The American Statistician, in press

The Gauss-Markov Theorem and Random Regressors

By

Juliet Popper Shaffer*

Technical Report No. 234
January 1990
Revised June 1991

Requests for reprints should be sent to Juliet P. Shaffer,
Department of Statistics, University of California, Berkeley, CA 94720.

Department of Statistics
University of California
Berkeley, California

THE GAUSS-MARKOV THEOREM AND RANDOM REGRESSORS

Juliet Popper Shaffer*

ABSTRACT

In the standard linear regression model with independent, homoescedastic errors, the Gauss-Markov theorem asserts that $\hat{\beta} = (X'X)^{-1} (X'y)$ is the best linear unbiased estimator of β, and furthermore that $c'\hat{\beta}$ is the best linear unbiased estimator of $c'\beta$ for all $p \times 1$ vectors c. In the corresponding random regressor model, X is a random sample of size n from a p-variate distribution. If attention is restricted to linear estimators of $c'\beta$ which are conditionally unbiased, given X, the Gauss-Markov theorem applies. If, however, the estimator is required only to be unconditionally unbiased, the Gauss-Markov theorem may or may not hold, depending upon what is known about the distribution of X. The results generalize to the case in which X is a random sample without replacement from a finite population.

Key words: Linear regression; Unbiased estimators; Best linear unbiased estimators; Finite-population sampling.

Requests for reprints should be sent to Juliet P. Shaffer, Department of Statistics, University of California, Berkeley, CA 94720

THE GAUSS-MARKOV THEOREM AND RANDOM REGRESSORS

1. INTRODUCTION

Assume a sample of n observations $y_1, y_2, \dots, y_n$ from the standard linear regression model. The assumptions are

$$E(y) = X\beta, \quad \Sigma_y = \sigma^2 I, \tag{1}$$

where y is the random $n \times 1$ observation vector, $n \geq p$, X is a fixed $n \times p$ matrix of rank p, β is a $p \times 1$ parameter vector, and Σ_y is the covariance matrix of y. The Gauss-Markov theorem states that $c'\hat{\beta} = c'(X'X)^{-1}(X'y)$ is the unique best linear unbiased estimator (BLUE) of $c'\beta$. Under the more general assumption $\Sigma_y = \sigma^2 B$, with B a known positive definite matrix, it follows that $c'\hat{\beta} = c'(X'B^{-1}X)^{-1}(X'B^{-1}y)$ is the BLUE of $c'\beta$. The theorem has been extended to apply when X is of rank < p, when B is singular, and when the parameters satisfy linear constraints (e.g.,Goldman and Zelen, 1964; Rao, 1972, 1973a, 1979, Schonfeld and Werner, 1987). Generalized forms of the theorem have been derived to apply to estimators utilizing estimated covariance matrices and to more general classes of nonlinear estimators (e.g.,Toyooka, 1984, Kariya, 1985, Kariya and Toyooka, 1985, Toyooka, 1987). The theorem has also been extended in various ways to the case of stochastic regression coefficients β (e.g.,Chipman, 1964, Rao, 1965, 1973b, Duncan and Horn, 1972, Rosenberg, 1972, Sarris, 1973, Harville, 1976, Pfeffermann, 1984).

The least squares estimator $\hat{\beta}$ of β is widely used in practice, because of the ease of calculating linear estimators, and because the Gauss-Markov theorem assures us that $\hat{\beta}$ is the best unbiased estimator within the class of linear estimators. However, in applied work, regression analysis is widely used in cases where the predictor variables X as well as the predicted variable y are random. The question of whether the Gauss-Markov theorem is valid in this case does not appear to have been considered in the literature. The random regression model corresponding to (1) is written as

$$E(y|X) = X\beta, \quad \Sigma_{y|X} = \sigma^2 I, \tag{2}$$

where (X, y) is a random sample of n (p + 1)-vectors, and (2) specifies the first two moments of the conditional distribution of y given X. It will be assumed initially that the joint distribution of a vector X is arbitrary except for the simplifying restriction that X is continuous and nondegenerate, so that the probability of a singular $n \times p$ matrix of observed values X is zero. (These assumptions on X will be modified in Section 2.3.) The distribution of y is arbitrary except for the restrictions in (2). Let $\delta(X, y)$ be an estimator of β. The estimator δ is linear in y if

$$\delta(X, y) = C'(X)y \tag{3}$$

where $C(X)$ is an $n \times p$ matrix of functions of X. To allow estimates of β unrestricted by linear equalities, C(X) must be of rank p.

