FACULTY OF
ACTUARIAL SCIENCE
&
ECONOMETRICS

A & E REPORT

REPORT AE 7/85

ON MULTIVARIATE RIDGE REGRESSION

Y. Haitovsky (Hebrew University of Jerusalem)

University of Amsterdam

Address: University of Amsterdam

            Faculty of Actuarial Science and Econometrics

            Jodenbreestraat 23

            1011 NH Amsterdam

Abstract

A multivariate linear regression model with q responses as a linear function
of p independent variables $\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{B} + \underset{\sim}{E}$ is considered with a p x q parameter
matrix $\underset{\sim}{B}$. The least squares (or Maximum Likelihood for multivariate normal $\underset{\sim}{E}$)
estimator of $\underset{\sim}{B}$ is deficient in that it takes no account of the "across
regression" correlations, on the one hand, and ignores the famous Stein effect,
on the other hand. A remedy was offered by Brown and Zidek (1980) in the form
of a multivariate ridge estimator. A richer class of estimators is obtained here
by casting the model in a linear hierarchical framework, obtaining the Brown and
Zidek multivariate ridge estimators., Efron and Morris' estimators of several
normal mean vectors and Fearn's Bayesian estimators of growth curves as special
cases. The unknown covariance cases result in an identifiability problem which
is treated in a Bayesian fashion using conjugate priors. The method is then
applied to forecasting the final election results from partial returns obtained
at election night.

On Multivariate Ridge Regression

1. INTRODUCTION:

Brown and Zidek (1980) consider a multivariate problem with q responses and n observations, $\underset{\sim}{Y}$, assumed to satisfy the standard multivariate linear regression model

$$\underset{\sim}{Y} = \underset{\sim}{X} \underset{\sim}{B} + \underset{\sim}{E} \tag{1}$$

with $\underset{\sim}{X}$ an $(n \times p)$ full column rank matrix whose elements

are treated as fixed known constants

and $\underset{\sim}{B}$ is a $(p \times q)$ matrix of unknown coefficients to be estimated. With $\underset{\sim}{E} = (\underset{\sim}{\varepsilon}_1, \ldots, \underset{\sim}{\varepsilon}_q)$, the usual assumptions on the error are

$$E(\underset{\sim}{\varepsilon}_j) = \underset{\sim}{0}, \ \operatorname{cov}(\underset{\sim}{\varepsilon}_j, \underset{\sim}{\varepsilon}_\ell) = \gamma_{j\ell} \underset{\sim}{I}_n \qquad j, \ell = 1, \ldots, q \tag{2}$$

or, in short $E(\underset{\sim}{E}) = \underset{\sim}{0}$, $\operatorname{cov}(\underset{\sim}{E}) = \Gamma \otimes \underset{\sim}{I}_n$, where $\otimes$ denotes the usual Kronecker product of matrices. The least squares estimator of $\underset{\sim}{B}$ is given by

$$\hat{\underset{\sim}{B}} = (\underset{\sim}{X}^T \underset{\sim}{X})^{-1} \underset{\sim}{X}^T \underset{\sim}{Y} \tag{3}$$

which is also the maximum likelihood estimator when normality of the error distribution is assumed. Equivalently, writing $\hat{\underset{\sim}{B}} = (\hat{\underset{\sim}{\beta}}_1, \ldots, \hat{\underset{\sim}{\beta}}_q)$, $\underset{\sim}{Y} = (\underset{\sim}{Y}_1, \ldots, \underset{\sim}{Y}_q)$ so that the vectors $\underset{\sim}{Y}_j (n \times 1)$ and $\hat{\underset{\sim}{\beta}}_j (p \times 1)$ pertain to the j-th of the q responses, (3) can be rewritten as

$$\hat{\underset{\sim}{\beta}}_j = (\underset{\sim}{X}^T \underset{\sim}{X})^{-1} \underset{\sim}{X}^T \underset{\sim}{Y}_j \qquad (j = 1, \ldots, q) \ . \tag{3a}$$

Brown and Zidek cite Sclove's (1971) argument that the least squares estimator (3) is deficient in that it takes no account of $\underset{\sim}{\Gamma} = (\gamma_{ij})$, the between regressions covariance matrix. In order to remedy this shortcoming and also to

take advantage of possible improvements in the estimation of the β_i's even

when $\Gamma = I$ by using the ridge regression estimator for that case, Brown and

Zidek suggest the multivariate ridge regression estimator (MRRE) of the form

$$\hat{\beta}^*(K) = (I_q \otimes X^T X + K \otimes I_p)^{-1} (I_q \otimes X^T) y \qquad (4)$$

where $K(q \times q) > 0$ is the ridge matrix, $y = \text{vec}\{Y\}$ is the $(nq \times 1)$ vector

$y^T = (y_1^T, \ldots, y_q^T)$ and $\hat{\beta}^*(K)$ is a $(pq \times 1)$ vector of estimators of

$\beta = \text{vec}\{\beta\} = (\beta_1^T, \ldots, \beta_q^T)^T$ where each β_j is a $(p \times 1)$ column vector.[1]

The least squares estimators(3) have also a stacked matrix representation

$$\hat{\beta} = (I_q \otimes X^T X)^{-1} (I_q \otimes X^T) y . \qquad (3b)$$

Clearly $\hat{\beta}^*(0) = \hat{\beta}$.

Estimator (4) is recognized by Brown and Zidek as a Bayes estimator, which

can be directly obtained from Lindley and Smith's (1972) treatment of the linear

model. However, a more careful application of the Lindley-Smith approach will

yield a richer class of estimators, which is the aim of this paper.

It will be shown that the relationship between the resulting estimator

and the Brown-Zidek MRRE is similar to that between the Lindley-Smith

Exchangeability Within Regression estimator and the Ridge Regression estimator,

where the Ridge estimator is obtained as a special case when an exchangeable prior

around zero is assumed for the regression coefficients. Furthermore, it will be

shown that the Efron-Morris (1972) estimators, at least for the known covariance

cases, can be obtained as special cases of the estimator suggested here. Hence

the latter can be regarded as an extension of Stein's method to the multivariate

regression model.

---

[1] Brown and Zidek stack their matrices row by row. I find it more natural
to stack them by columns and hence the few apparent differences.

The unknown covariance case, even after imposing initial restrictions on its structure, is shown to be nonidentifiable, and hence further (nonstochastic) restrictions are required for estimating the model parameters. Thus, prior information on at least some of the covariances is imposed in the form of a natural conjugate distribution.

2. THE LINEAR HIERARCHICAL MULTIVARIATE REGRESSION ESTIMATOR

The approach here is that of Lindley and Smith (1972). A three-stage model will be assumed, where in the first stage, given the design matrix X,

$$y \mid \beta, \Gamma \sim N\{(I_q \otimes X)\beta, \ \Gamma \otimes I_n\} . \tag{5}$$

Equation (5) is Zellner's (1961) "seemingly unrelated regression equations" representation of the multivariate linear regression model in the special case of identical design matrices.

A second stage is added by supposing

$$\beta \mid \xi, \Omega \sim N(1_q \otimes \xi, \ \Omega) , \qquad \text{independent of } y , \tag{6}$$

that is, each $E\{\beta_j\} = \xi$ $(j=1,\ldots,q)$. A vague prior knowledge on ξ is assumed for the third stage.

The formal posterior distribution of $(\beta^T, \xi^T)^T$ given y and X is shown, for known variances, to be normal with mean vector [see Haitovsky (1985)] :

$$\begin{pmatrix} \beta* \\ \xi* \end{pmatrix} = (A^T C^{-1} A)^{-1} A^T C^{-1} z \tag{7}$$

and covariance matrix

$$\text{cov}\begin{pmatrix} \beta* \\ \xi* \end{pmatrix} = (A^T C^{-1} A)^{-1} \tag{8}$$

where
$$\underset{\sim}{A} = \begin{bmatrix} \underset{\sim}{I}_q \otimes \underset{\sim}{X} & \underset{\sim}{0} \\ -\underset{\sim}{I}_{pq} & \underset{\sim}{1}_q \otimes \underset{\sim}{I}_p \end{bmatrix} \quad , \quad \underset{\sim}{C} = \begin{pmatrix} \underset{\sim}{\Gamma} \otimes \underset{\sim}{I}_n & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{\Omega} \end{pmatrix} \quad ,$$

and
$$\underset{\sim}{z}^T = (\underset{\sim}{y}^T , \underset{\sim}{0}^T) \quad .$$

