# LINEAR LEAST SQUARES REGRESSION[1]

By Geoffrey S. Watson

*The Johns Hopkins University*

**0. Summary.** The paper gives a self-contained account of linear least squares regression when the errors have an arbitrary error covariance matrix. The finite sample size case is treated algebraically by methods which are entirely analogous to those used for the asymptotic study of the same problem by spectral analysis when the errors are generated by a covariance stationary process. The algebraic methods and results are of interest in themselves and may also be useful as an introduction to the difficult analysis involved in the asymptotic treatment.

**1. Introduction.** The paper gives a self-contained account of linear least squares when the errors have an arbitrary error covariance matrix, bringing together the distinct literatures of analysis of variance and time series analysis.

Section 2 sets up the problem and the notation. Section 3 is concerned with conditions for least squares estimates to be efficient and with a lower bound to their efficiency when they are not. The condition used here had its origin in time series work by T. W. Anderson (1948). The most general set up with a regression matrix and an error covariance matrix possibly not of full rank is considered, as is common in analysis of variance but not in time series analysis. While it is comforting to know that one's least squares estimates are, in some circumstances, best linear unbiased estimates it will, in general, be impossible to estimate their variance unless further assumptions are made. The problem of finding a good lower bound to the efficiency remains an open problem.

From Section 4 onwards the full rank model is used to study the efficiency of least squares estimates from the point of view of U. Grenander (1954) and U. Grenander and M. Rosenblatt (1957). They made an asymptotic study using the spectral analysis of the errors which were assumed to be from a covariance stationary process. The present finite sample size approach with its purely algebraic methods is of interest in itself and may also be useful as an introduction to the difficult analysis used in the asymptotic treatment. The correspondence of their results to the simpler ones in Section 3 is shown. The residuals are examined in Section 5 because only from their study can one learn empirically about the error model. In Section 6, a brief correspondence is made between the earlier sections and the literature and method of spectral analysis of regression problems.

**2. Linear least squares.** It will be supposed that the dependent variable $y$ is a linear function of $k$ independent (or regression or regressor) variables $x_1, \cdots, x_k$ plus a random disturbance $u$. If $n$ values of $y$ are given, this model may be written

Received 8 August 1966; revised 15 June 1967.

[1] This research was sponsored by the Office of Naval Research on Contract Nonr 4010(09) awarded to the Department of Statistics, The Johns Hopkins University.

in matrix notation as

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u} \tag{2.1}$$

where $\mathbf{y}$ is $n \times 1$, $\mathbf{X}$ is $n \times k$, $\beta$ is $k \times 1$ and $\mathbf{u}$ is $n \times 1$. The $\mathbf{X}$ matrix will be regarded as fixed; initially we will not assume that the rank of $\mathbf{X}$ is $k$. The error vector $\mathbf{u}$ has the properties

$$E(\mathbf{u}) = \mathbf{0}, \qquad E(\mathbf{u}\mathbf{u}') = \mathbf{\Gamma} \tag{2.2}$$

where the prime denotes matrix transposition.

A least squares estimator $\mathbf{b}$ of $\beta$ is a value that makes $(\mathbf{y} - \beta\mathbf{X})'(\mathbf{y} - \beta\mathbf{X})$ least. The geometrical solution of this problem is well known. If $\mathbf{X}$ is written as a row vector whose elements are its columns $\mathbf{x}_1, \cdots, \mathbf{x}_k$, called regression, or regressor, vectors, we have

$$\mathbf{X}\beta = \beta_1\mathbf{x}_1 + \cdots + \beta_k\mathbf{x}_k \tag{2.3}$$

as a point in the linear subspace $\mathfrak{M}(\mathbf{X})$ of Euclidean $n$-space spanned by $\mathbf{x}_1, \cdots, \mathbf{x}_k$ and often called the regression space. If a perpendicular is dropped from the point in $n$-space with position $\mathbf{y}$ onto the regression space, the foot of the perpendicular is $\mathbf{Xb}$, where $\mathbf{b}$ is a least squares estimate of $\beta$. Let the perpendicular be denoted by $\mathbf{x}$. Then $\mathbf{z}$ lies in $\mathfrak{M}(\mathbf{X})^{\perp}$, the orthogonal complement of $\mathfrak{M}(\mathbf{X})$ and

$$\mathbf{y} = \mathbf{Xb} + \mathbf{z}. \tag{2.4}$$

$\mathfrak{M}(\mathbf{X})^{\perp}$ is often called the error-space. The error $\mathbf{u}$ can be uniquely written

$$\mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2 \tag{2.5}$$

where $\mathbf{u}_1 \varepsilon \mathfrak{M}(\mathbf{X})$, $\mathbf{u}_2 \varepsilon \mathfrak{M}(\mathbf{X})^{\perp}$. It is clear that

$$\mathbf{X}\beta + \mathbf{u}_1 = \mathbf{Xb}, \qquad \mathbf{u}_2 = \mathbf{z}. \tag{2.6}$$

Thus in least squares $\mathbf{z}$ is used to estimate the errors in $\mathbf{b}$ which is, in fact, affected not by $\mathbf{u}_2$ but by $\mathbf{u}_1$.

When $X$ is not of rank $k$, it is necessary to introduce the notion of linear estimable functions. We now summarize and extend the results in e.g. Rao (Section 4a, 1965) for later uses. A linear combination $\mathbf{p}'\beta$ is estimable if, and only if, there exists an $n$ vector $\mathbf{L}$ such that $\mathbf{L}'\mathbf{X} = \mathbf{p}'$. Hence $\mathbf{p}$ must belong to $\mathfrak{M}(\mathbf{X}')$, the space spanned by the rows of $\mathbf{X}$. Further $\mathfrak{M}(\mathbf{X}') = \mathfrak{M}(\mathbf{X}'\mathbf{X})$, so that $\mathbf{p} = \mathbf{X}'\mathbf{L} = \mathbf{X}'\mathbf{X}\lambda$ for some $k$-vector $\lambda$. The least squares estimate (LSE) of $\mathbf{p}'\beta$ is $\mathbf{p}'\mathbf{b}$ where $\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\mathbf{b}$. When rank $(\mathbf{X})$ = rank$(\mathbf{X}'\mathbf{X}) = r < k$, only the part of $\mathbf{b} \varepsilon \mathfrak{M}(\mathbf{X}'\mathbf{X})$ is determined. Any component of $\mathbf{b}$ in $\mathfrak{M}(\mathbf{X}'\mathbf{X})^{\perp}$ is annihilated by $\mathbf{p} \varepsilon \mathfrak{M}(\mathbf{X}'\mathbf{X})$ so that $\mathbf{p}'\mathbf{b} = \lambda'\mathbf{X}'\mathbf{y}$ is the (determinate) LSE of $\mathbf{p}'\beta$. Any part of $\lambda$ orthogonal to $\mathfrak{M}(\mathbf{X}')$ does not contribute to $\mathbf{p}'\mathbf{b}$ so we may assume $\lambda \varepsilon \mathfrak{M}(\mathbf{X}')$. But $\mathfrak{M}(\mathbf{X}') = \mathfrak{M}(\mathbf{X}'\mathbf{X})$ is spanned by the eigenvectors of $\mathbf{X}'\mathbf{X}$ not associated with zero eigenvalues. So if there were two ways of writing the (unbiased) LSE, $\lambda'\mathbf{X}'\mathbf{y}$, $\mu'\mathbf{X}'\mathbf{y}$ with $\lambda$, $\mu \varepsilon \mathfrak{M}(\mathbf{X}')$, then it would follow that $\mathbf{X}'\mathbf{X}(\lambda - \mu) = 0$ which implies $\lambda = \mu$. Hence the LSE may be uniquely written as $\lambda'\mathbf{X}'\mathbf{y}$, $\lambda \varepsilon$

$\mathfrak{M}(X')$. Any linear unbiased estimator $\mathbf{L}'\mathbf{y}$ of $\mathbf{p}'\boldsymbol{\beta}$ may be written as $\mathbf{L}'\mathbf{y} = \mathbf{L}_1'\mathbf{y} + \mathbf{L}_2'\mathbf{y}$, $\mathbf{L}_1 \varepsilon \mathfrak{M}(\mathbf{X})$, $\mathbf{L}_2 \mathfrak{M}(\mathbf{X})^{\perp}$. The requirement $\mathbf{X}'\mathbf{L} = \mathbf{p}$ leaves $\mathbf{L}_2$ free but fixes $\mathbf{L}_1$. For if $\mathbf{L}_1$, $\mathbf{L}_1^* \varepsilon \mathfrak{M}(\mathbf{X})$ and $\mathbf{X}'\mathbf{L}_1 = \mathbf{X}'\mathbf{L}_1^* = \mathbf{p}$, then $X'(\mathbf{L}_1 - \mathbf{L}_1^*) = 0$ implies $\mathbf{L}_1 = \mathbf{L}_1^*$. Finally if we have a linear unbiased estimator $\mathbf{L}'\mathbf{y}$, with $\mathbf{L} \varepsilon \mathfrak{M}(\mathbf{X})$, for any estimable functions $\mathbf{p}'\boldsymbol{\beta}$, $\mathbf{L}'\mathbf{y}$ must be identical to the LSE, $\mathbf{p}'\mathbf{b} = \boldsymbol{\lambda}'\mathbf{X}'\mathbf{y}$ since $\mathbf{X}\boldsymbol{\lambda} \varepsilon \mathfrak{M}(\mathbf{X})$ i.e. $\mathbf{L} = \mathbf{X}\boldsymbol{\lambda}$. This last result will be referred to as Lemma 1 several times in proof of Theorem 1.

The background is completed by a spectral resolution of the error vector $\mathbf{u}$. If the eigenvectors and values of the $n \times n$ non-negative symmetric matrix $\boldsymbol{\Gamma}$ are denoted, respectively, by $\mathbf{g}_1, \cdots, \mathbf{g}_n$ and $\mathbf{f}_1, \cdots, \mathbf{f}_n$ then we may write

$$\boldsymbol{\Gamma} = \sum_1^n f_i \mathbf{g}_i \mathbf{g}_i'. \tag{2.7}$$

There is no loss of generality in supposing that $0 \leqq f_1 \leqq f_2 \leqq \cdots \leqq f_n$. When the $f_i$ are not all unequal, the $\mathbf{g}_i$ are not all uniquely defined. It will often be enough to suppose that some orthonormal choice has been made. In Section 4 it is essential to write $\boldsymbol{\Gamma}$ uniquely in terms of orthogonal projectors onto its spectral subspaces; the treatment then becomes less intuitive. The resolution of $\mathbf{u}$, corresponding to (2.7), is

$$\mathbf{u} = \eta_1 \mathbf{g}_1 + \eta_2 \mathbf{g}_2 + \cdots + \eta_n \mathbf{g}_n \tag{2.8}$$

where $\eta_1, \cdots, \eta_n$ are uncorrelated random variables with zero means and variances $f_1, \cdots, f_n$ respectively. For any vector $\mathbf{c}$ of unit length, var $(\mathbf{c}'\mathbf{u})$ could be called the variance (or power) of $\mathbf{u}$ in the direction $\mathbf{c}$. Then $f_i$ is the variance of $\mathbf{u}$ in the direction $\mathbf{g}_j$.

