This PDF is a selection from an out-of-print volume from the National Bureau of Economic Research

Volume Title: Annals of Economic and Social Measurement, Volume 2, number 4

Volume Author/Editor: Sanford V. Berg, editor

Volume Publisher:

Volume URL: http://www.nber.org/books/aesm73-4

Publication Date: October 1973

Chapter Title: Criteria Constraints And Multicollinearity In Random Coefficient Regression Models

Chapter Author: P. A. V. B. Swamy

Chapter URL: http://www.nber.org/chapters/c9935

Chapter pages in book: (p. 429 - 450)

Annals of Economic and Social Measurement, 2/4, 1973

CRITERIA, CONSTRAINTS AND MULTICOLLINEARITY IN RANDOM COEFFICIENT REGRESSION MODELS

BY P. A. V. B. SWAMY*

This paper analyzes six alternative estimators for random coefficient regression models: (1) minimum variance linear unbiased estimator (MVLU), (2) the Stein-like estimator, (3) the ridge regression estimator, (4) minimum conditional mean square error estimator (MCMSE), (5) the mixed regression estimator, and (6) a maximum likelihood estimator (ML). Attention is focused on the criteria of estimation and parametric constraints in RCR models.

1. INTRODUCTION

It has been recognized by many econometricians that the usefulness of the conventional fixed-parameter regression model in the analysis of cross-section data is limited because individuals differ greatly in their behavior, and the diversity of individual decision units implies parameter variation across units, see Swamy (1971) and the references cited therein. In recent years, econometric models, which permit different schemes of parameter variation, have been developed. All these different schemes have been compared by Swamy (1972) who developed an asymptotically efficient procedure of estimating the parameters in a general random coefficient regression (RCR) model. Application of these estimation methods in the analysis of real world data is just beginning, see Feige and Swamy (1972). It has been observed that the use of RCR methods can result in more fruitful and meaningful econometric analyses of micro panel data. In the present paper we analyze alternative estimators with purely algebraic tools. Attention is focused on the criteria of estimation and parametric constraints in RCR models.

The plan of the paper is as follows. Section 2 sets out the estimation rules for random coefficient regression models with and without an unbiasedness condition. Constraints on the parameters and partial prior information are introduced in Section 3 and it is indicated how their presence can help estimation. Methods of using sample data in conjunction with the first two moments of a prior distribution are reviewed in Section 4. The maximum likelihood method of estimating the parameters of a random coefficient model is discussed in Section 5. Summary and Conclusions of the study are presented in Section 6.

2. RANDOM COEFFICIENT REGRESSION MODEL

2.1. *The Model*

Swamy (1971) considers the problem of estimating the following equation from a time series of cross-sections,

$$\mathbf{y}_i = X_i\boldsymbol{\beta}_i + \mathbf{u}_i \qquad (i = 1, 2, \ldots, n) \tag{1}$$

* I am grateful to Professor A. Zellner and Dr. Richard D. Porter for helpful comments.

where $\mathbf{y}_i \equiv (y_{i1}, y_{i2}, \ldots, y_{iT})'$ is a $TX1$ vector of observations on a dependent variable, $X_i \equiv \{x_{ikt}\}$ $(k = 1, 2, \ldots, K; t = 1, 2, \ldots, T)$ is a TXK matrix of observations on K independent variables, $\boldsymbol{\beta}_i$ is a $KX1$ vector of coefficients, and $\mathbf{u}_i \equiv (u_{i1}, u_{i2}, \ldots, u_{iT})'$ is a $TX1$ vector of disturbances.

Observations on y's and x's for n individuals taken over T periods of time are available. These temporal cross-section data are obtained by assembling cross-sections of T years, with the same n cross-section units appearing in all years. The individuals here may be firms, consumers or regions. The subscript i indexes cross-section observations and the subscript t indexes time series observations.

In (1) both $\boldsymbol{\beta}_i$ and $\mathbf{u}_i$ are regarded as realizations of random vectors,[1] and the following assumptions are made.

Assumption 1:

(1) The rank of X_i is K, $n > K$ and $T > K$;

(2) For $i, j = 1, 2, \ldots, n$: $E\mathbf{u}_i = \mathbf{0}$ and $E\mathbf{u}_i\mathbf{u}_j' = \sigma_{ij}\Omega_{ij}$ where

$$\Omega_{ij} = \frac{1}{1 - \rho_i\rho_j}\begin{bmatrix} 1 & \rho_i & \rho_i^2 & \cdots & \rho_i^{T-1} \\ \rho_j & 1 & \rho_i & \cdots & \rho_i^{T-2} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho_j^{T-1} & \rho_j^{T-2} & \rho_j^{T-3} & \cdots & 1 \end{bmatrix}, \quad |\rho_i| < 1;$$

(3) For $i, j = 1, 2, \ldots, n$: $E\boldsymbol{\beta}_i = \bar{\boldsymbol{\beta}}$,

$$E(\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}})(\boldsymbol{\beta}_j - \bar{\boldsymbol{\beta}})' = \begin{cases} \Delta & \text{if } i = j \\ 0 & \text{otherwise,} \end{cases}$$

$\Delta = \{\delta_{kk'}\}$ $(k, k' = 1, 2, \ldots, K)$ is positive definite;

(4) $\boldsymbol{\beta}_i$ is independent of $\mathbf{u}_j$ for $i, j = 1, 2, \ldots, n$;

(5) The x_{ikt} are exogenous variables distributed independently of $\boldsymbol{\beta}_i$ and $\mathbf{u}_i$.[2] Furthermore, X_i is nonstochastic.

The implications of Assumption 1 are discussed by Swamy (1972). If we arrange the observations on each variable first by individual and then according to period, we may represent eq. (1) by

$$\text{(2)} \qquad \mathbf{y} = X\bar{\boldsymbol{\beta}} + D_x\boldsymbol{\xi} + \mathbf{u}$$

where $\mathbf{y} \equiv (\mathbf{y}_1', \mathbf{y}_2', \ldots, \mathbf{y}_n')'$, $X \equiv [X_1', X_2', \ldots, X_n']'$, $\bar{\boldsymbol{\beta}} \equiv (\bar{\beta}_1, \bar{\beta}_2, \ldots, \bar{\beta}_K)'$, $D_x \equiv \operatorname{diag}[X_1, X_2, \ldots, X_n]$, $\boldsymbol{\xi} \equiv [\boldsymbol{\xi}_1', \boldsymbol{\xi}_2', \ldots, \boldsymbol{\xi}_n']'$, $\boldsymbol{\beta}_i = \bar{\boldsymbol{\beta}} + \boldsymbol{\xi}_i$ and $\mathbf{u} \equiv (\mathbf{u}_1', \mathbf{u}_2', \ldots, \mathbf{u}_n')'$.

For given X the random vector $\mathbf{y}$ is distributed with mean $X\bar{\boldsymbol{\beta}}$ and variance-covariance (V-C) matrix of the form

$$\text{(3)} \qquad \Sigma = \begin{bmatrix} X_1\Delta X_1' + \sigma_{11}\Omega_{11} & \sigma_{12}\Omega_{12} & \cdots & \sigma_{1n}\Omega_{1n} \\ \sigma_{21}\Omega_{21} & X_2\Delta X_2' + \sigma_{22}\Omega_{22} & \cdots & \sigma_{2n}\Omega_{2n} \\ \vdots & \vdots & & \vdots \\ \sigma_{n1}\Omega_{n1} & \sigma_{n2}\Omega_{n2} & \cdots & X_n\Delta X_n' + \sigma_{nn}\Omega_{nn} \end{bmatrix}.$$

[1] With an abuse of notation we use the same symbol to denote a random quantity and its value.
[2] This assumption is partly relaxed in Swamy (1972).

The objective is to estimate the parameter vector $\boldsymbol{\theta} = (\bar{\boldsymbol{\beta}}', \boldsymbol{\omega}')'$ where $\boldsymbol{\omega}$ is a $[n^2 + K^2 + n] \times 1$ vector containing all the elements of $\{\sigma_{ij}\}$, Δ and ρ_i $(i = 1, 2, \ldots, n)$ arranged in any order.

Model (1) contains a sample space Y of elements $\mathbf{y}$. The distribution of $\mathbf{y}$ over Y can be taken as known to belong to a continuously parameterized family of distributions with probability density function (pdf), $p(\mathbf{y}|X, \boldsymbol{\theta})$, the parameter vector $\boldsymbol{\theta}$ ranging over a well-defined parameter space $\Theta = \{\boldsymbol{\theta}: -\infty < \bar{\beta}_k < \infty, 0 < \delta_{kk} < \infty$ for $k = 1, 2, \ldots, K: \delta_{kk'}^2 < \delta_{kk}\delta_{k'k'}, \delta_{kk'} = \delta_{k'k}$ for $k \neq k' = 1, 2, \ldots, K$: $\sigma_{ij}^2 < \sigma_{ii}\sigma_{jj}$ $\sigma_{ij} = \sigma_{ji}$ for $i \neq j = 1, 2, \ldots, n$: $0 < \sigma_{ii} < \infty$, $0 \leq |\rho_i| < 1$ for $i = 1, 2, \ldots, n\}$. We assume that the unknown true value of $\boldsymbol{\theta}$ belongs to Θ.

2.2. *Criteria of Estimation*

Suppose that the seriousness of sampling errors, $\hat{\bar{\boldsymbol{\beta}}} - \bar{\boldsymbol{\beta}}$, is indicated by the loss matrix $(\hat{\bar{\boldsymbol{\beta}}} - \bar{\boldsymbol{\beta}})(\hat{\bar{\boldsymbol{\beta}}} - \bar{\boldsymbol{\beta}})'$ and we wish to find an estimator $\hat{\bar{\boldsymbol{\beta}}}$ for which

$$\mathbf{l}'E(\hat{\bar{\boldsymbol{\beta}}} - \bar{\boldsymbol{\beta}})(\hat{\bar{\boldsymbol{\beta}}} - \bar{\boldsymbol{\beta}})'\mathbf{l} \tag{4}$$

is minimum for every $\bar{\boldsymbol{\beta}} \in \Theta$ and every arbitrary vector $\mathbf{l} \neq \mathbf{0}$.

We assume that the loss matrix which expresses the demerit of the estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is separable in its components $\bar{\boldsymbol{\beta}}$ and $\boldsymbol{\omega}$. We do not specify the loss function involving $\boldsymbol{\omega}$. It is worth noting that in the problem of estimating $\boldsymbol{\omega}$ a quadratic loss function does not seem to be appropriate, see Ferguson (1967, p. 179). For each fixed $\boldsymbol{\theta}$, the expected value of $(\hat{\bar{\boldsymbol{\beta}}} - \bar{\boldsymbol{\beta}})(\hat{\bar{\boldsymbol{\beta}}} - \bar{\boldsymbol{\beta}})'$ relative to the distribution of $\mathbf{y}$ determined by $\boldsymbol{\theta}$ is called the risk matrix or the matrix of second order moments of $\hat{\bar{\boldsymbol{\beta}}}$ around $\bar{\boldsymbol{\beta}}$. $E(\hat{\bar{\beta}}_k - \bar{\beta}_k)^2$ is called the mean square error of $\hat{\bar{\beta}}_k$.

A moment's reflection will reveal that it is not possible to find an estimator $\hat{\bar{\boldsymbol{\beta}}}$ which minimizes (4) for every $\bar{\boldsymbol{\beta}} \in \Theta$ and every $\mathbf{l} \neq \mathbf{0}$, see Silvey (1970, p. 24). For example, if we take $\hat{\bar{\boldsymbol{\beta}}} = \mathbf{a}$ (a vector of constants) for all $\mathbf{y}$, this estimator will have zero risk when $\bar{\boldsymbol{\beta}} = \mathbf{a}$ and thus to have a better estimator in the sense of (4), an estimator $\hat{\bar{\boldsymbol{\beta}}}$ must have zero risk for every $\bar{\boldsymbol{\beta}}$. This is obviously not possible. So we must modify our criterion of estimation.

As is well-known, if we restrict ourselves to a class of linear unbiased estimators of $\bar{\boldsymbol{\beta}}$, we can find an estimator which minimizes the risk in (4) for every $\bar{\boldsymbol{\beta}} \in \Theta$ and every $\mathbf{l} \neq \mathbf{0}$. Such an estimator is the minimum variance linear unbiased (MVLU) estimator

$$\bar{\mathbf{b}}(\boldsymbol{\omega}) = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\mathbf{y}. \tag{5}$$

In the practical situation in which $\boldsymbol{\omega}$ is unknown, an estimate $\hat{\boldsymbol{\omega}}$ of $\boldsymbol{\omega}$ developed by Swamy (1972) can be used in place of the known value used in (5). We can offer an asymptotic justification for this procedure.

It has been emphasized by many statisticians that there is an element of arbitrariness in the criterion of MVLU, particularly with regard to unbiasedness. Consequently, in what follows we modify the criterion of MVLU.

