IN PURSUIT OF A SUITABLE ALTERNATIVE TO LEAST SQUARES ESTIMATION IN NORMAL LINEAR MODELS

A THESIS PRESENTED IN PARTIAL FULFILMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN STATISTICS
AT MASSEY UNIVERSITY

ROBERT HUGH FLETCHER

1979

ABSTRACT

A search for an estimator of β in the Normal Linear Model which has better mean squared error properties than the usual least squares estimator is undertaken. The properties of some classical techniques such as restricted least squares, which includes the selection of a subset of the independent variables, are examined, along with more recent techniques such as ridge regression and Bayesian estimators. Most of these can be shown analytically to improve over least squares only when the true parameter vector β is in some subspace of the parameter space. Empirical Bayes estimators are in general difficult to handle analytically, and so several of these are studied by Monte Carlo methods. A particular modification of one of these empirical Bayes estimators is found to improve over least squares over a large region of the parameter space, and it's use is demonstrated on a small data set. Some suggestions for further improvement of this estimator are given and some techniques for further study of estimators by Monte Carlo methods are recommended.

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Dr. Richard Brook for his guidance and encouragement during the preparation of this thesis, and also my wife Carol, for bearing with me over the last few months.

TABLE OF CONTENTS

		PAGE
	ACKNOWLEDGEMENT	
	TABLE OF CONTENTS	
	LIST OF TABLES	
1.	INTRODUCTION	1
2.	BIASED ESTIMATORS	3
2.1	Linear Estimators	3
2.2	Non-Linear Estimators	3
2.21	Restricted Least Squares or Preliminary Test Estimators	4
2.211	Biased Linear Estimators as Restricted Least Squares Estimators	4
2.212	Preliminary Test Estimators	7
2.22	Incorporation of Prior Information	8
2.221	Ridge Regression as a Bayesian Estimator	10
2.222	Another "Bayesian" Estimator	10
2.223	Numerical Example	13
2.224	Bayesian Estimator for Unscaled y	14
3.	SIMULATION OF THE mse PERFORMANCE OF THE VARIOUS ESTIMATORS	17
3.1	Experimental Design for the Simulation Study	17
3.11	Method	18
3.2	Analysis of the Relative Mean Squared Errors From the Simulation Study	19
3.3	Results of Study 2	22
3.4	Testing the Size and Direction of γ	25
3.41	Testing the Direction of γ	25
3.42	Testing the Length of γ	28
3.43	Modified Estimator	30
3.44	Performance of the Modified Estimator	33
3.5	Mean Squared Error for Non-Stochastic Ridge Regression	31
3.51	Modified Estimator	33
3.52	Performance of the Modified Estimator	33
3.53	Discussion of TABLE 3	35
3.6	Recommended Estimator	35
3.61	Recommended Estimation Rule	36

PAGE

4. NUMERICAL EXAMPLE 37
4.1 Analysis 37
4.2 Pass 1 38
4.21 Interpretation of TABLE 4 40
4.211 Variable Selection 40
4.3 Interpretation of TABLE 5 41
4.4 Interpretation of TABLE 6 42
4.5 Conclusions 43

5. SUMMARY AND DISCUSSION 45

6. APPENDIX 47
6.1 Eigenvalues and True Coefficients γ Used in the Simulation Studies 47
6.2 Random Unit-Normal Generator 49

BIBLIOGRAPHY

LIST OF TABLES

		PAGE
TABLE 1	Relative Mean Squared Errors	20
TABLE 2a.	mse Performance of g4 Relative to Least Squares	23
TABLE 2	Relative mse's at $\gamma \propto u$	24
TABLE 3	Relative Mean Squared Errors	34
TABLE 3a.	Percentage of Times Least Squares Was Used	34
TABLE 4	Summary Statistics From Pass 1	39
TABLE 5	Summary Statistics From Pass 2	41
TABLE 6	Summary Statistics From Pass 3	42
TABLE 7	Eigenvalues and Coefficients Used in the Simulation Studies	47

1. INTRODUCTION

This thesis is concerned with estimators of β in the Normal Linear Model,

$$y = X\beta + \xi$$

where y is an nxl vector of observed variables, X is an nxp matrix of known constants, β is an unobservable pxl vector of coefficients and ξ is an nxl vector of unobservable random errors assumed to be independently and identically Normally distributed with zero mean and constant, but generally unknown variance σ^2. This is written

$$\xi \sim N(0, \sigma^2 I).$$

The more general case where $\xi \sim N(0, \sigma^2 V)$ will not be explicitly discussed here since, if V is of full rank, the model can be reparameterised to conform to the simpler Normal Linear Model.

It is well known that among unbiased estimators of β, the least squares estimate

$$b_0 = (X^T X)^{-1} X^T y$$

has minimum variance. However this does not guarantee that the variance of the least squares estimate will be small and it is for this reason that some biased estimation techniques are considered here. The criterion adopted as a measure of the goodness of an estimator will be mean squared error , mse or MSE, where for a particular estimator b,

$$\text{mse}(b) = E(b - \beta)^T (b - \beta)$$

and

$$\text{MSE}(b) = E(b - \beta)(b - \beta)^T.$$

E has the usual meaning of "the expected value of". It is hoped to find an estimator which has good mean squared error properties when

compared with least squares.

Various biased estimators, including some commonly used variants of least squares, are discussed in Chapter 2. The emphasis is on their mean-squared error performance over different regions of the parameter space for β and σ^2. Then in Chapter 3, several simulation experiments are reported. These experiments investigate the mean-squared error performance of two Stein-type estimators and several stochastic ridge estimators over a wide range of the parameters. As the experiments progress a heuristically modified stochastic ridge estimator is developed which appears to have good mean-squared error properties. A numerical example demonstrating the use of this estimator is given in Chapter 4. Finally a summary, including suggestions for further study of stochastic ridge estimators and a discussion of several aspects of simulation studies and mean-squared error performance of estimators is presented in Chapter 5.

2. BIASED ESTIMATORS

2.1 Linear Estimators:

For any fixed p.s.d. matrix Z, I - Z p.s.d., the estimator Zb_0 of β in the Normal Linear Model $y \sim N(X\beta, \sigma^2 I)$ where b_0 is the usual least squares estimator , has mean squared error matrix

$$MSE(Zb_0) = \sigma^2 Z(X^T X)^{-1} Z^T + (Z - I)\beta\beta^T(Z - I)^T.$$

Now $$MSE(b_0) = \sigma^2 (X^T X)^{-1}$$

and hence $MSE(b_0) - MSE(Zb_0)$ is positive semi-definite if, and only if

$$\beta^T(Z - I)^T[(X^T X)^{-1} - Z(X^T X)^{-1} Z^T]^{-1}(Z - I)\beta \leq \sigma^2$$

(by the appendix in R. W. Farebrother (1976)). That is to say Zb_0 has better mean squared error properties than b_0 only for β and σ^2 in a certain region of the parameter space. Alternatively, a suitable choice of Z for improved mean squared error depends on the unknown β and σ^2 if we wish to guarantee reduced mean squared error. Therefore, for fixed or non-stochastic Z, we can only hope to improve uniformly upon b_0 if we have some prior knowledge about β and σ^2. If we have no prior knowledge then we can have no guarantee that any biased linear estimator will do better (or worse) than b_0. Under this situation of no prior knowledge it seems that we must either use b_0 or look to non-linear estimators.

2.2 Non-Linear Estimators:

Under this heading are included the Stein type estimators of the form $(I - A/(b_0^T C b_0))b_0$ where A and C are pxp positive definite matrices, and estimators of the form Zb_0 where the matrix Z is chosen after inspection of the data. The Stein estimators are

already well documented, so we shall have a brief look at some "data-chosen" linear estimators.

2.21 Restricted Least-Squares or Preliminary Test Estimators:

In the Normal Linear Model the unrestricted least-squares estimator is b_0 and the restricted least-squares estimator (subject to the restriction $H\beta = h$, where H is mxp of rank m) is

$$b_{RLS} = b_0 - S^{-1}H^T(HS^{-1}H^T)^{-1}(Hb_0 - h)$$

Now $$MSE(b_{RLS}) = \sigma^2 ZS^{-1}Z^T + (Z - I)\beta\beta^T(Z - I)^T + u\beta^T(Z - I)^T + (Z - I)\beta u^T + uu^T$$

where $Z = I - S^{-1}H^T(HS^{-1}H^T)^{-1}H$

$u = S^{-1}H^T(HS^{-1}H^T)^{-1}h$

and $S = X^TX$.

That is, $MSE(b_{RLS}) = \sigma^2 ZS^{-1}Z^T + [(Z - I)\beta + u][(Z - I)\beta + u]^T$

and therefore $MSE(b_0) - MSE(b_{RLS})$ is positive definite

$$\iff [(Z - I)\beta + u]^T[S^{-1} - ZS^{-1}Z^T]^{-1}[(Z - I)\beta + u] < \sigma^2$$

or, in this case

$$\iff [H\beta - h]^T(HS^{-1}H^T)^{-1}[H\beta - h] < \sigma^2$$

which is only true for certain values of β and σ^2.