When $\mathbf{X}$ has full rank $k$, $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ so that we may write

$$\mathbf{b} - \boldsymbol{\beta} = \sum_1^n \eta_i (\mathbf{X}'\mathbf{X})^{-} X^{1\prime} g_i \tag{2.9}$$

and

$$\mathbf{z} = \sum_1^n \eta_i (I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') g_i . \tag{2.10}$$

**3. The efficiency of least squares estimates.** When rank $\mathbf{X} = k$ and rank $\boldsymbol{\Gamma} = n$, the problem may be stated very briefly. The best linear estimator $\hat{\boldsymbol{\beta}}$ in (2.1) is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Gamma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Gamma}^{-1}\mathbf{y}, \tag{3.1}$$

with

$$\text{var } (\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\boldsymbol{\Gamma}^{-1}\mathbf{X})^{-1}. \tag{3.2}$$

These must be compared with $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and

$$\text{var } (\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Gamma}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \tag{3.3}$$

Conditions for var $(\mathbf{b})$ = var $(\hat{\boldsymbol{\beta}})$ (or equivalently, $\mathbf{b} = \hat{\boldsymbol{\beta}}$ almost surely) are required on $\mathbf{X}$ given $\boldsymbol{\Gamma}$ or, on $\boldsymbol{\Gamma}$ given $\mathbf{X}$. When var $(\mathbf{b})$ > var $(\hat{\boldsymbol{\beta}})$, it is of interest to have an inequality on the ratio of generalized variances $|\text{var } (\hat{\boldsymbol{\beta}})|/|\text{var } (\mathbf{b})|$.

In the full rank case, sufficient conditions arose in a paper by T. W. Anderson (1948) of most powerful tests for serial correlation in the errors $\mathbf{u}$ of (2.1). Anderson was only able to find such tests when the $k$ regression vectors were eigenvectors of $\mathbf{\Gamma}$. The result was used in R. L. Anderson and T. W. Anderson (1950) to test for circular serial correlation when a Fourier Series is fitted. It was also the starting point for the papers of J. Durbin and G. S. Watson (1950), (1951) on testing for serial correlation in the general regression model. The papers by G. S. Watson (1952), (1955), G. S. Watson and E. J. Hannan (1956) discuss the behavior and efficiency of least squares; the inequality given in the first two papers shows the necessity of the condition for $k = 1$. Anderson's condition was rediscovered by T. A. Magness and J. B. McGuire (1962) who showed it to be sufficient in the full rank case with general $k$. As will be shown in Section 4, U. Grenander (1954) and U. Grenander and M. Rosenblatt (1957) had, in effect, proved earlier an equivalent result. All these papers are concerned with conditions on $X$, given $\mathbf{\Gamma}$. Rao (1965) found conditions on $\mathbf{\Gamma}$, given $\mathbf{X}$, by asking (see (3.2), (3.3)) for a solution of

$$(\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{X})^{-1} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Gamma}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

for $\mathbf{\Gamma}$. He found that one must have

$$\mathbf{\Gamma} = \mathbf{X}\mathbf{\Sigma}\mathbf{X}' + \mathbf{Z}\theta\mathbf{Z}' + \sigma^2\mathbf{I} \tag{3.4}$$

where $\mathbf{Z}$ is $h \times 1$ such that $\mathbf{Z}'\mathbf{X} = 0$, $\mathbf{\Sigma}$, $\theta$, $\sigma^2$ arbitrary. The sufficiency of (3.4) is implied, not proved.

When $\mathbf{X}$ is not of full rank, Muller and Watson (1959) showed the sufficiency and went on to consider the difficulties in error estimation. The application in this paper was to experimental design. In some 1962 correspondence with Dr. M. E. Muller and the author, Professor W. Kruskal indicated a coordinate-free proof of the necessity and sufficiency when $\mathbf{X}$, but not $\mathbf{\Gamma}$, is possibly not of full rank. This result is particularly simple to prove because, instead of working with $\hat{\boldsymbol{\beta}}$ and $\mathbf{b}$ he uses $\hat{\mathbf{u}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}} = \mathbf{X}\mathbf{b}$. He states that "$\hat{\mathbf{u}} = \tilde{\mathbf{u}}$ if and only if $\mathbf{\Gamma}\mathfrak{M}(\mathbf{X}) = \mathfrak{M}(\mathbf{X})$." The author hopes that Professor Kruskal's result will appear in the near future. Zyskind (1962) announced a theorem with $\mathbf{\Gamma}$ non-singular and Zyskind (1967) gives the first general theorem with $\mathbf{\Gamma}$ possibly singular. In our notation it reads: "A necessary and sufficient condition for all simple least squares linear estimators to be also best linear unbiased estimators of the corresponding estimable parametric functions $\mathbf{p}'\boldsymbol{\beta}$ in the model (2.1) is that there exist a subset of $r$ orthogonal eigenvectors of $\mathbf{\Gamma}$ which form a basis for $\mathfrak{M}(\mathbf{X})$." Rao (1965) remarks that his result is true for $\mathbf{\Gamma}$ singular and rank $X = r < k$. (Then rank $\mathbf{Z} = r = n-n$ is implied.) Some skill with generalized inverses might show the proof is still valid. Neither Zyskind or Rao go on to discuss error estimation or are fully aware of the times series literature on this problem. Zyskind goes onto the example of randomized blocks, given earlier by Muller and Watson (1959). The essence of the four theorems and their three corollaries given by Zyskind are contained in the remarks in Section 2 on linear

estimable functions and the following:

THEOREM 1. *Let* $\mathbf{p}'\beta$ *be an estimable function for the linear model* $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$, $E(\mathbf{u}) = 0$, $E(\mathbf{uu}') = \mathbf{\Gamma}$. *Then* $\mathbf{\Gamma}\mathfrak{M}(\mathbf{X}) = \mathfrak{M}(\mathbf{X})$ *implies that the* BLUE *and* LSE's *of* $\mathbf{p}'\beta$ *are identical. If the* LSE $\mathbf{p}'\mathbf{b} = \lambda'\mathbf{X}'\mathbf{y}$, *is a* BLUE *of* $\mathbf{p}'\beta$, *then* $\mathbf{\Gamma}\mathbf{X}\lambda \varepsilon \mathfrak{M}(\mathbf{X})$; *in particular if the* LSE *is a* BLUE *for all estimable functions,* $\mathbf{\Gamma}\mathfrak{M}(\mathbf{X}) = \mathfrak{M}(\mathbf{X})$.

PROOF. Given an estimable function $\mathbf{p}'\beta$, assume that $\mathfrak{M}(\mathbf{X})$ is spanned by $r$ eigenvectors of $\mathbf{\Gamma}$, where $r = \text{rank } (\mathbf{X})$ i.e. assume that $\mathbf{\Gamma}\mathfrak{M}(\mathbf{X}) = \mathfrak{M}(\mathbf{X})$. Write any unbiased linear estimator $\mathbf{L}'\mathbf{y}$ as $\mathbf{L}_1'\mathbf{y} + \mathbf{L}_2'\mathbf{y}$ where, by Lemma 1, $\mathbf{L}_1$ is the (unique) part of $\mathbf{L}$ in $\mathfrak{M}(\mathbf{X})$ and $\mathbf{L}_2$ is the part in $\mathfrak{M}(\mathbf{X})^{\perp}$. Then

$$\text{var } (\mathbf{L}'\mathbf{y}) = \text{var } (\mathbf{L}_1'\mathbf{u}) + 2 \text{ cov } (\mathbf{L}_1'\mathbf{u}, \mathbf{L}_2'\mathbf{u}) + \text{var } (\mathbf{L}_2'\mathbf{u}).$$

But cov $(\mathbf{L}_1'\mathbf{u}, \mathbf{L}_2\mathbf{u}) = \mathbf{L}_2'\mathbf{\Gamma}\mathbf{L}_1 = 0$ since $\mathbf{\Gamma}\mathbf{L}_1 \varepsilon \mathfrak{M}(\mathbf{X})$ and $\mathbf{L}_2 \varepsilon \mathfrak{M}(\mathbf{X})^{\perp}$. Hence var $(\mathbf{L}'\mathbf{y}) \geqq$ var $(\mathbf{L}_1'\mathbf{y})$. Hence $\mathbf{L}_1'\mathbf{y}$ is the BLUE of $\mathbf{p}'\beta$. But $\mathbf{L}_1$ is unique in $\mathfrak{M}(\mathbf{X})$ so that $\mathbf{L}_1'\mathbf{y}$ must be identical with the LSE estimator $\mathbf{p}'\mathbf{b} = \lambda'\mathbf{X}'\mathbf{y}$ since $\mathbf{X}\lambda \varepsilon \mathfrak{M}(\mathbf{X})$. This proves the first part of the theorem.

To prove the converse we consider first a *given* estimable function $\mathbf{p}'\beta$ and assume that the LSE, $\lambda'\mathbf{X}'\mathbf{y}$, has the same variance as the BLUE. Then for any unbiased estimator $\mathbf{L}'\mathbf{y}$, with $\mathbf{L}'\mathbf{X} = \mathbf{p}'$, we may assume that

$$\lambda'\mathbf{X}'\mathbf{\Gamma}\mathbf{X}\lambda \leqq \mathbf{L}'\mathbf{\Gamma}\mathbf{L}. \tag{3.5}$$

Writing as before $\mathbf{L} = \mathbf{L}_1 + \mathbf{L}_2$, we know further by Lemma 1 that $\mathbf{L}_1 = \mathbf{X}\lambda$ so that $\mathbf{L} = \mathbf{X}\lambda + \mathbf{L}_2$. Thus for all $\mathbf{L}_2 \varepsilon \mathfrak{M}(\mathbf{X})^{\perp}$,

$$\mathbf{L}'\mathbf{\Gamma}\mathbf{L} = \lambda'\mathbf{X}'\mathbf{\Gamma}\mathbf{X}\lambda + 2\lambda'\mathbf{X}'\mathbf{\Gamma}\mathbf{L}_2 + \mathbf{L}_2\mathbf{\Gamma}\mathbf{L}_2$$

which, in conjunction with (3.5) implies that $\lambda'\mathbf{X}'\mathbf{\Gamma}\mathbf{L}_2 = 0$. For suppose e.g. $\lambda'\mathbf{X}'\mathbf{\Gamma}\mathbf{L}_2 = a > 0$. Then using $-\epsilon\mathbf{L}_2$ instead of $\mathbf{L}_2$, we could make $\mathbf{L}'\mathbf{\Gamma}\mathbf{L} < \lambda'\mathbf{X}'\mathbf{\Gamma}\mathbf{X}\lambda$ for sufficiently small $\epsilon$. But $\mathbf{L}_2'\mathbf{\Gamma}\mathbf{X}\lambda = 0$ for all $\mathbf{L}_2 \varepsilon \mathfrak{M}(\mathbf{X})^{\perp}$ implies that $\mathbf{\Gamma}\mathbf{X}\lambda = 0$ or $\mathbf{\Gamma}\mathbf{X}\lambda \varepsilon \mathfrak{M}(\mathbf{X})$. This proves the second part of the theorem. The third part is immediate since the last condition must hold for all $\lambda$.

We turn now to a brief study of what may be learned from the least squares residuals $\mathbf{z}$ about the variability of least squares estimates—the subject will be taken up in a different way in Section 5. It was clear in (2.9), that even if these estimates were optimal, their variance might not be estimable from $\mathbf{z}$. Of course, when $\mathbf{\Gamma}$ is singular, some optimal estimates have zero variance and the problem does not arise with them! To show the problem it is best to take the simple case of rank $\mathbf{X} = k$. Then for all $\mathbf{p}$, $\mathbf{p}'\mathbf{b}$ is optimal if and only if the columns of $\mathbf{X}$ are linear combinations of $k$ eigenvectors of $\mathbf{\Gamma}$ say $\mathbf{g}_1, \cdots, \mathbf{g}_k$.

COROLLARY 1. *If* $\mathbf{X} = [\mathbf{g}_1, \cdots, \mathbf{g}_k]\mathbf{P}$, $\mathbf{P}$ $k \times k$ *non-singular,* $\mathbf{b} - \beta$ *depends on* $\eta_1, \cdots, \eta_k$ *and* $\mathbf{z}$ *depends on* $\eta_{k+1}, \cdots, \eta_n$.