2.3. *Stein-like Estimators*

Following one approach of Zellner and Vandaele (1971), we consider the problem of estimating $\bar{\boldsymbol{\beta}}$ when the loss function is quadratic. Specifically, let the quadratic loss function be $(\hat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}})'Q(\hat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}})$ where Q is a known positive definite matrix. Since the range of each $\bar{\beta}_k$ is bounded, the risk function $E(\hat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}})'Q(\hat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}})$ is bounded, provided $\hat{\boldsymbol{\beta}}$ has finite V-C matrix. Zellner and Vandaele (1971) show that among all estimators of the form $c\bar{\mathbf{b}}(\boldsymbol{\omega})$, where c is a scalar lying between 0 and 1, the estimator

$$c^*\bar{\mathbf{b}}(\boldsymbol{\omega}) \equiv \left[1 - \frac{\operatorname{tr}(X'\Sigma^{-1}X)^{-1}Q}{\operatorname{tr}(X'\Sigma^{-1}X)^{-1}Q + \bar{\boldsymbol{\beta}}'Q\bar{\boldsymbol{\beta}}}\right]\bar{\mathbf{b}}(\boldsymbol{\omega}) \tag{6}$$

has the smallest risk. That is, $E[c\bar{\mathbf{b}}(\boldsymbol{\omega}) - \bar{\boldsymbol{\beta}}]'Q[c\bar{\mathbf{b}}(\boldsymbol{\omega}) - \bar{\boldsymbol{\beta}}]$ takes the smallest value for every $\bar{\boldsymbol{\beta}} \in \Theta$ when $c = c^*$. Also,

$$E[c^*\bar{\mathbf{b}}(\boldsymbol{\omega}) - \bar{\boldsymbol{\beta}}]'Q[c^*\bar{\mathbf{b}}(\boldsymbol{\omega}) - \bar{\boldsymbol{\beta}}] \leq E[\bar{\mathbf{b}}(\boldsymbol{\omega}) - \bar{\boldsymbol{\beta}}]'Q[\bar{\mathbf{b}}(\boldsymbol{\omega}) - \bar{\boldsymbol{\beta}}] \;\forall \bar{\boldsymbol{\beta}} \in \Theta. \tag{7}$$

Since $c^*\bar{\mathbf{b}}(\boldsymbol{\omega})$ involves parameters with unknown values, it cannot be computed. Therefore, as in Zellner and Vandaele (1971) we may approximate $c^*\bar{\mathbf{b}}(\boldsymbol{\omega})$ by

$$\hat{c}^*\bar{\mathbf{b}}(\hat{\boldsymbol{\omega}}) \equiv \left[1 - \frac{\operatorname{tr}(X'\hat{\Sigma}^{-1}X)^{-1}Q}{\bar{\mathbf{b}}'(\hat{\boldsymbol{\omega}})Q\bar{\mathbf{b}}(\hat{\boldsymbol{\omega}})}\right]\bar{\mathbf{b}}(\hat{\boldsymbol{\omega}}) \tag{8}$$

where $\hat{\Sigma}$ and $\bar{\mathbf{b}}(\hat{\boldsymbol{\omega}})$ are as shown in Swamy (1972).

The estimator $\hat{c}^*\bar{\mathbf{b}}(\hat{\boldsymbol{\omega}})$ is in the form of an estimator developed by Stein for the mean vector of a K-dimensional normal population, see Zellner and Vandaele (1971).

Following Mehta and Srinivasan (1971) we may approximate c^* by an exponential function with two adjustable parameters and write

$$\hat{f}(\gamma)\bar{\mathbf{b}}(\hat{\boldsymbol{\omega}}) = [1 - \gamma_1 \exp\{-\gamma_2\bar{\mathbf{b}}'(\hat{\boldsymbol{\omega}})X'\hat{\Sigma}^{-1}X\bar{\mathbf{b}}(\hat{\boldsymbol{\omega}})\}]\bar{\mathbf{b}}(\hat{\boldsymbol{\omega}}) \tag{9}$$

where $0 < \gamma_1 < 1$ and $\gamma_2 > 0$.

Notice that the factor $\hat{c}^*$ multiplying $\bar{\mathbf{b}}(\hat{\boldsymbol{\omega}})$ in (8) can take on negative values with positive probability. Baranchik's analysis of simpler situations (see Stein, 1966) indicates that the estimator in (8) can be improved upon by restricting $\hat{c}^*$ to be nonnegative. The factor $\hat{f}(\gamma)$ multiplying $\bar{\mathbf{b}}(\hat{\boldsymbol{\omega}})$ in (9) can be made positive by suitably choosing the values of γ_1 and γ_2. Experience in simpler situations (Mehta and Srinivasan, 1971) has shown that by judicious choice of γ_1 and γ_2 one can make the risk associated with $\hat{f}(\gamma)\bar{\mathbf{b}}(\hat{\boldsymbol{\omega}})$ smaller than that associated with $\bar{\mathbf{b}}(\hat{\boldsymbol{\omega}})$ or with $\hat{c}^*\bar{\mathbf{b}}(\hat{\boldsymbol{\omega}})$ for a range of values of $\bar{\boldsymbol{\beta}}$ around $\mathbf{0}$. Since the estimators in (8) and (9) provide only approximations to the optimal linear estimator $c^*\bar{\mathbf{b}}(\boldsymbol{\omega})$, neither of them is an estimator which has minimum average risk within the class of linear or nonlinear estimators of $\bar{\boldsymbol{\beta}}$, see Strawderman and Cohen (1971). Consequently, there are other ways of obtaining linear or nonlinear estimators which have smaller risks than $\hat{c}^*\bar{\mathbf{b}}(\hat{\boldsymbol{\omega}})$ and $\hat{f}(\gamma)\bar{\mathbf{b}}(\hat{\boldsymbol{\omega}})$ (see Section 4 below).

The estimator in (8) takes $\bar{\mathbf{b}}(\hat{\boldsymbol{\omega}})$ and pulls it towards a central value $\mathbf{0}$ or past $\mathbf{0}$ if $\bar{\mathbf{b}}'(\hat{\boldsymbol{\omega}})Q\bar{\mathbf{b}}(\hat{\boldsymbol{\omega}}) < \operatorname{tr}(X'\hat{\Sigma}^{-1}X)^{-1}Q$.[3] Since all elements of $\bar{\mathbf{b}}(\hat{\boldsymbol{\omega}})$ are shrunk by the same factor towards 0, the extreme values experience most shift. The estimators in (8) and (9) may do very poorly in estimating those elements of $\bar{\boldsymbol{\beta}}$ with unusually large

[3] If we knew *a priori* that the true values of the elements of $\bar{\boldsymbol{\beta}}$ lay closely to a value other than zero, we could easily modify the formulae in (8) and (9) to shrink the estimated value of $\bar{\boldsymbol{\beta}}$ towards that value, see Zellner and Vandaele (1971), and Mehta and Srinivasan (1971).

or small values. Unless the true values of all the elements of $\bar{\boldsymbol{\beta}}$ lie closely in almost the same interval around 0, the estimators in (8) and (9) may not yield good estimates of all the elements of $\bar{\boldsymbol{\beta}}$. It may happen that for some values of $\bar{\boldsymbol{\beta}}$ the total risk associated with (8) is smaller than that associated with $\bar{\mathbf{b}}(\hat{\boldsymbol{\omega}})$ but the risk associated with an element of (8) is larger than that associated with the corresponding element of $\bar{\mathbf{b}}(\hat{\boldsymbol{\omega}})$. To put it differently, the estimator $\hat{c}^*\bar{\mathbf{b}}(\hat{\boldsymbol{\omega}})$ may have good ensemble properties but not good component properties. This is also true of $\hat{f}(\gamma)\bar{\mathbf{b}}(\hat{\boldsymbol{\omega}})$.

To guard against this bad property of Stein-like estimators, Efron and Morris (1972) develop a "limited translation estimator" which is a compromise between Stein's estimator and the maximum likelihood estimator (MLE). The compromise consists of following the Stein rule as closely as possible subject to a fixed constraint on how far the estimator is allowed to deviate from MLE. This procedure is sensible if the probability that an ML estimator of $\bar{\boldsymbol{\beta}}$ will be far removed from the true value of $\bar{\boldsymbol{\beta}}$ is small. Indeed, this probability is large if $X'\Sigma^{-1}X$ is close to singularity.

The average value of the squared distance from $\bar{\mathbf{b}}(\boldsymbol{\omega})$ to $\bar{\boldsymbol{\beta}}$ is given by

$$(10) \qquad E[\bar{\mathbf{b}}(\boldsymbol{\omega}) - \bar{\boldsymbol{\beta}}]'[\bar{\mathbf{b}}(\boldsymbol{\omega}) - \bar{\boldsymbol{\beta}}] = \operatorname{tr}(X'\Sigma^{-1}X)^{-1} = \sum_{i=1}^{K} \lambda_i^{-1}$$

where λ_i is a latent root of $X'\Sigma^{-1}X$. Consequently, if the set of independent variables is such that reasonable data collection results in an $X'\Sigma^{-1}X$ with one or more latent roots close to 0, then the average distance from $\bar{\mathbf{b}}(\boldsymbol{\omega})$ to $\bar{\boldsymbol{\beta}}$ will be large. In this case the Efron–Morris procedure of pulling an estimate of $\bar{\boldsymbol{\beta}}$ towards $\bar{\mathbf{b}}(\boldsymbol{\omega})$ amounts to pulling an estimate away from $\bar{\boldsymbol{\beta}}$, which is not desirable. If the least squares estimates $\bar{\mathbf{b}}(\boldsymbol{\omega})$ lie far away from the true value of $\bar{\boldsymbol{\beta}}$ as a result of high multicollinearity, then so will be the estimates given by $\hat{f}(\gamma)\bar{\mathbf{b}}(\hat{\boldsymbol{\omega}})$ and $\hat{c}^*\bar{\mathbf{b}}(\hat{\boldsymbol{\omega}})$. Typically, $X'X$ will not be close to a diagonal matrix in applications of economic relevance. In the next section we discuss procedures which are specifically designed to minimize the bad effects of significant departures of $X'X$ from I. In order to guarantee good component properties we say that $\hat{\bar{\boldsymbol{\beta}}}$ is "uniformly" better than $\bar{\boldsymbol{\beta}}^*$ if

$$(11) \qquad \mathbf{l}'E(\hat{\bar{\boldsymbol{\beta}}} - \bar{\boldsymbol{\beta}})(\hat{\bar{\boldsymbol{\beta}}} - \bar{\boldsymbol{\beta}})'\mathbf{l} \leq \mathbf{l}'E(\bar{\boldsymbol{\beta}}^* - \bar{\boldsymbol{\beta}})(\bar{\boldsymbol{\beta}}^* - \bar{\boldsymbol{\beta}})'\mathbf{l}$$

for every $\mathbf{l} \neq \mathbf{0}$ and every $\bar{\boldsymbol{\beta}} \in \Theta$, with strict inequality for some $\bar{\boldsymbol{\beta}}$. In this way we avoid the specification of Q. An estimator, $\bar{\boldsymbol{\beta}}^*$, is "inadmissible" if there exists another estimator of $\bar{\boldsymbol{\beta}}$ which completely dominates $\bar{\boldsymbol{\beta}}^*$ in the sense of (11); otherwise it is "admissible". Notice that $\hat{\bar{\boldsymbol{\beta}}}$ is uniformly better than $\bar{\boldsymbol{\beta}}^*$ in the sense of (11), if and only if, $E(\bar{\boldsymbol{\beta}}^* - \bar{\boldsymbol{\beta}})(\bar{\boldsymbol{\beta}}^* - \bar{\boldsymbol{\beta}})'$ exceeds $E(\hat{\bar{\boldsymbol{\beta}}} - \bar{\boldsymbol{\beta}})(\hat{\bar{\boldsymbol{\beta}}} - \bar{\boldsymbol{\beta}})'$ by a positive semi-definite matrix for every $\boldsymbol{\theta} \in \Theta$.

3. Suggested Procedure of Estimation in Cases of Partial Prior Information

3.1. *Ridge Regression*

For the model in the present paper, let $\bar{\boldsymbol{\beta}}$ be constrained to be in a hypersphere of radius r. Let the estimation criterion be the minimum residual sum of squares $(\mathbf{y} - X\bar{\boldsymbol{\beta}})'\Sigma^{-1}(\mathbf{y} - X\bar{\boldsymbol{\beta}})$ subject to the condition $\bar{\boldsymbol{\beta}}'\bar{\boldsymbol{\beta}} = r^2 < \infty$. The value of $\bar{\boldsymbol{\beta}}$ that minimizes the function

$$(12)\qquad (\mathbf{y} - X\bar{\boldsymbol{\beta}})'\Sigma^{-1}(\mathbf{y} - X\bar{\boldsymbol{\beta}}) + \mu(\bar{\boldsymbol{\beta}}'\bar{\boldsymbol{\beta}} - r^2)$$

is

$$(13)\qquad \bar{\mathbf{b}}_\mu(\boldsymbol{\omega}) = (X'\Sigma^{-1}X + \mu I)^{-1}X'\Sigma^{-1}\mathbf{y}.$$

This is the ridge estimator developed by Hoerl and Kennard (1970a).

Unlike the Stein procedure, the above procedure takes into account the restrictions on the ranges of $\bar{\boldsymbol{\beta}}$. The estimation procedure based on the matrix $(X'\Sigma^{-1}X + \mu I)$ with $\mu > 0$ rather than $X'\Sigma^{-1}X$ can be used to circumvent many of the difficulties associated with the multicollinearity problem, and it can be used to obtain a point estimate of $\bar{\boldsymbol{\beta}}$, which is on the average closer to $\bar{\boldsymbol{\beta}}$ than is $\bar{\mathbf{b}}(\boldsymbol{\omega})$. The average value of the squared distance from $\bar{\mathbf{b}}_\mu(\boldsymbol{\omega})$ to $\bar{\boldsymbol{\beta}}$ is

$$(14)\qquad \begin{aligned} E[\bar{\mathbf{b}}_\mu(\boldsymbol{\omega}) - \bar{\boldsymbol{\beta}}]'[\bar{\mathbf{b}}_\mu(\boldsymbol{\omega}) - \bar{\boldsymbol{\beta}}] &= \operatorname{tr}[I + \mu(X'\Sigma^{-1}X)^{-1}]^{-1}(X'\Sigma^{-1}X)^{-1} \\ &\quad \cdot[\mu(X'\Sigma^{-1}X)^{-1} + I]^{-1} + \mu^2\bar{\boldsymbol{\beta}}'(X'\Sigma^{-1}X + \mu I)^{-2}\bar{\boldsymbol{\beta}} \\ &= \sum_{i=1}^{K} \lambda_i(\lambda_i + \mu)^{-2} + \mu^2\bar{\boldsymbol{\beta}}'(X'\Sigma^{-1}X + \mu I)^{-2}\bar{\boldsymbol{\beta}}. \end{aligned}$$

This can be compared with (10). If a λ_i is close to zero, (14) will be substantially smaller than (10) depending on the value of μ. That is, when $X'\Sigma^{-1}X$ is ill-conditioned, the estimates of $\bar{\boldsymbol{\beta}}$ based on $\bar{\mathbf{b}}(\hat{\boldsymbol{\omega}})$ (but not on $\bar{\mathbf{b}}_\mu(\hat{\boldsymbol{\omega}})$) have a high probability of being far removed from $\bar{\boldsymbol{\beta}}$. Hoerl and Kennard show that there exists a range of values of μ for which the average distance from $\bar{\mathbf{b}}_\mu(\boldsymbol{\omega})$ to $\bar{\boldsymbol{\beta}}$ is smaller than that from $\bar{\mathbf{b}}(\boldsymbol{\omega})$ to $\bar{\boldsymbol{\beta}}$.

