The Annals of Statistics
1975, Vol. 3, No. 4, 982-990

HOW MUCH DO GAUSS-MARKOV AND LEAST SQUARE ESTIMATES DIFFER? A COORDINATE-FREE APPROACH[1]

BY SHELBY J. HABERMAN

University of Chicago

A simple expression is developed for the difference between the least squares and minimum variance linear unbiased estimators obtained in linear models in which the covariance operator of the observation vector is nonsingular. Bounds and series expansion for this difference are obtained, and bounds for the efficiency of least squares estimates are also obtained.

1. Introduction. Kruskal (1968) has used coordinate-free methods to establish a necessary and sufficient condition for the least squares estimator in a linear model to be the minimum variance linear unbiased (Gauss–Markov) estimator. In this paper, similar methods are used to obtain a general formula for the difference between the least squares estimator and the minimum variance unbiased estimator when the covariance operator of the observations is nonsingular. This formula is useful in examining the effects of departures from underlying assumptions in regression models. It can also be helpful in computation of the minimum variance linear unbiased estimator in cases in which the least squares estimator has already been determined or is easily computed.

2. Least squares estimators and Gauss–Markov estimators. In this section, the general model considered by Kruskal (1961, 1968) is employed. In this model, Y is a random vector in an n-dimensional inner product space W with inner product $(\cdot, \cdot)$. It is assumed that Y has an expectation μ and a covariance operator $\sum$. In other words, for all vectors x and z in the space,

$$E(x, Y) = (x, \mu)$$

and

$$\text{Cov}\,[(x, Y), (z, Y)] = (x, \textstyle\sum z)\,.$$

The vector μ is assumed to lie in a linear manifold Ω, and $\sum = \sigma^2 V$, where V is a known symmetric positive definite linear transformation.

Given the inner product $(\cdot, \cdot)$, the orthogonal projection P on Ω is defined for any $x \in W$ to be the unique element Px of Ω such that $(x - Px, z) = 0$ for all $z \in \Omega$. If $\mu^* = PY$, then μ^* is the least squares estimator of μ. Thus, μ^* is

Received September 1972; revised November 1974.

[1] Support for this research has been provided in part by Research Grant No. NSF GP 32037X from the Division of Mathematical, Physical, and Engineering Sciences of the National Science Foundation, and in part by the Department of Statistics, University of Chicago.

AMS 1970 *subject classifications.* Primary 62J05, 62J10.

Key words and phrases. Gauss–Markov estimates, least squares, efficiency, linear models.

the unique element of Ω such that

$$(Y - \mu^*, z) = 0 \qquad \text{for all} \quad z \in \Omega .$$

If $((\cdot, \cdot))$ is the inner product defined for x and z in the space by

$$((x, z)) = (x, V^{-1}z) ,$$

then the projection Q on Ω orthogonal with respect to $((\cdot, \cdot))$ is defined for any $x \in W$ to be the unique element Qx of Ω such that

$$((x - Qx, z)) = 0 \qquad \text{for all} \quad z \in \Omega .$$

If $\hat{\mu} = QY$, then $\hat{\mu}$ is the Gauss–Markov estimate of μ; that is, $\hat{\mu}$ is the unique element of Ω such that

$$((Y - \hat{\mu}, z)) = 0 \qquad \text{for all} \quad z \in \Omega .$$

To compare μ^* and $\hat{\mu}$, some preliminary observations are needed concerning the relationships between adjoints and projections. The adjoint A^* of a linear transformation A on W is the unique linear transformation such that for all x and z in W,

$$(Ax, z) = (x, A^*z) .$$

If A and B are linear transformations on W, then

$$(AB)^* = B^*A^*$$

and

$$A^{**} = A .$$

The projections P and Q then satisfy the relationships

$$P^2 = P = P^* ,$$
$$Q^2 = Q ,$$
$$Q^*V^{-1} = V^{-1}Q ,$$
$$PQ = Q ,$$
$$QP = P .$$

The relationship between P and Q is described by means of the following theorem.