PROOF. From (2.6)

$$\mathbf{b} - \beta = \textstyle\sum_1^n \eta_i(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}\begin{bmatrix} \mathbf{g}_1' \mathbf{g}_i \\ \vdots \\ \mathbf{g}_k' \mathbf{g}_i \end{bmatrix},$$

$$= \sum_1^n \eta_i (\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}\begin{bmatrix}0\\ \vdots\\ 1\\ \vdots\\ 0\end{bmatrix},$$

the first assertion. In (2.10), $\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is well-known to be the orthogonal projector onto the regression space which is here spanned by $\mathbf{g}_1, \cdots, \mathbf{g}_k$. Hence $\mathbf{z} = \sum_{k+1}^n \eta_i \mathbf{g}_i$.

It follows that $\mathbf{z}$ can tell us nothing about the random variables in $\mathbf{b}$, and hence its variance. If some further assumptions are made relating $f_1, \cdots, f_k$ to $f_{k+1}, \cdots, f_n$, it may be possible. We could say in this case that

(3.6) var ($\mathbf{z}$) in directions $\mathbf{g}_1, \cdots, \mathbf{g}_k$ is zero;

var ($\mathbf{z}$) in directions $\mathbf{g}_{k+1}, \cdots, \mathbf{g}_n$ is $f_{k+1}, \cdots, f_n$.

In general this will not be so and $b \neq \hat{\beta}$ since each $\mathbf{x}_i$ will have a component on every eigenvector. We return to this point in Section 5.

For the rest of the paper we will assume that rank ($\mathbf{X}$) = $k$, rank ($\mathbf{\Gamma}$) = $n$.

In Watson (1952), (1955), the argument was continued by finding the bias in the usual estimate of the error variance, the disturbances in the usual tests of linear hypotheses about the regression coefficients, and a study of how inefficient $\mathbf{b}$ could be by finding a lower bound to the efficiency of $\mathbf{b}$. We take up again here only this latter point.

The natural measure of the efficiency of $\mathbf{b}$ is

$$\text{Eff } (\mathbf{b}) = |\text{var } (\hat{\beta})|/|\text{var } (\mathbf{b})|,$$

which from (3.2) and (3.3) is

$$\text{Eff } (\mathbf{b}) = |\mathbf{X}'\mathbf{X}|^2/|\mathbf{X}'\mathbf{\Gamma}\mathbf{X}|\,|\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{X}|. \tag{3.7}$$

Clearly $0 \leqq \text{Eff } (\mathbf{b}) \leqq 1$. For Eff ($\mathbf{b}$) $\geqq 0$ by definition and Eff ($\mathbf{b}$) $\leqq 1$ by a Cauchy inequality for determinants. To prove the latter consider $n \times k$ matrices $\mathbf{A}$, $\mathbf{B}$ where $\mathbf{B}'\mathbf{B}$ is non-singular. Then $\mathbf{A}'(\mathbf{I}_n - \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}')\mathbf{A}$ is non-negative or, in a familiar notation, $\mathbf{A}'\mathbf{A} \geqq \mathbf{A}'\mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{A}$. It follows that $|\mathbf{A}'\mathbf{A}| \geqq |\mathbf{A}'\mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{A}|$ i.e. that

$$|\mathbf{A}'\mathbf{B}|^2 \leqq |\mathbf{A}'\mathbf{A}|\,|\mathbf{B}'\mathbf{B}|.$$

Setting $\mathbf{A} = \mathbf{\Gamma}^{\frac{1}{2}}\mathbf{X}$, $\mathbf{B} = \mathbf{\Gamma}^{-\frac{1}{2}}\mathbf{X}$ shows that Eff ($\mathbf{b}$) in (3.7) does not exceed unity, as claimed. Theorem 1 shows when the upper bound may be attained. To find the lower bound, use was made of a converse of Cauchy's inequality to be found in Theorem 71 of Hardy, Littlewood and Polya (1934), that is now often attributed to a later writer, Kantorovich (1948). If $0 < f_1 \leqq \cdots \leqq f_n$ and $w_1, \cdots, w_n \geqq 0$, $\sum_1^n w_i = 1$, the inequality in question is

$$1 \leqq (\sum f_i w_i)(\sum f_i^{-1} w_i) \leqq \tfrac{1}{4}[(f_1/f_n)^{\frac{1}{2}} + (f_n/f_1)^{\frac{1}{2}}]^2 = (f_1 + f_n)^2/4f_1 f_n. \tag{3.8}$$

The simplest proofs of (3.8) draw on convexity results. The upper bound is only

attained when $w_1 = w_n = \frac{1}{2}$, $w_2 = \cdots = w_{n-1} = 0$; the lower bound is attained only when $f_i$ constant for non-zero $w_i$. Its use for (3.7) with $k = 1$ is immediate, when it is realized that there is no loss of generality in assuming $\mathbf{X}'\mathbf{X} = \mathbf{I}_k$. For, if $\mathbf{H}$ is an $n \times n$ non-singular matrix,

$$\text{Eff } (\mathbf{b}) = |\mathbf{H}'\mathbf{X}'\mathbf{X}\mathbf{H}|^2/|\mathbf{H}'\mathbf{X}'\boldsymbol{\Gamma}\mathbf{X}\mathbf{H}|\ |\mathbf{H}'\mathbf{X}'\boldsymbol{\Gamma}^{-1}\mathbf{X}\mathbf{H}|$$

because $|\mathbf{H}|^4$ may be cancelled. Since $\mathbf{X}$ has rank $k$, $\mathbf{H}$ may be chosen so that $\boldsymbol{\Xi} = \mathbf{X}\mathbf{H}$ is composed of $k$ orthonormal columns i.e. $\boldsymbol{\Xi}'\boldsymbol{\Xi} = \mathbf{I}_k$. When $k = 1$, $\boldsymbol{\Xi} = \sum_1^n \xi_i \mathbf{g}_i$ where $\sum_1^n \xi_i^2 = 1$. Thus

$$\text{Eff } (\mathbf{b}) = 1/(\sum f_i \xi_i^2)(\sum f_i^{-1} \xi_i^2). \tag{3.9}$$

By identifying $\xi_i^2$ with $w_i$, (3.8) immediately gives

$$4/[(f_1/f_n)^{\frac{1}{2}} + (f_n/f_1)^{\frac{1}{2}}]^2 \leqq \text{Eff } (\mathbf{b}) \leqq 1, \tag{3.10}$$

the lower bound being attained when $x = 2^{-\frac{1}{2}}(\mathbf{g}_1 + \mathbf{g}_n)$. The upper bound is attained only when all $f_i$ associated with non-zero $\xi_i$ are equal. If Eff $(\mathbf{b})$ is to be unity for all $\boldsymbol{\Gamma}$, we have the case where the eigenvalues $f_1, \cdots, f_n$ may not be equal and the only way to guarantee Eff $(\mathbf{b}) = 1$ is to have only one $\xi_i^2 \neq 0$ i.e. for $\mathbf{x}$ to be eigenvector of $\boldsymbol{\Gamma}$. Hence we have

THEOREM 2. *A necessary and sufficient condition when* $k = 1$ *that* Eff $(\mathbf{b}) = 1$, *for all regression vectors* $\mathbf{x}$, *is that all eigenvalues of* $\boldsymbol{\Gamma}$ *be equal* i.e. *that* $\boldsymbol{\Gamma} = \sigma^2\mathbf{I}$.

A necessary and sufficient condition that Eff $(\mathbf{b}) = 1$ for a given regression vector $\mathbf{x}$ and for all $\boldsymbol{\Gamma}$ with eigenvectors $\mathbf{g}_1, \cdots, \mathbf{g}_n$ is that the regression vector be an eigenvector of $\boldsymbol{\Gamma}$.

The lefthand side of (3.10) is the minimum, for all $\mathbf{x}$, of Eff $(\mathbf{b})$ when the least and greatest eigenvalues of $\boldsymbol{\Gamma}$ are $f_1$ and $f_n$. It is clear from the derivation that if $\mathbf{x}$ is restricted to a subspace spanned $\mathbf{g}_{i_1}, \cdots, \mathbf{g}_{i_l}$, $i_1 < i_2, \cdots < i_l$, then the minimum of Eff $(\mathbf{b})$ will increase to

$$4/[(f/F)^{\frac{1}{2}} + (F/f)^{\frac{1}{2}}]^2$$

where $f = \min(f_{i_1}, \cdots, f_{i_l}) = f_{i_1}$ $F = \max(f_i, \cdots, f_{i_l}) = f_{i_l}$. It will be convenient to refer to the (invariant) subspace spanned by $\mathbf{g}_{i_1}, \cdots, \mathbf{g}_{i_l}$ as the eigen subspace $s(i_1, \cdots, i_l)$. Thus we have shown

$$\min_{\mathbf{x} \varepsilon s(i_1, \cdots, i_l)} \text{Eff } (\mathbf{b}) = 4/[(f/F)^{\frac{1}{2}} + (F/f)^{\frac{1}{2}}]^2. \tag{3.11}$$

The eigen values $f$ and $F$ are the least and greatest variances of $\mathbf{u}$ for directions $\mathbf{c}$ in the eigen subspace $s(i_1, \cdots, i_l)$. For $\mathbf{c} = c_1\mathbf{g}_{i_2}^a + \cdots + c_l\mathbf{g}_{i_l}$ and $\mathbf{c}'\mathbf{c} = 1$ implies $\sum_1^l c_a^2 f_{i_a}$, we have, as asserted, that $f \leqq \text{var } (\mathbf{c}'\mathbf{u}) \leqq F$.

TABLE I

| $f/F$ | .1 | .2 | .3 | .4 | .5 | .6 | .7 | .8 | .9 |
|---|---|---|---|---|---|---|---|---|---|
| min Eff(b) | .364 | .556 | .710 | .816 | .889 | .938 | .969 | .989 | .997 |

For a single regressor restricted to an eigen subspace, min Eff (b) is given for various values of the ratio of the least to greatest error variances in this subspace.

COROLLARY 2. *A necessary and sufficient condition that, for* $k = 1$, Eff $(\mathbf{b}) = 1$ *for all* $\mathbf{x}$ *in the eigen subspace* $s(i_1, \cdots, i_l)$ *is that the variance of* $\mathbf{u}$ *for all directions in this subspace be constant.*

Theorem 1 for $k = 1$ is an immediate consequence of both Theorem 2 and Corollary 2 since $\mathbf{x}$ must, in this situation, be an eigen vector of $\mathbf{\Gamma}$. Some numerical values of the righthand side of (3.11) are given in Table 1. Thus we see for example that even with a two to one ratio of $f$ to $F$ the efficiency of $\mathbf{b}$ cannot drop by more than 10%.

Equation (3.11) is thus useful for an evaluation of the robustness of the least squares estimate.

For $k > 1$ and $n > 2k - 1$, Watson (1952), (1955) gave the inequality, for all $\mathbf{X}$ of rank $k$,

$$\text{(3.12)}\quad \text{Eff } (\mathbf{b}) \geqq [4f_1f_n/(f_1 + f_n)^2][4f_2f_{n-1}/(f_2 + f_{n-1})] \cdots [4f_nf_{n-k+1}/(f_k + f_{n-k+1})^2].$$

The proof however contains a flaw and the writer has been unable to establish or disprove (3.12). True weaker inequalities will be given below. Eff $(\mathbf{b})$ is certainly equal to the righthand side of (3.12) when

$$\text{(3.13)}\quad \mathbf{x}_1 = 2^{-\frac{1}{2}}(\mathbf{g}_1 + \mathbf{g}_n),\ \mathbf{x}_2 = 2^{-\frac{1}{2}}(\mathbf{g}_2 + \mathbf{g}_{n-1}),\ \cdots,\ \mathbf{x}_k = 2^{-\frac{1}{2}}(\mathbf{g}_k + \mathbf{g}_{n-k+1}).$$

While searching for proofs of (3.12), two Monte Carlo experiments were run with $k = 2, n = 10$ and $f_1 = 1, f_2 = 2, \cdots, f_{10} = 10$. The elements of $x$ in this first experiment were independent standard normal variables and in the second were chosen independently as $+1$ and $-1$ with equal probability. 5000 repetitions were made in each case. In no case did the efficiency attain or fall below the bound given by (3.12). In attempting to find a fresh approach, the following different interpretation of Eff $(\mathbf{b})$ was noticed—it is probably "well known."

