = x \text{ for } x \in M\}. \tag{4.1}$$

Then, a linear estimator of μ, say AY with $A \in L(V,V)$, is unbiased iff $A \in A$ since $EAY = A\mu$ when $EY = \mu$. Let H be any non-negative definite bilinear function on $V \times V$ and set

$$\Psi(A) = EH(AY - \mu, AY - \mu),\ A \in A. \tag{4.2}$$

The expectation in (4.2) is computed under the assumption that $EY = \mu$. The problem is to choose $A \in A$ to minimize Ψ, but, of course, $\Psi(A)$ depends on $\Sigma = \mathrm{Cov}(Y)$. Basically, the G.M. Theorem tells us how to choose A to minimize (4.2) when Σ is fixed. Here is the G.M. Theorem in the present context.

Theorem 4.1 (G.M. Theorem): Let $\Sigma = \mathrm{Cov}(Y)$ be fixed and let $A_1 \in A$ satisfy $N(A_1) \supseteq \Sigma(M^o)$. Then for any H in (4.2) and $A \in A$,

$$\Psi(A) = \Psi(A_1) + EH(A - A_1)Y, (A - A_1)Y) \tag{4.3}$$

so A_1 minimizes Ψ. If Σ is non-singular and H is positive definite, the unique minimizer is the projection on M along $\Sigma(M^o)$.

Remark 4.1: Such an A_1 always exists since $M \cap \Sigma(M^o) = \{0\}$ by Lemma 2.1. In fact, the usual choice of a minimizer is any projection P on M along N where $N \supseteq \Sigma(M^o)$. When Σ is non-singular, then such a specification uniquely determines P.

Proof: Set $B = A - A_1$ for $A \in \mathcal{A}$ so $N(B) \supseteq M$. Hence $R(B') = (N(B))^{\circ} \subseteq M^{\circ}$ which entails $R(\Sigma B') \subseteq \Sigma(M^{\circ})$. Therefore, $A_1 \Sigma B' = 0$ since $N(A_1) \supseteq \Sigma(M^{\circ})$ by assumption. This implies that $A_1 Y$ and BY are uncorrelated. Since $EBY = 0$, Corollary 2.1 yields

$$EH(A_1 Y - \mu,\ BY) = 0 \tag{4.4}$$

for all $\mu \in M$. Therefore

$$\Psi(A) = EH(A_1 Y - \mu + BY,\ A_1 Y - \mu + BY) =$$

$$EH(A_1 Y - \mu,\ A_1 Y - \mu) + 2EH(A_1 Y - \mu,\ BY) + EH(BY, BY) =$$

$$\Psi(A_1) + EH(BY, BY)$$

so (4.3) holds and A_1 minimizes Ψ. Clearly A_1 is the unique minimizer of Ψ if the equation

$$EH(BY, BY) = 0 \tag{4.5}$$

implies $B = 0$. But (4.5) implies $H(BY, BY) = 0$ a.e. which implies $BY = 0$ a.e. when H is positive definite. Thus $\mathrm{Cov}(BY) = B\Sigma B' = 0$ which entails $B = 0$ when Σ is non-singular. ■

Remark 4.2: The minimizing A_1 depends on Σ but not on the quadratic loss defined by H.

Henceforth, P without a subscript will denote a projection on M along N with $N \supseteq \Sigma(M^{\circ})$. The dependence of P on Σ is suppressed notationally, but should be remembered by the reader. Any such P minimizes Ψ in (4.2). The proof of Theorem 4.1 shows that if $T \in L(V,W)$ satisfies $N(T) \supseteq M$, then PY and TY are uncorrelated and the equation

$$EH_1(PY, TY) = 0 \tag{4.6}$$

holds for any bilinear function H_1. In particular, $Q \equiv I - P$ is such a T so

$$\begin{matrix} PY \text{ and } QY \text{ are uncorrelated} \\ \text{and (4.6) holds} \end{matrix} . \qquad (4.7)$$

In fact, (4.7) suggests the following very useful sufficient condition that $A_1 \in A$ minimize Ψ.

Theorem 4.2: Suppose $A_1 \in A$ is a projection on M along N. If A_1Y and $(I - A_1)Y$ are uncorrelated, then A_1 minimizes Ψ.

Proof: To say that A_1Y and $(I - A_1)Y$ are uncorrelated is to say that

$$A_1\Sigma(I - A_1)' = 0. \qquad (4.8)$$

But, $N(I - A_1) = M$ so $R((I - A_1)') = M^o$ which entails $R(\Sigma(I - A_1)') = \Sigma(M^o)$. This and (4.8) show that $N(A_1) \supseteq \Sigma(M^o)$. Now, apply Theorem 4.1. ■

Remark 4.3: Here is a partial converse to Theorem 4.2. If A_1 minimizes Ψ for all non-negative H, then A_1Y and $(I - A_1)Y$ are uncorrelated. The proof of this is not hard and is omitted (see Kruskal (1968) for an inner product version).

In what follows, the estimator $\hat{\mu} = PY$ is called the Gauss-Markov estimator of μ. Also, $Y - \hat{\mu} = QY$ is called the residual vector and as noted above, $\hat{\mu}$ and QY are uncorrelated. Before discussing the dependence of P, and hence $\hat{\mu}$, on $\Sigma = \text{Cov}(Y)$, we first make a remark concerning the estimation of linear transformations of μ.

Remark 4.4: In some cases one desires to estimate $B\mu$ rather than μ where B is a known linear transformation on V to W. In this instance, one considers estimators of the form CY, $C \in L(V,W)$ where C satisfies $C\mu = B\mu$ for all $\mu \in M$; thus, CY is an unbiased estimator of $B\mu$. Let $\mathcal{C}$ be this set of C's. Let H_1 be any non-negative definite bilinear function on $W \times W$ and

set

$$\Psi_1(C) = EH_1(CY - B\mu, CY - B\mu) \qquad (4.9)$$

for $C \in \mathcal{C}$. Consider any A_1 given in Theorem 4.1. That $C_1 = BA_1$ minimizes Ψ_1 is proved in the same way Theorem 4.1 is proved. For this reason, $B\hat{\mu} = BPY$ is called the Gauss-Markov estimator of $B\mu$.