THEOREM 1. *If A is a linear transformation on W such that*

(1) $$Ax = PV^{-1}x , \qquad x \in \Omega ,$$

then

(2) $$AQ = AP + PV^{-1}(I - P)$$

and

(3) $$A\hat{\mu} = A\mu^* + PV^{-1}(Y - \mu^*) .$$

PROOF. It suffices to note that

(4) $$\begin{aligned} AQ &= PV^{-1}Q = PQ^*V^{-1} = (QP)^*V^{-1} = P^*V^{-1} = PV^{-1} \\ &= PV^{-1}P + PV^{-1}(I - P) = AP + PV^{-1}(I - P) . \end{aligned}$$

As a corollary to Theorem 1, one has the following result.

COROLLARY 1. *If*

$$R = (I - P) + PV^{-1}$$

and

$$S = (I - P) + PV^{-1}P ,$$

then the following equations are satisfied:

$$\begin{aligned} Q &= P + R^{-1}PV^{-1}(I - P) \\ &= P + [I - R^{-1}](I - P) \\ &= P + S^{-1}PV^{-1}(I - P) . \\ \hat{\mu} &= \mu^* + R^{-1}PV^{-1}(Y - \mu^*) \\ &= \mu^* + [I - R^{-1}](Y - \mu^*) \\ &= \mu^* + S^{-1}PV^{-1}(Y - \mu^*) . \end{aligned}$$

2.1. *Conditions for identity of $\hat{\mu}$ and μ^*.* Theorem 1 may be used to provide the following necessary and sufficient conditions for the least squares and Gauss–Markov estimators to be identical.

THEOREM 2. *The estimators $\hat{\mu}$ and μ^* are identical if and only if one of the following equivalent conditions is satisfied*:

(5) $$PV^{-1}(I - P) = 0 .$$

(6) $$(I - P)V^{-1}P = 0 .$$

(7) $$V^{-1}\Omega \subset \Omega .$$

(8) $$V\Omega \subset \Omega .$$

(9) $$V^{-1}\Omega^{\perp} \subset \Omega^{\perp} .$$

(10) $$V\Omega^{\perp} \subset \Omega^{\perp} .$$

REMARKS. Equations (7) and (8) are given by Kruskal (1968). In (9) and (10), $\Omega^{\perp}$ is the orthogonal complement of Ω (see Halmos (1958, page 123)).

PROOF. By Corollary 1, $\hat{\mu}$ and μ^* are identical if and only if (5) holds. Condition (5) is equivalent to the assertion that the range of $V^{-1}(I - P)$ is in the null space of P. Since $V^{-1}(I - P)$ has range $V^{-1}\Omega^{\perp}$ and P has null space $\Omega^{\perp}$ (see Halmos (1958, pages 74 and 146)), (5) and (9) are equivalent. Since V is nonsingular, (9) and (10) are equivalent. Since

$$[PV^{-1}(I - P)]^* = (I - P)V^{-1}P ,$$

(5) and (6) are equivalent.

It may be the case that $\hat{\mu}$ and μ^* are not identical, but the least squares estimator $\beta^* = (c, \mu^*)$ and the Gauss–Markov estimator $\hat{\beta} = (c, \hat{\mu})$ of a linear functional $\beta = (c, \mu)$ of μ may still be identical. Without loss of generality, one may assume that $c \in \Omega$ (see Halmos (1958, page 130)). In this case, we have the following theorem:

THEOREM 3. *The estimators $\hat{\beta}$ and β^* are identical if and only if $Vc \in \Omega$.*

REMARK. This theorem is also given in Zyskind (1967). The proof presented here is somewhat simpler than Zyskind's proof.

PROOF. The estimators $\hat{\beta}$ and β^* are equal if and only if

$$(c, Px) = (c, Qx), \qquad x \in W.$$

This condition is equivalent to the assertion that

$$(c, x) = (Q^*c, x), \qquad x \in W.$$

In turn, this condition is equivalent to the assertion that

$$Vc = VQ^*c = QVc.$$

Consequently, $\hat{\beta} = \beta^*$ if and only if $Vc \in \Omega$.