2.211 Biased Linear Estimators as Restricted Least-Squares Estimators:

Consider estimators of the form Zb_0 where Z commutes with X^TX, i.e. has the same eigenvectors as X^TX, and Z has all it's eigenvalues in the range (0,1). This set of estimators includes many of the biased linear estimators which have been proposed in the literature but not Principal Components or Generalized Inverse as these have

one or more eigenvalues equal to zero or one.(see Reynolds (1977) Tables 4.1 and 4.2)

Let A be any pxp matrix such that $A^TA = X^TX(Z^{-1} - I)$ then, because of the above restrictions on the eigenvalues of Z, A^TA is p.d. (non-singular). For example in ordinary ridge regression $A = \sqrt{k}I$.

Now consider the augmented model

$$\begin{pmatrix} y \\ 0 \end{pmatrix} = \begin{pmatrix} X & 0 \\ 0 & A \end{pmatrix} \begin{pmatrix} \beta \\ \theta \end{pmatrix} + \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} \qquad \text{where} \quad \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} \right)$$

subject to the restriction

$$(I \quad -I) \begin{pmatrix} \beta \\ \theta \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

The unrestricted least-squares estimators are

$$\begin{pmatrix} \hat{b} \\ \hat{q} \end{pmatrix}_{l.s.} = \begin{pmatrix} b_0 \\ 0 \end{pmatrix}$$

and it is easy to show that the restricted estimates are

$$\begin{pmatrix} \hat{b} \\ \hat{q} \end{pmatrix}_{r.l.s.} = \begin{pmatrix} Zb_0 \\ Zb_0 \end{pmatrix}$$

It is also clear that the unbiased estimate of σ^2 from the unrestricted model is

$$s^2 = (y - Xb_0)^T(y - Xb_0)/(n - p)$$

i.e. the same as in the non-augmented model.

The usual F-ratio used to test the restriction is

$$F = (RSS_{\text{restricted}} - RSS_{\text{unrestricted}})/ps^2$$

$$= [(y - XZb_0)^T(y - XZb_0) + b_0{}^T Z^T A^T A Z b_0 - (n-p)s^2]/ps^2$$
$$= [b_0{}^T(I - Z)^T X^T X(I - Z)b_0 + b_0{}^T A Z b_0]/ps^2$$

Note that since A^TA is p.d. this F ratio is greater than (for $b_0 \neq 0$) the F ratio, $b_0{}^T(I - Z)^T X^T X(I - Z)b_0/ps^2$, for the increase in RSS due to Zb_0 in the non-augmented model which is used by Obenchain (1977) in his "Associated Probability" of Zb_0.

Under the augmented model, F has a non-central F-distribution with p and n-p degrees of freedom and non-centrality parameter

$$\phi = (H\beta^* - h)^T[HS^{*-1}H^T]^{-1}(H\beta^* - h)/\sigma^2$$

where here $H = [\, I \;\; -I]$, $\beta^{*T} = (\beta, \theta)$, $h = 0$ and

$$S^* = \begin{bmatrix} X^TX & 0 \\ 0 & A^TA \end{bmatrix}$$

If the restriction is true then $\phi = 0$. As was shown in the previous section MSE(l.s.) - MSE(r.l.s.) is p.s.d. if, and only if, $\phi \leq 1$. Fomby and Johnson (1977) suggest testing whether $\phi \leq 1$ by using the tables of Toro-Vizcarrondo and Wallace, (1969). But this is too strict a test in the sense that

$$MSE\ \underset{(\hat{q})\text{l.s.}}{(\hat{b})} \quad - \quad MSE\ \underset{(\hat{q})\text{r.l.s.}}{(\hat{b})}$$

involves θ, a vector of parameters in which we have no real interest unless $\theta = \beta$ in which case F has a central F-distribution and Zb_0 is an unbiased estimate of β!

2.212 Preliminary Test Estimators:

These are of the form

$$b_p = \begin{cases} b_{RLS} & \text{if } F \leq c \\ b_{LS} & \text{if } F > c \end{cases}$$

where F is the F-ratio used to test the restriction $H\beta = h$, c is some pre-chosen critical value against which F is compared, and b_{RLS} is the restricted least-squares estimator, b_{LS} the unrestricted estimator. As we have seen, many linear biased estimators can be considered as restricted least-squares estimators in an augmented model and hence could be candidates for a preliminary test estimator.

Brook (1976) has shown (equation 4.3) that the mean-squared-error matrix of a preliminary test estimator is still a function of the unknown parameters unless $c = 0$ and does not uniformly dominate least-squares. He suggests however that by using a suitable value of c, about 2, that the preliminary test estimator will improve on least-squares in certain regions of the parameter space and will not fare too badly over the rest of the parameter space. Once again, in the somewhat contrived situation where we can regard a linear biased estimator as a restricted least-squares estimator portion of a preliminary-test estimator, a value for c of about 2 may be too strict in the sense that we are not interested in the performance of any estimator of θ. However the approach taken by Brook could be followed in the case where our preliminary-test estimator is of the form

$$b_p = \begin{cases} Zb_0 & \text{if } F \leq c \\ b_0 & \text{if } F > c \end{cases}$$

with F a suitably defined function of b_0, Z and s^2.

Indeed this approach could be extended to a "continuous" preliminary test estimator such as

$$b_{cp} = (X^TX + kI)^{-1}X^TXb_0$$

where k is chosen so that F is identically equal to c, and for example

$$F = (b_{cp} - b_0)^TX^TX(b_{cp} - b_0)/ps^2$$

2.22 Incorporation of Prior Information:

Suppose we have a prior belief (or prior information) that β is in some pre-specified region R. Classical statistics would incorporate this prior belief by the use of a preliminary-test estimator. That is the estimate b_{RLS} in R which maximises the likelihood function, (minimises the residual sum of squares) would be calculated and then a choice made between the restricted and unrestricted estimates on the basis of the value of an appropriate function of both - usually the likelihood-ratio statistic, ℓ, or a monotonic function of ℓ. In the case where the region is defined by a certain number of independent linear restrictions the statistic used is the F-ratio, as we have seen. It is also clear that for any pre-specified region R which is independent of σ^2 the likelihood ratio statistic is a monotonic function of

$$(RSS_{restricted} - RSS_{unrestricted})/RSS_{unrestricted}.$$

Once again it would seem unlikely that the use of prior information in this manner would improve upon least-squares for all possible values of β, but certainly we could find appropriate critical values for the likelihood ratio statistic in each case so that the improvement over least squares in certain regions of the parameter space is balanced against the deficit in other regions.

Another method of incorporating prior information is to put a measure on our belief that β is at any particular point in the parameter space. The usual way to do this is to assign to β a "probability density function" over the parameter space. This does not necessarily mean that we believe β is a random variable, but in order to proceed it is convenient to treat β as though it were a random variable with a "user given" prior distribution and that the true random variable y has a conditional distribution given β which is normal with expected value $X\beta$ and variance $\sigma^2 I$. This approach, especially if one does consider β a random variable, is known as the Bayesian approach.

We could assume that β has a uniform prior distribution over a certain region of the parameter space, but this would utterly exclude β from the rest of the parameter space. It is convenient to assign to β a normal prior distribution with mean μ, the value we consider most likely for β, and with variance matrix Σ reflecting our degree of belief in μ as the most likely value for β.

We proceed by finding the conditional distribution for β given y. By Bayes Theorem

$$\begin{aligned} f(\beta:y) &= f(y:\beta).f(\beta)/f(y) \\ &\propto f(y:\beta).f(\beta) \end{aligned}$$

This last expression can also be viewed as the joint likelihood function for y and β. In the case where $f(\beta)$ is $N(\mu,\Sigma)$ with μ and Σ given,

$$\begin{aligned} f(y:\beta) &\propto \exp[-(y - X\beta)^T(y - X\beta)/2\sigma^2] \\ f(\beta) &\propto \exp[-(\beta - \mu)^T\Sigma^{-1}(\beta - \mu)/2] \end{aligned}$$

and so

$$\begin{aligned} f(\beta:y) &\propto f(y:\beta).f(\beta) \\ &\propto \exp[-(\beta - b_B)^T(X^TX/\sigma^2 + \Sigma^{-1})(\beta - b_B)/2] \end{aligned}$$

where $b_B = (X^TX/\sigma^2 + \Sigma^{-1})^{-1}(X^Ty/\sigma^2 + \mu)$ is the conditional

expectation of β given y. That is, with respect to the prior information, b_B is an unbiased estimate of β. Similarly we can regard $f(y:\beta).f(\beta)$ as a likelihood function:

$$L(y,\beta:\mu,\Sigma,\sigma^2) \propto \exp[-(y - X\beta)^T(y - X\beta)/2\sigma^2 + (\beta - \mu)^T\Sigma^{-1}(\beta - \mu)/2]$$

and it is clear that b_B maximises the likelihood function for β.

2.221 Ridge Regression as a Bayesian Estimator:

It can be shown that ordinary ridge regression where

$$b_k = (X^TX + kI)^{-1}X^Ty \text{ for } k > 0$$

is equivalent to assigning to β a prior normal distribution with mean 0 and variance $(\sigma^2/k)I$. That is, we could assume that firstly β is generated by the mechanism of the prior distribution, and then, with this value of β fixed, the observed vector y is generated by the conditional distribution for y given β. Under this assumption, it seems unnatural to me that the prior (unconditional) variance for β should be related to the variance of the as yet unobserved y through σ^2. It also leaves the choice of a suitable value of k undetermined, although there have been many suggestions for finding a suitable value of k proposed in the literature.