The relationship of a ridge estimator to the Aitken estimator $\bar{\mathbf{b}}(\boldsymbol{\omega})$ is given by the alternative form

$$(15)\qquad \bar{\mathbf{b}}_\mu(\boldsymbol{\omega}) = [I + \mu(X'\Sigma^{-1}X)^{-1}]^{-1}\bar{\mathbf{b}}(\boldsymbol{\omega}).$$

We may rewrite (13) as

$$(16)\qquad \bar{\mathbf{b}}_\mu(\boldsymbol{\omega}) = (X'\Sigma^{-1}X + \mu I)^{-1}X'\Sigma^{-1}X\bar{\mathbf{b}}(\boldsymbol{\omega}).$$

The estimator $\bar{\mathbf{b}}_\mu(\boldsymbol{\omega})$ will be recognized as a "matrix weighted average" of the vectors $\bar{\mathbf{b}}(\boldsymbol{\omega})$ and $\mathbf{0}$. Like $c^*\bar{\mathbf{b}}(\boldsymbol{\omega})$, it also shrinks the estimated value of $\bar{\boldsymbol{\beta}}$ a fixed percentage away from $\bar{\mathbf{b}}(\boldsymbol{\omega})$ towards $\mathbf{0}$. But the shrinkage factor is not the same for all the elements of $\bar{\mathbf{b}}(\boldsymbol{\omega})$. Thus, the ridge regression technique, by utilizing the restriction on the range of $\bar{\boldsymbol{\beta}}$, leads to an estimator which does not suffer from the limitations of $c^*\bar{\mathbf{b}}(\boldsymbol{\omega})$. The estimator in (15) is insensitive to multicollinearity. On the other hand, when $X'\Sigma^{-1}X = I$, the matrix factor multiplying $\bar{\mathbf{b}}(\boldsymbol{\omega})$ in (16) reduces to a scalar times identity matrix. In this case, by appropriately defining μ we can equate $\bar{\mathbf{b}}_\mu(\boldsymbol{\omega})$ to $c^*\bar{\mathbf{b}}(\boldsymbol{\omega})$.

The second order moment matrix of $\bar{\mathbf{b}}_\mu(\boldsymbol{\omega})$ around $\bar{\boldsymbol{\beta}}$ is

$$(17)\qquad \begin{aligned} E[\bar{\mathbf{b}}_\mu(\boldsymbol{\omega}) - \bar{\boldsymbol{\beta}}][\bar{\mathbf{b}}_\mu(\boldsymbol{\omega}) - \bar{\boldsymbol{\beta}}]' &= [I + \mu(X'\Sigma^{-1}X)^{-1}]^{-1}(X'\Sigma^{-1}X)^{-1} \\ &\quad \cdot[I + \mu(X'\Sigma^{-1}X)^{-1}]^{-1} + \mu^2(X'\Sigma^{-1}X + \mu I)^{-1}\bar{\boldsymbol{\beta}}\bar{\boldsymbol{\beta}}'(X'\Sigma^{-1}X + \mu I)^{-1}. \end{aligned}$$

The first term on the r.h.s. of (17) is the V–C matrix of $\bar{\mathbf{b}}_\mu(\boldsymbol{\omega})$ and the second term is the matrix of squares and cross products of the biases of the elements of $\bar{\mathbf{b}}_\mu(\boldsymbol{\omega})$.

As is well-known, $(X'\Sigma^{-1}X)^{-1}$ is the V–C matrix of $\bar{\mathbf{b}}(\boldsymbol{\omega})$. The matrix $(X'\Sigma^{-1}X)^{-1} - [I + \mu(X'\Sigma^{-1}X)^{-1}]^{-1}(X'\Sigma^{-1}X)^{-1}[I + \mu(X'\Sigma^{-1}X)^{-1}]^{-1}$ is non-negative definite so that for some values of μ and $\bar{\boldsymbol{\beta}}$ in a neighborhood of $\mathbf{0}$ there is a possibility of $E[\bar{\mathbf{b}}(\boldsymbol{\omega}) - \bar{\boldsymbol{\beta}}][\bar{\mathbf{b}}(\boldsymbol{\omega}) - \bar{\boldsymbol{\beta}}]' - E[\bar{\mathbf{b}}_\mu(\boldsymbol{\omega}) - \bar{\boldsymbol{\beta}}][\mathbf{b}_\mu(\boldsymbol{\omega}) - \boldsymbol{\beta}]'$ being positive semi-definite. However, the mean square error of an element of $\bar{\mathbf{b}}_\mu(\boldsymbol{\omega})$ may not be substantially smaller than that of the corresponding element of $\bar{\mathbf{b}}(\boldsymbol{\omega})$, if the true value of $\bar{\boldsymbol{\beta}}$ is not sufficiently close to $\mathbf{0}$.

An approximate ridge regression estimator is

$$\bar{\mathbf{b}}_\mu(\hat{\boldsymbol{\omega}}) = (X'\hat{\Sigma}^{-1}X + \mu I)^{-1}X'\hat{\Sigma}^{-1}\mathbf{y}. \tag{18}$$

In Hoerl and Kennard (1971b) some recommendations for choosing a $\mu > 0$ are given.

3.2. *Minimum Conditional Mean Square Error Estimator of* $\bar{\boldsymbol{\beta}}$

Recall that the second order moment matrix of a linear estimator $A\mathbf{y} + \mathbf{a}$ around $\bar{\boldsymbol{\beta}}$ is

$$A\Sigma A' + [(AX - I)\bar{\boldsymbol{\beta}} + \mathbf{a}][(AX - I)\bar{\boldsymbol{\beta}} + \mathbf{a}]'. \tag{19}$$

The quantity in (19) cannot be minimized unless it is bounded, see Barnard (1963). Since the range of $\boldsymbol{\theta}$ is bounded, the elements of (19) are bounded. Let $\bar{\boldsymbol{\beta}}^*$ be a guessed value of $\bar{\boldsymbol{\beta}}$. Using $\bar{\boldsymbol{\beta}}^*$ in place of $\bar{\boldsymbol{\beta}}$, we obtain

$$A\Sigma A' + [(AX - I)\bar{\boldsymbol{\beta}}^* + \mathbf{a}][(AX - I)\bar{\boldsymbol{\beta}}^* + \mathbf{a}]'. \tag{20}$$

If (20) is chosen as a criterion of estimation, the optimum choice of $\mathbf{a}$ is $\mathbf{0}$ and that of A is (see Rao, 1971, p. 389)

$$A^* = \bar{\boldsymbol{\beta}}^*\bar{\boldsymbol{\beta}}^{*\prime}X'(X\bar{\boldsymbol{\beta}}^*\bar{\boldsymbol{\beta}}^{*\prime}X' + \Sigma)^{-1}. \tag{21}$$

Consequently, the optimal estimator of $\bar{\boldsymbol{\beta}}$, given $\bar{\boldsymbol{\beta}}^*$, is

$$\bar{\mathbf{b}}^*(\boldsymbol{\omega}) = \bar{\boldsymbol{\beta}}^*\bar{\boldsymbol{\beta}}^{*\prime}X'(X\bar{\boldsymbol{\beta}}^*\bar{\boldsymbol{\beta}}^{*\prime}X' + \Sigma)^{-1}\mathbf{y}. \tag{22}$$

(Henceforth we shall refer to $\bar{\mathbf{b}}^*(\boldsymbol{\omega})$ as the minimum conditional mean square error (MCMSE) estimator of $\bar{\boldsymbol{\beta}}$. The result in (22) is given as an exercise in Theil (1971, p. 125, Problem 4.3).) Notice that the estimator $\bar{\mathbf{b}}^*(\boldsymbol{\omega})$ exists even when the rank of X is less than K. In cases where the rank of X is K, we can write

$$\begin{aligned}(X\bar{\boldsymbol{\beta}}^*\bar{\boldsymbol{\beta}}^{*\prime}X' + \Sigma)^{-1} &= \Sigma^{-1} - \Sigma^{-1}X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1} \\ &\quad + \Sigma^{-1}X(X'\Sigma^{-1}X)^{-1}[\bar{\boldsymbol{\beta}}^*\bar{\boldsymbol{\beta}}^{*\prime} + (X'\Sigma^{-1}X)^{-1}]^{-1} \\ &\quad \times (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1},\end{aligned} \tag{23}$$

see Rao (1965, p. 29, Problem 2.9).

Inserting this back into (22) gives

$$\bar{\mathbf{b}}^*(\boldsymbol{\omega}) = \bar{\boldsymbol{\beta}}^*\bar{\boldsymbol{\beta}}^{*\prime}[\bar{\boldsymbol{\beta}}^*\bar{\boldsymbol{\beta}}^{*\prime} + (X'\Sigma^{-1}X)^{-1}]^{-1}\bar{\mathbf{b}}(\boldsymbol{\omega}). \tag{24}$$

In the practical situation in which Σ is unknown, the estimator $\bar{\mathbf{b}}^*(\boldsymbol{\omega})$ can be approximated by

$$\begin{aligned}\bar{\mathbf{b}}^*(\hat{\boldsymbol{\omega}}) &= \bar{\boldsymbol{\beta}}^*\bar{\boldsymbol{\beta}}^{*\prime}X'(X\bar{\boldsymbol{\beta}}^*\bar{\boldsymbol{\beta}}^{*\prime}X' + \hat{\Sigma})^{-1}\mathbf{y} \\ &= \bar{\boldsymbol{\beta}}^*\bar{\boldsymbol{\beta}}^{*\prime}[\bar{\boldsymbol{\beta}}^*\bar{\boldsymbol{\beta}}^{*\prime} + (X'\hat{\Sigma}^{-1}X)^{-1}]^{-1}\bar{\mathbf{b}}(\hat{\boldsymbol{\omega}}) \qquad [\text{by } (23)]\end{aligned} \tag{25}$$

where $\hat{\Sigma}$ is as defined in (8).

In the Appendix to the paper, it will be shown that a sufficient condition for $(X'\Sigma^{-1}X)^{-1} - E\{[\bar{\mathbf{b}}^*(\omega) - \bar{\boldsymbol{\beta}}][\bar{\mathbf{b}}^*(\omega) - \bar{\boldsymbol{\beta}}]'|\bar{\boldsymbol{\beta}}^*\}$ to be positive definite is

$$\sup_k \bar{\boldsymbol{\beta}}'\mathbf{p}_k\mathbf{p}_k'\bar{\boldsymbol{\beta}} < 1 \tag{26}$$

where $\mathbf{p}_k$ is the kth column of P. P is a nonsingular matrix such that $P'(X'\Sigma^{-1}X)^{-1}P = I$, $P'\bar{\boldsymbol{\beta}}^*\bar{\boldsymbol{\beta}}^{*\prime}P = \lambda_1^*\mathbf{i}_1\mathbf{i}_1'$, and $\mathbf{i}_1$ is the first column of an identity matrix of order K.

It is clear from (A.4) in the Appendix that the conditional variance of an element of $\bar{\mathbf{b}}^*(\omega)$, given $\bar{\boldsymbol{\beta}}^*$, is substantially smaller than the variance of the corresponding element of the Aitken estimator $\bar{\mathbf{b}}(\omega)$ for every $\boldsymbol{\theta}$. But, for some values of $\boldsymbol{\theta}$, due to high magnitude of bias the conditional mean square error of an element of $\bar{\mathbf{b}}^*(\omega)$, given $\bar{\boldsymbol{\beta}}^*$, exceeds the variance of the corresponding element of $\bar{\mathbf{b}}(\omega)$. Condition (26) indicates the values of $\boldsymbol{\theta}$ for which $\bar{\mathbf{b}}^*(\omega)$ based on given $\bar{\boldsymbol{\beta}}^*$ is better than $\bar{\mathbf{b}}(\omega)$. Consequently, the approximate MCMSE estimator $\bar{\mathbf{b}}^*(\hat{\omega})$ cannot completely dominate the approximate MVLU estimator $\bar{\mathbf{b}}(\hat{\omega})$ in the sense of (11). When $K = 1$, condition (26) is satisfied if the square of the coefficient of variation of the MVLU estimator $\bar{\mathbf{b}}(\omega)$ is greater than one. In the general case condition (26) is likely to be satisfied if $X'\Sigma^{-1}X$ is close to singularity. Under these conditions, one can improve upon the MVLU estimator by relaxing the unbiasedness condition as in (20).