2.2. *Equality of P and $PV^{-1}P$.* If $P = PV^{-1}P$, then $S = I$ and Corollary 1 implies that

$$\hat{\mu} = \mu^* + PV^{-1}(Y - \mu^*). \tag{11}$$

This formula can be useful if P and V^{-1} are easily evaluated. The following theorem provides further insight into this formula:

THEOREM 4. *The following statements are equivalent*:

$$P = PV^{-1}P. \tag{12}$$

$$(x, z) = ((x, z)) \quad \textit{for all} \quad x \in \Omega \quad \textit{and} \quad z \in \Omega. \tag{13}$$

PROOF. Equation (12) holds if and only if for all $x, z \in W$

$$\begin{aligned}(Px, Pz) &= (x, Pz)\\ &= (x, PV^{-1}Pz)\\ &= (Px, V^{-1}Pz)\\ &= ((Px, Pz)).\end{aligned}$$

This condition holds if and only if (13) holds.

COROLLARY 2. *If β is a linear combination (c, μ) and if* (11), (12), *or* (13) *holds, then the least square estimate $\beta^* = (c, \mu^*)$ and the Gauss–Markov estimate $\hat{\beta} = (c, \hat{\mu})$ satisfy*

$$\hat{\beta} = \beta^* + (Pc, V^{-1}(Y - \mu^*)).$$

If $c \in \Omega$, then

$$\hat{\beta} = \beta^* + (c, V^{-1}(Y - \mu^*)).$$

PROOF. By Corollary 1,

$$\begin{aligned}\hat{\beta} &= (c, \mu^*) + (c, PV^{-1}(Y - \mu^*))\\ &= \beta^* + (Pc, V^{-1}(Y - \mu^*)).\end{aligned}$$

If $c \in \Omega$, then $Pc = c$.

3. The effect of small differences between P and $PV^{-1}P$. When the difference between P and $PV^{-1}P$ is small, S^{-1} is approximately equal to I and the difference between $\hat{\mu}$ and μ^* is approximately $PV^{-1}(Y - \mu^*) = PV^{-1}(I - P)Y$. To make this statement more precise, let $||\cdot||$ be a norm on W. Define the norm $||A||$ of a linear transformation A on W to be

$$||A|| = \sup_{x \in W; ||x||=1} ||Ax|| .$$

Given this definition, if A and B are two linear transformations, then

$$||AB|| \leqq ||A||\,||B|| \tag{14}$$

and

$$||A + B|| \leqq ||A|| + ||B|| . \tag{15}$$

If $||A|| < 1$, then $I - A$ is invertible and

$$||(I - A)^{-1}|| \leqq 1/(1 - ||A||) \tag{16}$$

(see Loomis and Sternberg (1968, page 224)).

Given these results, the following theorem is a simple consequence of Corollary 1.

THEOREM 5. *For any nonnegative integer k,*

$$\begin{aligned}\hat{\mu} - \mu^* = {} & PV^{-1}(Y - \mu^*) + \textstyle\sum_{j=1}^{k} (P - PV^{-1}P)^j PV^{-1}(Y - \mu^*) \\ & + S^{-1}(P - PV^{-1}P)^{k+1} PV^{-1}(Y - \mu^*) .\end{aligned} \tag{17}$$

If $||P - PV^{-1}P|| < 1$, then

$$\begin{aligned}&||S^{-1}(P - PV^{-1}P)^{k+1} PV^{-1}(Y - \mu^*)|| \\ &\qquad \leqq \frac{||P - PV^{-1}P||^{k+1}||PV^{-1}(I - P)||\,||Y - \mu^*||}{1 - ||P - PV^{-1}P||} .\end{aligned} \tag{18}$$

PROOF. Note that

$$\begin{aligned}S^{-1} &= [I - (P - PV^{-1}P)]^{-1} \\ &= I + \textstyle\sum_{j=1}^{k} (P - PV^{-1}P)^j + S^{-1}(P - PV^{-1}P)^{k+1}\end{aligned} \tag{19}$$