2.222 Another "Bayesian" Estimator:

Suppose we plan to centre and scale X and y so that the elements of X^TX and X^Ty are simple correlations between the variables. This goes a long way towards reducing unnecessary ill-conditioning in the model, and the coefficients in β could be regarded as the partial derivative of the expected change in y in standard deviations for a unit standard deviation change in the corresponding input variable, in the presence of the other input variables.

Under this situation a natural prior distribution for β in the absence of any better information would seem to be $\beta \sim N(0,I)$. Under this prior distribution the coefficients are independent of each other and the expected length of β is

$$E(\beta^T \beta) = tr(I) = p.$$

That is, if all the input variables in X are equally important we would expect a priori that the p coefficients would all be about plus or minus one.

The resulting estimators for β and σ^2 are:

$$b = (X^T X + s^2 I)^{-1} X^T y$$

and $$s^2 = (y - Xb)^T (y - Xb)/(n - p - 1).$$

This is a form of ridge regression where the ridge constant k is chosen by the data in a readily identifiable form and hence eliminates the problems of "guessing" a suitable value of k. In practice the estimates would have to be calculated by iteration, but that is no problem on modern computers.

What about the performance of the estimator? If we are true Bayesians, there is nothing further to say, as b is the a posteriori mean for β. It is instructive, however, to look at the performance of b if we regard β as a fixed but unknown vector, and σ^2 known. In this case

$$b = (X^T X + \sigma^2 I)^{-1} X^T y$$

and $$MSE(b) = (X^T X + \sigma^2 I)^{-1} (\sigma^2 X^T X + \sigma^4 \beta\beta^T)(X^T X + \sigma^2 I)^{-1}$$

while $$MSE(b_0) = \sigma^2 (X^T X)^{-1}.$$

Hence, as in Farebrother (1976), $MSE(b_0) - MSE(b)$ is p.s.d

$$\iff \beta^T (2/\sigma^2 I + (X^T X)^{-1})^{-1} \beta \leq \sigma^2$$

which is guaranteed by $\beta^T\beta \leq 2$. However $MSE(b_0) - MSE(b)$ can still be p.s.d. for $\beta^T\beta$ as large as $2 + \sigma^2/\lambda_{min}$, where λ_{min} is the smallest eigenvalue of X^TX, if β is proportional to the eigenvector corresponding to λ_{min}. Under these conditions every component of b has mean-squared error at least as small as b_0. If we look at average mean-squared error then

$$mse(b) \leq mse(b_0)$$
$$\iff \beta^T(X^TX + \sigma^2I)^{-2}\beta \leq tr[(X^TX)^{-1} - X^TX(X^TX + \sigma^2I)^{-2}]/\sigma^2.$$

Let X^TX have eigenvalues $\lambda_{max} = \lambda_1 \geq \lambda_2, \ldots, \geq \lambda_p = \lambda_{min} > 0$. Then sufficient conditions for the average mean-squared error of b to be less than or equal to that of b_0 are

$$\beta^T\beta \leq 2p\lambda_{min}^2/\lambda_{max}^2 \qquad (1)$$
$$\beta^T\beta \leq 2 + \sigma^2/\lambda_{min} \qquad (2)$$

Proof:

Maximising $\beta^T\beta + L(\beta^T(X^TX + \sigma^2I)^{-2}\beta - d)$ where L is a Lagrange multiplier and

$$d \leq tr[(X^TX)^{-1} - X^TX(X^TX + \sigma^2I)^{-2}]/\sigma^2$$

$\Rightarrow \beta = l.p_i$, where l is a constant and p_i is the i'th eigenvector of X^TX.

Hence $\beta^T\beta = l^2$ and

$$\beta^T(X^TX + \sigma^2I)^{-2}\beta = l^2/(\lambda_i + \sigma^2)^2$$
$$\leq d \qquad \text{for all } i$$
$$\iff l^2 \leq (\lambda_{min} + \sigma^2)^2.tr[(X^TX)^{-1} - X^TX(X^TX + \sigma^2I)^{-2}]/\sigma^2 \qquad (3)$$

But the right-hand sides of inequalities 1 and 2 are both less than the right-hand side of 3.

2.223 Numerical Example:

The data of Longley (1967) with p = 6 regressors, n = 16 observations and response y = "total derived employment" were centred and scaled to correlation form. Iteration was done by using an ordinary ridge regression program with initial value for k of zero, and successive values of k were calculated by dividing the residual sums of squares from the previous fit by the degrees of freedom 9. This converged to 3 significant figures in just six iterations, with s^2 (least-squares) = .0005 converging to a final value for s^2 = .000595. The maximum "variance inflation factor" (diagonal element of $(X^TX + kI)^{-1}$ when X^TX is in correlation form) dropped from 1788 to 274. The coefficients for the least squares fit and the final fit are given below:

b_0	: .04	-1.0	-.53	-.20	-.10	2.5
b_k	: -.009	-.07	-.40	-.17	-.30	1.7

The first has changed sign, both the first and second have gone almost to zero, the fifth has tripled in magnitude, while the other three have been reduced in size by up to 30 percent.

The eigenvalues of X^TX were calculated and ranged from 4.6 down to .00038. This suggests that b_k improves over b_0 in average mean-squared error, for values of $\beta^T\beta$ upto at least 3.44, using σ^2 = .0005 in equation (3). But $b_0^Tb_0$ = 7.6 while $b_k^Tb_k$ = 3.2, so it is perhaps questionable whether or not b_k improves over b_0 in mean-squared error for all components.

Discussion:

A problem arises in the study of this estimator due to the scaling of y to have unit length. Thus, for known σ^2, the centred but unscaled model is $y \sim N(X\beta, \sigma^2(I - 11^T/n)$, while the estimators from

the scaled model are:

$$b_0 = (X^TX)^{-1}X^Ty/\sqrt{(y^Ty)}$$
and $$b(\sigma^2) = (X^TX + \sigma^2I)^{-1}X^Ty/\sqrt{(y^Ty)}.$$

These can both be said to be estimates of $\beta/\sqrt{(y^Ty)}$ and hence we should be looking at

$$E(b - \beta/\sqrt{(y^Ty)})^T(b - \beta/\sqrt{(y^Ty)}) \quad \text{for } b = b_0 \text{ or } b(\sigma^2).$$

This is difficult since the estimates no longer have a normal distribution.

We can circumvent this problem by centering but not scaling y, (scaling y does not affect the conditioning of the problem), but then it is no longer natural to assign a known value to the length of β in its prior distribution, since multiplication of y by an arbitrary constant would also multiply β by the same constant.

2.224 Bayesian Estimator for Unscaled y:

Suppose we assign to β the prior distribution $N(0, \tau^2I)$ with τ^2 unknown. This, from a Bayesian point of view, gives

$$E(\beta : y) = (X^TX + \sigma^2/\tau^2I)^{-1}X^Ty.$$

Regarding $f(y:\beta).f(\beta)$ as a likelihood function and maximising this for the unknown parameters gives:

$$b = (X^TX + s^2/t^2I)^{-1}X^Ty$$
$$s^2 = (y - Xb)^T(y - Xb)/n$$
$$t^2 = b^Tb/p.$$

This system of equations could be solved iteratively or else we could use the least-squares values:

$$b_0 = (X^TX)^{-1}X^Ty$$
$$s_0^2 = (y - Xb_0)^T(y - Xb_0)/(n - p - 1)$$
$$t_0^2 = b_0^Tb_0/p$$
$$b = (X^TX + (s_0^2/t_0^2)I)^{-1}X^Ty.$$

Comparison of b with Stein's estimator:

Writing $(n - p - 1) = \nu$, $\nu.s_0^2 = S$, $c = p/\nu$, and $x = S/b_0^Tb_0$ we see that b can be written as

$$b = (X^TX + cxI)^{-1}X^Ty$$
$$= (X^TX + cxI)^{-1}X^TXb_0.$$

When $X^TX = I$

$$b = (1 + cx)^{-1}b_0$$
$$= (1 - cx/(1 + cx))b_0$$

whereas Stein's estimator, for $X^TX = I$ is

$$bs = (1 - c_0x)b_0 \quad \text{with } c_0 = (p - 2)/(\nu + 2).$$

Stein (1966) showed that for any constant c between 0 and $2c_0$ the estimator $b = (1 - cx)b_0$ uniformly dominates b_0 in average mean-squared error when $X^TX = I$ and $p \geq 3$, and that the choice of $c = c_0$ minimises this average mean-squared error independently of β. This suggests that perhaps we should modify the constant in the empirical Bayes estimator above to c_0.

In 1964 Baranchik showed that the estimator

$$bs^+ = (1 - c_0x)^+b_0$$

dominates bs in average mean-squared error for all β, where

$$a^{+} = \begin{cases} a, & a > 0 \\ 0, & a \leqslant 0. \end{cases}$$

Noting that

$$(1 - x)^{+} \leqslant (1 - x/(1 + x)) \leqslant 1 \quad \text{for all } x \geqslant 0$$

we might suspect that our estimator (using c_0), should have average mean-squared error lying between that of bs^{+} and b_0, i.e. it should uniformly dominate b_0 for $X^TX = I$. This is difficult to guarantee analytically, and would be even more so for X^TX non-orthogonal. Accordingly it was decided to perform a Monte-Carlo study to investigate the mean-squared-error performance of b.

3. SIMULATION OF THE mse PERFORMANCE OF THE VARIOUS ESTIMATORS

3.1 Experimental Design For the Simulation Study:

Since the proposed estimators have the form of a ridge estimator, but with k stochastic, it is instructive to look at the performance of a non-stochastic ridge estimator with k fixed.