Now, consider a linear model $Y = \mu + \varepsilon$ with a known regression subspace $M \subseteq V$ and a known set γ of possible covariances for Y (and hence ε). Although the following result is an easy consequence of Theorem 4.1, it is a useful and important tool for dealing with linear models which have non-trivial covariance sets γ.

Theorem 4.3: Assume there exists a subspace $N \subseteq V$ such that $M \cap N = \{0\}$, and

$$\Sigma(M^o) \subseteq N \quad \text{for all } \Sigma \in \gamma. \qquad (4.10)$$

Then any $A_2 \in \mathcal{A}$ which satisfies $N(A_2) \supseteq N$ minimizes Ψ in (4.2) for all $\Sigma \in \gamma$.

Proof: Just apply Theorem 4.1 for each $\Sigma \in \gamma$. ■

In most applications of Theorem 4.3, each $\Sigma \in \gamma$ is non-singular and the A_2 of this Theorem is the projection on M along $\Sigma(M^o)$. When each Σ is non-singular, then the assumption of Theorem 4.3 is that

$$\Sigma_1(M^o) = \Sigma_2(M^o), \quad \Sigma_1, \Sigma_2 \in \gamma \qquad (4.11)$$

That (4.11) holds for the multivariate linear model of Section 3 is well known and easily verified (for example, see Eaton (1970)). When (4.11) holds, then $\hat{\mu} = PY$ can be calculated under any fixed $\Sigma \in \gamma$ as all $\Sigma \in \gamma$ give the same value for $\hat{\mu}$. However, the value of

$$\mathrm{Cov}(\hat{\mu}) = P\Sigma P' \qquad (4.12)$$

does depend on Σ even when (4.11) holds.

Remark 4.5: When all the Σ's in γ are non-singular, (4.11) is basically the well known necessary and sufficient condition that G.M. estimate and least squares estimates be the same. For coordinate space versions, see Rao (1967) and Zyskind (1967), and for an inner product space version, see Kruskal (1968).

Remark 4.6: One consequence of Theorem 4.1 is that for any $A \in \mathcal{A}$, $\mathrm{Cov}(AY) \geqslant \mathrm{Cov}(PY)$ in the sense of positive definiteness (the Loewner ordering). In other words,

$$[\xi, A\Sigma A'\xi] \geqslant [\xi, P\Sigma P'\xi] \qquad (4.13)$$

for all $\xi \in V'$ and all $A \in \mathcal{A}$.

Remark 4.7: Consider the case when $\mu = F\beta$ where F is a known non-singular linear transformation on W to V. Thus, $M = \{\mu \mid \mu = F\beta,\ \beta \in W\}$. Assume that Σ is non-singular. It is easily verified that

$$P = F(F'\Sigma^{-1}F)^{-1}F'\Sigma^{-1} \qquad (4.14)$$

is the projection on M along $\Sigma(M^{o})$. Hence $\hat{\mu} = PY$ in this case. For the case when Σ is singular, see Rao (1973) or Takeuchi, Yanai, and Mukherjee (1982).

Again consider a linear model $Y = \mu + \varepsilon$ with a regression subspace M and a covariance set γ. There are many interesting and useful situations where $\hat{\mu} = PY$ depends on $\Sigma \in \gamma$ in a way that precludes the calculation of $\hat{\mu}$. In other words, $\hat{\mu}$ can not be calculated since Σ is unknown, even though the set γ is known. This situation will be discussed further in Section 6.

§5 Prediction

Consider $Y_i \in V_i$ with $EY_i = \nu_i$ and $\mathrm{Cov}(Y_i) = \Sigma_{ii}$, $i = 1,2$. Also, let Σ_{12} be the cross covariance between Y_1 and Y_2 (see (2.8)). Assume that ν_i and Σ_{ij} are known, $i,j = 1,2$. If we observe Y_1, the following result tells us how to predict Y_2 on the basis of affine functions of Y_1 - no matter what the quadratic measure of error. In what follows, the linear transformation C_0 will be as in Proposition 2.2. The following is well known.

Proposition 5.1: Let H be any non-negative definite bilinear function on $V_2 \times V_2$ and for $A \in L(V_1,V_2)$, $b \in V_2$, let

$$\Psi(A,b) = EH(Y_2 - AY_1 - b, Y_2 - AY_1 - b). \qquad (5.1)$$

With $\hat{b} = \nu_2 - C_0\nu_1$, we have

$$\Psi(A,b) = \Psi(C_0,\hat{b}) + H(\nu_2 - A\nu_1 - b, \nu_2 - A\nu_1 - b)$$
$$+ EH((A - C_0)(Y_1 - \nu_1),\ (A - C_0)(Y_1 - \nu_1)).$$

Hence $(C_0,\hat{b})$ minimizes Ψ over all (A,b).

Proof: The proof is similar to the proof of Theorem 4.1. The key is to observe that $Y_2 - C_0Y_1$ and Y_1 are uncorrelated; then use Proposition 2.2 and Proposition 2.1. ∎

Proposition 5.1 is similar to Theorem 4.1 in at least two respects:

(i) the solution to the minimization does not depend on H;

(ii) the proof of each depends on Proposition 2.1 for uncorrelated random vectors.

