SOME ALTERNATIVES

TO

LEAST SQUARES ESTIMATION

IN

LINEAR MODELLING

A thesis presented in partial fulfilment
of the requirements for the degree of
Master of Science in Statistics
at Massey University

John Reynolds

1977

ABSTRACT

The effects of non-standard conditions on the application of the Gauss-Markov Theorem are discussed and methods proposed in the literature for dealing with these effects are reviewed. The multicollinearity problem, which is typified by imprecise least squares estimation of parameters in a multiple linear regression and which arises when the vectors of the input or predictor variables are nearly linearly dependent, is focussed upon and a class of alternative biased estimators examined. In particular several members of the class of biased linear estimators or linear transformations of the Gauss-Markov least squares estimator are reviewed. A particular generalized ridge estimator is introduced and its relation to other techniques already existing in the literature is noted. The use of this estimator and the simple ridge regression estimator is illustrated on a small data set. Further comparisons of the estimator, the ridge estimator and other generalized ridge estimators are suggested.

ACKNOWLEDGEMENTS

I wish to record my gratitude to my supervisor Dr Richard J. Brook for his encouragement and guidance in the preparation of this thesis. Grateful thanks are owed to Miss Maryanne Collins who patiently typed the tedious manuscript.

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION

The solution of a system of overdetermined or overidentified linear equations requires some kind of approximation method. The most common method of arriving at a solution for B in the overdetermined system of linear equations,

$$XB = Y$$

where X is an $n\chi p$ matrix of full column rank p, B is an unknown $p\chi 1$ vector of parameters and y is an $n\chi 1$ vector, and in which $n > p$, is the method of least squares. The least squares solution identifies the $p\chi 1$ vector which minimizes the Euclidean norm of Y-XB.

The source of the overdetermination or inconsistency in the system of linear equations is usually attributed to the presence of some kind of error component in the n realizations of the p+1 variables which form X and Y. Statisticians often make very specific assumptions about the error content of the n realizations. Errors are usually assumed, in the lack of any knowledge concerning their origin, to be generated by some sort of random device which may be represented by a probability density. The realizations of the p variables which make up the matrix X and which are often controllable are usually assumed to be measurable without error whereas the vector variable Y is usually assumed to contain the randomly generated errors. Thus statisticians have concerned themselves with the linear model,

$$y = X\beta + \epsilon$$

where ϵ is an $n\chi 1$ vector of stochastic errors which are independent of the measurements of the p variables which make up the matrix S, and, have used the method of least squares to extract an approximation to, or an estimate of, the unknown vector of parameters, β. Under various assumptions about X,y and ϵ, and under various restrictions on possible methods of approximation, the method of least squares has other optimal features besides the norm minimization property

mentioned above. If, however, these assumptions are not met in practice the other optimal features may disappear.

The purpose of this thesis is to review some of the work which has been completed or is currently in progress, concerning the effect of the relaxation of these assumptions and restrictions on the optimality properties of least squares and to review some of the alternatives to least squares which have been developed in response to these effects. The conditions of the Gauss-Markov Theorem, which are presented in Chapter 2, form the framework for the review and it is the effect of the relaxation of these conditions on the least squares procedure which is presented in Chapter 3. In Chapter 3 it is established that multicollinearity in the matrix X is one non-standard condition which can have serious effects on least squares estimation of the parameter vector. A class of alternatives to the least squares estimator, namely biased estimators, is focussed upon in Chapter 4. These estimators were designed originally to tackle the multicollinearity problem but many variants of these biased estimation procedures have been constructed with different goals in mind. A particular member of a subclass of these biased estimators, a doubly ridged estimator, is introduced in Chapter 5. The doubly ridged estimator, which is a generalized ridge estimator, displays many of the advantages and disadvantages of the well known ridge estimator. The application of the ridge and doubly ridged estimators to a small but well known test problem - the Longley data - is undertaken in Chapter 6. A summary, Chapter 7, which also includes suggestions for further investigations in the search for alternatives to least squares, completes the thesis.