Let $X^TX = P\Lambda P^T$ where P is the matrix of eigenvectors of X^TX and let $u = \Lambda^{-1}1$ be the vector of reciprocals of the eigenvalues of X^TX. Also let $\gamma = P^T\beta$. Then, as in Farebrother (1976),

$$\begin{aligned} & mse(b_k) < mse(b_0) \\ \Leftrightarrow \quad & k^2(\gamma/\sigma)^T(\Lambda + kI)^{-2}(\gamma/\sigma) < tr[\Lambda^{-1} - \Lambda(\Lambda + kI)^{-2}] \end{aligned}$$

Therefore we would expect b_k to perform poorly with respect to b_0 when

i) $\gamma^T\gamma/\sigma^2 = \beta^T\beta/\sigma^2$ is large

ii) the correlation squared between γ and u, $(\gamma^Tu)^2/(\gamma^T\gamma.u^Tu)$, is close to 1

Since σ appears only with γ, we can without loss of generality set $\sigma^2 = 1$ and look at the effects of i) and ii), under varying degrees of non-orthogonality. For the purposes of this study it will be presumed that X^TX has been scaled to correlation form so that $tr(\Lambda) = tr(X^TX) = p$ and the non-orthogonality of X^TX will be defined by the size of $tr(\Lambda^{-1})$.

The design is a full factorial with: $p = 4$, $\sigma^2 = 1$, and
3 levels of non-orthogonality: $tr(\Lambda^{-1}) = 4, 100, 1000$
3 levels of corr.-squared$(\gamma,u) = 0, .5, 1$
4 levels of $\gamma^T\gamma = 0, 4, 100, 1000$
2 levels of $\nu = 2, 10$
- the degrees of freedom for S ($= \nu.s^2$).

The low levels of p and ν are chosen in the hope that we can identify whether it is best to use p/ν or $(p - 2)/(\nu + 2)$ in the calculation of k. The actual values of γ used are given in the APPENDIX.

3.11 Method:

For any estimator b, let $g = P^Tb$, then

$$\begin{aligned} mse(b) &= E(b - \beta)^T(b - \beta) \\ &= E(g - \gamma)^T P^T P(g - \gamma) \\ &= E(g - \gamma)^T(g - \gamma) \end{aligned}$$

It can be seen that, for the purposes of this experiment, we do not need to know the matrix P.

It is proposed to look at 5 estimators:

$$\begin{aligned} g_0 &= P^Tb_0 \\ gs &= (1 - x)g_0 \\ gs^+ &= (1 - x)^+g_0 \\ g1 &= (\Lambda + xI)^{-1}\Lambda g_0 \\ g2 &= (\Lambda + x1I)^{-1}\Lambda g_0 \end{aligned}$$

where $x = (p - 2)S/((\nu + 2)g_0^Tg_0)$
and $x1 = pS/(\nu g_0^Tg_0)$.

Since, with $\sigma^2 = 1$, $g_0 \sim N(\gamma, \Lambda^{-1})$, we can generate each element $g_0(i)$ of g_0 using a unit-normal Random number generator (given in

the APPENDIX), dividing by $\sqrt{\lambda(i)}$ and adding $\gamma(i)$. Similarly we can generate the random chi-squared variable S with ν degrees of freedom, by taking the sum of squares of ν further independent unit-normal random variables. For each cell in the experiment, 1000 random vectors g_0 were generated and from each of these the various estimators were calculated. The mean squared errors of the estimators were estimated by calculating the mean of the 1000 values of $(g - \gamma)^T(g - \gamma)$ for each estimator, including g_0.

Since the object of the experiment was to compare the mse's of the various biased estimators with the mse of least-squares, the actual values of the mse's are of no direct interest to us. Rather it is the mse performance of the estimators relative to least-squares that we are interested in. Accordingly for each cell in the experiment the mse of each estimator was divided by the mse for least-squares in that cell. The resulting ratio is called the "relative (to least-squares) mean squared error" of an estimator. Thus relative mean squared errors less than 1 indicate that an estimator has outperformed least-squares in mse, while values larger than 1 indicate the reverse.

It is recommended that the results of future simulation studies on biased estimators should report the relative mean squared errors rather than the actual mse's, as it is much easier to assess the overall performance of an estimator.

3.2 Analysis of the Relative Mean-Squared Errors from the Simulation Study:

Since the biased estimators were calculated from g_0 in each cell, a split-plot analysis was performed, using the four biased estimators as the subplots. The factors in the subplot analysis were estimators, E, and all the first and second order interactions of E with the other main effects: non-orthogonality N; correlation-squared, C; length-squared, L; and degrees of freedom for chi-squared, D. The error term in the ANOVA had 120 degrees of

freedom. Apart from L x D x E and C x D x E (both non-significant) and N x D x E (significant at 5% level) everything else was highly significant at the .1% level. This suggests that a linear, additive model is not a good one for the relative mse's. Nevertheless, the various means from the analysis of variance do tell their own story. For example, increasing the degrees of freedom for chi-squared from 2 to 10 improved all the biased estimators in every case (though not by a constant amount). This agrees with intuition.

The effect of most interest was the N x L x E interaction for which the means are given in Table 1.

TABLE 1

RELATIVE MEAN-SQUARED-ERRORS.

($\sigma^2 = 1$, $p = 4$)

Estimators	$L = \gamma^T\gamma/\sigma^2$				$N = tr(\Lambda^{-1})$
	0	4	100	1000	
gs	.67	.85	.99	1.00	
gs^+	.56	.83	.99	1.00	
g1	.65	.83	.99	1.00	4
g2	.41	.70	1.01	1.00	
gs	.98	.98	.99	1.00	
gs^+	.97	.98	.99	1.00	
g1	.49	.49	.69	1.05	100
g2	.27	.28	.54	1.22	
gs	1.00	1.00	1.00	1.00	
gs^+	1.00	1.00	1.00	1.00	
g1	.50	.50	.52	.69	1000
g2	.28	.27	.30	.54	

From the table we see that the mse of each estimator approaches that of least-squares from below as the length of $\gamma^T\gamma$ increases relative to σ^2. This is to be expected from the form of the estimator as both x and xl tend to zero with probability one as $\gamma^T\gamma/\sigma^2$ tends to infinity. Both of the ridge-type estimators were better than the two Stein estimators on the non-orthogonal models, except for N = 100, L = 1000, which was to be expected since they take account of the eigenvalue structure of X^TX. However, it was surprising to see that g2 also outperformed both of the Stein estimators when X^TX = I, at least for the range of parameters considered here. The only exception was at L = 100 where g2 performed slightly worse than the Stein estimators. However the figure of 1.22 for g2 at $tr(\Lambda^{-1})$ = 100 and $\gamma^T\gamma/\sigma^2$ = 1000 suggest that perhaps g2, and to a lesser extent g1, is overshrinking the estimates.

It was decided to investigate the behaviour of the two ridge-type estimators more closely, and also to include another variant.

Returning to the Bayesian derivation of these two estimators where we assumed for β a normal prior distribution $N(0, \tau^2I)$, we see that there is theoretically a better estimate available for τ^2.

In the canonical form of the model we assume:

$$\gamma \sim N(0, \tau^2 I)$$
$$g_0 : \gamma \sim N(\gamma, \sigma^2\Lambda^{-1})$$

The estimate we have been using for τ^2 is $b_0^Tb_0/p = g_0^Tg_0/p$.
The Bayesian expectation for $g_0^Tg_0$ is :

$$EE(g_0^Tg_0) = E(\gamma^T\gamma + \sigma^2.tr(\Lambda^{-1}))$$
$$= p\tau^2 + \sigma^2.tr(\Lambda^{-1}).$$

When X^TX is ill-conditioned we could be grossly over-estimating τ^2 by using $g_0^Tg_0/p$.

However,

$$EE(g_0^T \Lambda g_0) = E(\gamma^T \Lambda \gamma + p\sigma^2)$$
$$= tr(\Lambda).\tau^2 + p\sigma^2$$
$$= p(\tau^2 + \sigma^2) \quad \text{when } X^TX \text{ is in correlation form.}$$

Thus $g_0^T \Lambda g_0/p$ ($= \hat{y}^T\hat{y}/p$) should be a much better estimate of τ^2.

We shall use $S/g_0^T \Lambda g_0$ multiplied by $(p - 2)/(\nu + 2)$ and p/ν as the additive eigenvalue "constant" for our two new ridge-type estimators g3 and g4. Of course, when $X^TX = I$ these estimators are identical to g1 and g2 respectively. The estimator g4 is actually the same as the estimator proposed by Lawless and Wang (1976).

This second study looked at estimators g1 to g4 under the same eigenvalue structures and true coefficients γ as Study 1, but with the degrees of freedom for S fixed at 2, and five lengths-squared for $\gamma^T\gamma/\sigma^2$. These were 0, 10, 100, 1000, 10,000. As in the first study $\sigma^2 = 1$ and the three levels 0, 1/2 and 1 were retained for the correlation squared between γ and the vector of reciprocals of the eigenvalues, u.

3.3 Results of Study 2:

As in study 1, for each cell in the experiment, the mean-squared errors for each estimator over 1000 values were calculated and divided by the mean-squared error for least-squares in that cell, to give the "relative mean squared errors".

Unfortunately there were occasions when g4 was up to 46 times worse than least-squares! Specifically, g4 was found to be very sensitive to the correlation between γ and the reciprocals of the eigenvalues, as is evident in Table 2.