To combine Proposition 5.1 and Theorem 4.1, now assume that $EY_1 = \mu \in M$ as in Section 4 and assume that $EY_2 = T\mu$, $\mu \in M$ where $T:V_1 \to V_2$ is a known linear transformation. Here, μ is unknown but Σ_{ij} is known for $i,j = 1,2$. The

traditional approach to the problem of predicting Y_2 from Y_1 is to consider unbiased linear predictors - that is, predictors AY_1 with A mapping V_1 to V_2 such that

$$A\mu = EAY_1 = EY_2 = T\mu, \quad \mu \in M.$$

Thus, we look at the class

$$\mathcal{A} = \{A | A \in L(V_1, V_2),\ A\mu = T\mu,\ \mu \in M\}. \tag{5.2}$$

Again, let H be a non-negative definite bilinear function on $V_2 \times V_2$ and set

$$\Psi(A) = EH(Y_2 - AY_1, Y_2 - AY_1) \tag{5.3}.$$

One version of the following appeared in Goldberger (1962).

Proposition 5.2: The linear transformation

$$A_0 \equiv C_0 - (C_0 - T)P \tag{5.4}$$

minimizes Ψ over $\mathcal{A}$. Here, P is the projection on M along N where $N \supseteq \Sigma_{11}(M^0)$ (see Remark 4.1).

Proof: First, set $Z_1 = Y_1 - \mu$ and $Z_2 = Y_2 - T\mu$. Since $A \in \mathcal{A}$,

$$\Psi(A) = EH(Z_2 - AZ_1, Z_2 - AZ_1).$$

Using the fact that $Z_2 - C_0Z_1$ and Z_1 are uncorrelated and have mean 0, it follows that

$$\begin{aligned}\Psi(A) &= EH(Z_2 - C_0Z_1, Z_2 - C_0Z_1) + \\ &\quad EH(C_0 - A)Z_1, (C_0 - A)Z_1) \\ &= EH(Z_2 - C_0Z_1, Z_2 - C_0Z_1) + \\ &\quad EH((C_0 - A)Y_1 - (C_0 - T)\mu, (C_0 - A)Y_1 - (C_0 - T)\mu)\end{aligned}$$

where the last equality uses the assumption that $A\mu = T\mu$. Since $(C_0 - T)$ is known, Remark 4.4 shows that the final expression in (5.5) is minimized by choosing $(C_0 - A)$ to be $(C_0 - T)P$. Hence choosing $A = A_0$ yields a minimum for (5.3) since $A_0 \in \mathcal{A}$. ■

The predictor of Y_2 given by

$$\hat{Y}_1 = C_0 Y_1 - (C_0 - T)PY_1 = C_0 QY_1 + T\hat{\mu} \tag{5.6}$$

is often called the Gauss-Markov predictor of Y_2. In some situations, Y_1 and Y_2 are uncorrelated in which case the Gauss-Markov predictor is just $T\hat{\mu}$ since $C_0 = 0$. To calculate $\hat{Y}_2$, both Σ_{11} and Σ_{12} need to be known up to a common scalar. The case when Σ_{11} and Σ_{12} are estimated from data (i.e. Y_1) is taken up in Section 6.

§6 Estimated Covariances

In this section, we consider a linear model $Y = \mu + \varepsilon$, $\mu \in M$ where $\Sigma = \mathrm{Cov}(Y) \in \gamma$. The set γ is arbitrary so $\hat{\mu} = PY$ cannot be calculated as P depends on the unknown $\Sigma \in \gamma$. In such situations, it is common statistical practice to use Y to estimate Σ and then use the estimated Σ, say $\tilde{\Sigma}$, to estimate P by $\tilde{P}$ where $\tilde{P}$ is the projection on M along $\tilde{\Sigma}(M^o)$. One typical procedure is iteratively reweighted least squares which can be described as follows: Make an initial guess $\tilde{\Sigma}_0$ for Σ to get $\tilde{P}_0$ (projection on M along $N_0 \supseteq \tilde{\Sigma}_0(M^o)$) and then calculate $\tilde{\mu}_0 = \tilde{P}_0 Y$. Based on the residual $Y - \tilde{\mu}_0$, calculate a new estimate $\tilde{\Sigma}_1$ of Σ to get $\tilde{P}_1$ (projection on M along $N_1 \supseteq \tilde{\Sigma}_1(M^o)$) and $\tilde{\mu}_1 = \tilde{P}_1 Y$. Then, based on the new residual $Y - \tilde{\mu}_1$, calculate a new estimate $\tilde{\Sigma}_2$ of Σ to get $\tilde{P}_2$ and $\tilde{\mu}_2 = \tilde{P}_2 Y$. This process is iterated a certain number of times to yield final estimates $\tilde{\Sigma}$, $\tilde{P}$ and $\tilde{\mu} = \tilde{P}Y$. Naturally, the estimate $\tilde{\Sigma}$ will depend on the assumed form of $\Sigma \in \gamma$.

Rather than focus attention on any particular estimation procedure, we will establish some results which hold for a wide class of estimators of Σ which include all of the covariance estimates of which I am aware.

Definition 6.1: A function $\tilde{\Sigma}: V \to \gamma$ is a residual type estimator if

(i) $\tilde{\Sigma}(y+x) = \tilde{\Sigma}(y)$, for $y \in V$, $x \in M$

(ii) $\tilde{\Sigma}(y) = \tilde{\Sigma}(-y)$, for $y \in V$.

Throughout the remainder of this paper, $\tilde{\Sigma}$ denotes a residual type estimator for Σ. In most cases, the argument of the function $\tilde{\Sigma}$ is suppressed, but at times $\tilde{\Sigma}$ is written $\tilde{\Sigma}(Y)$ to emphasize the randomness of $\tilde{\Sigma}$. In the same vein, $\tilde{P}$ denotes the projection on M along any subspace N such that $N \supseteq \tilde{\Sigma}(M^o)$, and $\tilde{\mu}$ is defined by

$$\tilde{\mu} \equiv \tilde{P}Y \qquad (6.1).$$

In most cases of interest, $N = \tilde{\Sigma}(M^o)$ since $\tilde{\Sigma}$ is usually non-singular. The dependence (usually non-linear) of $\tilde{P}$ and $\tilde{\mu}$ on Y is ordinarily suppressed.