After some algebra this reduces to the means and variance, conditional on $\underset{\sim}{y}$ and $\underset{\sim}{X}$:

$$\underset{\sim}{\beta}* = \{ \underset{\sim}{I}_q \otimes \underset{\sim}{X}^T \underset{\sim}{X} + (\underset{\sim}{\Gamma} \otimes \underset{\sim}{I}_p) \underset{\sim}{Q} \}^{-1} (\underset{\sim}{I}_q \otimes \underset{\sim}{X}^T) \underset{\sim}{y} \tag{9}$$

with

$$\text{cov} (\underset{\sim}{\beta}*) = (\underset{\sim}{\Gamma}^{-1} \otimes \underset{\sim}{X}^T \underset{\sim}{X} + \underset{\sim}{Q})^{-1} \tag{10}$$

and

$$\underset{\sim}{\xi}* = \left\{ (\underset{\sim}{1}_q \otimes \underset{\sim}{I}_p)^T \underset{\sim}{\Omega}^{-1} (\underset{\sim}{1}_q \otimes \underset{\sim}{I}_p) \right\}^{-1} (\underset{\sim}{1}_q \otimes \underset{\sim}{I}_p)^T \underset{\sim}{\Omega}^{-1} \underset{\sim}{\beta}* \tag{11}$$

with

$$\text{cov}(\underset{\sim}{\xi}*) = \left[(\underset{\sim}{1}_q^T \otimes \underset{\sim}{I}_p) \{ \underset{\sim}{\Gamma} \otimes (\underset{\sim}{X}^T \underset{\sim}{X})^{-1} + \underset{\sim}{\Omega} \}^{-1} (\underset{\sim}{1}_q \otimes \underset{\sim}{I}_p) \right]^{-1} \tag{12}$$

where

$$\underset{\sim}{Q} = [\underset{\sim}{I}_{pq} - \underset{\sim}{\Omega}^{-1} (\underset{\sim}{1}_q \otimes \underset{\sim}{I}_p) \{ (\underset{\sim}{1}_q \otimes \underset{\sim}{I}_p)^T \underset{\sim}{\Omega}^{-1} (\underset{\sim}{1}_q \otimes \underset{\sim}{I}_p) \}^{-1} (\underset{\sim}{1}_q \otimes \underset{\sim}{I}_p)^T] \underset{\sim}{\Omega}^{-1} \quad . \tag{13}$$

Equations (9) and (10) can be regarded as ridge estimators of the multi-variate regression model. Furthermore, the MRRE reduces to the least squares estimator (3) when $\underset{\sim}{\Omega}^{-1} = \underset{\sim}{0}$, as is the case with the univariate ridge.

It will be instructive at this stage to impose some structure on the general covariance matrix $\underset{\sim}{\Omega}$. Let,

$$\underset{\sim}{\Omega} = \underset{\sim}{V} \otimes \underset{\sim}{\Sigma} \tag{14}$$

so that $\underset{\sim}{V}$ (q x q) > 0 is the covariance (except for an overall scale factor) between any two rows in $\underset{\sim}{B}$ and $\underset{\sim}{\Sigma}$ (p × p) > 0 is the covariance (except for

an overall scale factor between any two columns in $\underset{\sim}{B}$. As a result $\underset{\sim}{Q}$

reduces to $\underset{\sim}{Q} = \underset{\sim}{V}^{-1}\underset{\sim}{W} \otimes \underset{\sim}{\Sigma}^{-1}$, and hence eqns. (9) - (13) reduce

respectively to

$$\underset{\sim}{\beta^*} = (\underset{\sim q}{I} \otimes \underset{\sim}{X}^T\underset{\sim}{X} + \underset{\sim w}{K} \otimes \underset{\sim}{\Sigma}^{-1})^{-1}(\underset{\sim q}{I} \otimes \underset{\sim}{X}^T)\underset{\sim}{y} \tag{15}$$

$$\text{cov}(\underset{\sim}{\beta^*}) = (\underset{\sim q}{I} \otimes \underset{\sim}{X}^T\underset{\sim}{X} + \underset{\sim w}{K} \otimes \underset{\sim}{\Sigma}^{-1})^{-1}(\underset{\sim}{\Gamma} \otimes \underset{\sim p}{I}) \tag{16}$$

$$\underset{\sim}{\xi^*} = (\underset{\sim q}{1}^T\underset{\sim}{V}^{-1}\underset{\sim q}{1})^{-1}(\underset{\sim q}{1}^T\underset{\sim}{V}^{-1} \otimes \underset{\sim p}{I})\underset{\sim}{\beta^*} \qquad \text{and} \tag{17}$$

$$\text{cov}(\underset{\sim}{\xi^*}) = [(\underset{\sim q}{1}^T \otimes \underset{\sim p}{I})\{\underset{\sim}{\Gamma} \otimes (\underset{\sim}{X}^T\underset{\sim}{X})^{-1} + \underset{\sim}{V} \otimes \underset{\sim}{\Sigma}\}^{-1}(\underset{\sim q}{1} \otimes \underset{\sim p}{I})]^{-1} \quad , \tag{18}$$

where $\underset{\sim w}{K} = \underset{\sim}{\Gamma}\underset{\sim}{V}^{-1}\underset{\sim}{W}$, $\underset{\sim}{W} = \underset{\sim q}{I} - (\underset{\sim q}{1}^T\underset{\sim}{V}^{-1}\underset{\sim q}{1})^{-1}\underset{\sim q}{J}\underset{\sim}{V}^{-1}$ and $\underset{\sim q}{J} = \underset{\sim q}{1}\underset{\sim q}{1}^T$.

Brown & Zidek (1980) require that MRRE obeys two rules:

(i)   that it should be a Bayes rule for fixed known $\underset{\sim}{K}$, which, being a

posterior mean, our estimator (13) certainly is;

(ii) for a suitably chosen estimator for $\underset{\sim}{K}$, in the special case of equal

information $(\underset{\sim}{X}^T\underset{\sim}{X} = \underset{\sim}{I})$, the MRRE should correspond to the Efron and

Morris (1972) multivariate extension of the James and Stein (1961)

estimator. It will be shown in the next section that this requirement is

met here as well.

3.   SPECIAL CASES

In this section we consider a few important subclasses of priors on

the regression coefficients. We start with those affecting the ridge

matrix $\underset{\sim w}{K}$.

3.1   The prior $\underset{\sim}{\xi} = \underset{\sim}{0}$ is of special interest in applied work, since it results in the multivariate regression analog to the Hoerl and Kennard (1970) "ordinary" ridge rule, in Thisted's (1976) terminology. Thus, as Lindley and Smith (1972, p. 11) show, the univariate (multiple) regression estimator, when the elements of $\underset{\sim}{\beta}_j$ are assumed exchangeable about an unknown value (i.e. a vague ξ), is $\underset{\sim}{\beta}_j^{LS} = (\underset{\sim}{X}^T\underset{\sim}{X} + k\underset{\sim}{U}_p)^{-1}\underset{\sim}{X}^T\underset{\sim}{Y}$, where $\underset{\sim}{U}_p = \underset{\sim}{I}_p - p^{-1}\underset{\sim}{J}_p$ and $k = \gamma/\sigma_\beta^2$, the ratio of the regression error variance to the prior variance of $\underset{\sim}{\beta}$. When exchangeability is assumed about 0 (i.e., $\xi = 0$, in our notation),

$\underset{\sim}{U}_p$ reduces to $\underset{\sim}{I}_p$ and the "ordinary" ridge is obtained. Similarly, in multivariate regression, $\underset{\sim}{\xi} = \underset{\sim}{0}$ is implemented by deleting the last column of submatrices from $\underset{\sim}{A}$ (eqn. 7), resulting in $\underset{\sim}{Q} = \underset{\sim}{\Omega}$ in eqn. (13). Or, when (14) is taken into account, in analogy to the univariate case, $\underset{\sim}{W}$ becomes $\underset{\sim}{I}$ and $\underset{\sim}{K}_W = \underset{\sim}{\Gamma}\underset{\sim}{V}^{-1}$ (in contrast to $\underset{\sim}{K}_W = \underset{\sim}{\Gamma}\underset{\sim}{V}^{-1}\underset{\sim}{W}$ when the prior on $\underset{\sim}{\xi}$ is vague). $\underset{\sim}{K}_W$ is an obvious multivariate generalization of k in the Lindley and Smith framework.

3.2   Multivariate Exchangeability Between Multiple Regression Equations is expressed by letting $\underset{\sim}{V} = \underset{\sim}{I}$, obtaining a ridge matrix $\underset{\sim}{K}_u = \underset{\sim}{\Gamma}\underset{\sim}{U}_q$.