TABLE 2a

Relative mse's at $\gamma \propto u$

Estimators	$\gamma^T\gamma/\sigma^2$					$tr(\Lambda^{-1})$
	0	10	100	1000	10,000	
g1	.73	.93	.99	1.00	1.00	
* g2	.46	.89	1.02	1.00	1.00	4
g1	.57	.61	.85	1.25	1.03	
g2	.30	.38	.76	1.90	1.20	100
g3	.07	.16	.84	6.4	19.2	
g4	.03	.13	.93	8.6	46.3	
g1	.59	.59	.63	.82	1.2	
g2	.32	.33	.39	.74	1.8	1000
g3	.01	.02	.11	.92	9.0	
g4	.002	.01	.10	.95	9.7	

* g3 and g4 are identical to g1 and g2 when $X^TX = I$.

From the present study, (Table 2) we see that g4 performed very well when γ was not oriented in the direction of u. The worst relative mse for g4 in other directions was 1.38 at correlation- squared $(\gamma,u) = 1/2$, $tr(\Lambda^{-1}) = 100$ and $\gamma^T\gamma = 10,000$. Taking a closer look at Table 2a we see that g4 performed satisfactorily even for γ in the direction of u provided $\gamma^T\gamma/\sigma^2$ was less than or equal to $tr(\Lambda^{-1})$, and under these conditions g4 was the best estimator in these studies. In view of the excellent performance of g4 in these regions perhaps we should use g4 after a preliminary test on the size and direction of γ.

3.4 Testing the size and direction of γ:

The following work is not to be confused with the ideas of R. L. Obenchain (1975), who provides a test of the hypothesis that the "optimal" p-parameter generalised ridge-estimate (Hoerl and Kennard, (1970 a.)) is in the one parameter family of Mayer and Wilke (1973), under which the $\gamma(i)^2 . \lambda(i)$ are all equal.

The aim here is to test whether or not the true coefficient vector γ lies in a region in which we could expect the stochastic ridge estimator g4 to improve upon least-squares in mean-squared error. From the present studies it appears that in this region g4 is not only superior to least-squares but also to the other biased estimators considered here, and outside this region the other estimators considered here perform about as well, or as poorly, as least-squares. That is to say the present studies indicate that it is only inside this region that we can expect any substantial improvement on least-squares from any of the estimators considered.

3.41 Testing the direction of γ:

Suppose that we set up the hypothesis

$$H_0: \quad \gamma = c.u \text{ for some unknown } c.$$

Under H_0, γ lies in the direction where, if $\gamma^T \gamma / \sigma^2$ is too large, we could expect g4 to perform poorly. Now, if H_0 is true,

$$\begin{aligned} g_0 &= \gamma + \xi, \qquad \xi \sim N(0, \sigma^2 \Lambda^{-1}) \\ &= c.u + \xi \\ \Rightarrow \quad \sqrt{\lambda(i)} . g_0(i) &= c/\sqrt{\lambda(i)} + \sqrt{\lambda(i)} . \xi(i) \end{aligned}$$

and we can rewrite this last equation in terms of a one-parameter vector linear model:

$$h = c.x + \psi, \qquad \psi \sim N(0, \sigma^2 I)$$

Thus the residual sum of squares from this model, with c estimated by $c_0 = (x^Tx)^{-1}x^Th$ is

$$RSS(c_0) = h^T(I - xx^T/x^Tx)h$$

which is distributed as $\sigma^2 \cdot \chi^2_{p-1}$.

Theorem:

The residual sum of squares from the model above, $RSS(c_0)$, is distributed independently of the residual sum of squares $S = y^T(I - X(X^TX)^{-1}X^T)y$, from the original model $y \sim N(X\beta, \sigma^2 I)$.

Proof:

With the eigenvalue decomposition $X^TX = P\Lambda P^T$ and $g_0 = P^Tb_0$, it is clear that $h = \Lambda^{1/2}P^Tb_0$, $x = \Lambda^{-1/2}1$. It follows that

$$\begin{aligned} & h^T(I - xx^T/x^Tx)h \\ &= y^TX(X^TX)^{-1}[X^TX - 11^T/(1^T(X^TX)^{-1}.1)](X^TX)^{-1}X^Ty \\ &= y^TAy \quad \text{say,} \end{aligned}$$

and since $(I - X(X^TX)^{-1}X^T)X = 0$, it follows that $(I - X(X^TX)^{-1}X^T)A = 0$ and hence the two residual sums of squares are independent. Note: this is true whether or not H_0 is true. If H_0 is true, $RSS(c_0)$ has a central χ^2 distribution, and otherwise it has a non-central χ^2 distribution with non-centrality parameter

$$\phi = (\gamma^T\Lambda\gamma - (1^T\gamma)^2/1^T\Lambda^{-1}1)/\sigma^2$$

Since $RSS(c_0)$ and S are independent, we can test H_0 at level α by comparing

$$[RSS(c_0)/(p-1)]/[S/(n-p)]$$

against $F_{p-1,n-p;\alpha}$. If we reject H_0 at probabilty level α, the

results of the previous simulation studies suggest that we can quite safely use g4 as our estimator of γ. On the other hand if we accept H_0 then we must find some means of testing whether or not $\gamma^T\gamma/\sigma^2 \geq tr(\Lambda^{-1})$. Before we proceed to find a test for the length of γ one could ask: "Given that we have accepted H_0 in any particular situation, that is we are prepared to beleve that $\gamma = c.u$ for some c, why not estimate γ by $c_0.u$ where c_0 is given above ?".

If H_0 is true, then $\gamma = c.u$

and $c_0 \sim N(c, \sigma^2(x^Tx)^{-1})$

$$\Rightarrow mse(c_0.u) = E(c_0.u - \gamma)^T(c_0.u - \gamma)$$
$$= u^Tu.E(c_0 - c)^2$$
$$= u^Tu.\sigma^2(x^Tx)^{-1}$$
$$= \sigma^2.tr(\Lambda^{-2})/tr(\Lambda^{-1})$$

Since $mse(g_0) = \sigma^2.tr(\Lambda^{-1})$ we have that the relative mse of $c_0.u$ under H_0 is:

$$rmse(c_0.u) = mse(c_0.u)/mse(g_0)$$
$$= tr(\Lambda^{-2})/(tr(\Lambda^{-1}))^2.$$

If $X^TX = I$, $rmse(c_0.u) = 1/p$ (provided H_0 is true).

Thus if $X^TX = I$, $c_0.u$ ($= c_0.1$ in this case) has a much smaller mean-squared error than least-squares or any of the estimators considered in these simulation studies, as shown in Table 1. However, for non-orthogonal X^TX, $rmse(c_0.u)$ approaches 1. For example, for the two non-orthogonal X^TX in the studies the formula above gives $rmse(c_0.u) = .94$ at $tr(\Lambda^{-1}) = 100$ and .83 at $tr(\Lambda^{-1}) = 1000$. Comparing these figures with those in Table 2a (H_0 is true for all of Table 2a), we see that g4 in particular can do much better than this, provided that $\gamma^T\gamma/\sigma^2 \leq tr(\Lambda^{-1})$.

3.42 Testing the Length of γ:

For the present purposes, we are interested in testing

$$H_1: \gamma^T\gamma/\sigma^2 \geq tr(\Lambda^{-1})$$

after having accepted

$$H_0: \gamma = c.u \text{ for some } c.$$

In terms of c, H_1 becomes

$$H_1: c^2.u^Tu/\sigma^2 \geq tr(\Lambda^{-1}) = x^Tx.$$

Under H_0 we have that $h \sim N(c.x,\sigma^2 I)$, thus $h^Th \sim \sigma^2.\chi^2(p,\phi)$ where $\phi = c^2.x^Tx/\sigma^2$.

Also,
$$\begin{aligned} h^Th &= g_0^T\Lambda g_0 \\ &= b_0^TX^TXb_0 \\ &= y^TX(X^TX)^{-1}X^Ty \end{aligned}$$

with y from the original model.

It follows that h^Th is distributed independently of

$$S = y^T(I - X(X^TX)^{-1}X^T)y$$

and therefore $(h^Th/p)/(S/\nu) \sim F'(p,\nu,\phi)$.

In terms of ϕ, H_1 becomes

$$\begin{aligned} H_1: \quad \phi &\geq (x^Tx)^2/u^Tu \\ &= (tr(\Lambda^{-1}))^2/tr(\Lambda^{-2}) \\ &= \phi_1 \quad \text{say.} \end{aligned}$$

Now, because of the poor performance of g4 when H_1 is true, (given H_0), we would like to safeguard ourselves by ensuring that the probability of rejecting H_1 when H_1 is true is small. Accepting H_1 when H_1 is false means that we will use least-squares instead of g4 and at least we will be no worse off than before.

Suppose that we reject H_1 if $\nu.h^Th/pS \leqslant C$ some critical value. Then,

$$\begin{aligned}
&\Pr[\text{ reject } H_1 : H_1 \text{ is true }] \\
&= \Pr[\ F'(p,\nu,\phi) \leqslant C : \phi \geqslant \phi_1\ ] \\
&\leqslant \Pr[\ F'(p,\nu,\phi_1) \leqslant C\ ] \qquad\qquad * \\
&\leqslant \Pr[\ F(p,\nu) \leqslant C\ ] \quad \text{since } \phi_1 \neq 0.
\end{aligned}$$

* By the properties of the non-central F-distribution
(Graybill (1976))

Thus if we choose $C = F'(p,\nu,\phi_1;1-\alpha)$ or $C = F(p,\nu;1-\alpha)$ then $\Pr[\text{reject } H_1 : H_1 \text{ true}] \leqslant \alpha$.