We now compare the moment matrices of $\bar{\mathbf{b}}_\mu(\omega)$ and $\bar{\mathbf{b}}^*(\omega)$. It is seen from (A.4) and (A.6) in the Appendix that since the rank of $\bar{\boldsymbol{\beta}}^*\bar{\boldsymbol{\beta}}^{*\prime}$ is unity, the conditional variance of an element of $\bar{\mathbf{b}}^*(\omega)$ is substantially smaller than the variance of the corresponding element of $\bar{\mathbf{b}}_\mu(\omega)$. However, for any reasonable values of μ and $\bar{\boldsymbol{\beta}}^*$, the magnitude of bias of an element of $\bar{\mathbf{b}}^*(\omega)$ is likely to be larger than that of bias of the corresponding element of $\bar{\mathbf{b}}_\mu(\omega)$. For certain values of parameters, $\bar{\mathbf{b}}^*(\hat{\omega})$ is better than $\bar{\mathbf{b}}_\mu(\hat{\omega})$.

Next, we note that, if a prior estimate of $\bar{\boldsymbol{\beta}}$ is not available, we may consider the following estimator:

$$\hat{\bar{\mathbf{b}}}(\hat{\omega}) = \bar{\mathbf{b}}_\mu(\hat{\omega})\bar{\mathbf{b}}'(\hat{\omega})X'[X\bar{\mathbf{b}}(\hat{\omega})\bar{\mathbf{b}}'(\hat{\omega})X' + \hat{\Sigma}]^{-1}\mathbf{y}. \tag{27}$$

When there is near-extreme multicollinearity, a precise estimation of $\bar{\boldsymbol{\beta}}$ is not possible, but a relatively precise estimation of $X\bar{\boldsymbol{\beta}}$ and Σ is possible, see Rao (1965, pp. 184–5) and Theil (1971, pp. 153–4). The estimator $\hat{\bar{\boldsymbol{\beta}}}(\hat{\omega})$ is based on the precise estimates of $\bar{\boldsymbol{\beta}}$, $X\bar{\boldsymbol{\beta}}$ and Σ.

The estimator $\bar{\mathbf{b}}^*(\omega)$ is based on a prior estimate of $\bar{\boldsymbol{\beta}}$, while the estimator $\bar{\mathbf{b}}_\mu(\omega)$ is based on a prior knowledge of the range of $\bar{\boldsymbol{\beta}}'\bar{\boldsymbol{\beta}}$. Since the rank of $\bar{\boldsymbol{\beta}}^*\bar{\boldsymbol{\beta}}^{*\prime}$ is unity, we cannot express $\bar{\mathbf{b}}^*(\omega)$ in the form of a matrix weighted average of the vectors $\bar{\mathbf{b}}(\omega)$ and $\mathbf{0}$. However, when $K = 1$, by appropriately defining μ we can equate $\bar{\mathbf{b}}_\mu(\omega)$ to $\bar{\mathbf{b}}^*(\omega)$, see Theil (1971, p. 126, Problem 4.4).

In summary, we have found that none of the estimators $\bar{\mathbf{b}}(\hat{\omega})$, $\hat{c}^*\bar{\mathbf{b}}(\hat{\omega})$, $\hat{f}(\gamma)\bar{\mathbf{b}}(\hat{\omega})$, $\bar{\mathbf{b}}_\mu(\hat{\omega})$, and $\bar{\mathbf{b}}^*(\hat{\omega})$ is uniformly better than the other in the sense of (11). Consequently, it is not possible to choose among them unless we know "where in the parameter space to look" for the most efficient estimates. When we are faced with an extreme multicollinearity situation, we may use $\bar{\mathbf{b}}^*(\hat{\omega})$ if a reliable prior estimate of $\bar{\boldsymbol{\beta}}$ is available and $\hat{\bar{\mathbf{b}}}(\hat{\omega})$ otherwise.

4. Estimating Parameters with the First Two Moments of a Prior Distribution

There are several situations in which extraneous information on some of the parameters of an equation is available. This information may arise from an analysis of past data and/or from theoretical and practical considerations; that is, from sources other than currently available sample. To incorporate such a prior information the following procedure was suggested by Durbin (1953) and developed further by Theil and Goldberger (1961) and Theil (1963).

4.1. *Mixed Estimation When* $\bar{\boldsymbol{\beta}}$ *is Regarded as Fixed*

Suppose that extraneous information of the following form is available.

$$\mathbf{r} = R\bar{\boldsymbol{\beta}} + \mathbf{v} \quad \text{with} \quad E\mathbf{v} = \mathbf{0} \quad \text{and} \quad E\mathbf{v}\mathbf{v}' = \tau^2\psi, \tag{28}$$

where $\mathbf{r}$ is a $q \times 1$ vector of prior estimates of $R\bar{\boldsymbol{\beta}}$. R is a $q \times K$ matrix of known constants, $\mathbf{v}$ is a $q \times 1$ vector of errors in $\mathbf{r}$ and $q \leq K$. We assume that $\mathbf{v}$ is uncorrelated with $\mathbf{u}$ and $\boldsymbol{\xi}$ in (2). We now combine equations (2) and (28) and apply the Aitken theorem to obtain the following estimator for $\bar{\boldsymbol{\beta}}$.

$$\hat{\bar{\boldsymbol{\beta}}}_p(\boldsymbol{\omega}) = \left(X'\Sigma^{-1}X + \frac{1}{\tau^2}R'\psi^{-1}R\right)^{-1}\left(X'\Sigma^{-1}\mathbf{y} + \frac{1}{\tau^2}R'\psi^{-1}\mathbf{r}\right). \tag{29}$$

The estimator $\hat{\bar{\boldsymbol{\beta}}}_p(\boldsymbol{\omega})$ is the MVLU estimator of $\bar{\boldsymbol{\beta}}$ where linear now means linear in $\mathbf{y}$ and $\mathbf{r}$. Here the distinction between $\bar{\mathbf{b}}(\boldsymbol{\omega})$ as a MVLU estimator of $\bar{\boldsymbol{\beta}}$ and $\hat{\bar{\boldsymbol{\beta}}}_p(\boldsymbol{\omega})$ as a MVLU estimator of the same $\bar{\boldsymbol{\beta}}$ is to be clearly understood. The linear function of $\mathbf{y}$, namely $\bar{\mathbf{b}}(\boldsymbol{\omega})$, is the MVLU estimator of $\bar{\boldsymbol{\beta}}$ in the sense that any other estimator of $\bar{\boldsymbol{\beta}}$ which is also linear in the vector $\mathbf{y}$ and unbiased has a V-C matrix which exceeds that of $\bar{\mathbf{b}}(\boldsymbol{\omega})$ by a positive semidefinite matrix. On the other hand, $\hat{\bar{\boldsymbol{\beta}}}_p(\boldsymbol{\omega})$ is the MVLU estimator of $\bar{\boldsymbol{\beta}}$ in the sense that any other estimator of $\bar{\boldsymbol{\beta}}$ which is linear in $\mathbf{y}$ and $\mathbf{r}$ and unbiased has a V-C matrix which exceeds that of $\hat{\bar{\boldsymbol{\beta}}}_p(\boldsymbol{\omega})$ by a positive semidefinite matrix. We shall refer to $\hat{\bar{\boldsymbol{\beta}}}_p(\boldsymbol{\omega})$ as the "mixed regression" estimator. We again remind the reader that the criterion of MVLU is defective in its premises, in that the condition of unbiasedness sometimes leads to inadmissible estimates, see Ferguson (1967, pp. 135–6).

As $\tau^2 \to 0$, the estimator $\hat{\bar{\boldsymbol{\beta}}}_p(\boldsymbol{\omega})$ approaches the restricted estimator of $\bar{\boldsymbol{\beta}}$ given by the normal equations (see Chipman, 1964, p. 1101)

$$\begin{bmatrix} X'\Sigma^{-1}X & R' \\ R & 0 \end{bmatrix}\begin{bmatrix} \bar{\boldsymbol{\beta}} \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} X'\Sigma^{-1}\mathbf{y} \\ \mathbf{r} \end{bmatrix}. \tag{30}$$

Eq. (30) is obtained by minimizing

$$\tfrac{1}{2}(\mathbf{y} - X\bar{\boldsymbol{\beta}})'\Sigma^{-1}(\mathbf{y} - X\bar{\boldsymbol{\beta}}) - \boldsymbol{\mu}'(\mathbf{r} - R\bar{\boldsymbol{\beta}}) \tag{31}$$

where $\boldsymbol{\mu}$ is a vector of Lagrangian multipliers. Theil and Goldberger (1961), solve eq. (30) under the assumption that the ranks of X and R are K and q respectively, while Rao and Mitra (1971, p. 147) solve the same equation without any restrictions on the ranks of X and R.

Chipman (1964, pp. 1101–2) points out an important special case of (29). If ψ is known, eq. (28) can be written as

$$\psi^{-1/2}\mathbf{r} = \psi^{-1/2}R\boldsymbol{\beta} + \psi^{-1/2}\mathbf{v} \tag{32}$$

$$\mathbf{r}^* = R^*\boldsymbol{\beta} + \mathbf{v}^*.$$

When the rank of X is less than K, and when $X^{*\prime} = X'\Sigma^{-1/2}$ and R^* are "complementary",[4] $X^{*-} = (X'\Sigma^{-1}X + (1/\tau^2)R'\psi^{-1}R)^{-1}X^{*\prime}$ is a generalized inverse of X^*, independently of $1/\tau^2$, as long as $0 < 1/\tau^2 < \infty$, because for all such $1/\tau^2$, $1/\tau R^*$ has the same row space as R^*. Similarly, $R^{*-} = (\tau^2 X'\Sigma^{-1}X + R'\psi^{-1}R)^{-1}R^{*\prime}$ is a generalized inverse of R^*, independently of $1/\tau^2$, as long as $0 < \tau^2 < \infty$, since for all such τ^2, τX^* has the same row space as X^*. Therefore the estimator $\hat{\bar{\boldsymbol{\beta}}}_p(\boldsymbol{\omega})$ is functionally independent of $1/\tau^2$ as long as $0 < 1/\tau^2 < \infty$ and R^* is complementary to X^*. In this case the estimator $\hat{\bar{\boldsymbol{\beta}}}_p(\boldsymbol{\omega})$ can be computed even when τ^2 is unknown.

To consider another case, let $q = K$ and $R = I$. Then $\hat{\bar{\boldsymbol{\beta}}}_p(\boldsymbol{\omega})$ becomes

$$\hat{\bar{\boldsymbol{\beta}}}_p(\boldsymbol{\omega}) = \left(X'\Sigma^{-1}X + \frac{1}{\tau^2}\psi^{-1}\right)^{-1}\left(X'\Sigma^{-1}\mathbf{y} + \frac{1}{\tau^2}\psi^{-1}\mathbf{r}\right). \tag{33}$$

It is easily seen that $\hat{\bar{\boldsymbol{\beta}}}_p(\boldsymbol{\omega})$ in (33) is a "matrix weighted average" of $\bar{\mathbf{b}}(\boldsymbol{\omega})$ and $\mathbf{r}$, with weights inversely proportional to their respective V–C matrices. Hence, an estimate of $\bar{\boldsymbol{\beta}}$ is pulled towards $\mathbf{r}$ away from $\bar{\mathbf{b}}(\boldsymbol{\omega})$. The estimator in (33) covers $\bar{\mathbf{b}}_\mu(\boldsymbol{\omega})$ in (13) as a special case. When $\mathbf{r} = \mathbf{0}$ and $\tau^2\psi = (1/\mu)I$, $\bar{\mathbf{b}}_\mu(\boldsymbol{\omega})$ is the same as (33).

Analysis of simpler situations has shown that the estimator

$$\hat{\bar{\boldsymbol{\beta}}}_p(\hat{\boldsymbol{\omega}}) = \left(X'\hat{\Sigma}^{-1}X + \frac{1}{\tau^2}\psi^{-1}\right)^{-1}\left(X'\hat{\Sigma}^{-1}\mathbf{y} + \frac{1}{\tau^2}\psi^{-1}\mathbf{r}\right), \tag{34}$$

with known $\tau^2\psi$, completely dominates $\bar{\mathbf{b}}(\hat{\boldsymbol{\omega}})$ in the sense of (11), provided $E\mathbf{v} = \mathbf{0}$ and $\boldsymbol{\xi}$ and $\mathbf{u}$ are normal, see Swamy and Mehta (1969), and Mehta and Swamy (1972b). In cases where $E(\mathbf{v} - \boldsymbol{\eta})(\mathbf{v} - \boldsymbol{\eta})' = \tau^2\psi$, $\boldsymbol{\eta}$ is unknown, $\tau^2\psi$ is known, and $\boldsymbol{\xi}$ and $\mathbf{u}$ are normal, $\hat{\bar{\boldsymbol{\beta}}}_p(\hat{\boldsymbol{\omega}})$ is better than $\bar{\mathbf{b}}(\hat{\boldsymbol{\omega}})$ if only the coefficient of variation of each element of $\mathbf{v}$ is sufficiently large in magnitude, see Swamy and Mehta (1972). Thus, if we misspecify the prior moments, there is no guarantee that each diagonal element of the second order moment matrix of $\hat{\bar{\boldsymbol{\beta}}}_p(\hat{\boldsymbol{\omega}})$ around $\bar{\boldsymbol{\beta}}$ will be less than or equal to the corresponding diagonal element of the second order moment matrix of $\bar{\mathbf{b}}(\hat{\boldsymbol{\omega}})$ around $\bar{\boldsymbol{\beta}}$.

The compatibility test statistic developed by Theil (1963) can be utilized to test whether prior information is in conflict with sample information. Mehta and Swamy (1972a) have derived the exact finite sample distribution of Theil's compatibility test statistic. They have also considered the consequences for estimation, in terms of mean square error, of making preliminary tests. The efficiency of preliminary testing procedures has been examined by comparison of the risk functions of preliminary test estimators with that of pure regression estimator, $\bar{\mathbf{b}}(\boldsymbol{\omega})$, which is an Aitken estimator when no prior information is used. The preliminary test estimator dominated the pure regression estimator over certain regions of the parameter space.