The canonical correlations $\rho_1, \cdots, \rho_k$ of the two vectors $\mathbf{b}$ and $\hat{\boldsymbol{\beta}}$ are easily defined since we know that var $(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Gamma}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$, cov $(\mathbf{b}, \hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{X})^{-1}$, var $(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{X})^{-1}$. The determinantal equation with roots $\rho_1^2, \cdots, \rho_k^2$ immediately leads to

$$\text{(3.14)}\quad \text{Eff } (\mathbf{b}) = \prod_1^k \rho_i^2,$$

where the righthand side of (3.14) is a product of $k$ factors like (3.12). When (3.13) is true, the factors in (3.14) are, in fact, the values of $\rho_1^2, \cdots, \rho_k^2$.

G. H. Golub (1963) rediscovered (3.10) and went on to the general case to obtain (3.18), an alternative proof of which (due to I. Olkin) will now be given. If $\mathbf{A}^{(k)}$ stands for the $k$th compound of a matrix $\mathbf{A}$ (its elements are the $k \times k$ minors of $\mathbf{A}$ arranged in lexicographical order: See MacDuffee (1946) p. 86 where $\mathbf{A}^{(k)}$ is called the $k$th adjugate of $\mathbf{A}$ and its main properties are stated), then it may be shown that

$$\text{(3.15)}\quad (\mathbf{X}'\mathbf{\Gamma}\mathbf{X})^{(k)} = \mathbf{X}^{(k)'}\mathbf{\Gamma}^{(k)}\mathbf{X}^{(k)}$$

where $\mathbf{X}^{(k)'}$ is a row vector of $\binom{n}{k}$ elements and, since $\mathbf{X}'\mathbf{\Gamma}\mathbf{X}$ is a $k \times k$ matrix

$$(\mathbf{X}'\mathbf{\Gamma}\mathbf{X})^{(k)} = |\mathbf{X}'\mathbf{\Gamma}\mathbf{X}|. \tag{3.16}$$

Thus (3.7) may be written

$$\text{Eff } (\mathbf{b}) = (\mathbf{X}^{(k)'}\mathbf{X}^{(k)})^2/\mathbf{X}^{(k)'}\mathbf{\Gamma}^{(k)}\mathbf{X}^{(k)'}\mathbf{\Gamma}^{(k)^{-1}}\mathbf{X}^{(k)}. \tag{3.17}$$

The eigen values of $\mathbf{\Gamma}^{(k)}$ are the $\binom{n}{k}$ products of the roots of $\mathbf{\Gamma}$, $k$ at a time. Applying (3.8) to (3.17) completes the proof that

$$\text{Eff } (\mathbf{b}) \geqq 4f_1f_2 \cdots f_k \cdot f_{n-k+1} \cdots f_n/(f_1f_2 \cdots f_k + f_{n-k+1} \cdots f_n)^2, \tag{3.18}$$

since it is assumed that $f_1 \leqq f_2 \leqq \cdots \leqq f_n$.

Inequalities (3.12) and (3.18) are the same for $k = 1$ but for $k \geqq 2$, the bound in (3.18) is smaller than the bound in (3.12) by elementary inequalities.

If the bound in (3.18) were attainable, it would represent the solution to our problem. To show that it is not we first derive another bound. Applying Hadamard's inequality to (3.7), in which no generality is lost by assuming $\mathbf{X}'\mathbf{X} = \mathbf{I}_k$, we see that

$$\text{Eff } (\mathbf{b}) \geqq 1/(\mathbf{x}_1'\mathbf{\Gamma}\mathbf{x}_1\mathbf{x}_1'\mathbf{\Gamma}^{-1}\mathbf{x}_1) \cdots (\mathbf{x}_k'\mathbf{\Gamma}\mathbf{x}_k)(\mathbf{x}_k'\mathbf{\Gamma}^{-1}\mathbf{x}_k).$$

Putting every factor in the denominator at its absolute maximum, we have

$$\text{Eff } (\mathbf{b}) \geqq (4f_1f_n/(f_1 + f_n)^2)^k. \tag{3.19}$$

A numerical example with $k = 2$, $n = 10$, $f_1 = 1$, $f_2 = 2$, $\cdots$, $f_{10} = 10$, gives the righthand side of (3.18) equal to 0.0157 which is less than the righthand side of (3.19), 0.1093. (3.12) is 0.1967. Thus (3.18) cannot be attained in this case. In other cases the relationship of the bounds (3.18) and (3.19) is reversed, e.g. for $n = 10$, $k = 2$, $f_1 = \cdots = f_9 = 1$, $f_{10} = 10$. Of course (3.19) is always smaller than (3.12). To see the reason why (3.18) is not in general attainable, consider the case $k = 2$ and $n = 4$ and write $\mathbf{X} = [\mathbf{X}_1', \mathbf{X}_2', \mathbf{X}_3', \mathbf{X}_4']$. Then $\mathbf{X}^{(k)}$ is a row of $2 \times 2$ determinants,

$$\mathbf{X}^{(k)'} = \left[\begin{vmatrix}\mathbf{X}_1'\\ \mathbf{X}_2'\end{vmatrix}, \begin{vmatrix}\mathbf{X}_1'\\ \mathbf{X}_3'\end{vmatrix}, \begin{vmatrix}\mathbf{X}_1'\\ \mathbf{X}_4'\end{vmatrix}, \begin{vmatrix}\mathbf{X}_2'\\ \mathbf{X}_3'\end{vmatrix}, \begin{vmatrix}\mathbf{X}_2'\\ \mathbf{X}_4'\end{vmatrix}, \begin{vmatrix}\mathbf{X}_3'\\ \mathbf{X}_4'\end{vmatrix}\right].$$

The condition for attaining the bound in (3.8) is that the first and last elements of this vector all equal and non-zero and that the rest are zero. The last condition requires all the rows of $\mathbf{X}$ to be proportional and so the first and last elements cannot be non-zero. This argument holds for $n \geqq 4$; for $n = 3$ the bound is attainable. Similar difficulties presumably occur for $k > 2$.

Thus the problem of finding an attainable lower bound for Eff $(\mathbf{b})$ when $k > 1$ remains open. We conclude this discussion with the remark that both (3.18) and (3.19) may be improved to match (3.11) if the columns of $\mathbf{X}$ are restricted to lie in $s(i_1, \cdots, i_l)$.

To conclude this section we note an interesting sidelight of this study. Suppose that we were able to predetermine the values of $\mathbf{x}_1 \cdots \mathbf{x}_n$ before observing $\mathbf{y}$. How then should $\mathbf{X}$ be chosen to obtain the most precise least squares estimate of $\boldsymbol{\beta}$, assuming $\mathbf{\Gamma}$ is known?

The obvious strategy is to make the choices so that the

$$\mathbf{x}_i = (\mathbf{g}_j \text{ corresponding to the least var } (\eta_j) = f_j).$$

Since $f_1 \leqq f_2 \leqq \cdots \leqq f_n$, this means $\mathbf{X} = [\mathbf{g}_1, \cdots, \mathbf{g}_k]$. Then

$$\mathbf{b} - \beta = \sum_1^k \eta_i \mathbf{g}_i, \qquad \mathbf{z} = \sum_{k+1}^n \eta_i \mathbf{g}_i,$$

from (2.9), (2.10) on noting that $\mathbf{X}'\mathbf{X} = \mathbf{I}_k$ and $\mathbf{X}'\mathbf{g}_j = 0$ if $j > k$. In communication theory language, we must put in the "signal" where there is least "noise," as is intuitively evident. However, as we observed in (3.6) the residuals are not, without further assumption, useful for constructing an estimate of var $(\mathbf{b})$.

**4. The regression spectrum.** In Section 3, the argument frequently called for the expression of the regressor vectors as linear combinations of the eigen vectors of $\mathbf{\Gamma}$. In this section, this technique will be formalized and be the center of attention. This approach was initiated by Grenander (1954) for the asymptotic study of regression analysis when the error $\mathbf{u}$ is generated by a stationary process. A finite dimensional version, for an arbitrary error covariance matrix $\mathbf{\Gamma}$, of the results in that paper and in Chapter 7 of the book by Grenander and Rosenblatt (1957) will now be developed. This may serve the secondary purpose of making their difficult asymptotic results more accessible. In this version, their main result is an immediate consequence of Theorem 1. The treatment would be much simpler if we used before the resolution $\mathbf{\Gamma} = \sum_1^n f_i \mathbf{g}_i \mathbf{g}_i'$, as we did above. It does however raise uniqueness problems so that we will work with the invariant subspaces of $\mathbf{\Gamma}$.

Suppose that the spectral resolution of $\mathbf{\Gamma}$ is given by

$$\mathbf{\Gamma} = \sum_{i=1}^m f_{E_i} \mathbf{E}_i \tag{4.1}$$

where

$$\mathbf{E}_i \mathbf{E}_j = \delta_{ij} \mathbf{E}_i, \qquad i, j = 1, \cdots, m, \tag{4.2}$$

$$\sum_1^m \mathbf{E}_i = \mathbf{I}_n.$$

Denote by $\mathbf{T}_i$ the spectral subspace of dimension $t_i$ onto which $\mathbf{E}_i$ projects, $i = 1, \cdots, m$, and $\sum_1^m t_i = n$. Then the $n \times k$ regression matrix $\mathbf{X}$ of rank $k$ may be written uniquely

$$\mathbf{X} = \sum_{i=1}^m \mathbf{E}_i \mathbf{X} \tag{4.3}$$

so that $\mathbf{E}_i\mathbf{X}$ is the part of the columns of $\mathbf{X}$ lying in $\mathbf{T}_i$. Then, if $(\mathbf{X}'\mathbf{X})^{-\frac{1}{2}}$ denotes the unique positive definite square root of $(\mathbf{X}'\mathbf{X})^{-1}$,

$$\mathbf{X}'\mathbf{X} = \sum_1^m \mathbf{X}'\mathbf{E}_i^2\mathbf{X},$$

$$\mathbf{I}_k = \sum_1^m \{(\mathbf{X}'\mathbf{X})^{-\frac{1}{2}}\mathbf{X}'\mathbf{E}_i'\}\{(\mathbf{X}'\mathbf{X})^{-\frac{1}{2}}\mathbf{X}'\mathbf{E}_i\}',$$

so that, defining the $k \times n$ matrix $\mathbf{W}_i$ by

$$\mathbf{W}_i = (\mathbf{X}'\mathbf{X})^{-\frac{1}{2}}\mathbf{X}'\mathbf{E}_i, \qquad i = 1, \cdots, m, \tag{4.4}$$

we have

$$\mathbf{W}_i\mathbf{W}_j' = \delta_{ij}\mathbf{W}_i\mathbf{W}_i', \tag{4.5}$$

$$\mathbf{I}_k = \sum_1^m \mathbf{W}_i\mathbf{W}_i'.$$

Then the least squares estimator of $\beta$ is given by

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sum_1^m \mathbf{E}_i)\mathbf{y}, \quad \text{i.e.} \tag{4.6}$$

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-\frac{1}{2}} \sum_1^m \mathbf{W}_i\mathbf{y},$$

where we note that $(\mathbf{X}'\mathbf{X})^{-\frac{1}{2}} \sum_1^m \mathbf{W}_i\mathbf{X} = \mathbf{I}_k$ since $E(\mathbf{b}) = \beta$. Also

$$\text{var } (\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-\frac{1}{2}} \sum_1^m \mathbf{W}_i\mathbf{\Gamma}\mathbf{W}_i'(\mathbf{X}'\mathbf{X})^{-\frac{1}{2}}, \tag{4.7}$$

$$\text{var } (\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-\frac{1}{2}} \sum_1^m f_{E_i}\mathbf{W}_i\mathbf{W}_i'(\mathbf{X}'\mathbf{X})^{-\frac{1}{2}},$$

by (4.1), (4.2). The least square residual vector $\mathbf{z}$ is given by

$$\mathbf{z} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{u},$$

or

$$\mathbf{z} = (\mathbf{I} - (\sum_1^m \mathbf{W}_i')(\sum_1^m \mathbf{W}_i))\mathbf{u}. \tag{4.8}$$

Finally the best linear unbiased estimator is given by

$$\hat{\beta} = (\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Gamma}^{-1}, \quad \text{i.e.}$$

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-\frac{1}{2}}(\sum_1^m f_{E_i}^{-1}\mathbf{W}_i\mathbf{W}_i')^{-1} \sum_1^m f_{E_i}^{-1}\mathbf{W}_i, \tag{4.9}$$

and

$$\text{Cov } (\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-\frac{1}{2}}(\sum_1^m f_{E_i}\mathbf{W}_i\mathbf{W}_i')^{-1}(\mathbf{X}'\mathbf{X})^{-\frac{1}{2}}. \tag{4.10}$$

The treatment which follows of necessary and sufficient conditions for $\mathbf{b} = \hat{\beta}$ is directly analogous to that of Grenander and Rosenblatt (1957).