2. THE METHOD OF LEAST SQUARES AND THE GAUSS-MARKOV THEOREM

The method of least squares has its origins in the writings of Gauss, Laplace and Legendre in the early 19th century. The allocation of credit for various justifications of the method has been attempted by Plackett (1949) who concludes that Gauss was the first to give a distribution-free proof that least squares provides minimum variance unbiased linear estimates of the parameters in a linear model. Since the 19th century, justifications for the method have been modified to allow for more general formulations of the linear model, namely, a not necessarily diagonal variance - covariance matrix, a design matrix of less than full column rank, and, constraints on the parameters in the model. The most recent attempts to provide a unified theory of least squares, embracing the formulations of the linear model mentioned above, have been made by Rao (1971, 1973) and have utilized the theory of generalised inverses of matrices.

The popularity of least squares, as either a method of parameter estimation in a linear model or a provider of a linear interpolation formula giving what could well be the best linear fit to a nonlinear model, has prompted Tukey (1975) to describe least squares as a "scientific idol". Tukey suggests that, "neither unquestioning acceptance or iconoclasm is a proper way to manage a scientific idol", but that the idol should be used as a point of embarkation for developing techniques which are more realistic and useful.

A particular "manifestation" of the least squares "scientific idol", which might serve as the embarkation point for developing techniques which have a wider application to more realistic situations, is the Gauss-Markov Theorem (with reference to the name of the theorem, Plackett (1949) comments, "Markov, who refers to Gauss's work, may perhaps have clarified assumptions implicit there but proved nothing new"). A nondefinitive formulation of the Gauss-Markov Theorem follows:

(i) $y = X\beta + \epsilon$ (a linear model)

where y is an $n\times 1$ vector of observations
X is a known $n\times p$ matrix of full column rank $p \leq n$
β is a $p\times 1$ vector of unknown parameters
ϵ is an $n\times 1$ vector of errors

(ii) X is known exactly, there are no stochastic or non-stochastic errors in X.
There are no non-stochastic errors in y.

(iii) The error vector ϵ has zero mean, i.e.

$$E(\epsilon) = 0, \quad E(y) = X\beta$$

The error vector has variance-covariance matrix $\sigma^2 I_n$ so that the variance-covariance matrix for y is also $\sigma^2 I_n$. The scalar quantity σ^2 is generally unknown.

(iv) The estimates of the parameters can only be linear functions of y, i.e.

$B = Ay + c$, where A is $p\times n$ and not a function of y and c is $p\times 1$ and not a function of y.

(v) The estimates of the parameters are unbiased in mean,

$$E(B) = \beta$$

(vi) The variance of the estimates of the unknown parameters is to be minimised.

If the conditions (i) to (vi) are met then the best estimate of the unknown parameter vector β is to be found in <u>least squares</u>, i.e.,

$$\min_B \|y-XB\|^2$$

where $\|...\|$ denotes the Euclidean norm of a vector.

The Gauss-Markov Theorem, as stated here, consists of six conditions and a conclusion. The six conditions restrict the choice of an estimator B, of the unknown parameter vector β, to a class of linear, unbiased-in-mean estimators and make assumptions about the first and second moments of the error vector ϵ. Under these conditions the least squares estimator is "best" (of minimum variance).

Tukey's suggestion, that the least squares scientific idol should be used as a takeoff point for developing statistical tools which have greater utility in the real world, can be implemented by relaxing singly or simultaneously the conditions of the Gauss-Markov Theorem and evaluating the consequent alternative estimation procedures or modified least squares procedures. Tukey has of course attempted this (see Tukey (1975)) but the same approach is used here as a means of surveying some of the developments (past and current) in the theory of estimation in linear statistical models.