[4] The matrices X^* and R^* are complementary if (1) rank (X^*) + rank $(R^*) = K$, (2) X^* and R^* have the same number of columns, and (3) the row spaces of X^* and R^* have only the origin in common.

Returning again to the case where $E\mathbf{v} = \mathbf{0}$ and $E\mathbf{v}\mathbf{v}' = \tau^2\psi$, it can be seen that the matrix

$$\left(X'\Sigma^{-1}X + \frac{1}{\tau^2}\psi^{-1}\right)^{-1} - \boldsymbol{\beta}^*\boldsymbol{\beta}^{*\prime}[\boldsymbol{\beta}^*\boldsymbol{\beta}^{*\prime} + (X'\Sigma^{-1}X)^{-1}]^{-1}(X'\Sigma^{-1}X)^{-1}$$
$$\times [\bar{\boldsymbol{\beta}}^*\bar{\boldsymbol{\beta}}^{*\prime} + (X'\Sigma^{-1}X)^{-1}]^{-1}\bar{\boldsymbol{\beta}}^*\bar{\boldsymbol{\beta}}^{*\prime} - (X'\Sigma^{-1}X)^{-1}[\bar{\boldsymbol{\beta}}^*\bar{\boldsymbol{\beta}}^{*\prime}$$
$$+ (X'\Sigma^{-1}X)^{-1}]^{-1}\bar{\boldsymbol{\beta}}\bar{\boldsymbol{\beta}}'[\bar{\boldsymbol{\beta}}^*\bar{\boldsymbol{\beta}}^{*\prime} + (X'\Sigma^{-1}X)^{-1}]^{-1}(X'\Sigma^{-1}X)^{-1}$$

is positive definite only for certain values of $\boldsymbol{\theta}$, $\bar{\boldsymbol{\beta}}^*$ and $\tau^2\psi$. Consequently, the estimator $\bar{\mathbf{b}}^*(\hat{\boldsymbol{\omega}})$ in (25) will not be uniformly better than $\hat{\bar{\boldsymbol{\beta}}}_p(\hat{\boldsymbol{\omega}})$ in (34) even when the first two moments of $\mathbf{r}$ are exactly known.

A particular case which can be solved exactly, and for which there is a complete and simpler treatment is the following. Let $K = 1$, and $\bar{\beta}^{*2} = |r^2 - \tau^2\psi|$. Notice that $Er^2 = \bar{\beta}^2 + \tau^2\psi$. We can use standard analytical and numerical methods (Mehta and Swamy, 1972a) to evaluate the unconditional mean square error of $\bar{b}^*(\boldsymbol{\omega})$ with respect to the distributions of $\bar{\beta}^{*2}$ and $\mathbf{y}$. If the square of the coefficient of variation of r, $(\tau^2\psi/\bar{\beta}^2)$, is greater than one and the square of the coefficient of variation of the MVLU estimator $\bar{b}(\boldsymbol{\omega})$ is greater than or equal to one, then $\bar{b}^*(\boldsymbol{\omega})$ is better than $\hat{\bar{\beta}}_p(\boldsymbol{\omega})$.

Formulae (25) and (34) provide two different ways of combining prior information with sample information. Neither one of them is better than the other regardless of the true values of parameters. It should be emphasized that the estimator $\bar{\mathbf{b}}^*(\hat{\boldsymbol{\omega}})$ should not be used unless $\bar{\boldsymbol{\beta}}^*\bar{\boldsymbol{\beta}}^{*\prime}$ is a reliable estimate of $\bar{\boldsymbol{\beta}}\bar{\boldsymbol{\beta}}'$. If the prior point estimates of the elements of $\bar{\boldsymbol{\beta}}$ are not reliable, then it is better to express the uncertainties associated with these estimates in the form of a distribution with mean $\bar{\boldsymbol{\beta}}$ and V–C matrix $\tau^2\psi$ and use the estimator $\hat{\bar{\boldsymbol{\beta}}}_p(\hat{\boldsymbol{\omega}})$. That the prior information be unbiased is a severe restriction on the nature of such information, see Zellner (1970, p. 189). This restriction will be eliminated in the next subsection.

4.2. *Bayesian Estimation When $\bar{\boldsymbol{\beta}}$ is Regarded As a Random Variable*

We now make the following "wide-sense" assumption.

Assumption 2: A probability distribution on a class of measurable sets in Θ exists. The variable $\bar{\boldsymbol{\beta}}$ is judged *a priori* to be distributed independently of $\boldsymbol{\omega}$ whose distribution is a point distribution with the whole mass of the distribution concentrated at one point. Furthermore, $E\bar{\boldsymbol{\beta}} = \mathbf{r}$ and $E(\bar{\boldsymbol{\beta}} - \mathbf{r})(\bar{\boldsymbol{\beta}} - \mathbf{r})' = \tau^2\psi$ which is positive definite.

Even if a purely pragmatic attitude is adopted it does seem to be true that for at least some inference problems, an approach which assumes the existence of a prior distribution of $\boldsymbol{\theta}$ is more appropriate than one which does not. However, it is very restrictive to assume that the distribution of $\boldsymbol{\omega}$ is a point distribution. If this assumption is relaxed, the analysis gets very complicated, see Lindley and Smith (1972).

Assuming that $\boldsymbol{\theta}$ is a random variable, Zellner and Vandaele (1971) discuss the Bayesian interpretations (attributable to Lindley and others) of the Stein-like estimator $c^*\bar{\mathbf{b}}(\boldsymbol{\omega})$. When $X'X = I$, $Q = I$, $\Sigma = \sigma^2 I$, and the prior distribution of $\bar{\boldsymbol{\beta}}$ has mean $\mathbf{0}$ and scalar V–C matrix, one can generate a Bayes estimator of the form

$c^*\bar{\mathbf{b}}(\omega)$. Notice that when $\mathbf{r}$ is regarded as a fixed parameter, $\hat{\bar{\boldsymbol{\beta}}}_p(\omega)$ is still a linear function of $\mathbf{y}$ but becomes a biased estimator of $\bar{\boldsymbol{\beta}}$. It is interesting to note that if $\bar{\boldsymbol{\beta}}$ is considered to be a random variable with mean equal to fixed $\mathbf{r}$ and fixed V–C matrix $\tau^2\psi$, then $\hat{\bar{\boldsymbol{\beta}}}_p(\omega)$ in (33) is the "best linear" predictor of $\bar{\boldsymbol{\beta}}$ in the sense that any other predictor of $\bar{\boldsymbol{\beta}}$ which is also linear in the vector $\mathbf{y}$ has an averaged second order moment matrix around $\bar{\boldsymbol{\beta}}$ which exceeds that of $\hat{\bar{\boldsymbol{\beta}}}_p(\omega)$ by a positive semidefinite matrix. In other words, if $\mathbf{r}$ and $\tau^2\psi$ are the mean and V–C matrix of $\bar{\boldsymbol{\beta}}$, then $\hat{\bar{\boldsymbol{\beta}}}_p(\omega)$ completely dominates every other linear (in $\mathbf{y}$) estimator (predictor) of $\bar{\boldsymbol{\beta}}$ in the sense of (11).[5] Proof of this important result is given in Chipman (1964, p. 1105) and Rao (1965, p. 192). If $\mathbf{r} \neq \mathbf{0}$, $\psi \neq I$ and Assumption 2 is true, the formulae $c^*\bar{\mathbf{b}}(\omega)$ and $\bar{\mathbf{b}}_\mu(\omega)$ are inappropriate. When $\bar{\boldsymbol{\beta}}$ is random, the procedure outlined in subsection 3.2 is also inappropriate because, under Assumption 2, (19) is not the second order moment matrix of $A\mathbf{y} + \mathbf{a}$ around $\bar{\boldsymbol{\beta}}$, see Chipman (1964, p. 1104). Notice that the estimators $\bar{\mathbf{b}}(\omega)$, $c^*\bar{\mathbf{b}}(\omega)$, $\bar{\mathbf{b}}_\mu(\omega)$ and $\bar{\mathbf{b}}^*(\omega)$ for given $\bar{\boldsymbol{\beta}}^*$, are all linear functions of $\mathbf{y}$. Hence, it follows from the Chipman–Rao theorem that they are inferior to the best linear estimator $\hat{\bar{\boldsymbol{\beta}}}_p(\omega)$ if Assumption 2 is true. Thus, the biased estimators generated through the Chipman–Rao procedure are better than those generated through the procedure outlined in subsection 3.2.

We called $\hat{\bar{\boldsymbol{\beta}}}_p(\omega)$ the best linear estimator of $\bar{\boldsymbol{\beta}}$. The qualification linear can be dropped if the prior distribution of $\bar{\boldsymbol{\beta}}$, given $\mathbf{r}$ and $\tau^2\psi$, is normal and the conditional distribution of $\mathbf{y}$, given X, Σ, and $\bar{\boldsymbol{\beta}}$, is also normal. This is because, under these normality assumptions, the estimator $\hat{\bar{\boldsymbol{\beta}}}_p(\omega)$ is the mean of the conditional posterior distribution of $\bar{\boldsymbol{\beta}}$, given Σ, $\mathbf{r}$, $\tau^2\psi$ and the data, see Zellner (1971, p. 76), and Zellner and Vandaele (1971). The posterior mean $\hat{\bar{\boldsymbol{\beta}}}_p(\omega)$ with known Σ, $\tau^2\psi$ and $\mathbf{r}$ is admissible with respect to a quadratic loss function, see Zellner (1971, p. 24). Thus, admissible estimates can be found if the prior distribution of $\boldsymbol{\theta}$ is completely known, see Ferguson (1967).

Even though the result in (33) is intuitively appealing, it has certain weaknesses. In (13) and (33) different posterior means have been obtained by combining two different priors with the same likelihood of parameters. These priors were therefore influential in deciding the posterior means in small samples. It is worth noting that if the Aitken estimate $\bar{\mathbf{b}}(\omega)$ and the prior mean $\mathbf{r}$ are very different, then the estimate (33) is a long way from $\bar{\mathbf{b}}(\omega)$. In this case it may happen that either the model specification is at fault or the prior information is incompatible with sample information, see Box and Jenkins (1970, p. 251). Efron and Morris (1971) also point out that the estimator $\hat{\bar{\boldsymbol{\beta}}}_p(\omega)$ must give bad estimates when $\mathbf{r}$ is far from $\bar{\boldsymbol{\beta}}$. Let $N_K(\mathbf{r}, \tau^2\psi)$ represent the true prior distribution of $\bar{\boldsymbol{\beta}}$.[6] Suppose that this distribution is actually a mixture of various other distributions, one of which is $N_K(\mathbf{r}_1, \tau_1^2\psi_1)$ such that $\tau^2\psi - \tau_1^2\psi_1$ is positive definite. For any fixed value of $\tau_1^2\psi_1$, the expected squared error risk of an element of $\hat{\bar{\boldsymbol{\beta}}}_p(\omega)$ with respect to the prior distribution $N_K(\mathbf{r}_1, \tau_1^2\psi_1)$ can be made arbitrarily large by moving $\mathbf{r}_1$ arbitrarily far from $\mathbf{r}$. That is, the estimator $\hat{\bar{\boldsymbol{\beta}}}_p(\omega)$ does well on the population, $N_K(\mathbf{r}, \tau^2\psi)$, as a whole, but may perform very poorly on a particular subpopulation, $N_K(\mathbf{r}_1, \tau_1^2\psi_1)$. The estimator $(X'\Sigma^{-1}X + (1/\tau_1^2)\psi_1^{-1})^{-1}(X'\Sigma^{-1}\mathbf{y} + (1/\tau_1^2)\psi_1^{-1}\mathbf{r}_1)$

[5] The requirement that an estimator of $\bar{\boldsymbol{\beta}}$ be linear arises from the absence, in our "distribution-free" formulation, of the assumption about the form of the prior distribution of $\bar{\boldsymbol{\beta}}$.

[6] $N_K(\mathbf{r}, \tau^2\psi)$ represents K-dimensional normal with mean $\mathbf{r}$ and V–C matrix $\tau^2\psi$.

does well on the subpopulation $N_K(\mathbf{r}_1, \tau_1^2\psi_1)$. If we knew that a particular $\bar{\boldsymbol{\beta}}$ belonged to the subpopulation $N_K(\mathbf{r}_1, \tau_1^2\psi_1)$, then we could use the estimator $(X'\Sigma^{-1}X + (1/\tau_1^2)\psi_1^{-1})^{-1}(X'\Sigma^{-1}\mathbf{y} + (1/\tau_1^2)\psi_1^{-1}\mathbf{r}_1)$ rather than $\hat{\bar{\boldsymbol{\beta}}}_p(\boldsymbol{\omega})$. Information on subpopulation distributions can be obtained by assessing r and $\tau^2\psi$ as precisely as possible. Now the relevant question is: How can we assess a prior distribution in practice?

Notice that the probability distribution on a class of measurable sets in Θ is viewed merely as a reflection of the belief of the statistician about where the true value of $\boldsymbol{\theta}$ lies prior to an observation being made. Conditions under which such a distribution exists are given in Ferguson (1967, Section 1.4). It has been shown by Savage and others that personal probabilities assessed in accordance with certain plausible behavioral postulates of "coherence" must conform mathematically to a probability measure, see Lindley (1971). Winkler (1967a,b: 1971) discusses the practical problem of the assessment of personal probabilities. An operational way of assessing a probability is through the study of relevant gambles. Methods such as scoring rules and bets are useful in leading individuals to make careful probability assessments.