DEFINITION. For the regression matrix $\mathbf{X} = \sum_1^m \mathbf{W}_i'(\mathbf{X}'\mathbf{X})^{\frac{1}{2}}$, the spectral distribution of $\mathbf{X}$ is $\{\mathbf{W}_1\mathbf{W}_1', \cdots, \mathbf{W}_m\mathbf{W}_m'\}$. The spectrum $S$ of $X$ is the set of subscripts for which $\mathbf{W}_i$ is non-null.

Since

$$\mathbf{E}_i\mathbf{X} = \mathbf{W}_i'(\mathbf{X}'\mathbf{X})^{\frac{1}{2}}, \tag{4.11}$$

the spectrum of $\mathbf{X}$ may be identified with spectral subspaces of $\mathbf{\Gamma}$ in which the columns of $\mathbf{X}$ have components. Off the spectrum, the $\mathbf{W}_i$ are null matrices. Hence the summations in (4.6) (4.7) (4.8) (4.9) (4.10) need only be over $S$; we will make this change now and all $\mathbf{W}_i$ appearing subsequently will be non-null.

Now since we wish to arrange the $k \times n$ matrices $\mathbf{W}_i$ into $l$ subsets $S_1, \cdots, S_l$, with $l$ as large as possible, so that if $\mathbf{W}_i$ comes from one subset and $\mathbf{W}_j$ from another, $\mathbf{W}_i'$, $\mathbf{W}_j$ is a null-matrix, i.e. the subsets are to be orthogonal, or to put it another way, all the columns of all $\mathbf{W}_i$'s in one subset are to be orthogonal to all the columns of all the $\mathbf{W}_i$'s in other subsets. There may only be one subset

(i.e. all the $\mathbf{W}_i$) or there may be more but certainly $l \leqq k$ because the columns are vectors in $k$-space. This *maximal* subdivision of $S$ into subsets of $S_1, \cdots, S_l$ of subscripts of $\mathbf{W}_i$'s belonging to the sets is unique. For suppose there is a distinct subdivision $S_1^*, \cdots, S_l^*$. Then distinct means that there exists a set $S_j$ and two sets $S_p^*$, $S_q^*$ $(p \neq q)$ such that some of the members of $S_j$ belong to $S_p^*$, and some to $S_p^*$. By the orthogonal construction of these subsets, the remaining elements of $S_j$ are neither in $S_p^*$ or $S_q^*$ and the members of $S_j$ fall into their orthogonal classes and are orthogonal to the elements of all other sets $S_i$ $(i \neq j)$. Hence $S_j$ may be subdivided into 3 sets so $l$ is not maximal—this contradiction proves the uniqueness.

This subdivision of $S$ corresponds to a subdivision of $k$ space into $l$ orthogonal subspaces $S_i(k)$ where $S_i(k)$ is the space spanned by the columns of the $\mathbf{W}_i$ belonging to $S_i$, $i = 1, \cdots, l$. The union of these subspaces is all of $k$-space—this follows from (4.5) and rank $(\mathbf{X}) = k$. The $k \times k$ orthogonal projectors onto these subspaces are defined by

$$\mathbf{N}_j = \sum_{i \varepsilon S_j} \mathbf{W}_i \mathbf{W}_i' . \tag{4.12}$$

From the construction of the subdivision $\mathbf{N}_i\mathbf{N}_j =$ null matrix, $i \neq j$, and from (4.5)

$$\sum_1^m \mathbf{W}_i \mathbf{W}_i' = \mathbf{I}_k = \sum_{j=1}^l \mathbf{N}_j .$$

Thus $\mathbf{N}_i\mathbf{I}_k = \mathbf{N}_i = \sum_{j=1}^l \mathbf{N}_i\mathbf{N}_j = \mathbf{N}_i^2$ i.e. the $\mathbf{N}_i$ are idempotent. Hence $\mathbf{N}_1, \cdots, \mathbf{N}_l$ are symmetric orthogonal idempotents i.e.

$$\mathbf{N}_i = \mathbf{N}_i', \quad \mathbf{N}_i\mathbf{N}_j = \delta_{ij}\mathbf{N}_i , \quad \sum_1^l \mathbf{N}_i = \mathbf{I}_k . \tag{4.13}$$

Corresponding to the subspaces $S_j(k)$ of $k$-space, there are orthogonal subspaces $S_j(n)$ in $n$-space. For to each $\mathbf{W}_i$, there is an invariant subspace $\mathbf{T}_i$ in $n$-space. Hence

$$S_j(n) = \{ \cup \mathbf{T}_i \mid i \varepsilon S_j \}, \qquad j = 1, \cdots, l. \tag{4.14}$$

The projectors onto $S_j(n)$ is given by

$$\mathbf{G}_j = \mathbf{I}_n - \prod_{i \varepsilon S_j} (\mathbf{I}_n - \mathbf{E}_i), \qquad j = 1, \cdots, l. \tag{4.15}$$

These projectors in $n$-space are related to the projectors in $k$-space by the formula

$$\mathbf{G}_j\mathbf{X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\frac{1}{2}}\mathbf{N}_j(\mathbf{X}'\mathbf{X})^{\frac{1}{2}}. \tag{4.16}$$

To establish (4.16), (4.15) applied to (4.3) gives

$$\mathbf{G}_j\mathbf{X} = \mathbf{G}_j \sum_{i \varepsilon s} \mathbf{E}_i\mathbf{X} = \sum_{i \varepsilon s_j} \mathbf{E}_i\mathbf{X} = \sum_{i \varepsilon s_j} \mathbf{W}_i'(\mathbf{X}'\mathbf{X})^{\frac{1}{2}}. \tag{4.17}$$

But

$$\begin{aligned} \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\frac{1}{2}}\mathbf{N}_j(\mathbf{X}'\mathbf{X})^{\frac{1}{2}} &= \sum_1^m \mathbf{E}_i\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\frac{1}{2}}\mathbf{N}_j(\mathbf{X}'\mathbf{X})^{\frac{1}{2}} \\ &= \sum_S \mathbf{W}_i'\mathbf{N}_j(\mathbf{X}'\mathbf{X})^{\frac{1}{2}} \\ &= \sum_{i \varepsilon S_j} \mathbf{W}_i'(\mathbf{X}'\mathbf{X})^{\frac{1}{2}} \end{aligned} \tag{4.18}$$

since $\mathbf{N}_j$ projects orthogonally onto the space spanned by $\{\mathbf{W}_i \mid i \varepsilon S_j\}$. The equality of (4.18) and (4.17) is (4.16).

Finally if we define

$$\mathbf{M}_j = (\mathbf{X}'\mathbf{X})^{-\frac{1}{2}}\mathbf{N}_j(\mathbf{X}'\mathbf{X})^{\frac{1}{2}} \tag{4.19}$$

then

$$\mathbf{G}_j\mathbf{X} = \mathbf{X}\mathbf{M}_j, \quad \mathbf{M}_i\mathbf{M}_j = \delta_{ij}\mathbf{M}_i, \quad \sum_1^l \mathbf{M}_i = \mathbf{I}_n ; \tag{4.20}$$

however the $\mathbf{M}_j$ are not necessarily symmetric.

Thus many results follow from the constructions based upon the maximal subdivision $S_1, \cdots, S_l$ of the regression spectrum $S$, and we make the following:

DEFINITION. The subsets $S_1, \cdots, S_l$ of the regression spectrum used in (4.12) to define the orthogonal idempotents are called the elements of the regression spectrum.

The analogue of the main theorem of Grenander and Rosenblatt (Chapter 7, 1957) may now be proved.

THEOREM 3. *For a given* $\mathbf{X}$ *and* $\mathbf{\Gamma}$ *of full rank a necessary and sufficient condition for* $\mathbf{b} = \hat{\boldsymbol{\beta}}$ *for all observations* $\mathbf{y}$ *is that the eigen values of* $\mathbf{\Gamma}$ *be constant on the elements of the regression spectrum.*

PROOF. Since $\mathbf{b} = \hat{\boldsymbol{\beta}}$ implies var $(\mathbf{b})$ = var $(\hat{\boldsymbol{\beta}})$, the necessity part of the theorem may use the consequences of equating (4.7) and (4.10) which yields

$$\sum_s f_{E_i}\mathbf{W}_i\mathbf{W}_i' = (\sum_s f_{E_i}^{-1}\mathbf{W}_i\mathbf{W}_i')^{-1}, \tag{4.21}$$

or

$$(\sum_s f_{E_i}\mathbf{W}_i\mathbf{W}_i')(\sum_s f_{E_i}^{-1}\mathbf{W}_i\mathbf{W}_i') = \mathbf{I}_k . \tag{4.22}$$

From (4.21), there exists an orthogonal transformation $\mathbf{H}$ diagonalizing both factors on the lefthand side of (4.22). Each of these factors may be written as a sum,

$$\sum_s f_{E_i}\mathbf{W}_i\mathbf{W}_i' = \sum_{j=1}^l \sum_{s_j} f_{E_i}\mathbf{W}_i\mathbf{W}_i', \tag{4.23}$$

$$\sum_s f_{E_i}^{-1}\mathbf{W}_i\mathbf{W}_i' = \sum_{j=1}^l \sum_{s_j} f_{E_i}^{-1}\mathbf{W}_i\mathbf{W}_i'. \tag{4.24}$$