The estimator is conditionally unbiased, given X, if $E[C'(X)y|X] = E[C'(X)X\beta|X] = \beta$, identically in β, implying

$$C'(X)X = I. \tag{4}$$

It is unconditionally unbiased if $E[C'(X)y] = \beta$, identically in β, implying

$$E[C'(X)X] = I. \tag{5}$$

There are differences of opinion in the statistical literature on the importance of unbiasedness and the appropriateness of unconditional as opposed to conditional inference. While some statisticians consider conditional unbiasedness desirable, others feel that unconditional unbiasedness is a sufficient requirement for estimators, especially if this less stringent requirement permits more efficient estimators. Therefore, both conditional and unconditional unbiasedness will be considered.

II. RESULTS

2.1 Conditional Unbiasedness

In the class of conditionally unbiased estimators, the Gauss-Markov theorem obviously holds, since $\hat{\beta}$ is then BLUE for every realized matrix X, and therefore is BLUE unconditionally. The interesting case is that in which the class is expanded to the larger class of all unconditionally unbiased estimators.

2.2 Unconditional Unbiasedness

Three cases will be considered: (a) (X, y) multivariate normal with unknown parameters, (b) Distribution of X completely unknown, and (c) $E(X'X)$ known.

(a) (X, y) multivariate normal with unknown parameters μ_X, μ_y, Σ_X, σ_y^2, and ρ_{Xy}.

For (X, y) multivariate normal, $\hat{\beta}$ is a function of the complete sufficient statistics, and therefore, by the Lehmann-Scheffé theorem, is the UMVU estimator of β. Thus, the Gauss-Markov theorem holds; in fact, $\hat{\beta}$ is best among all unbiased estimators, not only those linear in y.

(b) Distribution of X completely unknown; distribution of y arbitrary except for the assumptions (2).

Definition The c_j-order statistics of a sample of vectors are the vectors arranged in increasing order according to their jth components.

Lemma 1. Let $(X|y)$ be the n x p+1 matrix resulting from adjoining the y-vector to the matrix X. The c_j-order statistics of the row vectors of (X, y) for any $j = 1, \ldots, p + 1$ are sufficient statistics for (X, y) (except that if the first component is a constant, the c_1-order statistics would be excluded here and in the following discussion.)

Proof: Since the n row vectors (x_i, y_i) are a random sample, it follows that the conditional probability of (X, y), given the c_j-order statistics for any j, is $1/n!$, independent of the joint distribution of X and y.

Lemma 2: The c_j-order statistics of the row vectors of X for any $j = 1, \ldots, p$ are complete sufficient statistics for X.

Proof: For a sample of scalars the proof is given in Lehmann (1986, p. 173, Problem 12). That proof generalizes directly to a sample of p-vectors, replacing $I_{a_i}(x) F(x)$ by $I_{(a_1, a_2, \ldots, a_p)}(x_1, x_2, \ldots, x_p) F(x_1, x_2, \ldots, x_p)$.

Theorem 1. If the distribution of X is unknown, $\hat{\beta}$ is the BLUE of β.

Proof: The proof proceeds by demonstrating that the requirement $E[C'(X)X] = I$ and minimum variance implies $C'(X)X = I$ for almost all observed matrices X, i.e. that unconditional unbiasedness implies conditional unbiasedness, and therefore the Gauss-Markov theorem holds.

An estimator $T(X, y)$ is a symmetric function of (X, y) if it is a function of the c_j-order statistics of (X, y) for any j, $j = 1, \ldots, p + 1$. Suppose an estimator δ is not a symmetric function of (X, y). By Lemma 1, the c_j-order statistics are sufficient. Then by the Rao-Blackwell theorem the expected value of δ given the sufficient statistics, which is a symmetric function of (X, y), would have the same expected value as δ, but a smaller variance. Therefore, we can restrict attention to estimators δ which are symmetric functions of (X, y). It follows that $C'(X)X$ is a symmetric function of X. By Lemma 2 the order statistics of X are complete sufficient statistics. Consequently, by the Lehmann-Scheffe theorem, if $E(C'(X)X - I) = 0$, $C'(X)X - I$ is identically zero. Then δ must be conditionally unbiased, so the Gauss-Markov theorem holds. .

(c) $E(X'X)$ known.