It should be emphasized, however, that in many economic situations there remains the practical difficulty of assessing a prior distribution to reflect one's degree of belief. If the parameter space contains a finite number of points, then by sufficient introspection one can arrive at the prior odds at which one would just accept a bet on this parameter value rather than that, and so eventually find the prior distribution appropriate for a particular problem. If Θ is continuous, as it usually is, it is not clear whether any reasonable consideration of the way in which inferences cohere leads to the existence of the prior distribution, see Lindley (1971, pp. 7–8). The difficulty of choosing a prior distribution is highlighted, when the parameter space is infinite-dimensional as in Sims (1971). Efron and Morris (1971, p. 808) argue that in the realistic situations there is seldom any one prior distribution that is "true" in an absolute sense. There are only more or less relevant priors. If a distribution with mean $\mathbf{r}$ and V–C matrix $\tau^2\psi$ is at all in doubt, it would be well to modify the estimator $\hat{\bar{\boldsymbol{\beta}}}_p(\boldsymbol{\omega})$.

In large samples the situation improves. With a reasonably informative experiment, the values $\mathbf{r}$ and $\tau^2\psi$ adequate for describing rather imprecise knowledge can be changed quite considerably without affecting the final result all that much. This is the consequence of the fact that, under general conditions, sample information dominates prior information in fairly large samples. In fact, Lindley (1971, p. 62) has shown that if the pdf $p(\mathbf{y}|X, \boldsymbol{\theta})$ satisfies certain regularity conditions (see Silvey, 1961 and Perlman, 1972), the method of maximum likelihood is shown to be a reasonably "coherent" technique in large samples. We, therefore, turn to a study of this topic.

5. Maximum Likelihood Method

In this section we assume the following:

Assumption 3: Given X, $\bar{\boldsymbol{\beta}}$, and Σ, $\mathbf{y}$ is normally distributed with mean $X\bar{\boldsymbol{\beta}}$ and V–C matrix Σ, i.e., $\mathbf{y} \sim N_{nT}(X\bar{\boldsymbol{\beta}}, \Sigma)$.

For simplicity, we let $\sigma_{ij} = 0$ if $i \neq j$ and $\rho_i = 0$ for every i. Now $\boldsymbol{\theta} = (\bar{\boldsymbol{\beta}}', \boldsymbol{\omega}')'$ where $\boldsymbol{\omega}$ is a $(n + K^2)X1$ vector. $\boldsymbol{\omega}$ denotes the vector presentation of the σ_{ii}'s and all elements of Δ in which $\sigma_{11}, \ldots, \sigma_{nn}$ appear in order first, then the elements of the first column of Δ, the elements of the second column and so on.

The pdf of $\mathbf{y}$, given X, is

$$(35)\quad p(\mathbf{y}|X, \boldsymbol{\theta}) = (2\pi)^{-nT/2} \prod_{i=1}^{n} \{\sigma_{ii}^{-\frac{1}{2}(T-K)} |X_i'X_i|^{-1/2} |\Delta + \sigma_{ii}(X_i'X_i)^{-1}|^{-1/2}\}$$

$$\cdot \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left[\frac{(T-K)s_{ii}}{\sigma_{ii}} + (\mathbf{b}_i - \bar{\boldsymbol{\beta}})'\right.\right.$$

$$\left.\left.\cdot [\Delta + \sigma_{ii}(X_i'X_i)^{-1}]^{-1}(\mathbf{b}_i - \bar{\boldsymbol{\beta}})\right]\right\}$$

where

$$s_{ii} = \mathbf{y}_i'M_i\mathbf{y}_i/(T - K), \qquad M_i = I - X_i(X_i'X_i)^{-1}X_i'$$

and

$$\mathbf{b}_i = (X_i'X_i)^{-1}X_i'\mathbf{y}_i,$$

see Swamy (1971, pp. 111–12).

Now, given the data $\mathbf{y}$, X, $p(\mathbf{y}|X, \boldsymbol{\theta})$ in (35) may be regarded as a function of $\boldsymbol{\theta}$. When so regarded, it is called the likelihood function of $\boldsymbol{\theta}$ for given $\mathbf{y}$ and X. The likelihood function is defined up to a multiplicative constant. The likelihood expresses the relative plausibilities of different parameter values after we have observed the data $\mathbf{y}$ and X, see Barnard (1967). Methods of eliminating nuisance parameters from the likelihood function so that inferences can be made about the parameters of interest are considered by Kalbfleisch and Sprott (1970). In this regard "marginal" and "conditional" likelihoods are introduced. These can be computed if only the likelihood function factors into two parts, one of which contains a parameter of interest, say $\bar{\beta}_k$, only and the other being uninformative about $\bar{\beta}_k$ in the absence of knowledge of other parameters. It is clear from (35) that the likelihood function has the form (apart from irrelevant constants)

$$(36)\quad l(\boldsymbol{\theta}|\mathbf{y}, X) \propto \left[\prod_{i=1}^{n} \sigma_{ii}^{-\frac{1}{2}(T-K)}\right] \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{(T-K)s_{ii}}{\sigma_{ii}}\right\}$$

$$\cdot \left[\prod_{i=1}^{n} |\Delta + \sigma_{ii}(X_i'X_i)^{-1}|^{-1/2}\right] \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} (\mathbf{b}_i - \bar{\boldsymbol{\beta}})'[\Delta\right.$$

$$\left. + \sigma_{ii}(X_i'X_i)^{-1}]^{-1}(\mathbf{b}_i - \bar{\boldsymbol{\beta}})\right\}.$$

Each of the first n factors on the right hand side of (36) contains one of the σ_{ii} only. It contains no available information concerning $\bar{\boldsymbol{\beta}}$ and Δ in the absence of knowledge of the σ_{ii}. Unfortunately, the last factor contains available information about every element of $\boldsymbol{\theta}$, see Kalbfleisch and Sprott (1970, p. 200). However, as $T \to \infty$ since $(X_i'X_i)^{-1} \to 0$, the last factor gives less and less information about the σ_{ii}'s. $\bar{\boldsymbol{\beta}}$ and Δ are the parameters of our interest and we cannot derive their marginal likelihoods from (36). It is meaningless to integrate $l(\boldsymbol{\theta}|\mathbf{y}, X)$ in an attempt to obtain the marginal likelihoods of the elements of $\bar{\boldsymbol{\beta}}$, see Box and Tiao (1973, p. 73). However, a close study of the likelihood function is always desirable. In certain instances, the data will contain no information regarding certain parameters. It is important to study the likelihood function's properties to determine when this is the case, see, for example, Box and Jenkins (1970, pp. 225–6), Silvey (1970, pp. 81–2), Swamy and Mehta (1971), and Swamy and Rao (1971). A general method for obtaining a reasonable estimate of $\boldsymbol{\theta}$ in most situations is the well-known maximum likelihood method, see Rao (1962). In this section we try to verify the conditions which ensure the consistency and asymptotic normality of an ML estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. First, we indicate a method of obtaining $\hat{\boldsymbol{\theta}}$.

An ML estimate $\hat{\boldsymbol{\theta}}$ is any element of Θ such that $p(\mathbf{y}|X, \hat{\boldsymbol{\theta}}) = \sup_{\boldsymbol{\theta} \in \Theta} p(\mathbf{y}|X, \boldsymbol{\theta})$. $\hat{\boldsymbol{\theta}}$ belongs to the set which is most plausible after we have observed $\mathbf{y}$ and X. At this point it should be appreciated that the ML method always estimates the entire underlying distribution from given data. Successful estimation of the entire underlying distribution is the maximum of objectives attainable by any statistical method. Since Θ is an open set, it may happen that no ML estimate of $\boldsymbol{\theta}$ exists. However, a neighborhood ML estimate of $\boldsymbol{\theta}$, which is defined by Kiefer and Wolfowitz (1956, p. 892), exists in some cases where an ML estimate does not. Usually, ML estimates emerge as a solution of the likelihood equations $\partial \log l(\boldsymbol{\theta}|\mathbf{y}, X)/\partial\boldsymbol{\theta} = \mathbf{0}$ shown in Swamy (1971, p. 112). These equations are nonlinear in the unknowns and have to be solved numerically. A convenient method of solving the likelihood equations is the method of scoring described in Rao (1965, p. 302), see also Silvey (1970, 70–1). This method requires an explicit derivation of information matrix which is given by (see Swamy, 1971, p. 114)

$$(36) \qquad I(\boldsymbol{\theta}) = \begin{bmatrix} -\dfrac{E\partial^2 \log l}{\partial\bar{\boldsymbol{\beta}}\partial\bar{\boldsymbol{\beta}}'} & -\dfrac{E\partial^2 \log l}{\partial\bar{\boldsymbol{\beta}}\partial\boldsymbol{\omega}'} \\ & \dfrac{E\partial^2 \log l}{\partial\bar{\boldsymbol{\beta}}\partial\bar{\boldsymbol{\beta}}'} \end{bmatrix}$$

where

$$-\frac{E\partial^2 \log l}{\partial\bar{\boldsymbol{\beta}}\partial\bar{\boldsymbol{\beta}}'} = \sum_{i=1}^{n} [\Delta + \sigma_{ii}(X_i'X_i)^{-1}]^{-1}, \qquad -\frac{E\partial^2 \log l}{\partial\bar{\boldsymbol{\beta}}\partial\boldsymbol{\omega}'} = 0,$$

$$-\frac{E\partial^2 \log l}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}'} = \begin{bmatrix} \left\{-\dfrac{E\partial^2 \log l}{\partial\sigma_{ii}\partial\sigma_{jj}}\right\} & \left\{-\dfrac{E\partial^2 \log l}{\partial\sigma_{ii}\partial\Delta_r}\right\} \\ & \left\{-\dfrac{E\partial^2 \log l}{\partial\Delta_r\partial\Delta_r}\right\} \end{bmatrix}.$$

Δ_r denotes the vector presentation of all elements of Δ in which the elements of the first row appear in order first, then the elements of the second row and so on:

$$-\frac{E\partial^2 \log 1}{\partial\sigma_{ii}\partial\sigma_{ii}} = \frac{1}{2}\frac{(T-K)}{\sigma_{ii}^2} + \tfrac{1}{2}\operatorname{tr}[\Delta + \sigma_{ii}(X_i'X_i)^{-1}]^{-1}(X_i'X_i)^{-1}$$

$$\cdot[\Delta + \sigma_{ii}(X_i'X_i)^{-1}]^{-1}(X_i'X_i)^{-1} \quad (i = 1, 2, \ldots, n)$$

$$-\frac{E\partial^2 \log 1}{\partial\sigma_{ii}\partial\sigma_{jj}} = 0 \quad \text{if } i \neq j,$$

$$-\frac{E\partial^2 \log 1}{\partial\sigma_{ii}\partial\Delta} = \{\tfrac{1}{2}[\Delta + \sigma_{ii}(X_i'X_i)^{-1}]^{-1}(X_i'X_i)^{-1}[\Delta + \sigma_{ii}(X_i'X_i)^{-1}]^{-1}\}$$

$$(i = 1, 2, \ldots, n)$$

$$-\frac{E\partial^2 \log 1}{\partial\Delta_r\partial\Delta_r} = \frac{1}{2}\sum_{i=1}^{n}[\Delta + \sigma_{ii}(X_i'X_i)^{-1}]^{-1} \otimes [\Delta + \sigma_{ii}(X_i'X_i)^{-1}]^{-1}.$$

$\otimes$ denotes the Kronecker product, see Tracy and Dwyer (1969, pp. 1580, 88–89).

5.1. *Consistency of An ML Estimator of* $\boldsymbol{\theta}$

The pdf $p(\mathbf{y}|X, \boldsymbol{\theta})$ in (35) depends on an unknown parameter vector $\boldsymbol{\theta}$ belonging to a metric space Θ which is a subset of $[K + n + \frac{1}{2}K(K+1)]$-dimensional Euclidean space. In (35) there is a family of possible distributions given by different values of $\boldsymbol{\theta}$ in Θ and we do not know which one is appropriate. Let $\boldsymbol{\theta}_0$ be the unknown true value of $\boldsymbol{\theta}$. We shall denote by $E_0 \log p(\mathbf{y}|X, \boldsymbol{\theta})$ and $\text{var}_0 \log p(\mathbf{y}|X, \boldsymbol{\theta})$ the mean and variance respectively of the random variable $\log p(\mathbf{y}|X, \boldsymbol{\theta})$ on the sample space Y (of elements $\mathbf{y}$) with respect to the distribution of $\mathbf{y}$ determined by $\boldsymbol{\theta}_0$. Let N_0 be an open neighborhood of $\boldsymbol{\theta}_0$. To prove that $\hat{\boldsymbol{\theta}}$ is weakly consistent we have to show that $[\log p(\mathbf{y}|X, \boldsymbol{\theta}_0) - \sup_{\boldsymbol{\theta}\in\Theta - N_0} \log p(\mathbf{y}|X, \boldsymbol{\theta})] > 0$ in probability according to $p(\mathbf{y}|X, \boldsymbol{\theta}_0)$ see Silvey (1961, pp. 445–6). This means that the value of $\boldsymbol{\theta}$ which maximizes $1(\boldsymbol{\theta}|\mathbf{y}, X)$ belongs to N_0 in probability when $\boldsymbol{\theta}_0$ obtains. If, for every n, T and $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, we have $E_0 \log p(\mathbf{y}|X, \boldsymbol{\theta}_0) > E_0 \log p(\mathbf{y}|X, \boldsymbol{\theta})$, and $E_0\{\log p(\mathbf{y}|X, \boldsymbol{\theta}_0) - \log p(\mathbf{y}|X, \boldsymbol{\theta})\}$ is large relative to $[\text{var}_0\{\log p(\mathbf{y}|X, \boldsymbol{\theta}_0) - \log p(\mathbf{y}|X, \boldsymbol{\theta})\}]^{1/2}$, then it follows from Chebychev's inequality that the method of maximum likelihood will discriminate well between $\boldsymbol{\theta}_0$ and other $\boldsymbol{\theta}$. By putting certain regularity conditions on $1(\boldsymbol{\theta}|\mathbf{y}, X)$ we can guarantee that the method will discriminate well between $\boldsymbol{\theta}_0$ and, simultaneously, all other parameter values outside an open neighborhood of $\boldsymbol{\theta}_0$, for large enough n and T. This is the basis of consistency proofs given by Silvey and others.