In the following Proposition, the reader may find the notation slightly confusing. The expression $\tilde{P}(Y)Y$, which is the same as $\tilde{P}Y = \tilde{\mu}$, appears below. First, $\tilde{P}(Y)x$ means the random projection $\tilde{P}(Y)$ (which is a function of Y) evaluated (as a linear transformation) at $x \in V$. Taking $x = Y$, we obtain $\tilde{P}(Y)Y$ and other similar expressions. Also, any function f defined on V which satisfies

$$f(y+x) = f(y), \quad y \in V, \; x \in M$$

is called a residual function.

Proposition 6.1: In the linear model $Y = \mu + \varepsilon$, $\mu \in M$, $E\varepsilon = 0$ and $\mathrm{Cov}(y) = \Sigma \in \gamma$, assume that $L(\varepsilon) = L(-\varepsilon)$ and $E\tilde{P}(Y)Y$ exists. Then $E\tilde{\mu} = \mu$ so $\tilde{\mu}$ is an unbiased estimate of μ.

Proof: First note that $\tilde{P}(Y) = \tilde{P}(Y - \mu)$ since $\tilde{\Sigma}(Y) = \tilde{\Sigma}(Y-\mu)$ and $\tilde{P}$ is a function of $\tilde{\Sigma}$. Hence $\tilde{P}(Y) = \tilde{P}(\varepsilon)$. For the same reason, $\tilde{P}(\varepsilon) = \tilde{P}(-\varepsilon)$. Thus, to show that $E\tilde{P}Y = \mu$, it suffices to show that

$$E\tilde{P}(\varepsilon)\varepsilon = 0 \tag{6.2}$$

since $\tilde{P}x = x$ for all $x \in M$. However, using the assumption that $L(\varepsilon) = L(-\varepsilon)$ we have

$$E\tilde{P}(\varepsilon)\varepsilon = -E\tilde{P}(-\varepsilon)\varepsilon = -E\tilde{P}(\varepsilon)\varepsilon$$

so $E\tilde{P}(\varepsilon)\varepsilon = 0$. ■

To compare the Gauss-Markov estimator of μ, namely $\hat{\mu} = PY$, and $\tilde{\mu} = \tilde{P}Y$, the following decomposition is useful.

Proposition 6.2: In the notation above,

$$\tilde{\mu} = \hat{\mu} + \tilde{P}QY \tag{6.3}$$

where $Q = I - P$.

Proof: Since $\tilde{P}$ is the identity on M, we have

$$\tilde{\mu} = \tilde{P}Y = \tilde{P}PY + \tilde{P}QY = PY + \tilde{P}QY = \hat{\mu} + \tilde{P}QY. \quad \blacksquare$$

Here is a version of Proposition 6.1 for the prediction problem. In the notation of Section 5, consider the Gauss-Markov predictor of Y_2 given by $\hat{Y}_2 = A_0 Y_1$ where A_0 is given in (5.4). When Σ_{11} and Σ_{12} are unknown, let $\tilde{\Sigma}_{11}$ and $\tilde{\Sigma}_{12}$ denote residual type estimators (both satisfying (i) and (ii) of Definition 6.1) of Σ_{11} and Σ_{12}. Then $\tilde{\Sigma}_{11}$ and $\tilde{\Sigma}_{12}$ determine $\tilde{P}$ and $\tilde{C}_0$, and hence $\tilde{A}_0 = \tilde{C}_0 - (\tilde{C}_0 - T)\tilde{P}$ which satisfies (i) and (ii) of Definition 6.1.

Proposition 6.3: In the notation above, if $L(Y_1 - \mu) = L(-(Y_1 - \mu))$, then $E\tilde{A}_0 Y_1 = T\mu$ for $\mu \in M$.

Proof: The proof is essentially the same as that for Propsotion 6.1. $\blacksquare$

At this point, a version of Proposition 6.2 for the prediction problem would look rather artificial. The results of the next section which compare covariances for $\hat{\mu}$ and $\tilde{\mu}$ will provide the motiviation for decomposing $\tilde{A}_0 Y_1$ into $A_0 Y_1$ and another vector.

§7 Variance Comparisons

In a linear model context, a common statistical problem is to estimate some linear function of regression coefficients and obtain an estimate of the standard error of the estimator. For our linear model $Y = \mu + \varepsilon$ with $\mu \in M$, $E\varepsilon = 0$ and $\mathrm{Cov}(Y) = \Sigma \in \gamma$, the first part of the above problem translates into one of estimating $[\xi,\mu]$ for a known $\xi \in V'$. Of course, when γ has the right structure (see Theorem 4.3), one would use the best linear unbiased estimator, $[\xi,\hat{\mu}]$, of $[\xi,\mu]$. This estimator has variance given by

$$\mathrm{var}([\xi,\hat{\mu}]) = [\xi, P\Sigma P'\xi] \tag{7.1}$$

where $\mathrm{Cov}(Y) = \Sigma$ and $\hat{\mu} = PY$. When Theorem 4.3 is not applicable and Σ is estimated by a residual type estimator, $\tilde{\Sigma}$, it is common practice to use $[\xi,\tilde{\mu}]$ as an estimator for $[\xi,\mu]$. Under the assumptions of Proposition 6.1 (which are to hold through this section), $[\xi,\tilde{\mu}]$ is unbiased, but one would suspect that $\mathrm{var}[\xi,\tilde{\mu}]$ is larger than (7.1) because Σ has been estimated. This is not true in general (see Freedman and Peters (1982)), but is true in some generality. Below, $E(\cdot|QY)$ denotes conditional expectation given QY.

Proposition 7.1: Assume that

$$E(P(Y - \mu)\,|\,QY) = 0 \tag{7.2}$$

for each $\mu \in M$ and $\Sigma \in \gamma$. Then

$$\mathrm{var}([\xi,\tilde{\mu}]) = \mathrm{var}([\xi,\hat{\mu}]) + E[\xi,\tilde{P}QY]^2 \tag{7.3}$$

for each $\Sigma \in \gamma$.