3. SOME CONSEQUENCES OF MODIFICATIONS TO THE CONDITIONS OF THE GAUSS-MARKOV THEOREM

3.1 The Model Misspecified.

The first condition of the Gauss-Markov Theorem specifies a model linear in the components of the parameter vector β. Relaxing this linearity assumption leads to a consideration of nonlinear models which does not necessarily take one too far from the least squares idol. If the model is not linear but is intrinsically linear (the model can be transformed into a form which is linear in the parameters) then nonlinearity is not a problem and the method of least squares can be applied to the transformed model. If the model is nonlinear and intrinsically nonlinear (not able to be transformed into a linear form) an approximating linear expansion may be investigated using an iterative process involving least squares. An introduction to the estimation of parameters in nonlinear models is contained in Draper and Smith (1966).

Even if it is assumed that the linear model relationship between the variables is appropriate, the first condition of the Gauss Markov Theorem still contains requirements which may be difficult to meet in practice.

3.11 Latent or Lurking Variables.

The first condition of the Gauss-Markov Theorem requires the matrix of independent or predictor variables to be known. Box (1966) has drawn attention to the fact that in many regression analyses all of the predictor variables may not be known. The error vector ϵ, the undetermined or non-measurable component of the data, is usually dismissed as a random variable (see condition (iii) of the Gauss-Markov Theorem). Box points out that ϵ is a "catch-all" vector that actually contains the effects of other unknown "latent" or "lurking"

predictor variables. Thus the model,

$$y = X\beta+\varepsilon \tag{3.1}$$

in which the matrix X contains n values of p predictor variables may be a facade masking the true relationship,

$$y = X\beta + Z\gamma \tag{3.2}$$

in which the matrix Z contains n unknown values of some m-p latent predictor variables. Including only the X matrix of input variables in the least squares regression analysis gives the following for the predicted values of y:

$$\begin{aligned}\hat{y} = X\hat{\beta} &= X(X'X)^{-1}X'y \\ &= X\beta + X(X'X)^{-1}X'Z\gamma \\ &= X\beta + \hat{Z}\gamma\end{aligned} \tag{3.3}$$

The matrix $\hat{Z} = XA$, where $A = (X'X)^{-1}X'Z$ can be thought of as the matrix of estimated parameters of the regression of the m-p latent variables on the p known independent variables. Comparing the predicted values of y with the values of y in the true model, the expression for the predicted values is similar to the true model except that the unknown values of the matrix of latent variables Z are replaced by an estimate $\hat{Z}$.

However the estimates of β are biased,

$$\hat{\beta} = \beta + A\gamma \tag{3.4}$$

and in a sense because there are no estimates of γ the estimates of γ are biased too (they are shrunk to the value zero).

Box makes the point that if the matrix of independent variables X is passively observed or unplanned, the prediction equation (3.3) which results from fitting the model in equation (3.1) may be appropriate for prediction of y from further passive or unplanned observation of the independent variables but will not be appropriate for predicting

in a controlled situation how adjustments in the observable independent variables will affect the dependent variable y. The impropriety of the use of equation (3.3), based on unplanned data, for prediction of y from controlled data stems from the fact that the estimated coefficients (3.4) stand for combinations of effects due to the known independent variables and the unknown latent variables, so that $\hat{\beta}$ in (3.4) does not tell of the effect on y of unit changes in X.

Box further comments that planned observation, including randomization of the levels of the observable input variables, prevents the levels of these known input variables from being affected by the levels of the latent variables, so that predication of y for controlled values of X using least squares estimates of β is appropriate even though the known predictor variables may be producing the predicted changes in y through some latent variable.

Thus the message from Box is that the consequences of overlooking predictor variables in a regression analysis may not be great in situations where the data is unplanned and predictions are required for further unplanned observation of the known predictor variables in unaltered circumstances, and in situations where a randomized design has been used and predictions are required for controlled levels of the predictor variables within some region of interest. There are, however, dangers in using a prediction equation based on passive, unplanned, historical data to predict expected responses in situations where the levels of the input variables are being manipulated. This is because the relationships which obtain while the system is being passively observed may change drastically when the system is interfered with (the overlooked lurking variables may change wildly when the known input variables are adjusted and could cause responses quite different from the predicted responses). This situation is referred to in section 4.22.