If, furthermore, it is assumed that $\underset{\sim}{\Gamma} = \gamma\underset{\sim}{I}_q$ (15) reduces to

$$\underset{\sim}{\beta}_i^* = \underset{\sim}{H}\hat{\underset{\sim}{\beta}}_i + (\underset{\sim}{I}_p - \underset{\sim}{H}) q^{-1} \sum_{j=1}^{q} \hat{\underset{\sim}{\beta}}_j$$

with

$$\text{cov}(\underset{\sim}{\beta}_i^*) = \{\underset{\sim}{H} + q^{-1}(\underset{\sim}{I}_p - \underset{\sim}{H})\} \gamma(\underset{\sim}{X}^T\underset{\sim}{X})^{-1}$$

where $\underset{\sim}{H} = (\gamma^{-1}\underset{\sim}{X}^T\underset{\sim}{X} + \underset{\sim}{\Sigma}^{-1})^{-1}\gamma^{-1}\underset{\sim}{X}^T\underset{\sim}{X}$ and $\hat{\underset{\sim}{\beta}}_i$ as before are the least squares estimators. This is the Bayesian approach to growth curves with identical X_i matrices as developed by Fearn [1975, equations (10) and (11)].

It is worth mentioning here that if Zellner's seemingly unrelated regression

equations model is assumed instead of (5) , i.e., the design matrix Z , say, in (5) takes the form of a diagonal block matrix with different $X_i (i=1,...,q)$ submatrices along the main diagonal and exchangeability between the regression equations is assumed, then

$$\beta^* = \{Z^T(\Gamma^{-1} \otimes I_n)Z + U_q \otimes \Sigma^{-1}\}^{-1}Z^T(\Gamma^{-1} \otimes I_n)y .$$

See Lindley and Smith (1972, eq. 23) and Haitovsky (1979, §VI-1).

Here again if Γ is a diagonal matrix, Fearn's results for non-identical design matrices are easily obtained. See also Lindley and Smith (1972, p.10).

3.3 Multivariate Exchangeability Within Multiple Regression Equations will

require, in addition to letting $\Sigma = I_p$, a fourth stage in the Linear Hierarchical Model [cf. Haitovsky (1979)]. If only the former is assumed then,

$$\beta^* = (I_q \otimes X^TX + KW \otimes I_p)^{-1}(I_q \otimes X^T)y \tag{19}$$

The last two cases can be combined with the $\xi = 0$ prior assumption, resulting in $W = U_q = I_q$ and hence, for instance for the present case one obtains:

$$\beta^* = (I_q \otimes X^TX + K \otimes I_p)^{-1} (I_q \otimes X^T)y = \hat{\beta}^*(K) . \tag{20}$$

This is the Brown-Zidek MRRE .

3.4 The Efron-Morris Extension of Stein's Method can be reduced from the

hierarchical representation of the multivariate regression model (5) and (6) by letting n=p and $X = I_p$, $\Gamma = I_q$, $\xi = 0$ and $\Omega = I_q \otimes \Sigma$, thus obtaining

$$\beta^* = (I_{pq} + I_q \otimes \Sigma^{-1})^{-1}y = \{I_{pq} - (I_{pq} + I_q \otimes \Sigma^{-1})^{-1}(I_q \otimes \Sigma^{-1})\}y$$

or

$$\beta_i^* = \{I_p - (\Sigma + I_p)^{-1}\}y_i \qquad (i=1,..., q) .$$

Cf., Efron and Morris (1972, eqn. 2.4).

If we relax now the *a priori* assumption that $\underset{\sim}{\xi} = \underset{\sim}{0}$ and let it be unknown [Efron and Morris (1972, §7)], $\underset{\sim}{\xi}$ is estimated according to eqn. (17) by $\bar{\underset{\sim}{y}} = \frac{1}{p} \overset{p}{\underset{i}{\Sigma}} \underset{\sim}{y}_i$ and hence

$$\underset{\sim}{\beta}^* = (\underset{\sim}{I}_{pq} + \underset{\sim}{I}_q \otimes \underset{\sim}{\Sigma}^{-1})^{-1} (\underset{\sim}{y} + \underset{\sim}{1}_q \otimes \underset{\sim}{\Sigma}^{-1} \bar{\underset{\sim}{y}})$$

or

$$\underset{\sim}{\beta}_i^* = (\underset{\sim}{I}_p + \underset{\sim}{\Sigma}^{-1})^{-1}(\underset{\sim}{y}_i + \underset{\sim}{\Sigma}^{-1}\bar{\underset{\sim}{y}}) = (\underset{\sim}{I}_p + \underset{\sim}{\Sigma}^{-1})^{-1}\{\underset{\sim}{y}_i - \bar{\underset{\sim}{y}} + (\underset{\sim}{I}_p + \underset{\sim}{\Sigma}^{-1})\bar{\underset{\sim}{y}}\}$$

$$= \{\underset{\sim}{I}_p - (\underset{\sim}{I}_p + \underset{\sim}{\Sigma})^{-1}\}(\underset{\sim}{y}_i - \bar{\underset{\sim}{y}}) + \bar{\underset{\sim}{y}} \qquad (i=1,\ldots, q) \quad .$$

Cf. Efron and Morris (1972, eqn. 7.2) and recall that $(k-p-2)^{-1} \underset{\sim}{\tilde{S}}$ there estimates $\underset{\sim}{I}_p + \underset{\sim}{\Sigma}.$

Efron and Morris's more general formulation (1972, §4) requires a modification in our model, i.e., replacing the covariance matrix in (5) by $\underset{\sim}{I}_q \otimes \underset{\sim}{\Gamma}$ and setting $\underset{\sim}{X} = \underset{\sim}{I}_p$, $\underset{\sim}{V} = \underset{\sim}{I}_q$, one obtains

$$\underset{\sim}{\beta}^* = \{\underset{\sim}{I}_q \otimes (\underset{\sim}{\Gamma}^{-1} + \underset{\sim}{\Sigma}^{-1})^{-1}\underset{\sim}{\Gamma}^{-1}\}\underset{\sim}{y}$$

or

$$\underset{\sim}{\beta}_i^* = \{\underset{\sim}{I}_p - \underset{\sim}{\Gamma}(\underset{\sim}{\Gamma} + \underset{\sim}{\Sigma})^{-1}\}\underset{\sim}{y}_i \qquad (i=1,\ldots, q)$$

for the $\underset{\sim}{\xi} = \underset{\sim}{0}$ case [cf. Efron and Morris (1972, eqn. 4.6)] , and

$$\underset{\sim}{\beta}_i^* = \bar{\underset{\sim}{y}} + \{\underset{\sim}{I}_p - \underset{\sim}{\Gamma}(\underset{\sim}{\Gamma} + \underset{\sim}{\Sigma})^{-1}\}(\underset{\sim}{y}_i - \bar{\underset{\sim}{y}}) \qquad (i=1,\ldots, q)$$

for the unknown $\underset{\sim}{\xi}$ case.

4.   RESTRICTED LEAST SQUARES INTERPRETATION

Ridge regression estimators were originally obtained by either minimizing the residual error sum of squares subject to a positive (scalar) constraint on the length of the vector of regression parameters, or by minimizing the length of the parameter vector subject to a positive constraint on the residual error sum of squares. A generalization is possible if the minimization of the residual error sum of squares is performed subject to a constraint on the length, in a prespecified positive metric, of the *differences* between the vector of parameters and an *a priori* given vector of constants. See Haitovsky and Wax (1980). Similarly, multivariate ridge-regression-like estimates can be obtained as a solution to the problem:

$$\min_{\underset{\sim}{\beta}} \{\underset{\sim}{y} - (I_{\underset{\sim}{q}} \otimes \underset{\sim}{X})\underset{\sim}{\beta}\}^T (\underset{\sim}{\Gamma} \otimes I_{\underset{\sim}{n}})^{-1} \{\underset{\sim}{y} - (I_{\underset{\sim}{q}} \otimes \underset{\sim}{X})\underset{\sim}{\beta}\}$$

subject to the constraint

$$(\underset{\sim}{\beta} - 1_{\underset{\sim}{q}} \otimes \bar{\underset{\sim}{\beta}})^T \Omega^{-1} (\underset{\sim}{\beta} - 1_{\underset{\sim}{q}} \otimes \bar{\underset{\sim}{\beta}}) = h$$

yielding

$$\hat{\underset{\sim}{\beta}}*(k, \underset{\sim}{\Omega}^{-1}, \bar{\underset{\sim}{\beta}}) = (\underset{\sim}{\Gamma}^{-1} \otimes \underset{\sim}{X}^T\underset{\sim}{X} + k\Omega^{-1})^{-1}\{(\underset{\sim}{\Gamma}^{-1} \otimes \underset{\sim}{X}^T)\underset{\sim}{y} + k\Omega^{-1}(1_{\underset{\sim}{q}} \otimes \bar{\underset{\sim}{\beta}})\}$$