Proof: From Proposition 6.2, we have

$$\text{var}([\xi,\tilde{\mu}]) = \text{var}([\xi,\hat{\mu}]) + \text{var}([\xi,\tilde{P}QY]) + 2\,\text{cov}([\xi,\hat{\mu}],[\xi,\tilde{P}QY]).$$

The equalities $\hat{\mu} = PY$ and $E\hat{\mu} = \mu$ yield

$$\text{cov}([\xi,\hat{\mu}],[\xi,\tilde{P}QY]) = E([\xi,P(Y-\mu)][\xi,\tilde{P}QY]).$$

Since $\tilde{P}$ is a residual function,

$$\tilde{P}(Y) = \tilde{P}(Y-\hat{\mu}) = \tilde{P}(QY).$$

Conditioning on QY yields,

$$\text{cov}([\xi,\hat{\mu}],[\xi,\tilde{P}QY]) = E\{[\xi,\tilde{P}(QY)QY]\ E([\xi,P(Y-\mu)]\,|\,QY)\} = 0$$

since (7.2) is assumed. Also, the unbiasedness of $[\xi,\tilde{\mu}]$ and $[\xi,\hat{\mu}]$ implies $E[\xi,\tilde{P}QY] = 0$, which entails

$$\text{var}([\xi,PQY]) = E[\xi,PQY]^2.$$

Thus, (7.3) holds. ■

Remark 7.1: If the distribution of the error vector ε is normal, then (7.2) always holds. This follows since PY and QY are are uncorrelated and hence independent in the normal case. Of course, this argument shows that $\hat{\mu}$ and $\tilde{P}QY$ are independent when ε is normal. See Khatri and Shah (1981) for a comparable argument.

Remark 7.2: Since $Y-\mu = \varepsilon$ and $QY = Q\varepsilon$, condition (7.2) is a condition concerning the distribution of the error vector - namely, $E(P\varepsilon\,|\,Q\varepsilon) = 0$. To describe distributions other than normals for which (7.2) holds, first, let U be a random vector in V with $EU = 0$ and $\text{Cov}(U) = \Sigma$. Call U linear in conditional

expectation (l.c.e.) if for each vector space W and each $A \in L(V,W)$

$$E(U|AU) = BAU \tag{7.4}$$

where $B \in L(W,V)$ is any solution to the equation

$$BA\Sigma A' = \Sigma A'. \tag{7.5}$$

By Proposition 2.2, there is a solution to (7.5). Notice that if U is normal, then U is l.c.e. since $U - BAU$ and AU are uncorrelated and hence independent. Also, not that if U is l.c.e., then any linear transformation of U from V into W_1 is also l.c.e. To see that (7.4) implies (7.2), take $A = Q$ and take B to be the identity. That this B satisfies (7.5) follows from the equations $P \Sigma Q' = 0$ and $P = I - Q$.

As an example of a U which is non-normal and l.c.e., first take $V = R^n$ with the usual inner product. Suppose that the distribution of U is orthogonally invariant and $\mathrm{Cov}(U)$ exists. Necessarily, $EU = 0$ and $\mathrm{Cov}(U) = cI_n$ for some $c \geq 0$ where I_n is the $n \times n$ identity matrix. That (7.4) holds for such distributions can be established using arguments similar to those in Cambanis, Hwang, and Simons (1981). Thus, if

$$U_1 = GU \tag{7.6}$$

for some $n \times n$ matrix G, then U_1 is also l.c.e. Such distributions are sometimes called elliptical distributions. The above arguments show that for any such elliptical distribution for the error vector ε, equation (7.4) holds and hence

$$E(P\varepsilon|Q\varepsilon) = 0 \tag{7.7}$$

holds for any subspace M. Thus Proposition 7.1 holds for elliptical error

distributions. This completes Remark 7.2.

Remark 7.3: The results of Remark 7.2 provide an alternative proof of some recent results of Kariya and Toyooka (1983) on elliptical error distributions.

Remark 7.4: The obvious weakest condition for (7.3) to hold is that $\hat{\mu} = PY$ and $\tilde{P}QY$ be uncorrelated. However, since $\tilde{P}$ is more or less an arbitrary residual function (hence a function of QY), a sufficient condition for (7.3) which is conditional on QY is desirable. This is the reason for the form of (7.2). There are undoubtedly interesting linear models where (7.3) holds but (7.2) does not.

Remark 7.5: An alternative way to state (7.3) is

$$\mathrm{Cov}(\tilde{\mu}) = \mathrm{Cov}(\hat{\mu}) + \mathrm{Cov}(\tilde{P}QY).$$

This shows that $\mathrm{Cov}(\hat{\mu})$ is smaller than $\mathrm{Cov}(\tilde{\mu})$ in the positive semi-definite ordering of Loewner.

We now want to establish a version of Proposition 7.1 for the prediction problem discussed in Section 5. In that notation,

$$\hat{Y}_2 = A_0 Y_1 = C_0 Y_1 - (C_0 - T)PY_1 = C_0 QY_1 + T\hat{\mu} \qquad (7.8)$$

is the Gauss-Markov prediction for Y_2. As discussed in Section 6, residual type estimators $\tilde{\Sigma}_{11}$ and $\tilde{\Sigma}_{12}$ for Σ_{11} and Σ_{12} determine $\tilde{P}$, $\tilde{C}_0$ and

$$\tilde{A}_0 = \tilde{C}_0 - (\tilde{C}_0 - T)\tilde{P} = \tilde{C}_0\tilde{Q} + T\tilde{P} \qquad (7.9)$$

This in turn determines the predictor

$$\tilde{Y}_2 = \tilde{A}_0 Y_1 \qquad (7.10).$$

The conditions of Proposition 6.3 are assumed to hold so $E\tilde{Y}_2 = E\hat{Y}_2 = EY_2$. The following result allows a comparison of $\mathrm{Cov}(Y_2 - \hat{Y}_2)$ and $\mathrm{Cov}(Y_2 - \tilde{Y}_2)$.