There may be situations in which there is considerable previous information on the predictor variables, making it possible to assume a known value for $E(X'X)$. For example, in survey sampling, the demographic characteristics of a population, or the distribution of answers to some questions, may be known from census data, or known to a high degree of accuracy from a large sample. It may be desired to predict some new variable, such as the score on a test of scientific knowledge, from these demographic or other variables, assuming a regression model as in (2), and estimating the regression coefficients from a relatively small sample. Because the marginal distribution of X does not involve the parameters β and σ^2, it might be thought that the information on X would be irrelevant, and that the Gauss-Markov theorem would still apply, so that there would be no reason to consider alternative linear estimators of β. However, this turns out not to be the case.

Theorem 2. If $E(X'X)$ is known, no BLUE of β exists.

In comparing the covariance matrices of two vectors v_1 and v_2, the statement $\mathrm{Var}(v_1) > \mathrm{Var}(v_2)$ will mean that the difference $\mathrm{Var}(v_1) - \mathrm{Var}(v_2)$ between the two covariance matrices is positive definite. The proof of Theorem 2 consists of two parts:

(i) It will be shown that the variance of $\hat{\beta}$ is smaller than the variance of any conditionally biased linear estimator for sufficiently large β, so that no conditionally biased estimator can be BLUE, and the only possible BLUE is $\hat{\beta}$.

(ii) It will be shown that there exists a conditionally biased linear estimator with a smaller variance than $\hat{\beta}$ for sufficiently small values of β. Therefore, $\hat{\beta}$ cannot be BLUE.

These two facts taken together imply that no BLUE exists. The proof requires the following 2 lemmas.

Lemma 3. The set of $p \times p$ matrices $X'X$ is convex, where X is $n \times p$ of rank p.

Proof. Given two matrices $X_1'X_1$ and $X_2'X_2$, it is sufficient to prove that $\alpha(X_1'X_1) + (1-\alpha)(X_2'X_2) = X_3'X_3$, where X_3 is $n \times p$ of rank p.

Since $X_1'X_1$ and $X_2'X_2$ are positive definite, there exists a nonsingular $p \times p$ matrix M such that $X_1'X_1 = M'M$ and $X_2'X_2 = M'D_\theta M$, where $D_\theta = \mathrm{diag}(\theta_1, \theta_2, \ldots, \theta_p)$ and $\theta_1 \ge \theta_2 \ge \cdots \ge \theta_p > 0$ are the eigenvalues of $(X_2'X_2)(X_1'X_1)^{-1}$ (Marshall and Olkin, 1979). Then $\alpha(X_1'X_1) + (1-\alpha)(X_2'X_2) = M'[\alpha I + (1-\alpha)D_\theta]M = U'U$, where $U = [\alpha I + (1-\alpha)D_\theta]^{.5}$.

Adjoin an $(n-p) \times p$ matrix of zeros to U, calling the result X_3. Then $\alpha(X_1'X_1) + (1-\alpha)(X_2'X_2) = X_3'X_3$, where X_3 is an $n \times p$ matrix of rank p, as was to be shown.

Lemma 4. $E[(X'X)^{-1}] - [E(X'X)]^{-1}$ is positive definite.

Proof. Note that for $p = 1$, the lemma follows directly from Jensen's inequality, since the reciprocal is a convex function. Consider any convex set $\underline{A}$ of matrices. A function $\phi(A)$, $A \in \underline{A}$ is said to be strictly matrix-convex if, for any A_1, A_2 in $\underline{A}$, $[\alpha\phi(A_1) + (1-\alpha)\phi(A_2)] - \phi[\alpha A_1 + (1-\alpha)A_2]$ is positive definite. The function $\phi(A) = A^{-1}$ is strictly matrix-convex in $\underline{A}$ (Marshall and Olkin, 1979).

The set of $p \times p$ matrices $X'X$ is convex by Lemma 3. Then for any two such matrices $X_1'X_1$ and $X_2'X_2$,

$$\alpha (X_1{'}X_1)^{-1} \; + \; (1-\alpha)(X_2{'}X_2)^{-1} \; - \; [\alpha X_1{'}X_1 + (1-\alpha)X_2{'}X_2]^{-1} \tag{6}$$

is positive definite.