Equation (17) follows immediately. Equation (18) follows by application of (14), (16), the first equation in (19), and the observation that

$$(I - P)(Y - \mu^*) = Y - \mu^* .$$

In the particular case in which $||P - PV^{-1}P|| < 1$ and k is 0, Corollary 1, (19), and Theorem 5 lead to the following corollary:

COROLLARY 3. *If $||P - PV^{-1}P|| < 1$, then*

$$||\hat{\mu} - \mu^*|| \leqq \frac{||PV^{-1}(I - P)||\,||Y - \mu^*||}{1 - ||P - PV^{-1}P||}$$

and

$$||\hat{\mu} - \mu^* - PV^{-1}(Y - \mu^*)|| \leqq \frac{||P - PV^{-1}P||\,||PV^{-1}(I - P)||\,||Y - \mu^*||}{1 - ||P - PV^{-1}P||} .$$

4. A canonical analysis of $\hat{\mu} - \mu^*$. In this section, a description of $\hat{\mu} - \mu^*$ is obtained in terms of appropriately chosen bases $\{x_i : i = 1, \cdots, m\}$ and $\{z_j : j = 1, \cdots, n - m\}$ of Ω and $\Omega^\perp$, respectively, where m, the dimension of Ω, is assumed in this section to satisfy the inequality $0 < m < n$.

The basic observation required in this section is the result given in Dempster (1969, page 99) that $\{x_i : i = 1, \cdots, m\}$ and $\{z_j : j = 1, \cdots, n - m\}$ may be chosen so that $((x_i, x_{i'})) = 0$ if $i \neq i'$, $((x_i, x_i)) = 1$, $((z_j, z_{j'})) = 0$ if $j \neq j'$, $((z_j, z_j)) = 1$, and $((x_i, z_j)) = 0$ if $i \neq j$. If k is the minimum of m and $n - m$, then the vectors may be ordered so that if $\theta_i = ((x_i, z_i))$, $i = 1, \cdots, k$, then $|\theta_i| \geqq |\theta_{i'}|$ if $1 \leqq i < i' \leqq k$. Since $((x_i, x_i)) = ((z_i, z_i)) = 1$ and $x_i \neq z_i$, Schwarz's inequality implies that $|\theta_i| < 1$ for $i = 1, \cdots, k$.

Given this result, the following theorem may be proven:

THEOREM 6. *The difference between $\hat{\mu}$ and μ^* satisfies*

$$\hat{\mu} - \mu^* = \sum_{j=1}^{k} [\theta_j((z_j - \theta_j x_j, Y))/(1 - \theta_j^2)]x_j . \tag{20}$$

PROOF. Let $u \otimes v$ be defined for $u, v \in W$ as the linear transformation on W such that

$$(u \otimes v)w = u((v, w)) , \qquad w \in W .$$

It is well known that

$$\hat{\mu} = QY = \sum_{j=1}^{m} ((x_j, Y))x_j .$$

If

$$U = \sum_{j=1}^{k} \frac{x_j \otimes (x_j - \theta_j z_j)}{1 - \theta_j^2} + \sum_{j=k+1}^{m} x_j \otimes x_j ,$$

then $Ux_j = x_j$ for $j = 1, \cdots, m$ and $Uz_j = 0$ for $j = 1, \cdots, n$. Thus $Ux = x$ for $x \in \Omega$ and $Ux = 0$ for $x \in \Omega^\perp$. Consequently, $P = U$. It now follows that

$$\begin{aligned} \hat{\mu} - \mu^* &= QY - UY \\ &= \sum_{j=1}^{k} [((x_j, Y)) - ((x_j - \theta_j z_j, Y))/(1 - \theta_j^2)]x_j \\ &= \sum_{j=1}^{k} [\theta_j((z_j - \theta_j x_j, Y))/(1 - \theta_j^2)]x_j . \end{aligned}$$