where $\bar{\underset{\sim}{\beta}}$ is a p × 1 vector of constants, h is a given positive constant, and k > 0 is the Lagrange multiplier. If now instead of fixing $\bar{\underset{\sim}{\beta}}$ one takes it to be (in some sense) a mean vector of the estimated $\underset{\sim}{\beta}_j$'s, one obtains   (9) with $\underset{\sim}{Q}$ replaced by $k\underset{\sim}{Q}$, where $\underset{\sim}{Q}$ depends on how exactly $\bar{\underset{\sim}{\beta}}$ was derived. If, on the other hand, $\bar{\underset{\sim}{\beta}}$ is set to $\underset{\sim}{0}$

$$\hat{\underset{\sim}{\beta}}*(k, \underset{\sim}{\Omega}^{-1}, \underset{\sim}{0}) = (\underset{\sim}{\Gamma}^{-1} \otimes \underset{\sim}{X}^T\underset{\sim}{X} + k\Omega^{-1})^{-1}(\underset{\sim}{\Gamma}^{-1} \otimes \underset{\sim}{X}^T)\underset{\sim}{y} ,$$

i.e., (9) and (10) are obtained with $\underset{\sim}{Q}$ replaced by $k\underset{\sim}{\Omega}^{-1}$.

In a similar manner the ridge regression estimators are obtained from the generalized ridge estimators by setting the *a priori* given vector of constants equal to zero.

The properties of the generalized ridge regression estimators are explored by Haitovsky and Wax (1980).

5. COVARIANCES UNKNOWN

Turning now to the more realistic situation where the covariances are unknown, it is natural to choose maximum likelihood as the estimation principle. One may regard our three-stage hierarchical model as a mixed effect model, where the first stage parameters are assumed random and the second stage parameters are assumed fixed, in view of the widely known fact that classical (BLUE) estimators of fixed effects are the limiting posterior means corresponding to random effects with zero precision matrix (i.e. flat prior). Maximum likelihood estimators for the mixed effect models were reviewed by Harville (1977) for variance component models and were extended by Dempster, Rubin and Tsutakawa (1979) for covariance component models. We follow Haitovsky (1985) who writes the joint probability function, when assuming (14), as proportional to

$$\left| \Gamma \otimes I_n \right|^{-\frac{1}{2}} \left| V \otimes \Sigma \right|^{-\frac{1}{2}} \text{etr}\left\{ -\frac{1}{2}(Y - XB)\Gamma^{-1}(Y - XB)^T - \frac{1}{2}\Sigma^{-1}\Xi V^{-1}\Xi^T \right\} \tag{21}$$

where etr denotes the exponent of the matrix trace, and $\Xi = B - 1_q^T \otimes \xi$ is obtained by "rematricizing" the $\beta - 1_q \otimes \xi$ vector. Maximizing (21) with respect to Γ^{-1}, V^{-1}, Σ^{-1}, B and ξ, the following estimators are obtained:

$$\left. \begin{aligned} \hat{\Gamma} &= \frac{1}{n}(Y - X\hat{B}^*)^T(Y - X\hat{B}^*) \\[1em] \hat{V} &= \frac{1}{p}\hat{\Xi}^T\hat{\Sigma}^{-1}\hat{\Xi} \\[1em] \hat{\Sigma} &= \frac{1}{q}\hat{\Xi}\hat{V}^{-1}\hat{\Xi} \end{aligned} \right\} \tag{22}$$

and $\hat{\beta}^*$ and $\hat{\xi}^*$ are as in (15) and (17), but with the covariance matrices

replaced by their estimators.

There are two things to note here:

a)   in the traditional treatment of the multivariate regression, the matrix of coefficients $\underset{\sim}{B}$ is considered fixed, hence $\underset{\sim}{V}$ and $\underset{\sim}{\Sigma}$ are equal to zero, and $\underset{\sim}{\Gamma}$ is indeed estimated by $\hat{\underset{\sim}{\Gamma}}$ in (22). Likewise, if it is assumed that $\underset{\sim}{\Sigma} = \underset{\sim}{I}_p$ and $\underset{\sim}{\xi} = 0$, i.e., the Brown-Zidek MRRE (see Section 3.3 above), there is no need to estimate $\underset{\sim}{\Sigma}$, while $\underset{\sim}{V}$ is estimated by

$$\hat{\underset{\sim}{V}} = \frac{1}{p} \hat{\underset{\sim}{B}}*^T \hat{\underset{\sim}{B}}*$$
(23)

where $\hat{\underset{\sim}{B}}*$ is a matrix of estimators obtained by "rematricizing" $\hat{\underset{\sim}{\beta}}*$ with $\underset{\sim}{\xi} = \underset{\sim}{0}$ and $\underset{\sim}{\Sigma} = \underset{\sim}{I}_p$. Estimator (23) is the "natural estimator", suggested by Brown and Zidek (1980, eqn. 2.14), apart from some differences in the number of degrees of freedom, which the ML method so notoriously fails to account for. A correction for the number of degrees of freedom is suggested below which will make (23) identical to the Brown-Zidek estimator.

b)   the matrix of order $p \times q$

$$\hat{\underset{\sim}{\Xi}} = \hat{\underset{\sim}{B}}* - \underset{\sim}{1}_q^T \otimes \underset{\sim}{\xi} = \hat{\underset{\sim}{B}}*\{\underset{\sim}{I}_q - (\underset{\sim}{1}_q^T \hat{\underset{\sim}{V}}^{-1} \underset{\sim}{1}_q)^{-1} \hat{\underset{\sim}{V}}^{-1} \underset{\sim}{J}_q\}$$.

is of rank $\leq \min \{p, q-1\}$, since its rows lie in the $q - 1$ dimensional space. (Note that $\underset{\sim}{W}^T = \underset{\sim}{I}_q - (\underset{\sim}{1}_q^T \underset{\sim}{V}^{-1} \underset{\sim}{1}_q)^{-1} \underset{\sim}{V}^{-1} \underset{\sim}{J}_q$ is an idempotent (nonsymmetric) matrix with rank and trace equal to $q - 1$). Consequently, either $\hat{\underset{\sim}{\Sigma}}$ or $\hat{\underset{\sim}{V}}$ is a singular matrix. Alternatively, there are $p(p + 1)/2 + q(q + 1)/2 - 1$ independent parameters to be estimated in $\underset{\sim}{V} \otimes \underset{\sim}{\Sigma}$, but only pq ($< p(p + 1)/2 + q(q + 1)/2 - 1$) entries in $\hat{\underset{\sim}{B}}*$. That is, the structure imposed on $\underset{\sim}{\Omega}$ does not suffice to identify the second stage covariance matrix.

In order to overcome the singularity, a Bayesian solution is suggested.

To this end we specify prior distributions for the covariance matrices. It will be done through the conjugate family, which is the Wishart distribution. Assuming $\underset{\sim}{\Gamma}^{-1}$, $\underset{\sim}{V}^{-1}$ and $\underset{\sim}{\Sigma}^{-1}$ are independent, Wishart distributed with γ, ν and σ degrees of freedom and $\underset{\sim}{G}$, $\underset{\sim}{R}$, and $\underset{\sim}{S}$ matrices, respectively, the following densities (up to a proportionality factor) are obtained:

$$|\underset{\sim}{\Gamma}|^{-\frac{1}{2}(\gamma-q-1)} \; etr\{-\tfrac{1}{2}\underset{\sim}{\Gamma}^{-1}\underset{\sim}{G}\}$$

$$|\underset{\sim}{V}|^{-\frac{1}{2}(\nu-q-1)} \; etr\{-\tfrac{1}{2}\underset{\sim}{V}^{-1}\underset{\sim}{R}\}$$

and

$$|\underset{\sim}{\Sigma}|^{-\frac{1}{2}(\sigma-p-1)} \; etr\{-\tfrac{1}{2}\underset{\sim}{\Sigma}^{-1}\underset{\sim}{S}\} \quad .$$

The joint distribution of all the quantities is obtained by multiplying the joint probability function (21) by the three prior densities above. The integration with respect to $\underset{\sim}{\xi}$ is performed by rewriting the second term in the exponent in (21) as

$$tr\{\underset{\sim}{\Sigma}^{-1}\underset{\sim\sim}{\Xi}\underset{\sim}{V}^{-1}\underset{\sim}{\Xi}^{T}\} = (\underset{\sim}{\beta} - \underset{\sim q}{1} \otimes \underset{\sim}{\xi})^{T}(\underset{\sim}{V} \otimes \underset{\sim}{\Sigma})^{-1}(\underset{\sim}{\beta} - \underset{\sim q}{1} \otimes \underset{\sim}{\xi})$$

$$= (\underset{\sim}{\beta} - \underset{\sim q}{1} \otimes \underset{\sim}{\bar{\xi}})^{T}(\underset{\sim}{V} \otimes \underset{\sim}{\Sigma})^{-1}(\underset{\sim}{\beta} - \underset{\sim q}{1} \otimes \underset{\sim}{\bar{\xi}}) + (\underset{\sim q}{1}^{T}\underset{\sim}{V}^{-1}\underset{\sim q}{1})(\underset{\sim}{\xi} - \underset{\sim}{\bar{\xi}})^{T}\underset{\sim}{\Sigma}^{-1}(\underset{\sim}{\xi} - \underset{\sim}{\bar{\xi}})$$

where $\underset{\sim}{\bar{\xi}} = (\underset{\sim q}{1}^{T}\underset{\sim}{V}^{-1}\underset{\sim q}{1})^{-1}(\underset{\sim q}{1}^{T}\underset{\sim}{V}^{-1} \otimes \underset{\sim p}{I})\underset{\sim}{\beta}$ [the counterpart of (17)] ,
since the cross product term vanishes.