It follows by induction that, for any m, given a set of matrices $X_1{'}X_1$, $X_2{'}X_2, \ldots, X_m{'}X_m$, and a set of numbers $\alpha_1, \alpha_2, \ldots, \alpha_m$ such that $\sum_{i=1}^{i=m} \alpha_i = 1$,

$$\sum_{i=1}^{i=m} \alpha_i (X_i{'}X_i)^{-1} - [\sum_{i=1}^{i=m} \alpha_i X_i{'}X_i]^{-1} \tag{7}$$

is positive definite. By appropriate choice of the α_i and by taking a limit as $m \to \infty$, the conclusion follows.

We'll now turn to the proof of Theorem 2.

(i) We have

$$\mathrm{Var}(\hat{\beta}) \; = \; \sigma^2 E[(X'X)^{-1}]. \tag{8}$$

The unconditional variance of the linear unbiased estimator (3) is $E_X[\mathrm{Var}(\delta | X)] + \mathrm{Var}[E(\delta | X)]$, or

$$\mathrm{Var}(\delta) \; = \; \sigma^2 E[C'(X)C(X)] \; + \; \mathrm{Var}[C'(X)X\beta]. \tag{9}$$

Note that the second term of (9) will be zero if and only if the estimator is conditionally unbiased. Note also that the second term, if nonzero, increases without bound as the components of β increase. Therefore, the variance of any estimator δ which is conditionally biased will be larger than the variance of $\hat{\beta}$ for a sufficiently large β.

(ii) Let $\hat{\hat{\beta}} \; = \; [E(X'X)]^{-1}X'y$. The variance of $\hat{\hat{\beta}}$ is

$$\mathrm{Var}(\hat{\hat{\beta}}) \; = \; \sigma^2[E(X'X)]^{-1} \; + \; \mathrm{Var}[E(X'X)]^{-1}X'X\beta]. \tag{10}$$

For $\beta = 0$, $\mathrm{Var}(\hat{\beta}) - \mathrm{Var}(\hat{\hat{\beta}}) = \sigma^2\{E(X'X)^{-1} - [E(X'X)]^{-1}\}$, which is positive definite by Lemma 4, so that the variance of $\hat{\hat{\beta}}$ is smaller than the variance of $\hat{\beta}$. Furthermore, by continuity of $\hat{\beta}$ and $\hat{\hat{\beta}}$, the variance of $\hat{\hat{\beta}}$ is smaller than the variance of $\hat{\beta}$ in a neighborhood of $\beta = 0$.

This completes the proof of Theorem 2.

From a practical point of view, the possible magnitude of the difference between the variances of $\hat{\beta}$ and $\hat{\hat{\beta}}$ is of interest. This is easily determined in the very simple situation in which X_1 is identically one (in other words, the model contains an intercept), and X_2 through X_p have a joint multivariate normal distribution with arbitrary mean vector and covariance matrix. Assume a random sample of size n, let $\beta_{(-1)}$ be the (p–1)-vector of regression coefficients excluding β_1, and let $\hat{\beta}_{(-1)}$ and $\hat{\hat{\beta}}_{(-1)}$ be the estimators $\hat{\beta}$ and $\hat{\hat{\beta}}$ of the corresponding coefficients. Let S be the usual unbiased estimator of the covariance matrix Σ, $S = (s_{ij})$, where

$s_{ij} = \Sigma_{k=1}^{k=n} (X_{ik} - \overline{X}_i)(X_{jk} - \overline{X}_j)/(n-1)$. Using results on partitioned matrices (Rao, 1973b, p. 33), the variances when $\beta = 0$ can be expressed as follows:

$$\text{Var}(\hat{\beta}_{(-1)}) = \sigma^2 E\{[(n-1)S]^{-1}\} \tag{11}$$

and

$$\text{Var}(\hat{\hat{\beta}}_{(-1)}) = \sigma^2 [n\Sigma]^{-1} = [\sigma^2/n]\Sigma^{-1}. \tag{12}$$

Since $[(n-1)S]^{-1}$ has the inverse Wishart distribution, its expected value is $\Sigma^{-1}/(n-p-1)$ (Johnson & Kotz, 1972, p. 164), so

$$\text{Var}(\hat{\beta}_{(-1)}) = [\sigma^2/(n-p-1)]\Sigma^{-1}. \tag{13}$$

Therefore, for each corresponding component i of $\hat{\beta}_{(-1)}$ and $\hat{\hat{\beta}}_{(-1)}$ we have

$$\text{Var}(\hat{\hat{\beta}}_{(-1)i})/\text{Var}(\hat{\beta}_{(-1)i}) = (n-p-1)/n. \tag{14}$$