3.12 The Disaggregation Problem.

In some circumstances the general linear model form assumed in the first condition of the Gauss-Markov Theorem may not be appropriate for all combinations or subsets of the p independent variables. This disaggregation problem often occurs in econometrics after microrelations have been aggregated in a macrorelation, the macrorelation used to construct a fitted model, and the fitted model used for prediction in a particular disaggregate of the macrorelation. Rao (1975) discusses three colourful examples in which problems occur when the regression equation is disaggregated.

In general, disaggregation problems occur in the following way. Consider a fitted model,

$$\hat{y} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p \qquad (3.5)$$

Suppose predictions are required for the following two sets of observations of the independent variables,

$$(x_1=1,\ x_2=1,\ x_3=1,\ \ldots,\ x_p=1)$$

and

$$(x_1=1,\ x_2=0,\ x_3=0,\ \ldots,\ x_p=0).$$

In the first case the prediction is,

$$\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 + \ldots + \hat{\beta}_p$$

and in the second case the prediction is,

$$\hat{\beta}_1$$

Both predictions use the estimate $\hat{\beta}_1$. However, in the second set of observations of the independent variables, p-1 of the independent variables are not present. The underlying microrelation which would generate the dependent variable y in this case is,

$$E(y) = b_1 x_1 \qquad (3.6)$$

But in this case the predicted value, $\hat{\beta}_1$, results from the regression of y on all of the p independent variables in the macrorelation,

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon. \qquad (3.7)$$

If β_1 in the macrorelation is equal to b_1 in the microrelation then $\hat{\beta}_1$ happens to be an unbiased estimate of b_1. However, it may not be true that β_1 and b_1 are the same and using an estimate of β_1 to estimate b_1 makes little statistical sense and explains the sometimes unreasonable values (negative estimates for parameters which on physical grounds must be positive) which result. Rao calls this phenomenon "Poti effect" and seems to regard it as a pathological case in regression analysis.

However, the problem seems to arise from the aggregation process itself and the temptation to attribute specific meanings to the estimated parameters in the macrorelation. If it is not true that all of the coefficients of a particular independent variable in all of the microrelations are equal, then the aggregation process should not proceed and separate models should be investigated for separate microrelations. If an aggregation of microrelations is proceeded with and a fitted model, in the form of equation (3.5) is the result, then disaggregating the fitted model, by, for example, trying to predict y for the set of observations,

$$(x_1=1, x_2=0, x_3=0, \ldots, x_p=0)$$

involves attaching specific meanings to the estimated parameters (in this example, disaggregation involves assuming that $\hat{\beta}_1$ tells of the effect of x_1 on y). In a different context, Beaton and Tukey (1974) and Tukey (1975) have discussed the meanings of parameter estimates in fitted models. They point out that a particular parameter estimate, say $\hat{\beta}_1$, does not tell of the effect of the <u>carrier</u> to which it is attached, in this case the independent variable x_1, but tells of the apparent effect of the carrier, x_1, in the presence of a <u>subposse</u> composed of all the linear combinations of all the other carriers (in this case the independent variables

$x_2, x_3, \ldots, x_p$) in the fitted model, which give the same degree of fit.

In the given disaggregation example, $\hat{\beta}_1$ is supposed to tell of the effect of x_1 on y in the microrelation (3.6) when in fact $\hat{\beta}_1$ tells of the effect of x_1 on y in the presence of a particular set (a subposse) of linear combinations of the p-1 other independent variables in the fitted macrorelation (3.5). It really is little wonder that anomalous estimates and predictions result from the disaggregation of a regression equation when the prior aggregation of separate models itself may be unjustified, and, the estimated parameters in the regression equation are misinterpreted in the disaggregation.