The likelihood function in (36) contains terms of different orders, each containing a particular subvector of $\boldsymbol{\theta}$. Consequently, we proceed as follows: First, we assume that $\boldsymbol{\theta}_0 \in \Theta$. Second, we rewrite (35) as

$$p(\mathbf{y}|X, \boldsymbol{\theta}) = \left[\prod_{i=1}^{n} g(s_{ii}|\sigma_{ii})\right] f(\mathbf{b}|X, \boldsymbol{\theta}) \tag{37}$$

where

$$g(s_{ii}|\sigma_{ii}) \propto \sigma_{ii}^{-\frac{1}{2}(T-K)} \exp\left\{-\frac{1}{2}\frac{(T-K)s_{ii}}{\sigma_{ii}}\right\}$$

and

$$f(\mathbf{b}|X, \boldsymbol{\theta}) \propto \left[\prod_{i=1}^{n} |\Delta + \sigma_{ii}(X_i'X_i)^{-1}|^{-1/2}\right] \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{b}_i - \bar{\boldsymbol{\beta}})' \cdot [\Delta + \sigma_{ii}(X_i'X_i)^{-1}]^{-1}(\mathbf{b}_i - \bar{\boldsymbol{\beta}})\right\}.$$

By Jensen's inequality (Silvey, 1970, p. 75) we have

$$\text{(38)} \quad E_0\left[\frac{1}{nT}\sum_{i=1}^{n} \log g(s_{ii}|\sigma_{ii0}) + \frac{1}{n}\log f(\mathbf{b}|X, \boldsymbol{\theta}_0)\right] \geq E_0\left[\frac{1}{nT}\sum_{i=1}^{n} \log g(s_{ii}|\sigma_{ii}) + \frac{1}{n}\log f(\mathbf{b}|X, \boldsymbol{\theta})\right]$$

where σ_{ii0} is the true value of σ_{ii}. The inequality in (38) is strict unless $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ because, in view of Assumption 1, $\boldsymbol{\theta}$ is identified and the distributions corresponding to $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}$ are different.

There is a connection between "local" identifiability of a vector-valued parameter $\boldsymbol{\theta}$ and positive definiteness of the information matrix $I(\boldsymbol{\theta})$, see Rothenberg (1971) and Silvey (1970, pp. 81-2).

Assumption 4: The vectors $\mathbf{x}_{it} = (x_{i1t}, x_{i2t}, \ldots, x_{iKt})'$ are all contained in a compact subset of K-dimensional Euclidean space such that for each $i = 1, 2, \ldots, n$ the matrix $T^{-1}X_i'X_i$ converges to a finite positive definite matrix as $T \to \infty$.

Let $D = \text{diag}\,[nI_K, TI_n, nI_{K^2}]$. Now consider $D^{-1/2}I(\boldsymbol{\theta}_0)D^{-1/2}$ where $I(\boldsymbol{\theta}_0)$ is obtained from (36) by replacing $\boldsymbol{\theta}$ by $\boldsymbol{\theta}_0$. The positive definiteness of $\lim_{T\to\infty, n\to\infty} D^{-1/2}I(\boldsymbol{\theta}_0)D^{-1/2}$ which is necessary for the local identifiability of $\boldsymbol{\theta}_0$ follows from Assumption 4. Following the same argument as in Silvey (1970, pp. 81–2) we can show that for any $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$

$$\text{(39)} \quad \lim_{\substack{T\to\infty \\ n\to\infty}} E_0\left[\frac{1}{nT}\sum_{i=1}^{n} \log g(s_{ii}|\sigma_{ii0}) + \frac{1}{n}\log f(\mathbf{b}|X, \boldsymbol{\theta}_0)\right] > \lim_{\substack{T\to\infty \\ n\to\infty}} E_0\left[\frac{1}{nT}\sum_{i=1}^{n} \log g(s_{ii}|\sigma_{ii}) + \frac{1}{n}\log f(\mathbf{b}|X, \boldsymbol{\theta})\right].$$

It is easy to show that for every $\boldsymbol{\theta} \in \Theta$

$$\text{(40)} \quad E_0\left[\frac{1}{nT}\sum_{i=1}^{n} \log g(s_{ii}|\sigma_{ii}) + \frac{1}{n}\log f(\mathbf{b}|X, \boldsymbol{\theta})\right] = O(1),$$

$$\text{(41)} \quad \text{var}_0\left[\frac{1}{nT}\sum_{i=1}^{n} \log g(s_{ii}|\sigma_{ii})\right] = O\left(\frac{1}{nT}\right),$$

and

$$\text{(42)} \quad \text{var}_0\left[\frac{1}{n}\log f(\mathbf{b}|X, \boldsymbol{\theta})\right] = O(n^{-1}).$$

Let Θ_s be a compact subset of Θ.

Assumption 5: $\boldsymbol{\theta}_0 \in \Theta_s$.

In various practical situations it is often possible to rule out sufficiently extreme values of $\boldsymbol{\theta}$ on theoretical grounds and form Θ_s so that $\boldsymbol{\theta}_0 \in \Theta_s$. In cases where the maximum likelihood procedure outlined in the previous subsection leads to implausible estimates like negative estimates for the diagonal elements of Δ, Assumption 5 may not hold. In these cases we should examine Assumption 1 more closely. Under certain additional conditions we can replace Assumption 5 by a wider condition, see Perlman (1972).

The function $p(\mathbf{y}|X, \boldsymbol{\theta})$ is a pdf on the sample space Y, given X, for each $\boldsymbol{\theta}$ in Θ_s, and the function $l(\boldsymbol{\theta}|\mathbf{y}, X)$ is continuous on the metric space Θ_s for each $\mathbf{y}$, given X. Since $\Theta_s - N_0$ is compact, we can cover it by a finite number, say h, of open spheres of radius r_{N_0} having centers $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_h$, say. Let $\log p(\mathbf{y}|X, \boldsymbol{\theta}_m, r_{N_0})$ be the supremum of $\log p(\mathbf{y}|X, \boldsymbol{\theta}_j)$ with respect to $\boldsymbol{\theta}_j$ when $\|\boldsymbol{\theta}_m - \boldsymbol{\theta}_j\| < r_{N_0}$. For any $\boldsymbol{\theta}_m \in \Theta_s$ we have, $\lim E_0 \log p(\mathbf{y}|X, \boldsymbol{\theta}_m, r_{N_0}) < \infty$ as $r_{N_0} \to 0$ because $p(\mathbf{y}|X, \boldsymbol{\theta})$ is uniformly bounded in $\mathbf{y}$, $\boldsymbol{\theta}$ and $E_0 \log p(\mathbf{y}|X, \boldsymbol{\theta}_0) < \infty$. We can show that

$$
(43) \quad E_0\left[\frac{1}{nT}\sum_{i=1}^{n} \log g(s_{ii}|\sigma_{iim}, r_{N_0}) + \frac{1}{n}\log f(\mathbf{b}|X, \boldsymbol{\theta}_m, r_{N_0})\right] < E_0\left[\frac{1}{nT}\sum_{i=1}^{n} \log g(s_{ii}|\sigma_{ii0}) + \frac{1}{n}\log f(\mathbf{b}|X, \boldsymbol{\theta}_0)\right] \qquad (m = 1, 2, \ldots, h).
$$

The results in (38)–(43) are adequate to establish the consistency of an ML estimate of $\boldsymbol{\theta}$, see Swamy and Rao (1971), and Silvey (1961).

5.2. *Asymptotic Normality*

The standard method of establishing the asymptotic normality of an ML estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ utilizes the following results:

(a) Taylor's theorem in the expansion of $\partial \log l(\hat{\boldsymbol{\theta}}|\mathbf{y}, X)/\partial\boldsymbol{\theta}_0$;

(b) a central limit theorem applied to $D^{-1/2}(\partial \log l(\boldsymbol{\theta}_0|\mathbf{y}, X)/\partial\boldsymbol{\theta}_0)$;

(c) a law of large numbers applied to $D^{-1/2}(\partial^2 \log l(\boldsymbol{\theta}_0|\mathbf{y}, X)/\partial\boldsymbol{\theta}_0\, \partial\boldsymbol{\theta}_0')D^{-1/2}$.

Under Assumptions 1, 3, 4 and 5 we have enough regularity conditions to establish the above results, see Silvey (1971, pp. 77–8) and Swamy and Rao (1971). Consequently, $D^{-1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is asymptotically normal with mean $\mathbf{0}$ and V–C matrix $[\lim_{T\to\infty, n\to\infty} D^{-1/2}I(\boldsymbol{\theta}_0)D^{-1/2}]^{-1}$.

The argument just presented, combined with the fact that the prior distribution of $\boldsymbol{\theta}$ does not depend on n and T, shows that in large samples, when Assumptions 1–5 are satisfied, the posterior distribution of $\boldsymbol{\theta}$ is approximately normal with mean $\hat{\boldsymbol{\theta}}$ and V–C matrix $[-(\partial^2 \log l(\boldsymbol{\theta}|\mathbf{y}, X)/\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}')]^{-1}$ evaluated at $\hat{\boldsymbol{\theta}}$; see Lindley (1971, p. 62) and Zellner (1971, pp. 32–3). This result is true even when the prior distribution of $\boldsymbol{\omega}$ is not a point distribution, provided the above conditions are satisfied.

6. Summary and Conclusions

In this paper we considered six different estimators of the mean of a random coefficient vector. These are (1) the MVLU estimator $\bar{\mathbf{b}}(\boldsymbol{\omega})$, (2) the Stein-like estimator $c^*\bar{\mathbf{b}}(\boldsymbol{\omega})$, (3) the ridge regression estimator $\bar{\mathbf{b}}_\mu(\boldsymbol{\omega})$, (4) the MCMSE estimator $\bar{\mathbf{b}}^*(\omega)$, (5) the mixed regression estimator $\hat{\beta}_p(\omega)$, and (6) an ML estimator $\hat{\bar{\beta}}$ of $\bar{\beta}$. We also found feasible approximations to these estimators. None of the estimators $\bar{\mathbf{b}}(\omega)$, $c^*\bar{\mathbf{b}}(\omega)$, $\bar{\mathbf{b}}_\mu(\omega)$ and $\bar{\mathbf{b}}^*(\omega)$ is uniformly better than the other. Each of these estimators has its own weaknesses. In cases where *a priori* unbiased estimator r of $\bar{\beta}$ is available and its V–C matrix $\tau^2\psi$ is known, the estimator $\hat{\bar{\beta}}_p(\omega)$ is uniformly better than the estimator $\bar{\mathbf{b}}(\omega)$. Under these conditions, the estimator $\hat{\bar{\beta}}_p(\omega)$ is also better than $\bar{\mathbf{b}}^*(\omega)$ if $\bar{\beta}^*\bar{\beta}^{*\prime}$ is not a reliable estimate of $\bar{\beta}\bar{\beta}'$. The estimators $\bar{\mathbf{b}}_\mu(\omega)$, $\bar{\mathbf{b}}^*(\omega)$ and $\hat{\bar{\beta}}_p(\omega)$ are insensitive to extreme multicollinearity. The estimator $\hat{\bar{\beta}}_p(\omega)$ covers the estimators $c^*\bar{\mathbf{b}}(\omega)$ and $\bar{\mathbf{b}}_\mu(\omega)$ as special cases.

When $\bar{\beta}$ is regarded as a random variable, the formula $\bar{\mathbf{b}}^*(\omega)$ is inappropriate and the estimator $\hat{\bar{\beta}}_p(\omega)$ covers the estimators $\bar{\mathbf{b}}_\mu(\omega)$ and $c^*\bar{\mathbf{b}}(\omega)$ as special cases. The prior information utilized in obtaining the estimator $\hat{\bar{\beta}}_p(\omega)$ is likely to provide a better numerical approximation to the practical situation than those utilized in obtaining the estimators $c^*\bar{\mathbf{b}}(\omega)$ and $\bar{\mathbf{b}}_\mu(\omega)$. The estimator $\hat{\bar{\beta}}_p(\omega)$ is uniformly better than the estimators $\bar{\mathbf{b}}(\omega)$, $c^*\bar{\mathbf{b}}(\omega)$, $\bar{\mathbf{b}}_\mu(\omega)$ and $\bar{\mathbf{b}}^*(\omega)$ if $\bar{\beta}$ is distributed with mean $\mathbf{r}$ and V–C matrix $\tau^2\psi$. Furthermore, $\hat{\bar{\beta}}_p(\omega)$ has all the desirable properties of a posterior mean corresponding to a normal prior and normal likelihood. In small samples one cannot find a uniformly better estimator of $\bar{\beta}$ unless the prior distribution of $\bar{\beta}$ is proper and known.

Under certain regularity conditions, the maximum likelihood estimate $\hat{\bar{\beta}}$ is at least as good as any other estimator of $\bar{\beta}$ in large samples.

Federal Reserve System, Washington D.C.