For ill-conditioned X^TX the choice of $C = F(p,\nu;1-\alpha)$ will not be too conservative, since then ϕ_1 will be only slightly larger than 1.

3.43 Modified Estimator:

The modified estimation rule now consists of several steps, given g_0 and S, which are outlined below:

1. Calculate:

$$g_0^T \Lambda g_0 \qquad (= h^T h)$$

$$k4 = pS/\nu g_0^T \Lambda g_0$$

$$g4 = (\Lambda + k4I)^{-1} \Lambda g_0$$

$$1^T g_0 \qquad (= x^T h)$$

$$\bar{g}_0 = 1^T g_0 / p$$

$$tr(\Lambda^{-1}) \qquad (= x^T x)$$

$$tr(\Lambda^{-2}) \qquad (= u^T u)$$

$$\phi_1 = (tr(\Lambda^{-1}))^2 / tr(\Lambda^{-2})$$

$$F0 = \nu[g_0^T \Lambda g_0 - (1^T g_0)^2 / tr(\Lambda^{-1})]/(p-1)S$$

$$F1 = 1/k4$$

2. If $F0 > F(p-1,\nu;\alpha_0)$ then reject H_0 and use g4 as the estimate. Otherwise proceed to step 3.

3. If $F1 \geq F'(p,\nu,\phi_1;\ 1-\alpha_1)$
then accept H_1 and use g_0 as the estimate.
Otherwise proceed to step 4.

4. If $X^T X = I$ then use $\bar{g}_0 . 1$ $(= c_0 . u)$ as the estimate,
otherwise use g4.

3.44 Performance of the modified estimator:

This new estimator was labelled g5. Another simulation was done under the same conditions as previously, with α_0 set to .1 and α_1 to .5.

The worst relative (to least-squares) mean-squared error was 3.3 under the conditions which gave rise to a figure of 46.3 for g4. The preliminary tests were protecting the estimator to a reasonable extent, but were having an undesirable effect when γ was zero. The

figures for g5 at $\Upsilon = 0$ corresponding to those in Table 2, were .54 at $X^TX = I$, .55 and .45. These are much higher than the corresponding figures for g4.

Note that when $\Upsilon = 0$, the hypothesis H_0: $\Upsilon = c.u$ for some c, is true, and so for small Υ, the estimator g5 uses least-squares too often. Perhaps g5 could have been improved by testing the length of Υ first, and then the direction, but this was decided against after re-examining the mean-squared error for a non-stochastic ridge estimator with parameter k.

3.5 Mean Squared Error for Non-Stochastic Ridge Estimation:

For a given constant ridge parameter k, the mean-squared-error of a ridge estimator in canonical form is

$$mse(k) = \sigma^2.tr(\Lambda(\Lambda + kI)^{-2}) + k^2\Upsilon^T(\Lambda + kI)^{-2}\Upsilon$$

For fixed $\Upsilon^T\Upsilon = 1$, we find, by differentiation of mse(k) using a Lagrange multiplier, that mse(k) is a maximum when $\Upsilon = +/- e_j$;, where e_j is the jth column of the identity matrix. That is, there are 2p "worst directions", the worst of which are $\Upsilon = +/-e_p$, when λ_p is the smallest eigenvalue. From this, and the results of the previous simulation studies, it appears to be dangerous to use ridge regression when any component of Υ is large in absolute value compared with the others, particularly if that component corresponds to a small eigenvalue. Needless to say, directions intermediate between the 2p worst directions can be bad also. For instance, by direct substitution in the formula for mse(k) above, we see that mse(k) increases as Υ changes from e_i to $(e_i + e_j)/\sqrt{2}$ to e_j when $\lambda_j < \lambda_i$. We also note that when one component of Υ is much larger than the rest, that the elements of Υ look very little like a sample from an $N(0,\tau^2)$ population. However this observation on its own does not explain why the performance of ridge regression is poorest when it is the last component of Υ that is particularly large.

Several statistics were formulated and used to try to protect g4 from use in these "dangerous" situations. The best method so far, seems to be to estimate the mean-squared error by substituting in the formula for mse(k) the additive eigenvalue factor used for k, i.e. $ps^2/g_0^T\Lambda g_0$, and estimates of σ^2 and γ. Here again the problem arises as to whether to use the biased or unbiased estimate of γ. If we always use g_0 as our estimate of γ we tend to overestimate the mean-squared error and hence use least-squares in situations where it would be safer to use the biased estimates. Conversely, if we always used the biased estimates in our estimate of the mean-squared error, then, in the dangerous situations described above, we would tend to underestimate the mean-squared error and hence use the biased estimates when it would be safer to use least squares. In an attempt to overcome this difficulty a preliminary statistic, $\hat{R}$, was calculated and, on the basis of the value of $\hat{R}$, a choice was made to use g_0 or g4 in the estimate of the mean-squared error. The statistic $\hat{R}$ is:

$$\hat{R} = g_0^T g_0 / g_0^T \Lambda g_0$$

Now, when the components of γ are large in magnitude, the relative error in estimating γ by g_0, is small, and so $\hat{R}$ is then a reasonable estimate of $\gamma^T\gamma/\gamma^T\Lambda\gamma = R$. As the components of γ corresponding to small eigenvalues increase relative to the others, R will increase, reaching a maximum of $1/\lambda_p$ when all components but the last are zero. Similarly, if all the components of g_0 were about equally significant, then $\hat{R}$ would be approximately $tr(\Lambda^{-1})/p$.

3.51 Modified Estimator g6:

Using both of these statistics, i.e. $\hat{R}$ and an estimate of the mean squared error of g4 the new modification, g6, of the estimator g4 becomes:

1. if $\hat{R}$ ($= g_0^T g_0 / g_0^T \Lambda g_0$) $> tr(\Lambda^{-1})/p$

then use g_0 as the estimate of γ in the estimated mean-squared error, $\hat{mse}$, otherwise use g4 as the estimate of γ in $\hat{mse}$.

2. if $\hat{mse} \geq s^2.tr(\Lambda^{-1})$

(the estimated mse of least-squares) then use g_0 as the estimate of γ, otherwise use g4.

3.52 Performance of the Modified Estimator:

In Table 3 are displayed the mean-squared errors relative to least squares of the resulting estimator g6 for each point in the experimental design. Table 3a gives the corresponding percentage of times, out of 1000, that least-squares was used. The degrees of freedom for S were held fixed at 2, so that the figures in the table for a correlation-squared of 1 between γ and u can be compared directly with those for g4 in Table 2.

TABLE 3.

Relative Mean Squared Errors

c^2	$\gamma^T\gamma/\sigma^2$ 0	10	100	1000	10,000	$tr(A^{-1})$
0	.46	.89	1.03	1.00	1.00	
.5	.46	.90	1.02	1.00	1.00	4
1	.46	.88	1.05	1.00	1.00	
0	.70	.55	.24	.50	.87	
.5	.71	.58	.57	1.70	1.20	100
1	.67	.79	1.13	1.14	1.00	
0	.67	.18	.13	.39	.77	
.5	.67	.17	.16	.57	1.07	1000
1	.66	.69	.71	1.07	1.13	

TABLE 3a.

Percentage of times least-squares was used

c^2	$\gamma^T\gamma/\sigma^2$ 0	10	100	1000	10,000	$tr(A^{-1})$
0	0	0	0	0	0	
.5	0	0	0	0	0	4
1	0	0	0	0	0	
0	30	18	6	2	0	
.5	30	17	10	32	93	100
1	27	33	48	97	100	
0	31	5	5	6	0	
.5	29	5	4	8	11	1000
1	28	29	28	52	97	

3.53 Discussion of Table 3:

By comparison of Tables 2 and 3, we see that the worst relative mean-squared error has been reduced from 46 to 1.7 - a vast improvement. The worst value now occurs at a correlation-squared of .5, $tr(\Lambda^{-1}) = 100$, and $\gamma^T\gamma/\sigma^2 = 1000$. The preliminary tests seem to be protecting g6 fairly well, particularly when the correlation- squared is 1. The relative mean-squared errors at $\gamma = 0$ have increased (but are still less than 1), due to the fact that when $\gamma = 0$, the components of g_0 corresponding to small eigenvalues will tend to be larger in magnitude than the others, and this is reflected in the percentage of times least squares was used then. Inclusion of an F-test for $\gamma = 0$ was incorporated, using various α-levels, but although this reduced the use of least-squares at $\gamma = 0$, it had an undesirable effect elsewhere.

Increasing the degrees of freedom for S from 2 to 10 generally improved the results, giving a best figure of .03 at $\gamma^T\gamma/\sigma^2 = 100$, $c^2 = .5$, $tr(\Lambda^{-1}) = 1000$, but the worst figure remained at 1.7. This was again at $\gamma^T\gamma/\sigma^2 = 1000$, $c^2 = .5$, $tr(\Lambda^{-1}) = 100$, but the use of least-squares there had dropped from 32% of the time to 18%.

3.6 Recommended Estimator:

This last modification of the estimator g4, given in 3.51, was the final estimator studied by Monte-Carlo methods, and appears to be the best of those studied. It is the estimator recommended here, and for clarity the estimation rule is given below in non-canonical form.