Proposition 7.2: Assume that

$$E(Y_2 - \hat{Y}_2 | QY_1) = 0 \tag{7.11}$$

for each $\mu \in M$ and Σ_{11}, Σ_{12} and Σ_{22} in the linear model for Y_1 and Y_2. Then, for each $\xi \in V_2'$,

$$\mathrm{var}([\xi, Y_2 - \tilde{Y}_2]) = \mathrm{var}([\xi, Y_2 - \hat{Y}_2]) + E[\xi, Z]^2 \tag{7.12}$$

where

$$Z = C_0 QY_1 - \tilde{A}_0 QY_1 \tag{7.13}$$

Proof: First, write $Y_2 - \tilde{Y}_2 = Y_2 - \hat{Y}_2 + \hat{Y}_2 - \tilde{Y}_2$ and note that

$$\begin{aligned} \hat{Y}_2 - \tilde{Y}_2 = (A_0 - \tilde{A}_0)Y_1 &= (A_0 - \tilde{A}_0)PY_1 + (A_0 - \tilde{A}_0)QY_1 \\ &= (A_0 - \tilde{A}_0)QY_1 \\ &= C_0 QY_1 - \tilde{A}_0 QY_1 \\ &= Z. \end{aligned}$$

Since $\tilde{A}_0$ is a residual function, it follows that Z is a function of QY_1 (with Σ_{11}, Σ_{12} fixed). Thus,

$$\begin{aligned} &\mathrm{var}([\xi, Y_2 - \tilde{Y}_2]) = \mathrm{var}([\xi, Y_2 - \hat{Y}_2] + [\xi, Z]) = \\ &\mathrm{var}([\xi, Y_2 - \hat{Y}_2]) + \mathrm{var}([\xi, Z]) + 2\,\mathrm{cov}([\xi, Y_2 - \hat{Y}_2], [\xi, Z]). \end{aligned}$$

Since $E(Y_2 - \hat{Y}_2) = 0$ and Z is a function of QY_1, we have

$$\begin{aligned} \mathrm{cov}([\xi, Y_2 - \hat{Y}_2], [\xi, Z]) &= E([\xi, Y_2 - \hat{Y}_2], [\xi, Z]) \\ &= E([\xi, Z] E([\xi, Y_2 - \hat{Y}_2] | QY_1)) \\ &= 0 \end{aligned}$$

since (7.11) is assumed. Noting that $EZ = 0$, (7.12) follows. ■

Remark 7.6: When Y_1 and Y_2 are jointly normal, then (7.11) holds. To see this, write

$$Y_2 - \hat{Y}_2 = Y_2 - C_0 Y_1 - (C_0 - T)PY_1.$$

Observe that $Y_2 - C_0 Y_1$ and Y_1 are uncorrelated and hence independent. Now, decompose Y_1 into PY_1 and QY_1 which are uncorrelated and thus independent. This implies that $Y_2 - \hat{Y}_2$ and QY_1 are independent. Since $E(Y_2 - \hat{Y}_2) = 0$, (7.11) holds in the normal case.

Remark 7.7: To obtain a result for the present case similar to that given in Remark 7.2, first write the model for $Y_1 \in V_1$ and $Y_2 \in V_1$ as

$$\begin{aligned} Y_1 &= \mu + \underset{\sim}{\varepsilon}_1 \\ Y_2 &= T\mu + \varepsilon_2 \end{aligned} \tag{7.14}$$

where $\mu \in M \subseteq V_1$, $E\varepsilon_i = 0$, $\mathrm{Cov}\,\varepsilon_i = \Sigma_{ii}$, $i = 1,2$ and the cross covariance is Σ_{12}. In the direct sum space $V_1 \oplus V_2$ with elements written as $\{v_1, v_2\}$, it is easy to show that

$$\mathrm{Cov}\{Y_1, Y_2\} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \Sigma \tag{7.15}$$

where Σ is defined on $V_1' \oplus V_2'$ to $V_1 \oplus V_2$ by

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \{\xi_1, \xi_2\} = \{\Sigma_{11}\xi_1 + \Sigma_{12}\xi_2,\ \Sigma_{21}\xi_1 + \Sigma_{22}\xi_2\}.$$

Condition (7.11) can be written

$$E(\varepsilon_2 - C_0\varepsilon_1 + (C_0 - T)P\varepsilon_1 \,|\, Q\varepsilon_1) = 0. \tag{7.16}$$

Conditions so that $E(P\varepsilon_1|Q\varepsilon_1) = 0$ have been given in Remark (7.2) so we focus attention on conditions so that

$$E(\varepsilon_2 - C_0\varepsilon_1|Q\varepsilon_1) = 0 \tag{7.17}$$

However, a sufficient condition that (7.17) hold is that

$$E(\varepsilon_2 - C_0\varepsilon_1|\varepsilon_1) = 0 \tag{7.18}$$

since the left side of (7.17) is equal to

$$E(E(\varepsilon_2 - C_0\varepsilon_1|\varepsilon_1)|Q\varepsilon_1).$$

Define P^* on $V_1 \oplus V_2$ to $V_1 \oplus V_2$ by

$$P^*\{v_1, v_2\} = \{0, v_2 - C_0 v_1\}$$

so P^* is a projection with range $M^* = \{0\} \oplus V_2$. From the definition of C_0, it follows that

$$N(P^*) \supseteq \Sigma((M^*)^o).$$

Set $Q^* = I - P^*$ so

$$Q^*\{\varepsilon_1, \varepsilon_2\} = \{\varepsilon_1, C_0\varepsilon_1\}.$$

Hence, condition (7.18) is equivalent to

$$E(P^*\{\varepsilon_1, \varepsilon_2\}|Q^*\{\varepsilon_1, \varepsilon_2\}) = 0 \tag{7.19}$$

since ε_1 and $Q^*\{\varepsilon_1, \varepsilon_2\}$ are functions of each other. However, Remark 7.2 gives conditions for (7.19) to hold expressed in terms of the error vector $\{\varepsilon_1, \varepsilon_2\}$. Using this argument, one can give a proof of the Kariya-Toyooka (1983) result on prediction in the generality above.