The matrices $\sum_{s_j} f_{E_i}\mathbf{W}_i\mathbf{W}_i' = \mathbf{B}_j$, $\sum_{s_j} f_{E_i}^{-1}\mathbf{W}_i\mathbf{W}_i' = \mathbf{C}_j$, $j = 1, \cdots, l$, satisfy $\mathbf{B}_j = \mathbf{B}_j'$, $\mathbf{B}_i\mathbf{B}_j$ = null matrix = $\mathbf{B}_j\mathbf{B}_i$ $(i \neq j)$, $\mathbf{C}_j = \mathbf{C}_j'$, $\mathbf{C}_i\mathbf{C}_j$ = null-matrix = $\mathbf{C}_j\mathbf{C}_i$ $(i \neq j)$, by the orthogonality of the $\mathbf{W}_i$ in the elements of the spectrum. For the same reason $\mathbf{B}_i\mathbf{C}_j = \mathbf{C}_j\mathbf{B}_i$ = null matrix, $(i \neq j)$. But symmetric matrices which commute may be simultaneously diagonalized by the same orthogonal transformation. Thus each sum is separately diagonalized by the same matrix which we know to be $\mathbf{H}$. Moreover the orthogonality of the $\mathbf{B}_i$ and $\mathbf{C}_j$ implies that of $\mathbf{H}\mathbf{B}_i\mathbf{H}'$ and $\mathbf{H}\mathbf{C}_j\mathbf{H}'$, $i \neq j$. Hence (4.22) becomes

$$\sum_{j=1}^l (\sum_{s_j} f_{E_i}\mathbf{H}\mathbf{W}_i\mathbf{W}_i'\mathbf{H}')(\sum_{s_j} f_{E_i}^{-1}\mathbf{H}\mathbf{W}\mathbf{W}_i'\mathbf{H}') = \mathbf{I}_k \tag{4.25}$$

where $\sum_{s_j} f_{E_i}\mathbf{H}\mathbf{W}_i\mathbf{W}_i'\mathbf{H}'$ and $\sum_{s_j} f_E{}^{-1}\mathbf{H}\mathbf{W}_i\mathbf{W}_i'\mathbf{H}'$ are diagonal, all $j$, and

each diagonal product $(\sum_{S_j} f_{E_i}\mathbf{H}\mathbf{W}_i\mathbf{W}_i'\mathbf{H}')(\sum_{S_j} f_{E_i}^{-1}\mathbf{H}\mathbf{W}_i\mathbf{W}_i'\mathbf{H}')$ is orthogonal to every other. Hence these products have zeros and unities on their diagonal and the unities of one product correspond to the zeros of all others. For each $j$ there will be an $r$ such that the $(r, r)$ element of $\sum_{S_j} f_{E_i}\mathbf{H}\mathbf{W}_i\mathbf{W}_i'\mathbf{H}'$ $\sum_{S_j} f_{E_i}^{-1}\mathbf{H}\mathbf{W}_i\mathbf{W}_i'\mathbf{H}'$ is unity. Then orthogonality requires the (using $(\mathbf{A})rr$ for the $(r, r)$th element of $\mathbf{A}$)

$$\sum_{i\varepsilon S_j} f_{E_i}(\mathbf{H}\mathbf{W}_i\mathbf{W}_i'\mathbf{H}')_{rr} \sum_{i\varepsilon S_j} f_{E_i}^{-1}(\mathbf{H}\mathbf{W}_i\mathbf{W}_i'\mathbf{H}')_{rr} = 1. \tag{4.26}$$

Since all $\mathbf{W}_i$, $i \varepsilon S_j$, are non-null, the same is true of $\mathbf{H}\mathbf{W}_i$, $i \varepsilon S_j$ so that $(\mathbf{H}\mathbf{W}_i\mathbf{W}_i'\mathbf{H}')_{rr}$ is positive for every $i \varepsilon S_j$. Now $\sum_S \mathbf{W}_i\mathbf{W}_i' = \mathbf{I}_k$ so

$$\sum_S (\mathbf{H}\mathbf{W}_i\mathbf{W}_i'\mathbf{H}')_{rr} = 1, \qquad r = 1, \cdots, k.$$

But directly from (4.22) we have for each $r = 1, \cdots, k$

$$\sum \sum_{j,j'=1}^{l} (\sum_{S_j}(\mathbf{H}\mathbf{W}_i\mathbf{W}_i'\mathbf{H}')_{rr})(\sum_{S_{j'}} f_{E_i}^{-1}\mathbf{H}\mathbf{W}_i\mathbf{W}_i'\mathbf{H}')_{rr} = 1. \tag{4.27}$$

Combining (4.27) and (4.26) for the particular $r$ for which (4.26) is true we have

$$\sum \sum_{p,q=1,(p,q)\neq(j,j)}^{l} (\sum_{S_p} f_{E_i}(\mathbf{H}\mathbf{W}_i\mathbf{W}_i'\mathbf{H}')_{rr}) \cdot (\sum_{S_q} f_{E_i}^{-1}(\mathbf{H}\mathbf{W}_i\mathbf{W}_i'\mathbf{H}')_{rr}) = 0, \tag{4.28}$$

where all the factors on the lefthand side of (4.28) are positive. Hence for an $r$ for which (4.26) is true,

$$\sum_{S_p} f_{E_i}(\mathbf{H}\mathbf{W}_i\mathbf{W}_i'\mathbf{H}')_{rr} = \sum_{S_p} f_{E_i}^{-1}(\mathbf{H}\mathbf{W}_i\mathbf{W}_i'\mathbf{H}')_{rr} = 0 \qquad (p \neq j) \tag{4.29}$$

which implies that

$$\sum_{S_p}(\mathbf{H}\mathbf{W}_i\mathbf{W}_i'\mathbf{H}')_{rr} = 0, \qquad p \neq j. \tag{4.30}$$

Hence for $j$ and $r$ such that (4.26) is true

$$\sum_S(\mathbf{H}'\mathbf{W}_i\mathbf{W}_i'\mathbf{H})_{rr} = 1 = \sum_{S_j}(\mathbf{H}\mathbf{W}_i\mathbf{W}_i'\mathbf{H})_{rr} \tag{4.31}$$

so that (4.26) may be written

$$\sum_{S_j} f_{E_i}(\mathbf{H}\mathbf{W}_i\mathbf{W}_i'\mathbf{H})_{rr} \sum_{S_j} f_{E_i}^{-1}(\mathbf{H}\mathbf{W}_i\mathbf{W}_i'\mathbf{H}')_{rr} = \sum_S(\mathbf{H}\mathbf{W}_i\mathbf{W}_i\mathbf{H}')_{rr}. \tag{4.32}$$

The condition for equality in Cauchy's inequality then gives us that $f_{E_i}$ and $f_{E_i}^{-1}$ are proportional for $i \varepsilon S_j$, i.e. that $f_{E_i}$ is constant on the element of $S_j$ of the regression spectrum. This proves the necessity. To prove sufficiency, we have only to show that if $f_{E_i}$ is constant, $f_j$ say, for $i \varepsilon S_j$, $\mathbf{b} = \hat{\boldsymbol{\beta}}$ or, equivalently, that (4.22) is true. But (4.22) then reads

$$f_j^{-1}\mathbf{N}_j \sum_{S_j} \mathbf{W}_i = f_j^{-1} \sum_{S_j} \mathbf{W}_i, \qquad j = 1, \cdots, l,$$

which is obviously true since $\mathbf{N}_j\mathbf{W}_i = \mathbf{W}_i$, $i \varepsilon S_j$.

DISCUSSION. The conditions of Theorem 3 are not as easy to understand as those of Theorem 1 so it is helpful to see how each set implies the other.

If the conditions of Theorem 1 are satisfied, the $k$ columns of $\mathbf{X}$ are all linear

combinations of $\mathbf{g}_{i_1}, \cdots, \mathbf{g}_{i_k}$, some $k$ eigen vectors of $\mathbf{\Gamma}$. From (4.4),

$$\mathbf{W}_i = (\mathbf{X}'\mathbf{X})^{-\frac{1}{2}}\mathbf{X}'\mathbf{E}_i, \qquad i = 1, \cdots, m,$$

can be non-null for at most min $(k, m)$ values of $i$. Write

$$\mathbf{X} = [\mathbf{g}_{i_1}, \cdots, \mathbf{g}_{i_k}]\mathbf{C}$$

where $\mathbf{C}$ is a $k \times k$ non-singular matrix so that $\mathbf{X}'\mathbf{X} = \mathbf{C}'\mathbf{C}$ and

$$\mathbf{W}_i = (\mathbf{C}'\mathbf{C})^{-\frac{1}{2}}\mathbf{C}' \begin{bmatrix} \mathbf{g}_i' \mathbf{E}_i \\ \vdots \\ \mathbf{g}_{ik}' \mathbf{E}_i \end{bmatrix}$$

where the rows in the last factor are only non-null when the eigenvector is in $\mathbf{T}_i$. Thus the last factors in any two distinct $\mathbf{W}_i$'s can have no non-null rows in common because $\mathbf{T}$-subspaces are orthogonal. Hence

$$\mathbf{W}_i'\mathbf{W}_j = [\mathbf{E}_i\mathbf{g}_{i_1}, \cdots, \mathbf{E}_i\mathbf{g}_{i_k}]\mathbf{C}(\mathbf{C}'\mathbf{C})^{-1}\mathbf{C}' \begin{bmatrix} \mathbf{g}_{i1}' \mathbf{E}_j \\ \vdots \\ \mathbf{g}_{ik}' \mathbf{E}_j \end{bmatrix}$$

$$= [\mathbf{E}_i\mathbf{g}_{i_1}, \cdots, \mathbf{E}_i\mathbf{g}_{i_k}] \begin{bmatrix} \mathbf{g}_{i1}' \mathbf{E}_j \\ \vdots \\ \mathbf{g}_{ik}' \mathbf{E}_j \end{bmatrix}$$

$$= \mathbf{E}_i\mathbf{P}\mathbf{E}_j$$

where $\mathbf{P}$ is an idempotent that projects onto the regression subspace. For any vector $\mathbf{v}$ in $n$ space, $\mathbf{v}_j = \mathbf{E}_j\mathbf{v}$ is its component in $\mathbf{T}_j$, $\mathbf{P}\mathbf{v}_j$ is the component of $\mathbf{v}_j$ in the regression space but is still in $\mathbf{T}_j$. Therefore $\mathbf{E}_i\mathbf{P}\mathbf{E}_j\mathbf{v}$ is the null vector when $i \neq j$. Hence $\mathbf{W}_i'\mathbf{W}_j =$ null matrix. Hence the $l = \min(k, m)$ matrices $\mathbf{W}_i$ are individually the elements of the spectrum $S_1, \cdots, S_l$. The eigen values are trivially constant on these elements. Thus Grenander's conditions are satisfied.

The conditions of Theorem 3 will imply those of Theorem 1 if we can show that $k$ eigenvectors can be chosen out of the $l$ subspaces $S_j(n), j = 1, \cdots, l$, so that the columns of $\mathbf{X}$ are linear functions of these $k$ eigenvectors. Consider the part of $\mathbf{X}$ in $S_j(n)$, $\mathbf{G}_j\mathbf{X}$ whose columns span a space whose dimension is rank $(\mathbf{G}_j\mathbf{S})$. By (4.16) and a standard result on the rank of a product, rank $(\mathbf{G}_j\mathbf{X}) \leqq$ rank $(\mathbf{N}_j)$. Any set of rank $(\mathbf{N}_j)$ orthogonal vectors in $S_j(n)$ are eigen vectors of $\mathbf{\Gamma}$ for root $f_{S_j}$ (since $f_{E_i} = f_{S_j}, i \varepsilon S_j$) by Theorem 3. This is true for $j = 1, \cdots, l$ so that the columns of $\mathbf{X}$ may be expressed as linear functions of at most $\sum_{j=1}^{l}$ rank $(\mathbf{N}_j) = k$ eigen vectors of $\mathbf{\Gamma}$. In fact $k$ are needed because $\mathbf{X}$ must be of rank $k$. This proves the condition of Theorem 1.

Theorem 3 considers a given $\mathbf{X}$ and a given $\mathbf{\Gamma}$. Suppose we have a given $\mathbf{X}$ and consider a class of positive definite matrices $\mathbf{\Gamma}$ each member of which has the same eigen vectors (i.e. spectral subspaces) but may have any positive eigen values. For least squares to be efficient, or $\mathbf{b} = \hat{\beta}$ for every member of the class, the eigen values of $\mathbf{\Gamma}$ must be constant on each element of the regression spec-

trum, the latter being uniquely defined for $\mathbf{X}$ and the class of $\mathbf{\Gamma}$'s considered. Since $f_{E_1}, \cdots f_{E_m}$ may be any positive numbers, this will be so only if one $f_{E_i}$ is involved in each element. Hence the spectrum must consist of $l$ elements, each with only one $\mathbf{W}_i$ in it—the remaining $\mathbf{W}_i$ being null. The non-null $\mathbf{W}_i'$ are then orthogonal. Thus the spectrum must consist of $l \neq k$ points. Hence we have the analogue to another theorem of Grenander and Rosenblatt (1957) which may be stated as follows.