3.61 Recommended Estimation Rule:

In non-canonical form, the estimation rule is

1: Calculate $b_0 = (X^TX)^{-1}X^Ty$

$s^2 = (y - Xb_0)^T(y - Xb_0)/(n-p-1)$

$Freg = b_0^TX^TXb_0/ps^2$

$k = 1/Freg$

$b_k = (X^TX + kI)^{-1}X^Ty$

$\hat{R} = b_0^Tb_0/b_0^TX^TXb_0$

2: If $\hat{R} > tr(X^TX)^{-1}/p$ then set $b = b_0$,
otherwise set $b = b_k$.

3: Estimate the mean squared error of b_k by
$\hat{mse}(k) = s^2.tr(X^TX.(X^TX + kI)^{-2}) + k^2.b^T(X^TX + kI)^{-2}b$
and of b_0 by $\hat{mse}(0) = s^2.tr(X^TX)^{-1}$.

4: If $\hat{mse}(k) > \hat{mse}(0)$ then use b_0 as the estimate of β,
otherwise use b_k.

Note: The individual terms comprising the trace of $(X^TX)^{-1}$ or $X^TX(X^TX + kI)^{-2}$ when X^TX is in correlation form are termed the VIF's or variance inflation factors, after Marquardt and Snee (1975).

4. Numerical Example:

In order to demonstrate the techniques involved in the stochastic ridge estimator advocated here, we take the data given in Chapter 8 of Daniel and Wood (1971). The problem consists of estimating gasoline yields from various characteristics of the crude oil and a characteristic of the gasoline produced. Daniel and Wood, in their very careful analysis of the data, observed that the fourth independent variable was nested within the crude oils, and thus that there were only ten crudes rather than thirty-two. Their final equation involved only two of the original four variables, with error estimates for between and within crude variation in the response. In the analysis to follow, we shall purposely ignore the nesting and initially fit a full quadratic model in the four independent variables, so that as well as demonstrating the techniques, we shall get some idea of how well, or poorly, the estimator performs.

4.1 Analysis:

In order to fit the full quadratic model the original four variables were centred by subtracting their means before calculating the variables for the quadratic and cross-product (interaction) terms. This can greatly reduce the correlations between the independent variables.

Thus

x1 = crude oil gravity, ^{0}API - 39.25

x2 = crude oil vapour pressure, psi - 4.18

x3 = crude oil ASTM 10% point, ^{0}F - 241.5

x4 = gasoline end point, ^{0}F - 332.1

Variables 5 to 14 are then respectively $x1^2$, $x2^2$, $x3^2$, $x4^2$, x1.x2, x1.x3, x1.x4, x2.x3, x2.x4, x3.x4. The 14 independent variables were then all centred and scaled so that X^TX was a correlation matrix, and y was centred so that

y = gasoline yield as % of crude - 19.66.

4.2 Pass 1:

The model at this stage was $y = X\beta + \xi$, where X is 32 x 14. Since y has been centred and X is in correlation form, we are not directly fitting a constant term but nevertheless the degrees of freedom for error are 32 - 14 - 1 = 17. The stochastic ridge estimator uses as the ridge "constant",

$$k = ps^2/b_0^T X^T X b_0 = 1/Freg,$$

where s^2 is the estimate of σ^2 from the least-squares fit with n-p-1 = 17 degrees of freedom. The estimated mean-squared-error of the ridge estimator is

$$\begin{aligned}\hat{mse} &= s^2.tr(X^TX + kI)^{-1}X^TX(X^TX + kI)^{-1} + k^2.b^T(X^TX + kI)^{-2}b \\ &= s^2.(\text{ total of the ridge VIF's}) + \hat{\text{bias-squared}} \\ &= \hat{\text{variance}} + \hat{\text{bias-squared}}\end{aligned}$$

where $b = b_k$, the ridge estimator, if $b_0^T b_0 / b_0^T X^T X b_0$ is less than $tr(X^TX^{-1})/p$, otherwise b is b_0.

Table 4 gives the coefficients and summary statistics for the full 14-term model. The variance inflation factors in the table are given to two significant figures, the coefficients to three for clarity, although the computer program used gave at least 6 figure accuracy, using Householder transformations and iterative refinement (Fletcher (1975)).

TABLE 4.

	Variable	Coefficients		VIF's	
		L.S.	RIDGE	L.S.	RIDGE
1	x1	-5.27	6.45	17	4.9
2	x2	-360	8.55	2300	5.7
3	x3	-519	-28.3	4100	7.8
4	x4	61.6	58.5	1.6	1.4
5	$x1^2$	-47.6	1.89	91	5.9
6	$x2^2$	-294	-3.18	1400	7.9
7	$x3^2$	-721	2.64	9800	7.9
8	$x4^2$	5.43	4.61	2.0	1.7
9	x1.x2	-878	2.91	17,000	7.1
10	x1.x3	-992	1.32	22,000	4.1
11	x1.x4	-0.678	1.40	2.3	2.0
12	x2.x3	-966	0.594	16,000	3.8
13	x2.x4	-1.90	-1.76	11	5.8
14	x3.x4	-9.22	-6.35	17	8.6
			average:	5174	5.3

RSS:	52.3	118.2
R^2:	.985	.967
Freg:	81.5	$k = 1/Freg = .0123$
s^2:	3.08	

$$\hat{R} = b_0^T b_0 / b_0^T X^T X b_0 = 1,053 \quad < \quad tr((X^T X)^{-1})/p = 5,174$$

$\hat{var}$:	223,000	229
$\hat{bias}^2$:	0	78.3
$\hat{mse}$:	223,000	308

$$\hat{mse}(b_k)/\hat{mse}(b_0) = .001$$

4.21 Interpretation of Table 4:

Since $\hat{R} = b_0^T b_0 / b_0^T X^T X b_0$ is small compared with the average VIF for least-squares, $tr((X^TX)^{-1})/p$, we conclude that it is safe to use the stochastic ridge estimates in the estimate of the mean-squared error. Then, since the ratio of the estimated mse for ridge to that for l.s. is very much less than one, it would appear that there is much to be gained by using the ridge estimator as our estimate of β.

4.211 Variable selection:

The variance inflation factors for the ridge estimates are all relatively small and similar in size, and therefore the magnitudes of the coefficients should give an indication of the relative importance of the variables. On this basis, it is clear that variables 3 and 4 are easily the most important. At this stage we could remove all of the other variables, but, being of cautious nature, only variables 5, 10, 11, 12 and 13 were deleted for Pass 2. The results of Pass 2 are shown in Table 5.

TABLE 5.

	Variable	Coefficients		VIF's	
		L.S.	RIDGE	L.S.	RIDGE
1	x1	4.27	6.05	9.2	4.3
2	x2	5.80	9.19	16	6.3
3	x3	-33.5	-28.1	20	8.5
4	x4	60.0	58.6	1.4	1.3
6	$x2^2$	-.632	-2.12	12	7.1
7	$x3^2$	.946	.889	14	6.9
8	$x4^2$	4.77	4.36	1.7	1.6
9	x1.x2	2.91	3.04	18	9.1
14	x3.x4	-5.65	-5.48	2.3	2.1

RSS:	117.5	120.2	
R^2:	.967	.966	
Freg:	71.7		$k = 1/Freg = .014$
s^2:	5.34		

$$\hat{R} = b_0^T b_0 / b_0^T X^T X b_0 = 1.4 \ll tr((X^T X)^{-1})/p = 10.5$$

$\hat{var}$:	506	252	
$\hat{bias}^2$:	0	22.1	
$\hat{mse}$:	506	274.1	$\hat{mse}(b_k)/\hat{mse}(b_0) = .54$

4.3 Interpretation of Table 5:

The sum of the VIF's of the least-squares coefficients has dropped markedly from 223,000 to 506, while the residual mean-square has only increased from 3 to 5. Applying the same rules as before, we conclude once again that the ridge estimate is safer to use than least squares, although there is relatively less to be gained this time, as is also evidenced by the closer agreement between the two estimators.

Note also the close agreement between the ridge estimates in Tables 4 and 5 and the huge disparity, for most of the coefficients, between the least-squares estimates in Tables 4 and 5.

Variables 3 and 4 are still the most important, and so for completeness we will remove from the model all but these two variables. These are the two variables used by Daniel and Wood in their final equation. The results of this Pass 3 are shown in Table 6.

TABLE 6.

Variable		Coefficients		VIF's	
		L.S.	RIDGE	L.S.	RIDGE
3	x3	-43.75	-43.47	1.20	1.19
4	x4	60.52	60.19	1.20	1.19
	RSS:	170.6	170.7		
	R^2:	.952	.952		
	Freg:	288.4		$k = 1/\text{Freg} = .0035$	
	s^2:	5.88			

$$\hat{R} = b_0^T b_0 / b_0^T X^T X b_0 = 1.643 > \text{tr}((X^T X)^{-1})/p = 1.205$$

$\hat{var}$:	14.175	14.038	
$\hat{bias}^2$:	0	.1867	
$\hat{mse}$:	14.175	14.225	$\hat{mse}(b_k)/\hat{mse}(b_0) = 1.004$

4.4 Interpretation of Table 6:

This time, $\hat{R} = b_0^T b_0 / b_0^T X^T X b_0$ is greater than the average VIF for least-squares, and so we should use b_0 in our estimate of the squared bias for the ridge estimator, even though there is very

little difference between the two. Then we see that the estimated mse for ridge is worse (but only marginally) than that for least-squares and so we conclude that the best estimate of β is least-squares. This is not surprising considering that we are only dealing with two coefficients and that the VIF's for least-squares are nearly one, i.e. X^TX is well-conditioned.