This theorem implies that $\hat{\mu} - \mu^*$ depends only on the k random variables $((z_j - \theta_j x_j, Y))$, $j = 1, \cdots, k$. Each of these variables has mean

$$\begin{aligned} ((z_j - \theta_j x_j, \mu)) &= \sum_{i=1}^{m} ((x_i, \mu))((z_j - \theta_j x_j, x_i)) \\ &= 0 \end{aligned}$$

and variance

$$\sigma^2((z_j - \theta_j x_j, z_j - \theta_j x_j)) = \sigma^2(1 - \theta_j^2) .$$

Since

$$((z_j - \theta_j x_j, z_{j'} - \theta_{j'} x_{j'})) = 0$$

if $j \neq j'$, the random variables $((z_j - \theta_j x_j, Y))$ and $((z_{j'} - \theta_{j'} x_{j'}, Y))$ are uncorrelated when $j \neq j'$.

Given Theorem 6, a simple expression is available for the efficiency of a least

squares estimate β^* of a linear functional β of μ. As noted by Halmos (1958, page 130), to a linear functional β of μ corresponds a $c \in \Omega$ such that

$$\beta = ((c, \mu)) .$$

By Kruskal (1961, 1968), the corresponding Gauss–Markov estimate is

$$\begin{aligned} \hat{\beta} &= ((c, \hat{\mu})) \\ &= ((c, Y)) , \end{aligned}$$

and the corresponding least squares estimate is

$$\beta^* = ((c, \mu^*)) .$$

By (20),

$$\beta^* = \hat{\beta} - \textstyle\sum_{j=1}^{k} \theta_j((z_j - \theta_j x_j, Y))((x_j, c))/(1 - \theta_j^2) .$$

Since

$$((z_j - \theta_j x_j, c)) = 0 , \qquad j = 1, \cdots, k ,$$

$\hat{\beta}$ is uncorrelated with $((z_j - \theta_j x_j, Y))$, $j = 1, \cdots, k$. Thus

$$\operatorname{Var}(\beta^*) = \operatorname{Var}(\hat{\beta}) + \textstyle\sum_{j=1}^{k} ((x_j, c))^2 \theta_j^2/(1 - \theta_j^2) .$$

By Parseval's identity (see Halmos (1958, page 124)),

$$\operatorname{Var}(\hat{\beta}) = ((c, c)) = \textstyle\sum_{j=1}^{m} ((x_j, c))^2 .$$

Thus

$$\operatorname{Var}(\beta^*) = \textstyle\sum_{j=1}^{k} ((x_j, c))^2/(1 - \theta_j^2) + \sum_{j=k+1}^{m} ((x_j, c))^2 ,$$

where the second summation is 0 if $k \geqq m$. The efficiency is then

$$\frac{\operatorname{Var}(\hat{\beta})}{\operatorname{Var}(\beta^*)} = \frac{\sum_{j=1}^{m} ((x_j, c))^2}{\sum_{j=1}^{k} ((x_j, c))^2/(1 - \theta_j^2) + \sum_{j=k+1}^{m} ((x_j, c))^2} .$$

Since $1 - \theta_j^2 \leqq 1$, this equation for the efficiency is always less than or equal to 1, as is to be expected. At the other extreme, the efficiency is at least $1 - \theta_1^2$, with the lower bound achieved if $c = x_1$. Following Dempster (1969, page 99), it is useful to note that

$$\theta_1^2 = \sup_{x \in \Omega, x \neq 0; z \in \Omega^\perp, z \neq 0} \frac{((x, z))^2}{((x, x))((z, z))} .$$

Thus θ_1^2 is the square of the maximum correlation between a nonzero linear functional $((x, Y))$, $x \in \Omega$, and a nonzero linear functional $((z, Y))$, $z \in \Omega^\perp$.

The following bounds for the efficiency may be obtained in terms of the maximum eigenvalues γ and δ of V^{-1}:

THEOREM 7. *The variances of $\hat{\beta}$ and β^* satisfy the relationship*

$$1 \geqq \frac{\operatorname{Var}(\hat{\beta})}{\operatorname{Var}(\beta^*)} \geqq \frac{4\gamma\delta}{(\gamma + \delta)^2} = \frac{4\tau}{(\tau + 1)^2} ,$$

where $\tau = \gamma/\delta$ is the ratio of the largest and smallest eigenvalues of V^{-1} (or V).