THEOREM 4. *For a given* $\mathbf{X}$, $\mathbf{E}_1, \cdots, \mathbf{E}_m$ *and a class of matrices* $\mathbf{\Gamma}$, $\{\mathbf{\Gamma} \mid \mathbf{\Gamma} = \sum_1^m f_{E_i}\mathbf{E}_i, f_{E_1} > 0, \cdots, f_{E_m} > 0\}$, *a necessary and sufficient condition that* $\mathbf{b} = \hat{\beta}$ *for every* $\mathbf{\Gamma}$ *in the class is that the spectrum of* $\mathbf{X}$ *consists of* $l \leqq k$ *points.*

DISCUSSION. This theorem is a trivial consequence of Theorem 1 as may be seen from the discussion of Theorem 1.

**5. The residuals.** The residuals are used to find out the properties of the error distribution. The covariance matrix $\mathbf{\Gamma}$ of $\mathbf{u}$ is of most interest. For some unit vector $\mathbf{c}$ we will thus be interested in the relationship of var $(\mathbf{c}'\mathbf{z})$ to var $(\mathbf{c}'\mathbf{u})$. When the eigen vectors $\mathbf{g}_i$ of $\mathbf{\Gamma}$ are known, interest will largely be limited to comparisons of cov $(\mathbf{g}_i'\mathbf{z}, \mathbf{g}_j'\mathbf{z})$ with cov $(\mathbf{g}_i'\mathbf{u}, \mathbf{g}_j'\mathbf{u})$. Formulae for these purposes will now be derived.

When least squares is efficient and the $k$ columns of $\mathbf{X}$ are linear functions of $k$ eigen vectors, $\mathbf{g}_1, \cdots, \mathbf{g}_k$ say, of $\mathbf{\Gamma}$, it was seen in (3.5) that the variance of the residual vector $\mathbf{z}$ in the directions $\mathbf{g}_1, \cdots, \mathbf{g}_k$ was zero and in the direction $\mathbf{g}_i$ $(i = k + 1, \cdots, n)$ was $f_i$. To discuss the general case when the columns of $\mathbf{X}$ have components on more than $k$ eigen vectors, we have the general formula (4.8) for $\mathbf{z}$. The most revealing way to study (4.8) is to assume that the eigen vectors of $\mathbf{\Gamma}$ are uniquely defined, as was done in Sections 2 and 3.

Thus in the results of Section 4, $\mathbf{E}_i$ will be replaced by $\mathbf{g}_i\mathbf{g}_i'$ and $m$ by $n$ and (2.5) used to represent $\mathbf{u}$. Since the $k \times n$ matrix $\mathbf{W}_i$ now has the form $\mathbf{W}_i = (\mathbf{X}'\mathbf{X})^{-\frac{1}{2}}\mathbf{X}'\mathbf{g}_i\mathbf{g}_i'$, $\mathbf{W}_i\mathbf{g}_j$ is null if $i \neq j$ and the expression (4.8) for the residual vector $\mathbf{z}$ becomes

$$\mathbf{z} = \sum_{i=1}^n \eta_i(1 - (\sum_{j=1}^n \mathbf{W}_j')\mathbf{W}_i)\mathbf{g}_i . \tag{5.1}$$

Now $S$, the spectrum of $\mathbf{X}$, is the set of subscripts for which $\mathbf{W}_i$ is non-null so (5.1) can be written

$$\mathbf{z} = \sum_{i \notin S} \eta_i\mathbf{g}_i + \sum_{i \in S} \eta_i(1 - (\sum_{j \in S} \mathbf{W}_j')\mathbf{W}_i)\mathbf{g}_i . \tag{5.2}$$

Hence if $\mathbf{c}$ is a direction belonging to the spectral subspace spanned by $\{\mathbf{g}_i/i \in S\}$, the variance of $\mathbf{z}$ in the direction $\mathbf{c}$, var $(\mathbf{c}'\mathbf{z})$, is identical with var $(\mathbf{c}'\mathbf{u})$ i.e. it is unaffected by the regression. Introducing the elements $S_1, \cdots, S_l$ of the regression spectrum, the second term in (5.2) may be rewritten so that

$$\mathbf{z} = \sum_{i \notin S} \eta_i\mathbf{g}_i + \sum_{j^*=1}^l \sum_{i \in S_{j^*}} \eta_i(1 - (\sum_{j \in S_{j^*}} \mathbf{W}_i')\mathbf{W}_i)\mathbf{g}_i . \tag{5.3}$$

Thus the variance of $\mathbf{z}$ in any direction in the spectral subspace spanned by $\{\mathbf{g}_i/i \in S_j\}$ depends only on the eigenvalues $f_i/i \in S_j$. In particular the variance

of $\mathbf{z}$ in the direction $i^*$, when $i^* \varepsilon S_{j^*}$, is given by

$$\text{(5.4)} \qquad \text{var } (\mathbf{g}_{i^*}'\mathbf{z}) = f_{i^*} - 2 f_{i^*}\mathbf{g}_{i^*}'\mathbf{W}_{i^*}'\mathbf{W}_{i^*}\mathbf{g}_{i^*} + \sum_{i \varepsilon S_{j^*}} f_i \mathbf{g}_{i^*}'\mathbf{W}_{i^*}'(\mathbf{W}_i\mathbf{g}_i\mathbf{g}_i'\mathbf{W}_i')\mathbf{W}_{i^*}\mathbf{g}_{i^*}.$$

If the condition of Theorem 3 is true i.e. $f_i = f_{S_{j^*}}$, $i \varepsilon S_{j^*}$,

$$\text{(5.5)} \quad \text{var } (\mathbf{g}_{i^*}'\mathbf{z}) = f_{S_{j^*}}\{1 - 2\mathbf{g}_{i^*}'\mathbf{W}_{i^*}'\mathbf{W}_{i^*}\mathbf{g}_{i^*} + \mathbf{g}_{i^*}'\mathbf{W}_{i^*}'(\sum_{i \varepsilon S_{j^*}} \mathbf{W}_i\mathbf{g}_i\mathbf{g}_i'\mathbf{W}_i')\mathbf{W}_{i^*}\mathbf{g}_{i^*}\}.$$

If further we define

$$\text{(5.6)} \qquad \mathbf{X}_i = \mathbf{X}'\mathbf{g}_i, \quad \mathbf{X} = \sum_1^n \mathbf{g}_i\mathbf{X}_i',$$

$$\mathbf{W}_i\mathbf{g}_i = (\mathbf{X}'\mathbf{X})^{-\frac{1}{2}}\mathbf{X}_i = \mathbf{V}_i, \quad \text{say}$$

then $\mathbf{X}'\mathbf{X} = \sum_1^n \mathbf{X}_i\mathbf{X}_i'$ or $\mathbf{I}_k = \sum_1^n \mathbf{V}_i\mathbf{V}_i'$ and it may be shown, as in the derivation of (4.13), that

$$\text{(5.7)} \qquad \mathbf{O}_j = \sum_{i \varepsilon S_j} \mathbf{V}_i\mathbf{V}_i'$$

also obey (4.13). Hence (5.5) becomes, when least squares is efficient,

$$\text{var } (\mathbf{g}_{i^*}'\mathbf{z}) = f_{S_{j^*}}\{1 - 2\mathbf{V}_{i^*}'\mathbf{V}_{i^*} + \mathbf{V}_{i^*}'\mathbf{O}_{j^*}\mathbf{V}_{i^*}\},$$

i.e.

$$\text{(5.8)} \qquad \text{var } (\mathbf{g}_{i^*}'\mathbf{z}) = f_{S_{j^*}}\{1 - \mathbf{V}_{i^*}'\mathbf{V}_{i^*}\}.$$

It will now be shown that $0 \leqq \mathbf{V}_{i^*}'\mathbf{V}_{i^*} \leqq 1$ so that

$$\text{(5.9)} \qquad \text{var } (\mathbf{g}_{i^*}'\mathbf{z}) \leqq f_{S_{j^*}} = \text{var } (\mathbf{g}_{i^*}'\mathbf{u}), \quad \text{all} \quad i^* \varepsilon S_{j^*}.$$

From (5.6), consider the $k \times n$ matrix

$$\mathbf{V} = [\mathbf{V}_1, \cdots, \mathbf{V}_n] = (\mathbf{X}'\mathbf{X})^{-\frac{1}{2}}\mathbf{X}'[\mathbf{g}_1, \cdots, \mathbf{g}_n].$$

Clearly $\mathbf{V}\mathbf{V}' = \mathbf{I}_k$, so that $\mathbf{V}$ is the first $k$ rows of an orthogonal $n \times n$ matrix with real elements. If it were completed, the sum of squares of the elements in each column would be unity. Hence the sum of squares of the first $k$ elements in any column, which is $\mathbf{V}_i'\mathbf{V}_i$, lies in (0, 1). Because of the identity of Theorems 1 and 3, it follows from (3.5) that the eigenvectors may in this case be chosen to make var $(\mathbf{g}_{i^*}'\mathbf{z}) = 0$, $\mathbf{g}_{i^*}$ being an eigen vector in spectral subspace associated with $S_{j^*}$.

When least squares is not necessarily efficient, var $(\mathbf{g}_i'\mathbf{z})$ for $i \varepsilon S$ may be greater or less than var $(\mathbf{g}_i'\mathbf{u}) = f_i$. The formula (5.4) suggests that, as a general rule, if $f_i$ is high, var $(\mathbf{g}_i'\mathbf{z}) < f_i$ and if $f_i$ is low, var $(\mathbf{g}_i'\mathbf{z}_i) > f_i$. For suppose $f^*$ and $F^*$ are the least and greatest roots in the set $\{f_i \uparrow i \varepsilon S_{j^*}\}$. Then since (5.4) gives, if $i^*$ corresponds to $F^*$

$$\text{var } (\mathbf{g}_{i^*}'\mathbf{z}) = F^*(1 - 2\mathbf{V}_{i^*}'\mathbf{V}_{i^*}) + \sum_{S_{j^*}} f_i(\mathbf{V}_{i^*}'\mathbf{V}_{i^*})^2, \qquad \text{i.e.}$$

$$\text{(5.10)} \qquad \text{var } (\mathbf{g}_{i^*}'\mathbf{z}) \leqq F^*(1 - \mathbf{V}_{i^*}'\mathbf{V}_{i^*}) \leqq F^* = \text{var } (\mathbf{g}_{i^*}'\mathbf{u}).$$

However if $i^*$ corresponds to $f^*$, the same argument only gives

$$\text{var } (\mathbf{g}_{i^*}'\mathbf{z}) \geqq f^*(1 - \mathbf{V}_{i^*}'\mathbf{V}_{i^*}).$$

There is an overall tendency to lose variance because

$$\begin{aligned}\textstyle\sum_{i^*\varepsilon S_{j^*}} \text{var } (\mathbf{g}_{i^*}'\mathbf{z}) &= \textstyle\sum_{i^*\varepsilon S_{j^*}} f_{i^*} - 2\sum_{i^*\varepsilon S_{j^*}} f_{i^*}\mathbf{V}_{i^*}'\mathbf{V}_{i^*} \\ &\quad + \textstyle\sum_{i\varepsilon S_{j^*}} f_i\mathbf{V}_i'(\sum_{i^*\varepsilon S_{j^*}} \mathbf{V}_{i^*}\mathbf{V}_{i^*}')\mathbf{V}_i \\ &= \textstyle\sum_{S_{j^*}} f_{i^*} - \sum_{S_{j^*}} f_{i^*}\mathbf{V}_{i^*}'\mathbf{V}_{i^*}, \qquad \text{i.e.}\end{aligned}$$

$$\text{(5.11)} \qquad \textstyle\sum_{i^*\varepsilon S_{j^*}} \text{var } (\mathbf{g}_{i^*}'\mathbf{z}) < \sum_{S_{j^*}} f_{i^*} = \sum_{i^*\varepsilon S_{j^*}} \text{var } (\mathbf{g}_{i^*}'\mathbf{u}).$$

This is attributable to the linear restrictions $\mathbf{X}'\mathbf{z} = 0$ on the residual vector.