Now the integration with respect to $\underset{\sim}{\xi}$ is straightforward, resulting in the joint posterior density for $\underset{\sim}{\beta}$, $\underset{\sim}{\Gamma}$, $\underset{\sim}{V}$ and $\underset{\sim}{\Sigma}$ which is proportional to:

$$|\underset{\sim}{\Gamma}|^{-\frac{1}{2}(n+\gamma-q-1)} |\underset{\sim}{V}|^{-\frac{1}{2}(p+\nu-q-1)} |\underset{\sim}{\Sigma}|^{-\frac{1}{2}(q+\sigma-p-2)} (\underset{\sim q}{1}^{T}\underset{\sim}{V}^{-1}\underset{\sim q}{1})^{-\frac{1}{2}}$$

$$\times \; etr[-\tfrac{1}{2}\{\underset{\sim}{\Gamma}^{-1}\underset{\sim}{G} + \underset{\sim}{V}^{-1}\underset{\sim}{R} + \underset{\sim}{\Sigma}^{-1}\underset{\sim}{S} + (\underset{\sim}{Y} - \underset{\sim\sim}{XB})\underset{\sim}{\Gamma}^{-1}(\underset{\sim}{Y} - \underset{\sim\sim}{XB})^{T} + \underset{\sim}{\Sigma}^{-1}\underset{\sim\sim}{\Xi}\underset{\sim}{V}^{-1}\underset{\sim}{\Xi}^{T}\}] \qquad (24)$$

where here $\underset{\sim}{\bar{\Xi}} = \underset{\sim}{B} - \underset{\sim q}{1} \textcircled{x} \underset{\sim}{\bar{\xi}}$. The integration with respect to $\underset{\sim}{\xi}$ resulted in the loss of one degree of freedom in the estimation of $\underset{\sim}{\Sigma}$.

Following Lindley and Smith (1972) we derive the modal Bayesian estimators by differentiating the log joint posterior density partially with respect to $\underset{\sim}{B}$, $\underset{\sim}{\Gamma}^{-1}$, $\underset{\sim}{V}^{-1}$ and $\underset{\sim}{\Sigma}^{-1}$ obtaining $\underset{\sim}{\hat{B}}*$ as before (eqn. 15) with $\underset{\sim}{\Gamma}$, $\underset{\sim}{\Sigma}$ and $\underset{\sim}{V}$ replaced by their estimators:

$$\underset{\sim}{\hat{\Gamma}} = (n + \gamma - q - 1)^{-1}\{\underset{\sim}{G} + (\underset{\sim}{Y} - \underset{\sim\sim}{X}\underset{\sim}{\hat{B}}*)^{T}(\underset{\sim}{Y} - \underset{\sim\sim}{X}\underset{\sim}{\hat{B}}*)\}$$

$$\underset{\sim}{\hat{\Sigma}} = (q + \sigma - p - 2)^{-1}\{\underset{\sim}{S} + \underset{\sim\sim}{\hat{\Xi}}\underset{\sim\sim}{\hat{V}}^{-1}\underset{\sim}{\hat{\Xi}}^{T}\}$$

$$\underset{\sim}{\hat{V}} = (p + \nu - q - 1)^{-1}\{(\underset{\sim q}{1}^{T} \underset{\sim}{\hat{V}}^{-1} \underset{\sim q}{1})^{-1}\underset{\sim q}{J} + \underset{\sim}{R} + \underset{\sim}{\hat{\Xi}}^{T}\underset{\sim}{\hat{\Sigma}}^{-1}\underset{\sim}{\hat{\Xi}}\}$$

$$= (p + \nu - q - 1)^{-1}[\{(p + \nu - q - 2)\underset{\sim q}{1}^{T}(\underset{\sim}{R} + \underset{\sim}{\hat{\Xi}}^{T}\underset{\sim}{\hat{\Sigma}}^{-1}\underset{\sim}{\hat{\Xi}})^{-1}\underset{\sim q}{1}\}^{-1}\underset{\sim q}{J} + \underset{\sim}{R} + \underset{\sim}{\hat{\Xi}}^{T}\underset{\sim}{\hat{\Sigma}}^{-1}\underset{\sim}{\hat{\Xi}}]$$

The estimators for $\underset{\sim}{\Sigma}$ and $\underset{\sim}{V}$ are natural in view of the following theorem:

<u>Theorem</u>: Let the $p \times q$ random matrix $\underset{\sim}{B}$ be multivariate normal with a common mean vector $\underset{\sim}{\xi}$ and variance $\underset{\sim}{V} \textcircled{\otimes} \underset{\sim}{\Sigma}$, $\underset{\sim}{V}$, $\underset{\sim}{\Sigma} > 0$, i.e., $\underset{\sim}{\beta} = vec\{\underset{\sim}{B}\} \sim N(\underset{\sim q}{1} \textcircled{\otimes} \underset{\sim}{\xi} , \underset{\sim}{V} \textcircled{\otimes} \underset{\sim}{\Sigma})$. Then, if $q \geq p + 1$

$$\underset{\sim\sim}{\Xi}\underset{\sim\sim}{V}^{-1}\underset{\sim}{\Xi}^{T} \sim W(\underset{\sim}{\Sigma}, p, q) ;$$

and if $p > q + 1$, then

$$\underset{\sim}{\Xi}^{T}\underset{\sim}{\Sigma}^{-1}\underset{\sim}{\Xi} \sim W(\underset{\sim}{V}, q, p)$$

where here $\underset{\sim}{\Xi} = \underset{\sim}{B} - \underset{\sim q}{1}^{T} \textcircled{\otimes} \underset{\sim}{\xi}$ is the matrix of columnwise deviations from the common mean factor.

<u>Proof</u>: vec $\{\underset{\sim}{\Xi}\} \sim N(0, \underset{\sim}{V} \otimes \underset{\sim}{\Sigma})$. Diagonalize $\underset{\sim}{V}$ by the orthogonal matrix $\underset{\sim}{P}$, i.e., $\underset{\sim}{P}^T \underset{\sim}{V} \underset{\sim}{P} = \underset{\sim}{\Lambda}$, a diagonal matrix, and define $\underset{\sim}{z} = (\Lambda^{-\frac{1}{2}} \otimes \underset{\sim p}{I})(\underset{\sim}{P}^T \otimes \underset{\sim p}{I})$ vec $\{\underset{\sim}{\Xi}\} = (\Lambda^{-\frac{1}{2}}\underset{\sim}{P}^T \otimes \underset{\sim p}{I})$ vec $\{\underset{\sim}{\Xi}\}$. Clearly, $\underset{\sim}{z} \sim N(0, \underset{\sim q}{I} \otimes \underset{\sim}{\Sigma})$, hence $\underset{\sim\sim}{ZZ}^T \sim W(\Sigma, p, q)$, where $\underset{\sim}{z} = $ vec $\{\underset{\sim}{Z}\}$, but $\underset{\sim\sim}{ZZ}^T = \underset{\sim\sim}{\Xi P}^T \Lambda^{-1} \underset{\sim\sim}{P \Xi}^T = \underset{\sim\sim}{\Xi V}^{-1} \underset{\sim}{\Xi}^T$. Similarly for $p \geq q + 1$.