If n is very much larger than p, the difference in efficiency between $\hat{\beta}_{(-1)}$ and $\hat{\hat{\beta}}_{(-1)}$ in the vicinity of $\beta = 0$ will be negligible. However, in applied work, n and p are often of the same order of magnitude, at least in an initial model, since regression models are often fitted to screen out variables low in predictive power. Freedman, Navidi, and Peters (1987) consider $n = 100$ and $p = 75$ not unrealistic in applications to econometric modelling. Breiman and Spector (1989) carry out simulations with $n = 60$ and $p = 40$, using $\hat{\beta}$ only. Using these latter values, we get for corresponding components of $\hat{\beta}_{(-1)}$ and $\hat{\hat{\beta}}_{(-1)}$

$$\text{Var}\,\hat{\hat{\beta}}_{(-1)i}/\text{Var}\,\hat{\beta}_{(-1)i} = 19/60 = 1/3.16, \tag{15}$$

so that there is a more than three-fold difference between the variances of the two estimators in the vicinity of $\beta = 0$.

Suppose $\beta \neq 0$. The variance of $\hat{\beta}$ is not affected by the value of β, while the variance of $\hat{\hat{\beta}}$ increases without bound with increasing absolute values of the components of β, as can be seen from (9). It may be of interest to see how large the components of β can be before the variance of $\hat{\hat{\beta}}$ becomes larger than that of $\hat{\beta}$. To compare the estimators for $\beta \neq 0$, assume X_1 is identically one, as above, and $X_2, X_3, \ldots, X_p$ have a joint multivariate normal distribution with mean vector identically zero and covariance matrix I. Then, again comparing variances of the corresponding components of $\hat{\hat{\beta}}_{(-1)}$ and $\hat{\beta}_{(-1)}$ we get

$$\text{Var}\,\hat{\hat{\beta}}_{(-1)i}/\text{Var}\,\hat{\beta}_{(-1)i} = [(n-p-1)/n]\,[1 + (\Sigma_{j=1}^{j=p} \beta_j^2 + \beta_i^2)/\sigma^2. \tag{16}$$

For the case $\beta_1 = 0$, and $\beta_2, \beta_3, \ldots, \beta_p = b$, and letting $c = b/\sigma$, (16) reduces to

$$\text{Var}\,\hat{\hat{\beta}}_{(-1)i}/\text{Var}\hat{\beta}_{(-1)i} = [(n-p-1)/n](1 + pc^2). \tag{17}$$

Considering again the values $n = 60$ and $p = 40$, the variance ratio becomes

$(19/60)(1 + 40c^2)$. In this case, for $c < .23$, the variance of $\hat{\hat{\beta}}_{(-1)i}$ is smaller than the variance of $\hat{\beta}_{(-1)i}$ for all i, while the reverse is true for $c > .23$.

When p is large relative to n, it is often the case that individual β values are small, due to the patterns of association among the X variables. Thus, $\hat{\hat{\beta}}$ may be much more efficient than $\hat{\beta}$ during initial variable selection. As the number of variables decreases and possibly some of the βs increase, there may be a point at which it is advisable to begin estimating some or all of the βs by $\hat{\beta}$. Further theoretical and empirical investigation will be necessary to evaluate these possibilities.

From (13) on, calculations are based on the assumption that $X_2, \ldots, X_p$ are jointly multivariate normal. Although Breiman and Spector (1989) didn't consider the use of $\hat{\hat{\beta}}$, they demonstrated that the properties of $\hat{\beta}$ are more sensitive to the X-fixed vs. X-random distinction under some non-normal X distributions than when X is normal, suggesting that the advantage of $\hat{\hat{\beta}}$ over $\hat{\beta}$ near $\beta = 0$ may be considerably greater for some nonnormal distributions of X.

2.3 Finite Populations

Under the model (2), the sample vectors (X, y) are independently distributed. However, in motivating the result in 2.2(c), reference is made to a sample survey from a finite population. In the classical finite population model, the vectors (X, y) would not be independently distributed. In order to subsume the application in 2.2(c) under the model (2), it is implicitly assumed that the (X, y) finite population values are generated from a multivariate normal superpopulation. It is then explicitly assumed that the finite population is large, and that there is sufficient previous information either from a complete census or a very large sample so that the estimate of $E(X'X)$ based on that information can be assumed approximately equal to the true value.