The covariances of the compounds $\mathbf{g}_i'\mathbf{z}$ may be examined in the same way, thus

$$\begin{aligned}\text{(5.12)} \qquad \text{cov } (\mathbf{g}_h'\mathbf{z}, \mathbf{g}_i'\mathbf{z}) &= 0, \qquad h, i \text{ not both in } S_{j^*}, \\ &= \mathbf{V}_h'(-f_h - f_i + \textstyle\sum_{q\varepsilon S_{j^*}} f_q\mathbf{V}_q\mathbf{V}_q')\mathbf{V}_i, \quad h, i\,\varepsilon\, S_{j^*}.\end{aligned}$$

By comparison, cov $(\mathbf{g}_h'\mathbf{u}, \mathbf{g}_i\mathbf{u}) = \delta_{hi}f_i$.

**6. Application to time series data.** In 1950, when the present writer became interested in regression analysis in economics, the reasons for assuming that the error process in economic time series regression was stationary were less than compelling. For example there seemed no reason why the variance should not alter. Little information could then be gained empirically because of short series and limited computing facilities. Thus this writer was unwilling to assume $\mathbf{\Gamma}$ had much structure and lead to develop a more general theory. Furthermore the $\mathbf{X}$ vectors were often time series of complex form and so it seemed unlikely that they would be simply related to the eigen vectors of $\mathbf{\Gamma}$, whatever they might be. Hence the conditions for Eff (**b**) to be very low might often be met and, as also shown in Watson (1955), the standard errors and tests might be very misleading. (One would rarely *know* this because $\mathbf{\Gamma}$ is rarely known.) Thus the use of least squares in these circumstances seemed to rest on optimism and convenience.

If the error process is assumed to be stationary in time, the exact eigen vectors of $\mathbf{\Gamma}$ vary from process to process, in general. However, with an approximation that improves with sample size, it is possible to think of $\mathbf{\Gamma}$ for stationary processes as having fixed eigen vectors and varying roots, equal in pairs, which correspond to the values of the spectral density at multiples of $2\pi/n$ (Whittle, (1951)). The finite analysis may then be made fairly satisfactorily. A slight variant of this was used in Watson (1952)—approximations to simple autoregressive and moving average processes were devised with covariance matrices with known roots and vectors. Grenander (loc. cit.) took the other way out of this dilemma—rigorous asymptotic treatment. *For practical purposes each is approximate.* The finite treatment is easier to understand in relation to least squares and analysis of variance. The asymptotic treatment is more naturally related to covariance stationary processes: the many uses made by Hannan (1957), (1963a), (1963b), (1965)) indicate it is perhaps the most profitable approach in the long run.

The approximate eigen vectors of the last paragraph are the Fourier or harmonic vectors defined by

$$(6.1)\quad \begin{bmatrix} \cos\ 2\pi h/n \\ \cos\ 4\pi h/n \\ \vdots \\ \cos\ 2\pi nh/n \end{bmatrix} \quad \begin{aligned} h &= 0, 1, \cdots, n/2, \qquad n \text{ even}, \\ &= 0, 1, \cdots, (n-1)/2, n \text{ odd}, \\ \text{root} &= f(2\pi h/n); \end{aligned}$$

$$\begin{bmatrix} \sin\ 2\pi h'/n \\ \sin\ 4\pi h'/n \\ \vdots \\ \sin\ 2\pi h'n/n \end{bmatrix} \quad \begin{aligned} h' &= 1, 2, \cdots, (n-2)/2, n \text{ even}, \\ &= 1, 2, \cdots, (n-1)/2, n \text{ odd}, \\ \text{root} &= f(2\pi h'/n). \end{aligned}$$

The result (6.1) follows from the fact that $\Gamma$ is, approximately, a linear combination of $n \times n$ circulants. The roots are equal in pairs except that corresponding to $h = 0$ (whose vector is associated with the mean) and to $h = n/2$ for $n$ even (whose vector oscillates faster than any other).

It is vital to notice that the ordinates of $f(\theta)$ for $\theta =$ low multiples of $2\pi/n$ are associated with Fourier vectors that oscillate slowly. For example if $\theta_n = 2\pi h/n$, the elements of the vector form the sequence $\cos\theta_n$, $\cos 2\theta_n$, $\cdots$, $\cos n\theta_n$. The sign of the cosine changes as its argument passes through $\pi/2$ and $3\pi/2$. Multiples of $\theta_n$ "go round the clock" faster, the larger $\theta_n$ is i.e. for larger $h$. Thus for small $h$, $h'$, the oscillation in sign is relatively slow.

The consequence of Grenander's work that has received the greatest attention is the following: *if the columns of $X$ are slowly varying and the spectral density $f(\theta)$ is almost flat for $\theta$ near zero, least squares is nearly* 100% *efficient.* This may be deduced from our arguments as follows. The assumption on $X$ really requires that the columns of $X$ be expressible in terms of the vectors in (6.1) that use small $h$ and $h'$. These correspond to roots almost equal to values of $f(\theta)$ for small $\theta$—say for values of $\theta$ on $(0, c)$.

By the appropriate form (3.19) (i.e. the $k$th power of the bound in (3.11))

$$(6.2)\qquad f = \min_{0 \leqq \theta \leqq c} f(\theta), \qquad F = \max_{0 \leqq \theta \leqq c} f(\theta),$$

we see that the assumption of $f(\theta)$ (that it be almost constant for $0 \leqq \theta \leqq c$) implies that Eff(**b**) is bounded below by a number only a little less than unity. The key fact is that the latent vectors at low frequencies here correspond to nearly equal roots.

Our analysis in Section 5 gives us more than this. We see that the $\mathbf{W}_i$ and $\mathbf{V}_i$ matrices are null unless they are associated with low frequency eigen vectors that define the regression spectrum. From (5.2) it now follows that the power or variance of the residuals is the same as the error process in the direction of the higher frequency eigen vectors (i.e. off the regression spectrum) but are generally reduced in the direction of the low frequency vectors. Hence a study of the residuals will give a biased picture of the power distribution—in particular an underestimate of the relevant power for the sampling variation of the estimates—unless this regression effect is corrected.

The study of the residuals for the purpose of estimating the spectral density $f(\theta)$ is usually made via the periodogram. It is now computationally feasible (Jones, (1965)) to transform the model (2.1) by premultiplication by the orthogonal matrix $[\mathbf{g}_1, \cdots, \mathbf{g}_n]$ with the $\mathbf{g}$'s defined as in (6.1). Since $\mathbf{g}_i'\mathbf{u} = \eta_i$ has variance $f_i$, the square of each term is related to an $f_i$. If they are assumed to follow a regular pattern, suitable simple averaging will lead to good estimates. Formula (5.4) suggests that the regression bias may be eliminated. This is discussed in Hannan (1960). More simply, if $f(\theta)$ is estimated this way for $\theta$ away from the origin it may sometimes be safe to extrapolate back to 0. Devices of this kind are discussed with examples in Duncan and Jones (1966), who attempt to utilize the estimates of $f(\theta)$ to improve on the least squares estimates. Their examples are however very favorable to their method.

More generally the case of $\mathbf{X}$ made of low frequency vectors is seen as a case where the spectral distribution of the regressors is concentrated near zero. Clearly it is not essential that the region be near zero but simply that it be concentrated in a region where $f(\theta)$ is varying slowly. And furthermore the region of concentration need not be a simply-connected interval—it may be the union of up to $k$ narrow intervals, or frequency bands (this is the simplest case of Theorem 1). Of course it is less likely in this case that $f(\theta)$ will have similar values for widely different values of $\theta$ but it is only required that $f(\theta)$ be nearly constant *within* these different intervals. This leads to the ideas in Hannan (1963a).

Another point in the literature that may be easily explained in the above terms is Hannan's (1957) observation that the upper bounding significance point given the Durbin and Watson ((1950), (1951)) statistic for testing for serial correlation is almost exact for low frequency regressors. For the upper bound is obtained by replacing the actual regressors by the more slowly moving latent vectors which here correspond to the least roots of the matrix $\mathbf{A}$.

**7. Acknowledgment.** I am grateful to Professor James Durbin for many discussions and to Dr. A. S. Householder and Professor I. Olkin for suggesting the inequality (3.18).

## REFERENCES

ANDERSON, T. W. (1948). On the theory of testing serial correlation. *Skand. Aktuarietidskr.* **31** 88–116.

DUNCAN, D. B. and JONES, R. H. (1966). Multiple regression with stationary errors. Technical Report #46, Johns Hopkins University.

DURBIN, J. and WATSON, G. S. (1950). Testing for serial correlation in least square regression I. *Biometrika* **37** 409–428.

DURBIN, J. and WATSON, G. S. (1951). Testing for serial correlation in least square regression II. *Biometrika* **38** 159–177.

GOLUB, G. H. (1963). Comparison of the variance of minimum variance and weighted least squares regression coefficients. *Ann. Math. Statist.* **34** 984–991.

GRENANDER, U. (1954). On the estimation of the regression coefficients in the case of an auto-correlated disturbance. *Ann. Math. Statist.* **25** 252–272.

GRENANDER, U. and ROSENBLATT M. (1957). *Statistical Analysis of Stationary Time Series*. Wiley, New York.

HANNAN, E. J. (1957). Testing for serial correlation in least squares regression. *Biometrika* **44** 57–66.

HANNAN, E. J. (1963a). Regression for Time Series. *Time Series Analysis* (ed. M. Rosenblatt). Wiley, New York.

HANNAN, E. J. (1963b). Regression for time series with errors of measurement. *Biometrika* **50** 293–302.

HANNAN, E. J. (1965). The estimation of relationships involving distributed lags. *Econometrika* **33** 206–224.

HANNAN, E. J. (1966). *Time Series Analysis*. Methuen, London.

HARDY, G. H., LITTLEWOOD, J. E. and PÓLYA, G. (1934). *Inequalities*. Cambridge Univ. Press.

JONES, R. H. (1965). A reappraisal of the periodogram in spectral analysis. *Technometrics* **7** 531–542.

KANTOROVICH, L. V. (1948). Functional analysis and applied mathematics. *Uspehi Mat. Nauk*. **3** 19–185.

MAGNESS, T. A. and MCGUIRE, J. B. (1962). Comparison of least squares and minimum variance estimates of regression parameters. *Ann. Math. Statist.* **33** 462–470.

MACDUFFE, C. C. (1946). *The Theory of Matrices*. Chelsea Publishing Co., New York.

RAO, C. R. (1965). Least squares theory using an estimated dispersion matrix and its application to measurement of signals. *Fifth Berkeley Symposium Math. Statist. Prob.*

RAO, C. R. (1965). *Linear Statistical Inference and Its Applications*. Wiley, New York.

WATSON, G. S. (1952). Serial correlation in regression analysis. Mimeo Ser. #49, Inst. of Statist., Univ. of North Carolina.

WATSON, G. S. (1955). Serial correlation in regression analysis, I. *Biometrika* **42** 327–341.

WATSON, G. S. and HANNAN, E. J. (1956). Serial correlation in regression analysis II. *Biometrika* **43** 436–448.

WHITTLE, P. (1951). *Hypothesis Testing in Time Series*. Almqvist & Wiksells, Uppsala.

ZYSKIND, G. (1962). On conditions for equality of best and simple least squares estimators (Abstract). *Ann. Math. Statist.* **33** 1502–1503.

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**.