<div align="right">Q.E.D.</div>

<u>Corollary</u>: For $q \geq p + 1$, $(\underset{\sim}{\Xi} \underset{\sim}{V}^{-1} \underset{\sim}{\Xi}^T)^{-1} \sim W^{-1}(\underset{\sim}{\Sigma}^{-1}, p, p+q+1)$ and hence, $(q-p-1) E\{(\underset{\sim\sim}{\Xi V}^{-1}\underset{\sim}{\Xi}^T)^{-1}\} = \underset{\sim}{\Sigma}^{-1}$. If $p \geq q+1$, then $(\underset{\sim}{\Xi}^T\underset{\sim}{\Sigma}^{-1}\underset{\sim}{\Xi})^{-1} \sim W^{-1}(\underset{\sim}{V}^{-1}, q, p+q+1)$ and hence $(p - q - 1) E\{(\underset{\sim}{\Xi}^T_{\sim}\underset{\sim}{\Sigma}^{-1}\underset{\sim}{\Xi})^{-1}\} = \underset{\sim}{V}^{-1}$.

The correction of the number of degrees of freedom in the estimation of $\underset{\sim}{\Gamma}$ differs from that in least squares theory; in the latter the divisor is $(n-p-1)$.

This brings up a related problem. O'Hagan (1976) proves that in joint posterior densities as in our (24), modal estimators obtained from the joint density will differ, often substantially, from that of the marginal densities; the <u>main</u> difference being the number of degrees of freedom, i.e., the divisors of the estimated variances and covariances. Thus, the problem of which mode to choose can be problematic, especially when the number of regression coefficients to be estimated is large relative to the sample size. However, note that this issue affects the estimation of the vector of regression coefficients only indirectly, via the estimation of the variances and covariances. It is our opinion that the question of "which mode" cannot be ignored, nor overstated in situations like ours, where the main objective is the estimation of the regression coefficients rather than the covariances. A few trials made with our model and the example in O'Hagan (1976) indeed support this contention.

The Bayesian modal estimates can be calculated using an iterative procedure: initial values for $\underset{\sim}{\hat{B}}*$, $\underset{\sim 0}{\hat{B}}*$, are computed using least squares multivariate regression estimates (eqn. 3), from which the initial values for $\underset{\sim}{\hat{\Gamma}}$ and for $\underset{\sim}{\xi}*$ (eqn. 17) are computed, using for the latter a simple average of the q columns of $\underset{\sim 0}{\hat{B}}*$ (i.e., letting $\underset{\sim}{\hat{V}} = \underset{\sim q}{I}$ in eqn. 17). Next, $\underset{\sim 0}{\hat{\Xi}}$ is computed, from which

$\hat{\underset{\sim}{\Sigma}}_0$ is obtained after suppressing $\tilde{\underset{\sim}{Y}}$ to $\tilde{\underset{\sim}{X}}_q$. Having computed $\hat{\tilde{\underset{\sim}{\Xi}}}_0$ and $\hat{\underset{\sim}{\Sigma}}_0$,

$\hat{\underset{\sim}{Y}}_0$ can be calculated, completing the first round, from which iterations can start.

Another possible iterative solution is the E-M algorithm. See Dempster et al. (1979)

6. APPLICATION TO SCOTTISH ELECTION.

Merely as an illustration, our method is applied to the Scottish election of

October 1974 in forecasting the final results on the basis of partial returns.

The forecasting model is that of Brown and Zidek (1979), which was used as purely illus-

trative as well. It is a modified version of a model actually used by Brown and

Payne (1975) for predicting the February 1974 British General election.

6.1 Description of the Data: The data are taken from Brown and Zidek (1979) and

are described there in detail. The forecasting problem is to forecast the winner

of the undeclared constituencies on the basis of the declared constituencies.

The forecasting model is a multivariate regression with the dependent variables

being the change in the party's share in the number of electorates in each consti-

tuency, relative to previous election. The independent variables are the party's

shares of votes in the previous election plus three dummy variables. The four

main parties who participated in the February and October 1974 elections were the

Conservative, Labour, Liberal, and Nationalist parties. They will be denoted by

$i = 1, 2, 3, 4$ in that order. Thus, let W_i and Z_i denote the October 1974

and February 1974 votes for the ith party, respectively, and E the electorate

figure, which stayed practically the same in the two elections. Then

$$X_i = Z_i/E \quad \text{and} \quad Y_i = W_i/E - X_i \quad\quad i=1,\ldots,4$$

denote the (non dummy) explanatory variables and the dependent variables, respectively.

The 3 dummy variables are

$$X_5 = \begin{cases} 0.5 & \text{Liberal intervenes, i.e. } w_3 > 0 \text{ and } X_3 = 0 ; \\ 0 & \text{otherwise;} \end{cases}$$

$$X_6 = \begin{cases} 0.5 & R = 5, 6 ; \\ 0 & \text{otherwise;} \end{cases}$$

$$X_7 = \begin{cases} 0.5 & \text{Labour or Nationalist top party in February 1974} \\ & \text{and } |X_2 - X_4| \leq 0.2 , \\ 0 & \text{otherwise.} \end{cases}$$

where R is a categorical variable defining region where

1 =	GLASGOW;	2 =	Rest of Clydeside conurbation;
3 =	EDINBURGH;	4 =	Rest of industrial centres;
5 =	HIGHLANDS;	6 =	Rest of Scotland.

The value of 0.5 employed in these three dummy variables X_5, X_6, X_7 is somewhat arbitrary but was chosen so that a priori coefficients for all seven variables would be of a similar magnitude.

Thus, $q = 4$ and $p = 8$, including the unit vector for the intercepts.

6.2 The Choice of Priors in the multivariate regression with a covariance structure (14) is being studied now and will be reported in a subsequent paper. The choice of the priors in the present paper motivated the research and, hence is partially based on it.

There are three covariance matrices $\underset{\sim}{G}$, $\underset{\sim}{R}$, and $\underset{\sim}{S}$ and three scalars γ, ν, and σ, which require prior specification. Vague priors are assumed for the regression error covariance, i.e., $\underset{\sim}{G}$ and γ are set to zero. The remaining prior values are related to the regression coefficients and their assessment should be based on their careful interpretation.

The forecasting equations discussed in the last section can be viewed as including an implied explanatory variable for the eligible voters who did not participate in the February election. Thus, rewriting slightly the first equation as:

$$\frac{W_{1j} - Z_{1j}}{E} = \beta_{01} + \beta_{11} \frac{Z_{1j}}{E} + \beta_{21} \frac{Z_{2j}}{E} + \beta_{31} \frac{Z_{3j}}{E} + \beta_{41} \frac{Z_{4j}}{E}$$

$$+ \gamma_1 \frac{E - Z_{1j} - Z_{2j} - Z_{3j} - Z_{4j}}{E} + \text{dummies} + \text{error} \qquad (j=1,\ldots,71)$$

or, collecting terms:

$$\frac{W_{1i}}{E} = (\beta_{01} + \gamma_1) + (1 + \beta_{11} - \gamma_1) \frac{Z_{1j}}{E} + (\beta_{21} - \gamma_1) \frac{Z_{2j}}{E} + (\beta_{31} - \gamma_1) \frac{Z_{3j}}{E}$$

$$+ (\beta_{41} - \gamma_1) \frac{Z_{4j}}{E} + \text{dummies} + \text{error}$$

and similarly for the three other parties, it can be seen that the coefficients
measure the shifts of votes from last election to the present, adjusted for the
fraction of abstainer, attributable to the various party supporters in the last
election. Thus, for instance, the negative β_4 in the "Conservative equation"
signifies a shift of those who voted Nationalist in the last election away from
the Conservative party in the present election, over and above the shift away
from the conservative party of nonvoters in the last election. If we disregard
now the problem of abstainers, obviously each of the regression coefficients
across equations must sum up to zero: in the absence of nonvoters the real-
location of last election voters for each party will result in compensating
shifts in the present election. In the absence of more specific information the
componentwise deviation of the $\underset{\sim}{\beta}$ -vectors around $\underset{\sim}{\xi}$ is justified in the present
application. Moreover, $\underset{\sim}{\xi} = \underset{\sim}{0}$ should be imposed if all eligible voters vote,
or if it is believed that there will be no (non-proportional) shifts among the
last election abstainers, or nearly so. Otherwise, let $\underset{\sim}{\xi}$ be freely estimated.

Now, as the case happened to be, competition was particularly fierce
between the Conservative and Nationalist parties, but less so between the Labour

and Nationalist. In the final result the Conservatives lost in the October election 28.35% of their voters in February, the Labour lost 5.39%, the Nationalist gained 32.61%, while the Liberals remained the same. The participation rate dropped from 77.87% in February to 74.60% in October.