An alternative treatment of the finite population allows an extension of the results in 2.2(b) and 2.2(c) to a different random regressor model which is more appropriate in many survey sampling applications. Assume a finite population of N X-vector values, in which the associated y-values are realizations of random variables satisfying the assumptions of the standard linear regression model (1). A simple random sample of size n is drawn, so that the values of X in the sample are a random sample without replacement from a finite population.

Theorem 3. Assume

A. The labels attached to the sample elements are disregarded (as they would be if they carry no information).

B. The probability of a singular $X'X$ matrix is negligible and can be ignored.

Then

(i) If each element in the population has the same possible set of X-vector values, and the set is independent of the values taken on by any other elements, then $\hat{\beta}$ is the BLUE of β.

(ii) If $E(X'X)$ is known, no BLUE of β exists.

Proof. The c_j-order statistics of the row vectors of (X, y) may be defined as in Section 2.2, with random or arbitrary ordering in the case of tied values. The sufficiency of the c_j-order statistics of the sample X row vectors for any j follows as in the proof of Lemma 1. Under the condition A, and the condition given in (i), a simple extension of the proof in Lehmann (1983, p. 207-210), replacing scalar values $a_1, a_2, \ldots, a_n$ with vector values, demonstrates that these order statistics are complete. The results (i) and (ii) then follow from the proofs in 2.2(b) and 2.2(c), respectively.

An example suggested by a referee leads to a comparison of the estimators $\hat{\beta}$ and $\hat{\hat{\beta}}$, given $E(X'X)$ known, in a situation which is at the opposite extreme from that described in Section 2.2 (c), in that instead of a large number p of X variables, here we have a simple regression through the origin, so that $p = 1$. Consider the superpopulation model often used in survey sampling when x-values are positive and bounded away from zero:

$$Y_i = X_i \beta + \varepsilon_i, \quad i = 1, \ldots, N \tag{18}$$

where $E(\varepsilon_i) = 0$, $E(\varepsilon_i^2) = \sigma^2 X_i$, and $E(\varepsilon_i \varepsilon_j) = 0$ for $i \neq j$. Making the transformations $Y_i^* = Y_i / (X_i)^{1/2}$, $X_i^* = (X_i)^{1/2}$, and $\varepsilon_i^* = \varepsilon_i / (X_i)^{1/2}$ the model (18) substituting the starred quantities Y_i^* for Y_i, X_i^* for X_i, and ε_i^* for ε_i is in the form (1), leading to the model-unbiased estimators based on the population:

$$\hat{\beta}_N = \Sigma_{i=1}^{i=N} Y_i / \Sigma_{i=1}^{i=N} X_i \quad . \tag{19}$$

Letting y_i and x_i refer to the ith sampled values, we obtain the two sample estimators

$$\hat{\beta}_n = \Sigma_{i=1}^{i=n} y_i / \Sigma_{i=1}^{i=n} x_i \tag{20}$$

and

$$\hat{\hat{\beta}}_n = \Sigma_{i=1}^{i=n} y_i / [\,(n/N)\, \Sigma_{i=1}^{i=N} X_i \,. \tag{21}$$

Using (8) and (10), we get

$$\mathrm{Var}(\hat{\beta}_n) - \mathrm{Var}(\hat{\hat{\beta}}_n) = (\sigma^2/n)[E(1/\bar{x}) - 1/\bar{X}] - \beta^2[\mathrm{Var}(\bar{x})/\bar{X}^2]. \tag{22}$$

Using the Taylor expansion of $1/\bar{x}$ around $1/\bar{X}$ gives the approximation

$$E(1/\bar{x}) - 1/\bar{X} \cong (1/\bar{X}^3)\,\mathrm{Var}(\bar{x}) \tag{23}$$

so that approximately

(24) $$\mathrm{Var}(\hat{\beta}_n) \; > \; \mathrm{Var}(\hat{\hat{\beta}}_n) \iff \beta^2 < \sigma^2/(n\bar{X}) \; = \; (N/n)\,\mathrm{Var}(\hat{\beta}_N),$$

or, finally,

(25) $$\mathrm{Var}(\hat{\beta}_n) > \mathrm{Var}(\hat{\hat{\beta}}_n) \iff \mathrm{Rel\,Var}(\hat{\beta}_N) > n/N\,.$$

(Note that the estimator of the population total $Y = \Sigma_{i=1}^{i=N} Y_i$ utilizing $\hat{\beta}$ is the ratio estimator $X(\bar{y}/\bar{x})$, while the estimator utilizing $\hat{\hat{\beta}}$ is the expansion estimator $N\bar{y}$.)