Appendix

Here we provide the proof of (26). The conditional second order moment matrix of $\bar{\mathbf{b}}^*(\omega)$ in (24) around $\bar{\beta}$, given $\bar{\beta}^*$, is

$$\text{(A.1)}\quad \bar{\beta}^*\bar{\beta}^{*\prime}[\bar{\beta}^*\bar{\beta}^{*\prime} + (X'\Sigma^{-1}X)^{-1}]^{-1}(X'\Sigma^{-1}X)^{-1}[\bar{\beta}^*\bar{\beta}^{*\prime} + (X'\Sigma^{-1}X)^{-1}]^{-1}\bar{\beta}^*\bar{\beta}^{*\prime}$$
$$+ (X'\Sigma^{-1}X)^{-1}[\bar{\beta}^*\bar{\beta}^{*\prime} + (X'\Sigma^{-1}X)^{-1}]^{-1}\bar{\beta}\bar{\beta}'[\bar{\beta}^*\bar{\beta}^{*\prime}$$
$$+ (X'\Sigma^{-1}X)^{-1}]^{-1}(X'\Sigma^{-1}X)^{-1}.$$

The first term in (A.1) is the conditional V–C matrix of $\bar{\mathbf{b}}^*(\omega)$ and the second term is the matrix of squares and cross-products of the biases of the elements of $\bar{\mathbf{b}}^*(\omega)$ for given $\bar{\beta}^*$. Subtracting (A.1) from the V–C matrix of $\bar{\mathbf{b}}(\omega)$ gives

$$\text{(A.2)}\quad (X'\Sigma^{-1}X)^{-1} - \bar{\beta}^*\bar{\beta}^{*\prime}[\bar{\beta}^*\bar{\beta}^{*\prime} + (X'\Sigma^{-1}X)^{-1}]^{-1}(X'\Sigma^{-1}X)^{-1}$$
$$\cdot[\bar{\beta}^*\bar{\beta}^{*\prime} + (X'\Sigma^{-1}X)^{-1}]^{-1}\bar{\beta}^*\bar{\beta}^{*\prime} - (X'\Sigma^{-1}X)^{-1}[\bar{\beta}^*\bar{\beta}^{*\prime} + (X'\Sigma^{-1}X)^{-1}]^{-1}$$
$$\cdot\bar{\beta}\bar{\beta}'[\bar{\beta}^*\bar{\beta}^{*\prime} + (X'\Sigma^{-1}X)^{-1}]^{-1}(X'\Sigma^{-1}X)^{-1}.$$

Let P be a nonsingular matrix such that $P'(X'\Sigma^{-1}X)^{-1}P = I$, and $P'\bar{\beta}^*\bar{\beta}^{*\prime}P = \lambda_1^*\mathbf{i}_1\mathbf{i}_1'$ where $\mathbf{i}_1$ is the first column of an identity matrix of order K. We pre and post multiply (A.2) by $P'^{-1}P'$ and PP^{-1} respectively to obtain

$$\text{(A.3)}\qquad P'^{-1}P^{-1} - P'^{-1}\{\lambda_1^*\mathbf{i}_1\mathbf{i}_1'(\lambda_1^*\mathbf{i}_1\mathbf{i}_1' + I)^{-2}\lambda_1^*\mathbf{i}_1\mathbf{i}_1'\}P^{-1}$$
$$- P'^{-1}(\lambda_1^*\mathbf{i}_1\mathbf{i}_1' + I)^{-1}\lambda_1\mathbf{0}_1\mathbf{0}_1'(\lambda_1^*\mathbf{i}_1\mathbf{i}_1' + I)^{-1}P^{-1}$$

where $\boldsymbol{\theta}_1$ is the characteristic vector corresponding to the nonzero root λ_1 of $P'\bar{\boldsymbol{\beta}}\bar{\boldsymbol{\beta}}'P$. Using an identity in Swamy (1971, p. 25, Lemma 2.2.2) we have

$$\text{(A.4)} \qquad P'^{-1}P^{-1} - P'^{-1}\mathbf{i}_1\mathbf{i}_1'P^{-1}\frac{\lambda_1^{*2}}{(1+\lambda_1^*)^2} - P'^{-1}\left(I - \frac{\lambda_1^*}{1+\lambda_1^*}\mathbf{i}_1\mathbf{i}_1'\right)\lambda_1\boldsymbol{\theta}_1\boldsymbol{\theta}_1' \cdot \left(I - \frac{\lambda_1^*}{1+\lambda_1^*}\mathbf{i}_1\mathbf{i}_1'\right)P^{-1}.$$

Consequently, given $\bar{\boldsymbol{\beta}}^*$,

$$\text{(A.5)} \qquad E[\bar{\mathbf{b}}(\omega) - \bar{\boldsymbol{\beta}}][\bar{\mathbf{b}}(\omega) - \bar{\boldsymbol{\beta}}]' - E[\bar{\mathbf{b}}^*(\omega) - \bar{\boldsymbol{\beta}}][\bar{\mathbf{b}}^*(\omega) - \bar{\boldsymbol{\beta}}]'$$
$$= P'^{-1}\left\{I - \mathbf{i}_1\mathbf{i}_1'\left[\frac{\lambda_1^{*2}}{(1+\lambda_1^*)^2} + \frac{\lambda_1^{*2}\lambda_1\theta_{11}^2}{(1+\lambda_1^*)^2}\right] - \lambda_1\boldsymbol{\theta}_1\boldsymbol{\theta}_1' + \frac{\lambda_1^*\lambda_1\theta_{11}}{(1+\lambda_1^*)} \cdot (\mathbf{i}_1\boldsymbol{\theta}_1' + \boldsymbol{\theta}_1\mathbf{i}_1')\right\}P^{-1}$$

where θ_{11} is the first element of $\boldsymbol{\theta}_1$.

Let the matrix within the curl brackets be B. The matrix in (A.5) is positive definite if B is positive definite. Since B is symmetric, B is positive definite if all its diagonal elements are positive. The first diagonal element of B is positive if $\bar{\boldsymbol{\beta}}'\mathbf{p}_1\mathbf{p}_1'\bar{\boldsymbol{\beta}} < 1 + 2\lambda_1^*$ where $\mathbf{p}_1$ is the first column of P. Every other diagonal element of B is positive if $\bar{\boldsymbol{\beta}}'\mathbf{p}_k\mathbf{p}_k'\bar{\boldsymbol{\beta}} < 1$ $k = 2, \ldots, K$.

Using P we may rewrite (17) as

$$\text{(A.6)} \qquad P'^{-1}\{P'P(P'P + \mu I)^{-2}P'P\}P^{-1} + P'^{-1}\{\mu^2(P'P + \mu I)^{-1}P'\bar{\boldsymbol{\beta}}\bar{\boldsymbol{\beta}}'P \cdot (P'P + \mu I)^{-1}\}P^{-1}.$$

REFERENCES

Barnard, G. (1963), "The Logic of Least Squares," *Journal of the Royal Statistical Society*, Series B, 25, pp. 124–7.

——. (1967), "The Use of the Likelihood Function in Statistical Practice," pp. 27–40, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, Berkeley and Los Angeles: University of California Press.

Box, G. E. P. and G. M. Jenkins (1970), *Time Series Analysis, Forecasting, and Control*, San Francisco: Holden-Day.

Box, G. E. P. and G. C. Tiao (1973), *Bayesian Inference in Statistical Analysis*, Massachusetts: Addison Wesley Publishing Company.

Chipman, J. S. (1964), "On Least Squares with Insufficient Observations," *Journal of the American Statistical Association*, 59, pp. 1078–111.

Durbin, J. (1953), "A Note on Regression When There is Extraneous Information About one of the Coefficients," *Journal of the American Statistical Association*, 48, pp. 799–808.

Efron, B. and C. Morris (1971), "Limiting the Risk of Bayes and Empirical Bayes Estimators, Part I: The Bayes Case," *Journal of the American Statistical Association*, 66, pp. 807–15.

——. (1972), "Limiting Risk of Bayes and Empirical Bayes Estimators—Part II: The Empirical Bayes Case," *Journal of the American Statistical Association*, 67, pp. 130–9.

Feige, E. L. and P. A. V. B. Swamy (1972), "A Random Coefficient Model of the Demand for Liquid Assets," Workshop Paper #7211, Social Systems Research Institute, The University of Wisconsin, Madison. Accepted for publication in *Journal of Money, Credit, and Banking*.

Ferguson, T. S. (1967), *Mathematical Statistics*, New York: Academic Press.

Hoerl, A. E. and R. W. Kennard (1970a), "Ridge Regression: Biased Estimation for Non-Orthogonal Problems," *Technometrics*, 12, pp. 55–67.

——. (1970b), "Ridge Regression: Applications to Non-Orthogonal Problems," *Technometrics*, 12, pp. 69–82.

Kalbfleisch, J. G. and D. A. Sprott (1970). "Application of Likelihood Methods to Models Involving Large Numbers of Parameters." *Journal of the Royal Statistical Society*, Series B, 32, pp. 175–94.
Kiefer, J. and J. Wolfowitz (1956), "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Nuisance Parameters." *Annals of Mathematical Statistics*, 27, pp. 887–906.
Lindley, D. V. (1971). *Bayesian Statistics, A Review*. SIAM, Philadelphia.
Lindley, D. V. and A. F. M. Smith (1972). "Bayes Estimates for the Linear Model." *Journal of the Royal Statistical Society*, Series B, 34, pp. 1–41.
Mehta, J. S. and R. Srinivasan (1971). "Estimation of the Mean by Shrinkage to a Point." *Journal of the American Statistical Association*, 66, pp. 86–90.
Mehta, J. S. and P. A. V. B. Swamy (1972a), "The Exact Finite Sample Distribution of Theil's Compatibility Statistic and its Application." Report 7204, Division of Economic Research, Department of Economics, Ohio State University, Columbus. Accepted for publication in *Journal of the American Statistical Association*.
———. (1972b). "Efficient Method of Estimating the Level of a Stationary First-Order Autoregressive Process." *Communications in Statistics* (forthcoming).
Perlman, M. D. (1972): "On the Strong Consistency of Approximate Maximum Likelihood Estimators." pp. 263–81 in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, 1, Berkeley and Los Angeles: California University Press.
Rao, C. R. (1962). "Apparent Anomalies and Irregularities in Maximum Likelihood Estimation (with discussion)," *Sankhyā*, Series A, 24, pp. 73–102.
———. (1965). *Linear Statistical Inference and its Applications*. New York: John Wiley & Sons.
———. (1971). "Unified Theory of Linear Estimation." *Sankhyā*, Series A, 33, pp. 371–94.
Rao, C. R. and S. K. Mitra (1971). *Generalized Inverse of Matrices and its Applications*. New York: John Wiley & Sons.
Rothenberg, T. J. (1971). "Identification in Parametric Models." *Econometrica*, 39, pp. 577–92.
Silvey, S. D. (1961). "A Note on Maximum-Likelihood in the Case of Dependent Random Variables." *Journal of the Royal Statistical Society*, Series B, 23, pp. 444–52.
———. (1970). *Statistical inference*. Baltimore: Penguin.
Sims, C. A. (1971). "Distributed Lag Estimation When the Parameter Space is Explicitly Infinite-Dimensional." *Annals of Mathematical Statistics*, 42, pp. 1622–36.
Stein, C. (1966). "An Approach to the Recovery of Inter-Block Information in Balanced Incomplete Block Designs." pp. 351–66, in F. N. David (editor), *Research Papers in Statistics*, Festschrift for J. Neyman. New York: John Wiley & Sons.
Strawderman, W. E. and A. Cohen (1971). "Admissibility of Estimators of the Mean Vector of a Multivariate Normal Distribution with Quadratic Loss." *Annals of Mathematical Statistics*, 42, pp. 720–96.
Swamy, P. A. V. B. (1971). *Statistical Inference in Random Coefficient Regression Models*. New York: Springer-Verlag.
———. (1972). "Linear Models with Random Coefficients." to appear in Paul Zarembka (editor), *Frontiers of Econometrics*. New York: Academic Press, forthcoming.
Swamy, P. A. V. B. and J. S. Mehta (1969). "On Theil's Mixed Regression Estimator." *Journal of the American Statistical Association*, 64, pp. 273–6.
———. (1972). "Robustness of Theil's Mixed Regression Estimators." Report 7218, Division of Economic Research, Department of Economics, Ohio State University, Columbus.
———. (1971). "Bayesian Analysis of Error Components Regression Models," Report 7203, Division of Economic Research, Department of Economics, Ohio State University, Columbus, to appear in *Journal of the American Statistical Association*.
Swamy, P. A. V. B. and J. N. K. Rao (1971). "On Consistency and Asymptotic Normality of Maximum Likelihood Estimators of Parameters in a Distributed Lag Model with Auto-correlated Errors." Technical Report #16, University of Manitoba, Winnipeg.
Theil, H. (1963). "On the Use of Incomplete Prior Information in Regression Analysis." *Journal of the American Statistical Association*, 58, pp. 401–14.
———. (1971). *Principles of Econometrics*. New York: John Wiley & Sons.
Theil, H. and A. S. Goldberger (1961). "On Pure and Mixed Statistical Estimation in Economics." *International Economic Review*, 2, pp. 65–78.
Tracy, D. S. and P. S. Dwyer (1969). "Multivariate Maxima and Minima with Matrix Derivatives." *Journal of the American Statistical Association*, 64, pp. 1576–94.
Winkler, R. L. (1967a). "The Assessment of Prior Distributions in Bayesian Analysis," *Journal of the American Statistical Association*, 62, pp. 776–800.
———. (1967b). "The Quantification of Judgment: Some Methodological Suggestions," *Journal of the American Statistical Association*, 62, pp. 1109–20.
———. (1971). "Probabilistic Prediction: Some Experimental Results." *Journal of the American Statistical Association*, 66, pp. 675–85.

Zellner, A. (1970), "The Bayesian Approach and Alternatives to Econometrics I," pp. 178–93 in M. D. Intriligator (editor), *Frontiers of Quantitative Economics*. Amsterdam: North-Holland Publishing Company.

——— (1971), *An Introduction to Bayesian Inference in Econometrics*. New York: John Wiley & Sons.

Zellner, A. and W. Vandaele (1971), "Bayes-Stein Estimators for K-Means, Regression and Simultaneous Equation Models," presented at the Third NBER-NSF Symposium on Bayesian Inference in Econometrics.