Assuming that the nature of competition between the Conservatives and the Nationalists on the one hand, and the Labour and the Nationalists on the other, was known before the election, we have chosen a priori values for R to reflect a prior belief that $\text{corr}(\beta_{i1}, \beta_{i4}) = -0.7$ and $\text{corr}(\beta_{i2}, \beta_{i4}) = -0.5$; all other correlations between the β 's across equations were set to zero. Furthermore, in order to be in line with $\Xi^T \Sigma^{-1} \Xi$, to which R is added, the diagonal elements of R were set to 0.1. Thus, $R = 0.1\, I_4 + R^*$, where R^* is a 4×4 matrix of zeros except that -0.07 appears in the $(1,4)$ and $(4,1)$ positions and -0.05 appears in the $(2,4)$ and $(4,2)$ positions.

As for the choice of S, it was felt that the Conservatives benefited from the Liberal intervention, but were adversely affected by the last two dummy variables, and hence we set $S = I_8 + S^*$ where S^* is an 8×8 matrix of zeros except for $\frac{1}{2}$ in the $(2,6)$ and $(6,2)$ positions, $-\frac{1}{3}$ in the $(2,7)$ and $(7,2)$ positions, and $-\frac{1}{2}$ in the $(2,8)$ and $(8,2)$ positions. In the six explanatory variable case the non-zero off-diagonal elements were slightly increased to 0.7 and -0.5 respectively, while the last row and column were deleted. Finally, $\nu = 10$ and $\sigma = 20$ produced divisors for the estimates of V and Σ of 13 and 14 respectively, and judged to reflect our confidence in our choice of R and S.

6.3 The Forecasting Performance Criteria: It is not obvious that forecasting the winner is best achieved by first forecasting the vote shares from which the winner can be picked out. However, there are countries with proportional representation where the *number* of votes matters, and, hence the multivariate

regression might be better suited there. We nevertheless follow here, merely as an illustration, Brown and Zidek (1979) and use the first 25 declared constituencies to forecast the results of the remaining $71 - 25 = 46$ constituencies, using the multivariate regression set-up with n , number of observations, equal to 25 . Using these 25 observations we first estimate $\hat{\underset{\sim}{\beta}}{}^{*}$ from which we can predict the Y_i 's and eventually the number of votes to be cast in each constituencies for the four parties. Hence,

$$\hat{W}_{ij} = (\hat{Y}_{ij} + X_{ij})E_i \qquad\qquad i = 26,\ldots,71; \quad j = 1,\ldots,4 \ .$$

Three criteria for goodness of prediction were chosen.

$$SD = \left\{ tr(\underset{\sim}{W} - \hat{\underset{\sim}{W}})\,(\underset{\sim}{W} - \hat{\underset{\sim}{W}})^T/(4 \times 46) \right\}^{\frac{1}{2}}$$

$$PRED = \text{number of incorrect predictions of winning party}$$

$$RSD = \left\{ \sum_i \sum_i (W_{ij} - \hat{W}_{ij})^2/(W_{ij}^2 \times 4 \times 46) \right\}^{\frac{1}{2}}$$

where $\underset{\sim}{W}$ is the 46×4 matrix with typical element W_{ij} .

The first two measures were used by Brown and Zidek (1979). The first is just the square root of the mean squared prediction error and it indeed estimates the standard deviation of prediction.

The second records the number of times the prediction of winners proved to be wrong. It is a very crude measure, but has the appeal of simplicity. It does not however take account of the closeness of a particular contest.

The last is the square root of the *relative* mean squared prediction error. It standardizes the prediction errors, in some sense, for ease of prediction. Brown and Zidek also used a third measure which we feel will a priori bias the results in favour of their method, since it was constructed to be in line with the loss function they used in developing their estimates, and thus is rejected here in favour of more objective measures.

6.4 Results: Two variants of the linear hierarchical model were tried and compared to the multivariate regression least-squares (maximum likelihood) estimates (eqn. 3) and to the Brown-Zidek ridge estimates (eqn. 4 with $\underset{\sim}{K}$ estimated by $\hat{\underset{\sim}{\Gamma}}^{-1}\underset{\sim}{V}$, where $\hat{\underset{\sim}{\Gamma}}$ and $\hat{\underset{\sim}{V}}$ are given on p. 13 below except that in the latter the divisor is p-q-1). The two variants are with $\underset{\sim}{\xi}$ vague and with $\underset{\sim}{\xi}$ a prior set to zero. The three estimates are reported in Tables I - III. The predictive record of all those three estimates and that of Brown-Zidek are summarized in Table IV.

TABLE IV

PREDICTION RECORDS OF THE DIFFERENT ESTIMATES

	SD	RSD	PRED*
Least Squares (M.L.)	1551.03	0.2972	7(0)
Ridge (Brown-Zidek)	1315.78	0.2189	6(1)
Linear Hierarchical Model			
$\underset{\sim}{\xi} = \underset{\sim}{0}$	1315.62	0.2271	6(1)
$\underset{\sim}{\xi}$ vague	1423.97	0.2651	5(1)

* The numbers in parentheses denote the number of wrong "predictions" of the winner party in the 25 declared constituencies.

Interestingly, the best prediction records is for the Ridge and the linear hierarchical model with $\underset{\sim}{\xi} = \underset{\sim}{0}$, which is closely related to Brown-Zidek's Ridge, since the latter essentially presupposes exchangeability about zero. See section 3. The reason for the more restrictive cases, namely $\underset{\sim}{\xi} = \underset{\sim}{0}$, to perform better might be due to the fact that the "sample" of the 25 declared

TABLE I

LEAST SQUARES (ML) ESTIMATES OF THE MATRIX OF COEFFICIENTS OF THE MULTIVARIATE
REGRESSION MODEL AND THEIR RATIO TO THE CORRESPONDING STANDARD DEVIATIONS

	Equation 1		Equation 2		Equation 3		Equation 4	
	$\hat{\beta}$	$\hat{\beta}$/Std. Dev.	$\hat{\beta}$	$\hat{\beta}$/Std. Dev.	$\hat{\beta}$	$\hat{\beta}$/Std. Dev.	$\hat{\beta}$	$\hat{\beta}$/Std. Dev.
β_0	-0.0119	0.22	-0.0007	0.01	0.0367	0.69	-0.1501	2.06
β_1	-0.0627	1.04	-0.0911	2.32	-0.0303	0.51	0.1793	2.21
β_2	-0.0297	0.32	-0.0212	0.36	-0.0712	0.78	0.3193	2.54
β_3	-0.1070	0.79	0.2768	3.13	-0.2864	2.15	0.4736	2.59
β_4	-0.2202	1.88	-0.0768	1.01	-0.0854	0.74	0.3321	2.10
β_5	-0.0160	0.63	0.0276	1.67	0.0763	3.06	-0.0250	0.73
β_6	0.0315	0.92	0.0307	1.37	0.0028	0.08	-0.0146	0.31
β_7	-0.0061	0.36	-0.0025	0.22	0.0042	0.25	-0.0278	1.20

TABLE II

LINEAR HIERARCHICAL MODEL ESTIMATES OF THE MATRIX OF COEFFICIENTS OF THE
MULTIVARIATE REGRESSION MODEL AND THEIR RATIO TO THE CORRESPONDING STANDARD DEVIATIONS

	Equation 1		Equation 2		Equation 3		Equation 4		ξ	
	β^*	$\dfrac{\beta^*}{\text{Std. Dev.}}$	β^*	$\dfrac{\beta^*}{\text{Std. Dev.}}$	β^*	$\dfrac{\beta^*}{\text{Std. Dev.}}$	β^*	$\dfrac{\beta^*}{\text{Std. Dev.}}$	ξ^*	$\dfrac{\xi^*}{\text{Std. Dev.}}$
β_0	-0.0768	3.75	-0.0234	1.22	-0.0486	2.33	0.0364	1.56	-0.0183	0.47
β_1	-0.0254	0.97	-0.0338	1.45	0.0301	1.13	-0.0124	0.38	-0.0198	0.28
β_2	0.0452	1.18	0.0274	0.79	0.0619	1.59	0.0606	1.33	0.0477	0.94
β_3	0.0861	1.45	0.1858	3.48	-0.0450	0.74	0.1966	2.73	0.1474	2.87
β_4	-0.0721	1.52	-0.0533	1.22	-0.0157	0.32	-0.0300	0.54	-0.0485	0.84
β_5	-0.0022	0.16	0.0025	0.21	0.1024	7.00	-0.0184	0.95	-0.0026	0.05
β_6	0.0135	0.69	0.0330	2.09	0.0162	0.81	0.0136	0.51	0.0184	0.34

TABLE III

LINEAR HIERARCHICAL ESTIMATES OF THE MATRIX OF COEFFICIENTS OF THE MULTIVARIATE
REGRESSION AND THEIR RATIOS TO THEIR CORRESPONDING STANDARD DEVIATIONS

$$\underset{\sim}{\xi} = \underset{\sim}{0}$$

	Equation 1		Equation 2		Equation 3		Equation 4	
	β^*	β^*/Std. Dev.	β^*	β^*/Std. Dev.	β^*	β^*/Std. Dev.	β^*	β^*/Std. Dev.
β_0	-0.0595	5.73	-0.0036	0.41	-0.0346	3.06	0.0574	4.99
β_1	-0.0166	0.92	-0.0208	1.35	0.0254	1.30	0.0083	0.41
β_2	0.0022	0.09	-0.0140	0.71	0.0218	0.87	0.0034	0.14
β_3	-0.0038	0.14	0.0236	1.01	-0.0538	1.79	0.0203	0.73
β_4	-0.0264	0.99	-0.0096	0.42	0.0083	0.28	0.0135	0.50
β_5	-0.0232	1.96	-0.0247	2.49	0.0935	7.60	-0.0345	2.17
β_6	-0.0075	0.47	0.0103	0.78	-0.0011	0.07	-0.0047	0.25
β_7	-0.0112	0.82	-0.0000	0.00	0.0084	0.59	-0.0135	0.78

constituencies, on the basis of which the regression parameters were estimated,
is not a random sample, and hence might bias the forecast in the same systematic
way which determines the speed of declaration. The added restriction might
have the effect of correcting for that bias.