Remark 7.8: An alternative way to write (7.12) is

$$\mathrm{Cov}(Y_2 - \tilde{Y}_2) = \mathrm{Cov}(Y_2 - \hat{Y}_2) + \mathrm{Cov}(Z) \qquad (7.20)$$

Thus, $\mathrm{Cov}(Y_2 - \tilde{Y}_2)$ is greater than or equal to $\mathrm{Cov}(Y_2 - \hat{Y}_2)$ in the Loewner ordering when (7.20) holds.

§8 Discussion

In the linear model $Y = \mu + \varepsilon$ with $\mu \in M$, $E\varepsilon = 0$ and $\mathrm{Cov}(\varepsilon) = \Sigma$, it was noted earlier that

$$\mathrm{Cov}(\hat{\mu}) = P\Sigma P' \tag{8.1}$$

so

$$\mathrm{var}[\xi, \hat{\mu}] = [\xi, P\Sigma P'\xi] \tag{8.2}$$

Recall that P depends on Σ, but the dependence is suppressed notationally. When Theorem 4.3 cannot be applied and $\tilde{\Sigma}$ is a residual type estimator of Σ, then $[\xi, \tilde{\mu}]$ is commonly used to estimate $[\xi, \mu]$. There are a number of asymptotic results which justify this practice as well as the approximation

$$\sigma^2 = \mathrm{var}[\xi, \tilde{\mu}] \doteq \mathrm{var}[\xi, \hat{\mu}] \tag{8.3}$$

(see Cox and Hinkley (1974), p. 308 or Arnold (1980), Chapter 10 for a discussion). In fact, it is common to see

$$\tilde{\sigma}^2 = [\xi, \tilde{P}\tilde{\Sigma}\tilde{P}'\xi] \tag{8.4}$$

used as an estimate for σ^2.

In some practical situations (see Freedman and Peters (1982), (1983)) $\tilde{\sigma}^2$ seems to underestimate σ^2 significantly. When ε has a normal distribution, Freedman and Peters (1982) show that in many models, $E\tilde{\sigma}^2 < \sigma^2$. The discussion below extends this result to other models and weakens the normality assumption.

Let S be the set of non-negative definite covariances on V' to V.

Proposition 8.1: For each $\xi \in V'$, the function $\Sigma \to [\xi, P\Sigma P'\xi]$ is concave on S to R.

Proof: Let $\hat{A}$ be given by (4.1) and set

$$\Psi(A,\Sigma) = E_{\Sigma}[\xi, AY - \mu]^2.$$

Since $A \in \hat{A}$, $AY - \mu = A(Y - \mu)$ so

$$\Psi(A,\Sigma) = [\xi, A\Sigma A'\xi].$$

This shows that $\Psi(A,\cdot)$ is an affine function defined on the convex set S for each $A \in \hat{A}$, so $\Psi(A,\cdot)$ is concave on S. Applying Theorem 4.1 with $H(x,y) = [\xi,x][\xi,y]$ yields

$$\inf_{A \in \hat{A}} \Psi(A,\Sigma) = [\xi, P\Sigma P'\xi]$$

for each $\xi \in V'$. Since the inf of a family of concave functions is concave, the proof is complete. ■

Corollary 8.1: Suppose $\tilde{\Sigma}$ is an unbiased estimate of $\Sigma \in \gamma$. Then $E_{\Sigma}\tilde{\sigma}^2 \leq [\xi, P\Sigma P'\xi]$ for each $\xi \in V'$.

Proof: Apply Jensen's inequality. ■

Remark 8.1: The result of Proposition 8.1 shows that when Σ is non-singular (so P is uniquely defined), the map $\Sigma \to P\Sigma P'$ is concave in the Loewner ordering. If one represents everything in terms of matrices, this result is given in Ylvisaker (1964), but the usual proofs are quite different than the one given here (for example, see Marshall and Olkin (1979), p. 469-472).

When equation (7.3) holds (as in the normal case), the discussion above shows that there are two sources of bias when one uses $\tilde{\sigma}^2$ to estimate σ^2 (assuming $\tilde{\Sigma}$ is unbiased for Σ). First, (7.3) states that

$$\sigma^2 = [\xi, P\Sigma P'\xi] + E[\xi, \tilde{P}QY]^2 \qquad (8.5)$$

If $\tilde{\sigma}^2$ is used to estimate σ^2, Corollary 8.1 shows that $E\tilde{\sigma}^2 \leqslant [\xi, P\Sigma P'\xi]$. Of course, Jensen's inequality is ordinarily strict, so one source of bias is that $\tilde{\sigma}^2$ tends to be less than $[\xi, P\Sigma P'\xi]$. However, the term $E[\xi, \tilde{P}QY]^2$ is being estimated by 0 - namely, if we substitute $\tilde{Q}$ for Q, then $[\xi, \tilde{P}\tilde{Q}Y]^2 = 0$. This is a second source of bias. Unfortunately, the above analysis is so crude that one does not obtain any idea of the magnitude of either source of bias. It seems that particular models will have to be studied in detail to obtain exact information about the bias. However, the work of Freedman and Peters (1983) suggests that the bias is substantial even in situations where one might believe the asymptotics are valid.

An analysis similar to that above is valid in the prediction problem. Here is a brief outline of the details for the situation treated in Proposition 5.2 and Proposition 7.2. For $\xi \in V_2'$, observe that

$$\tau^2 = \text{var}[\xi, Y_2 - \hat{Y}_2] = \text{var}[\xi, Y_2 - A_0 Y_1] \qquad (8.6)$$

where A_0 is given by (5.4). As in Remark 7.7 let

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

be the covariance of $\{Y_1, Y_2\}$ in $V_1 \oplus V_2$.