III. CONCLUSION

It has been shown that in the linear regression model with random regressors, the Gauss-Markov theorem remains true when the distribution of (X, y) is multivariate normal with unknown parameters, and also when the distribution of the random vectors X is continuous and nondegenerate but otherwise completely unknown. The theorem is also true if the vector X is the realization of a simple random sample from a finite population of X-values, under mild conditions on the population. Under both the random regressor and finite population models, the theorem is not true when $E(X'X)$ is known, and furthermore a BLUE doesn't exist in that case.

REFERENCES

Breiman, L., and Spector, P. (1989), "Submodel selection and evaluation in regression-The X-random case," Technical Report No. 197, Department of Statistics, University of California, Berkeley, CA.

Chipman, J.S. (1964), "On least squares with insufficient observations," Journal of the American Statistical Association, 59, 1078-1111.

Duncan, D.B., and Horn, S. D. (1972), "Linear dynamic recursive estimation from the viewpoint of regression analysis," Journal of the American Statistical Association, 67, 815-822.

Freedman, D. A., Navidi, W., and Peters, S. C. (1987), "On the impact of variable selection in fitting regression equations," Technical Report No. 87, Department of Statistics, University of California, Berkeley, CA.

Goldman, A.J., and Zelen, M. (1964), "Weak generalized inverses and minimum variance linear unbiased estimation," J. Research National Bureau of Standards B, 68, 151-172.

Harville, D. (1976), "Extension of the Gauss Markov theorem to include the estimation of random effects," Annals of Statistics, 4, 384-396.

Johnson, Norman L., and Kotz, S. (1972), Distributions in Statistics: Continuous Multivariate Distributions, New York: John Wiley.

Kariya, Takeaki (1985), "A nonlinear version of the Gauss-Markov theorem", Journal of the American Statistical Association, 80, 476-477.

Kariya, Takeaki, and Toyooka, Yasuyuki (1982), "The lower bound for the covariance matrix of GLSE and its application to regression with serial correlation," Discussion Paper 65, Hitotsubashi University.

Lehmann, E. L. (1983), Theory of Point Estimation, New York: John Wiley.

Lehmann, E. L. (1986), Testing Statistical Hypotheses, 2nd Edition, New York: John Wiley.

Marshall, Albert W., and Olkin, Ingram (1979), Inequalities: Theory of Majorization and its Applications, New York: Academic Press.

Pfeffermann, D. (1984), "On extensions of the Gauss-Markov theorem to the case of stochastic regression coefficients," Journal of the Royal Statistical Society B, 46, 139-

148.

Rao, C.R. (1965), "The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves", Biometrika, 52, 447-458.

Rao, C.R. (1972), "Some recent results in linear estimation", Sankhya B, 34, 369-378.

Rao, C.R. (1973a), "Representations of best linear unbiased estimators in the Gauss-Markoff model with a singular dispersion matrix", Journal of Multivariate Analysis, 3, 276-292.

Rao, C.R. (1973b), Linear Statistical Inference and its Applications (2nd Ed.), New York: John Wiley.

Rao, C.R. (1979), "Estimation of parameters in the singular Gauss-Markoff model", Communications in Statistics-Theory and Methods, A8, 1353-1358.

Rosenberg, B. (1972), "The estimation of stationary stochastic regression parameters re-examined," Journal of the American Statistical Association, 67, 650-654.

Sarris, A.H. (1973), "Kalman filter models, A Bayesian approach to estimation of time-varying regression coefficients," Annals of Economic and Social Measurement, 2, 501-523.

Schonfeld, Peter, and Werner, Hans Joachim (1987), "A note on C.R. Rao's wider definition BLUE in the general Gauss-Markov model," Sankhya B, 49, 1-8.

Toyooka, Yasuyuki (1984), "An iterated version of the Gauss-Markov theorem in generalized least squares estimation," American Statistical Association, Proceedings of the Business and Economic Statistics Section, 190-195.

Toyooka, Yasuyuki (1987), "An iterated version of the Gauss-Markov theorem in generalized least squares estimation," Journal of the Japan Statistical Society, 17, 129-136.