Thus the final equation is:

$$y - 19.66 = -43.75(x3) + 60.52(x4)$$

or, in terms of the original uncentred and unscaled variables, X3 and X4,

$$\begin{aligned} y &= 19.66 - 43.75(X3 - 241.5)/209 + 60.52(X4 - 332)/388.4 \\ &= 70.21 - .209(X3) + .156(X4 - 332) \end{aligned}$$

which compares well with the final equation of Daniel and Wood:

$$y = 70.84 - .212(X3) + .159(X4 - 332)$$

4.5 Conclusions:

Although there are arguments for and against removal of variables from a prediction equation, we see by this example that the stochastic ridge estimator proposed here, used carefully, can pinpoint the most influential variables very well even when the X^TX matrix is highly ill-conditioned, or as other authors would say, in the presence of a high degree of multicollinearity. In this example, we see that the ridge coefficients for x3 and x4 from the full model (Table 4) are both far, far closer to the coefficients in the (drastically) reduced final model than are the least-squares coefficients. Also, in three passes from a 14-term original model, we reached the same final conclusion as did Daniel and Wood in 10 passes from an original 4-term model. This is not intended to suggest that such things as nested data should be ignored, but

merely that the stochastic ridge estimator can help protect us from failure to observe nesting as well as from multicollinearity in the independent variables.

5. SUMMARY and DISCUSSION

It is a demonstrable fact that non-stochastic biased estimators of β in the Normal Linear Model can do worse than least-squares in mean-squared error for β in certain regions of the parameter space. Therefore it was decided to investigate some stochastic biased estimators over a wide range of parameters.

The Bayesian approach of assigning to β a prior distributibn provides a suitable framework for the formation and study of various stochastic estimators. It also introduces a two-layer outlook to the problem: firstly we have the generation of a particular β by the mechanism of the prior distribution, and, from that β, the observation vector y is assumed to have been generated by the Normal distribution for y given β. Similarly we can look at two losses for an estimator: firstly how well it performs for a given β, which is what Rao (1976) calls "Individual's Loss", and secondly how well it performs over the whole range of β's generated by the prior distribution - Rao's "Statistician's Loss".

Our interest has been in Individual's Loss (IL) rather than in Statistician's Loss (SL) for the following reason. Suppose we have an estimator which can be shown to have smaller SL than least squares. This does not guarantee that it will have smaller IL than least-squares, and although, as pointed out by Antoniak and Effron (1976), an individual can reap his own rewards by repeated patronage of a statistician using this estimator, he may never wish to see that statistician again if he ever discovers that he has been "had".

On the other hand if we have an estimator with uniformly smaller IL than least squares, it follows that the SL, which is the average IL, is also smaller, and thus both parties benefit.

Ordinary ridge regression, apart from being a non-stochastic biased estimator, for fixed k, can be viewed as arising from assigning to β, the normal prior distribution $N(0,(\sigma^2/k)I)$, but it would seem

more natural to allow the prior distribution of β to be unrelated to the posterior distribution of y, through σ^2, by assigning to β an $N(0,\tau^2 I)$ prior distribution. This leads to estimators of the form $(X^T X + (\sigma^2/\tau^2).I)^{-1} X^T y$, which are in general unoperational. Replacing σ^2 and τ^2 by estimates leads to what might be called empirical Bayes estimators. The Stein-James estimator is one of this class, and when $X^T X$ is orthogonal has uniformly smaller mean-squared error, or Individual's Loss, than least squares. However, when $X^T X$ is not orthogonal an analytic expression for the mean-squared error of this or any other empirical Bayes estimator has not yet been obtained. Thus it was decided to perform some simulation studies on the mean squared error (Individual's Loss) performance of several empirical Bayes estimators.

The simulation studies showed that one of the estimators considered, a form of ridge regression where the ridge parameter is chosen to be the inverse of the F-statistic for the signifance of the least-squares regression, performed very well unless the true coefficient vector β was large in the directions of least information. It was also found that this estimator could be reasonably well protected from use in these situations by a preliminary test on its expected performance, or mean-squared error.

It is envisaged that this estimator would be particularly useful when $X^T X$ is ill-conditioned, and the resulting least-squares estimates suffer from large variances.

In conclusion it would seem that we have not yet found an estimator that can be routinely used in every situation that can be guaranteed to outperform least-squares, but the stochastic ridge estimator recommended here comes close. Its performance could no doubt be further improved by better tests of when it can be safely used.

6. APPENDIX

6.1 Eigenvalues and True Coefficients γ Used in the Simulation Studies:

The γ's given below were scaled by an appropriate constant to give the various ratios of $\gamma^T\gamma$ to σ^2 (= 1) used in the experiments. The column headed C gives the correlation-squared between the elements of γ and the reciprocals of the eigenvalues.

$tr(\Lambda^{-1})$	C	eigenvalues	coefficients γ^T
	0		1, -1, 1, -1
4	.5	1, 1, 1, 1	1, 0, 1, 0
	1		1, 1, 1, 1
	0		2.5, -1, 117.7, 2.5
100	.5	2.5, 1, .4896, .0104	1, 2, 3, 3.57
	1		1, 2.5, 5.11, 240.4
	0		-12.69, -10, -10, 1
1000	.5	3.5, .4876, .0113, .0011	-3.72, 1, 2, 4
	1		1, 7.18, 309.7, 3182

6.2 Random Unit Normal Generator:

The random normal generator used throughout the simulation studies used an exact method given in Jansson (1966): If u_1 and u_2 are two independent random variables with a rectangular distribution on [0,1], then the variables y_1 and y_2 given by

$$y_1 = (-2\ln(u_1))^{1/2}.\cos(2\pi u_2)$$

and

$$y_2 = (-2\ln(u_1))^{1/2}.\sin(2\pi u_2)$$

are independently distributed as N(0,1).

The rectangular generator used was the internal random number generator on the D.S.I.R./ D.R.I. PDP11/45 computer at Palmerston North.

BIBLIOGRAPHY

ANTONIAK, C.E., and EFFRON, B., (1976) Dealing with many problems simultaneously.
On the history of Statistics and probability. STATISTICS. p. 232. Edited by D. B. Owen, MARCEL DEKKER, inc. New York.

BARANCHIK, A., (1964) Multiple regression and estimation of the mean of a multivariate normal distribution.
Unpublished Ph. D. Thesis, Stanford University.

BROOK, R.J., (1976) On the use of a regret function to set significance points in prior tests of estimation.
Journal of the American Statistical Association, 71, 353, 126-131.

DANIEL, C. and WOOD, F.S., (1971) Fitting Equations to data.
WILEY SERIES IN PROBABILITY AND MATHEMATICAL STATISTICS.
New York

FAREBROTHER, R. W., (1976) Further results on the mean square error of ridge regression.
Journal of the Royal Statistical Society, Series B, 38, 3, 248-250.

FLETCHER, R.H., (1975) On the iterative refinement of least squares solutions.
Journal of the American Statistical Association, 70, 349, 109-112.

FOMBY, T.B., and JOHNSON, S.R., (1977) MSE evaluation of ridge estimators based on stochastic prior information.
Communications in Statistics, A6, 13, 1245-1258.

GRAYBILL, F.A., (1976) Theory and application of the linear model, Theorem 4.3.2, p130, DUXBURY PRESS, Massachussetts.

HOERL, A.E., and KENNARD, R.W., (1970a) Ridge regression: biased estimation for non-orthogonal problems.
Technometrics 12, 1, 55-67.

JANSSON, B., (1966) Random number generators.
VICTOR PETTERSONS BOKINDUSTRI AKTIEBOLAG, Stockholm.

LAWLESS, J.F., and WANG, P., (1976) A simulation study of ridge and other regression estimators.
Communications in Statistics. A5. 4. 307-323.

LONGLEY, J.W., (1967) An appraisal of least squares programs for the electronic computer from the point of view of the user.
Journal of the American Statistical Association, 62, 819-841.

MARQUARDT, D.W., and SNEE, R.D., (1975) Ridge regression in practice.
The American Statistician, 29, 1, 3-20.

MAYER, L.S., and WILLKE, T.A., (1973) On biased estimation in linear models.
Technometrics 15, 497-508.

OBENCHAIN, R.L., (1975) Ridge analysis following a preliminary test of the shrunken hypothesis.
Technometrics 17, 4, 431-441.

OBENCHAIN, R.L., (1977) Classical F-tests and confidence regions for ridge regression.
Technometrics 19, 4, 429-439

RADHAKRISHNA RAO, C., (1976) Characterisation of Prior Distributions and solution to a compound decision problem.
Annals of Statistics, 4, 5, 823-835.

REYNOLDS, J., (1977) Some Alternatives to least squares estimation in linear modelling.
Occasional Publications in Mathematics, Number 4, Massey University Department of Mathematics.

STEIN, C., (1966) An approach to the recovery of inter-block information in balanced incomplete block designs.
Research papers in statistics, WILEY, New York, 351-366.

TORO-VIZCARRONDO, C.E., and WALLACE, T.D., (1969) Tables for the mean square error test for exact linear restrictions in regression.
Journal of the American Statistical Association, 64, 1649-1663.