Proposition 8.2: For Σ non-negative definite and $\xi \in V_2'$, the map $\Sigma \to \text{var}[\xi, Y_2 - A_0 Y_1]$ is concave.

Proof: With $\mathcal{A}$ given by (5.2), for $A \in \mathcal{A}$, the function

$$\Psi(A, \Sigma) = E_\Sigma [\xi, Y_2 - AY_1]^2$$

is an affine and hence concave function of Σ. By Proposition 5.2, the inf over $\mathcal{A}$ of $\Psi(A, \Sigma)$ is (8.6). Now repeat the proof of Proposition 8.1. ■

The estimation of τ^2 is more problematic than the estimation of σ^2 since τ^2 depends on Σ_{22}. Even though residual type estimators can often be constructed for Σ_{11} and Σ_{12}, it is only in special circumstances that such estimators can be constructed for Σ_{22}. However, when such estimators are at hand and are unbiased, the obvious plug-in type estimates of τ^2 will have the same type of bias that $\tilde{\sigma}^2$ has. In particular, when (7.12) holds, there will be two sources of bias - one from Jensen's inequality and one from estimating $E[\xi,Z]^2$ to be zero.

It is not clear what to do about the bias problem. In many situations, an asymptotic argument shows that the lowest order term in $\text{var}([\xi,\tilde{\mu}])$ is $[\xi,P\Sigma P'\xi]$, but the evidence suggests that the higher order terms matter even for moderate samples. Furthermore, $\tilde{\mu} = \tilde{P}Y$ is often a very complicated function of $\tilde{\Sigma}$ and this precludes the usual Taylor series arguments to pick up the higher order terms in $\text{var}([\xi,\tilde{\mu}])$. One possibility is the bootstrap, but at present, there is very little theoretical justification for its use except in the simplest situations (see Freedman (1981)).

References

[1] Anderson, T.W. (1958). An Introduction to Multivariate Statistical Analysis. Wiley, New York.

[2] Arnold, S.F. (1981). The Theory of Linear Models and Multivariate Analysis. Wiley, New York.

[3] Cambanis, S., Huang, S. and Simons, G. (1981). On the theory of elliptically contoured distributions. J. Mult. Anal. 11, 368-385.

[4] Cox, D.R. and Hinkley, D.V. (1974). Theoretical Statistics. Chapman Hall, London.

[5] Eaton, M.L. (1970). Gauss-Markov estimation for multivariate linear models: A coordinate free approach. Ann. Math. Statist. 41, 528-538.

[6] Eaton, M.L. (1972). Multivariate Statistical Analysis. Institute of Mathematical Statistics, University of Copenhagen.

[7] Eaton, M.L. (1972). A note on the Gauss-Markov Theorem. Preprint, Institute of Mathematical Statistics, University of Copenhagen.

[8] Eaton, M.L. (1978). A note on the Gauss-Markov Theorem. Ann. Inst. Statist. Math. 30, 181-184.

[9] Eaton, M.L. (1983). Multivariate Statistics: A Vector Space Approach. Wiley, New York.

[10] Freedman, D.A. (1981). Bootstrapping regression models. Ann. Statist. 9, 1218-1228.

[11] Freedman, D.A. and Peters, S.C. (1982). Bootstrapping a regression equation: Some empirical results. Technical Report No. 10, Department of Statistics, University of California, Berkeley (Revised March, 1983).

[12] Freedman, D.A. and Peters, S.C. (1983). Bootstrapping a regression equation: Some empirical results. Technical Report No. 21, Department of Statistics, University of California, Berkeley (Revision of Technical Report No. 10).

[13] Goldberger, A.S. (1962). Best linear unbiased prediction in the generalized linear regression model. J. Amer. Statist. Assoc. 57, 369-375.

[14] Halmos, P.R. (1958). Finite-Demensional Vector Spaces. Springer-Verlag, New York.

[15] Kariya, T. and Toyooka, Y. (1983). Nonlinear versions of the Gauss-Markov Theorem and the GLSE. Paper presented at the Sixth International Symposium on Multivariate Analysis, held at University of Pittsburgh, July, 1983.

[16] Khatri, C.G. and Shah, K.R. (1981). On the unbiased estimation of fixed effects in a mixed model for growth curves. Comm. Statist. A 10, 401-406.

[17] Kruskal, W. (1961). The coordinate free approach to Gauss-Markov estimation and its application to missing and extra observations, in Proc. Fourth Berkeley Symp. Math. Stat. Prob., Vol. 1. University of California Press, Berkeley, 435-461.

[18] Marshall, A.W. and Olkin, I. (1979). Inequalities: Theory of Majorization and Its Applications. Academic Press, New York.

[19] Rao, C.R. (1967). Least squares theory using an estimated dispersion matrix and its application to measurement of signals. Proc. Fifth Berkeley Symp. Math. Statist. and Prob. 1, 355-372. University of California Press.

[20] Rao, C.R. (1973). Linear Statistical Inference and its Application, second edition. Wiley, New York.

[21] Scheffé, H. (1959). The Analysis of Variance, Wiley, New York.

[22] Takeuchi, K., Yanai, H., and Mukherjee, B.N. (1982). The Foundations of Multivariate Analysis. Wiley Eastern Limited, New Delhi.

[23] Theil, H. (1971). Principles of Econometrics. Wiley, New York.

[24] Toyooka, Y. (1982). Prediction error in a linear model with estimated parameters. Biometrika, 69, 453-459.

[25] Williams, J.S. (1975). Lower bounds on convergence rates of weighted least squares to best linear unbiased estimators. In Survey of Statistical Design and Linear Models, ed. by Srivastava. 555-570. North Holland, Amsterdam.

[26] Ylvisaker, N.D. (1964). Lower bounds for minimum covariance matrices in time series regression problems. Ann. Math. Statist. 35, 362-368.

[27] Zyskind, G. (1967). On canonical forms, non-negative covariance matrices and best and simple least squares linear estimators in linear models. Ann. Math. Statist. 38, 1092-1109.