PROOF. By application of Cleveland's (1971) Theorem 5 to the span of x, one finds that if $x \in \Omega$, $((x, x)) = 1$, $z \in \Omega^{\perp}$, and $((z, z)) = 1$, then

$$\frac{((z, z))}{((z, z)) - ((x, z))^2((x, x))} = \frac{1}{1 - ((x, z))^2} \leqq \frac{(\gamma + \delta)^2}{4\gamma\delta}.$$

Thus $\text{Var}\,(\hat{\beta})/\text{Var}\,(\beta^*) \geqq 1 - \theta_1^2 \geqq 4\gamma\delta/(\gamma + \delta)^2$.

In the special case in which Y is an element of R^n, $(\cdot, \cdot)$ is the Euclidean inner product, $\text{Cov}\,(Y_i, Y_j) = \sigma^2 v_{ij}$, and $v_{ij} = 0$ if $i \neq j$, then

$$\delta = \min_{1 \leqq i \leqq n} \frac{1}{v_{ii}}$$

and

$$\gamma = \max_{1 \leqq i \leqq n} \frac{1}{v_{ii}}.$$

If the covariance structure is permutation-invariant, with $v_{ii} = 1$ and $v_{ij} = \rho$ for $i \neq j$, where $-1/(n-1) < \rho < 1$, then

$$\begin{aligned} 1/\delta &= 1 + (n-1)\rho \quad \text{if} \quad \rho \geqq 0, \\ &= 1 - \rho \quad \text{if} \quad \rho < 0, \end{aligned}$$

and

$$\begin{aligned} 1/\gamma &= 1 - \rho \quad \text{if} \quad \rho \geqq 0, \\ &= 1 + (n-1)\rho \quad \text{if} \quad \rho < 0. \end{aligned}$$

Thus the lower bound for the efficiency of β^* is

$$\frac{4\gamma\delta}{(\gamma + \delta)^2} = \frac{4(1 - \rho)[1 + (n-1)\rho]}{[2 + (n-2)\rho]^2}.$$

The canonical decomposition and efficiency analysis provided in this section is related to work by Watson (1967). Watson's results are based on coordinate systems and generalized variances and are not strictly comparable to those derived in this paper. Theorem 7 is essentially equivalent to a result in Golub (1963) which is expressed in terms of coordinate systems. Referees have made substantial contributions to the exposition. In particular, their help has contributed to the formulation of Theorem 1 and to simplifications in the proofs of Theorems 3, 6, and 7.

REFERENCES

[1] CLEVELAND, W. S. (1971). Projection with the wrong inner product and its application to regression with correlated errors and linear filtering of time series. *Ann. Math. Statist.* **42** 616–624.

[2] DEMPSTER, A. P. (1969). *Elements of Continuous Multivariate Analysis.* Addison-Wesley, Reading, Mass.

[3] GOLUB, G. H. (1963). Comparison of the variance of minimum variance and weighted least squares regression coefficients. *Ann. Math. Statist.* **34** 984–992.

[4j HALMOS, P. A. (1958). *Finite-dimensional Vector Spaces.* Van Nostrand, Princeton.

[5] KRUSKAL, W. H. (1961). The coordinate-free approach to Gauss-Markov estimation, and its application to missing and extra observations. *Fourth Berkeley Symp. Math. Statist. Prob.* **1** 435–451.

[6] KRUSKAL, W. H. (1968). When are Gauss-Markov and least squares estimators identical? A coordinate-free approach. *Ann. Math. Statist.* **39** 70–75.

[7] LOOMIS, L. and STERNBERG, S. (1968). *Advanced Calculus.* Addison-Wesley, Reading, Mass.

[8] WATSON, G. S. (1967). Linear least squares regression. *Ann. Math. Statist.* **38** 1679–1699.

[9] 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.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
1118 EAST 58TH STREET
CHICAGO, ILLINOIS 60637