Finally, the pattern of convergence of the regression coefficient is
traced throughout the iterations and depicted in Figures I and II. For all
practical purposes, convergence is achieved in four or five steps. It continued
to iterate to the 11th iteration only because our convergence criterion is strict.

7.   A CONCLUDING REMARK

It was shown that the linear hierarchical model yields a richer family
of estimates than the Brown-Zidek MRRE, but at the cost of requiring more
information in order to identify all the necessary parameters in the case of
unknown covariances. This results from the fact that Brown-Zidek's MRRE is an
Empirical Bayes type estimate and hence all the parameters can be estimated
from the sample itself. Ours, on the other hand, is a Bayes type estimate
which requires some prior knowledge. Hence, the relative merits of both
estimates must take into account the availability and reliability of the extra
information required for the implementation of the latter.

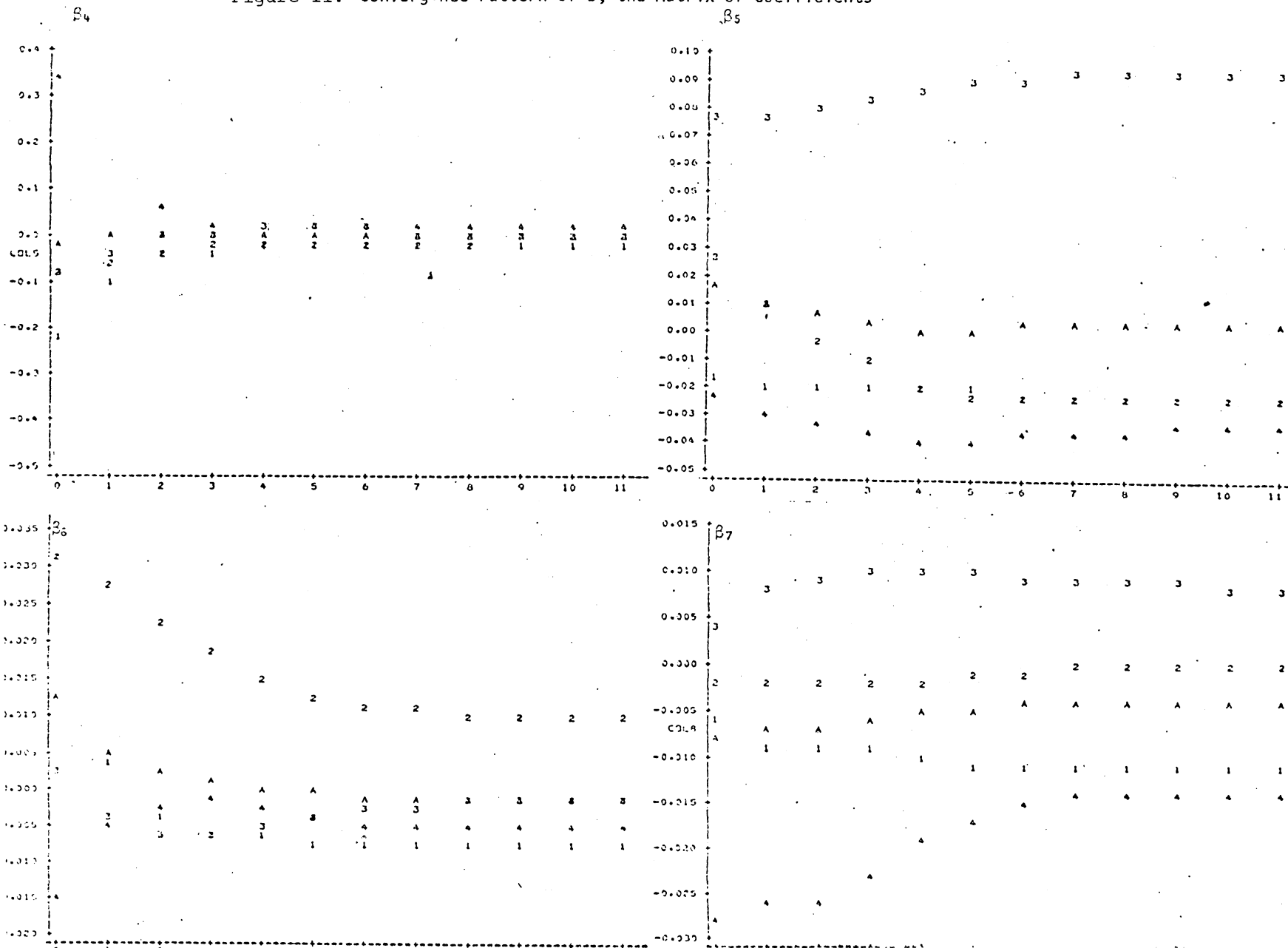Figure I: Convergence Pattern of B, the Matrix of Coefficients

Figure II: Convergence Pattern of B, the Matrix of Coefficients

ACKNOWLEDGEMENTS

REFERENCES

Brown, P.J. and Payne, C. (1975). Election night forecasting (with discussion). J.R. Statist. Soc. A. 138, 463-483.

Brown, P.J. and Zidek, J.V. (1979). Multivariate ridge regression with unknown covariance matrix. T.R. No. 79-11, Dept. of Mathematics, University of British Columbia, Vancouver, Canada.

Brown, P.J. and Zidek, J.V. (1980). Adaptive multivariate ridge regression. Ann. Statist., 6, 64-74.

Brown, P.J. and Zidek, J.V. (1982). Multivariate regression shrinkage estimators with unknown covariance matrix. Scand. J. Statist., 9, 209-215.

Dempster, A.P., Rubin, D.B. and Tsutakawa, R.K. (1979). Estimation in covanriance components models. J. Am. Statist. Assoc., 76, 341-353.

Efron, Bradely and Morris, Carl (1972). Empirical Bayes on vector observations: An extension of Stein's method. Biometrika, 59, 2, 335-347.

Fearn, T. (1975). A Bayesian approach to growth curves. Biometrika, 62, 89-100.

Haitovsky, Y. (1979). The linear hierarchical model: I. Estimation. Modelling Research Group, Dept. of Economics, USC, MRG 7902.

Haitovsky, Y. (1985). The linear hierarchical model and its applications in econometric analysis. In: Bayesian Inference and Decision Techniques with applications: Essays in Honor of Bruno de Finetti (ed. P.K. Goel & Arnold Zellner), forthcoming.

Haitovsky, Y. and Wax, Y. (1980). Generalized ridge regression, least squares with stochastic prior information and Bayesian estimators. Applied Mathematics and Computation, 7, 125-154.

Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. J. Am. Statist. Assoc., 72, 320-338.

Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: biased estimation for nonorthogonal problems. Technometrics, 12, 55-67.

Lindley, D.V. and Smith, A.F.M. (1972). Bayes estimates for the linear model (with discussion). J.R. Statist. Soc., B, 34, 1-41.

O'Hagan, A. (1976). On posterior joint and marginal modes. Biometrika, 63, 329-333.

Sclove, S.L. (1971). Improved estimation of parameters in multivariate regression. Sankhya, A, 33, 61-66.

Thisted, Ronald A. (1970). Ridge Regression, Minimax Estimation and Empirical Bayes Methods. PhD thesis, Stanford Un., Dept. of Statistics.

Zellner, Arnold (1962). An efficient method of estimating seemingly unrelated regressions and test for aggregation bias. J. Am. Statist. Assoc., 57, 348-368.