3.13 Deficiencies in the column rank of X.

The first condition of the Gauss-Markov Theorem also requires the matrix X of independent variables to have full column rank p. In specifying the matrix $X'X$ to be non-singular, something of a strawman has been set up. This strawman (the full rank model) has been knocked down and adequately treated (the non-full rank model) in, for example, Rao (1971, 1973) and Searle (1971). A related problem that has not, as yet, been adequately treated concerns near singularity of the matrix $X'X$ or multicollinearity of the independent variables. When the matrix $X'X$ is singular, the parameter vector β is non-identifiable and only some linear functions $t'\beta$ of the parameter vector are estimable. When two or more of the independent variables are highly correlated and the matrix $X'X$ is subsequently ill-conditioned, nearly-singular, non-orthogonal ($X'X$ may be thought of as having fractional rank), then the parameter vector β is identifiable and all linear combinations of the parameter vector β are estimable, but only some linear combinations are estimable with any precision. It is the relative lack of precision with which some linear functions of the parameter vector β are able to be estimated which characterizes the multicollinearity problem.

3.131 The Multicollinearity Problem

The usual least squares estimator of the parameter vector β may be written,

$$\hat{\beta} = \beta + (X'X)^{-1}X'\epsilon \tag{3.8}$$

and the corresponding estimator of the linear function $t'\beta$ may be written,

$$t'\hat{\beta} = t'\beta + t'(X'X)^{-1}X'\epsilon \tag{3.9}$$

Thus, depending on $(X'X)^{-1}X'\epsilon$, an estimate $\hat{\beta}$ may well be some distance away from β. If the usual distributional assumptions are made about ϵ, as in the third condition of the Gauss-Markov Theorem, then the variance of $\hat{\beta}$ and linear functions of $\hat{\beta}$ is given by,

$$\left.\begin{aligned} \text{var}(\hat{\beta}) &= \sigma^2(X'X)^{-1} \\ \text{var}(t'\hat{\beta}) &= \sigma^2 t'(X'X)^{-1}t \end{aligned}\right\} \tag{3.10}$$

If the matrix X can be predetermined or the collection of the data planned (X is truly a design matrix) then X can be chosen to make $(X'X)^{-1}$ as small as possible and hence $\hat{\beta}$ (or $t'\hat{\beta}$) as close to β (or $t'\beta$) as possible. One of the conditions for this kind of design optimality is that the columns of X should be orthogonal (see, for example, Rao (1973) p.235). With unplanned observations, if some of the columns of X are highly correlated or collinear and the matrix $X'X$ is subsequently ill-conditioned or nearly singular, then the matrix $(X'X)^{-1}$ is large and the estimate of β has a large variance matrix and may be described as unstable.

The presence of multicollinearity has been characterized in several ways. Marquardt (1970), Snee (1973) and Marquardt and Snee (1975) propose that the diagonal elements of the inverse of the correlation matrix be used as indicators of multicollinearity and instability of least squares estimates. These diagonal elements which they call variance inflation factors become infinite as the correlation of any independent variable with the others approaches unity. As

an example, consider the correlation matrix:

$$\frac{a}{b} I_{pxp} + \frac{(b-a)}{b} J_{pxp} \quad , \qquad b \geq a > 0$$

in which the matrix $J_{pxp} = 11'$. Then the inverse of this correlation matrix is

$$\frac{b}{a} I_{pxp} - \frac{b}{a} \left(\frac{b-a}{p(b-a)+a} \right) J_{pxp}.$$

If a=b the correlation matrix is the identity matrix (the design matrix X has orthogonal columns) and the inverse of the correlation matrix is the identity matrix. The variance inflation factors in this case are all equal to one and this suggests that the least squares estimates of β and linear functions of β are relatively precise (for fixed σ^2) from equation (3.10). If b > a the variance inflation factors are greater than one. If, for example, $b=10^3$, a=1 and p=3 then $X'X$ in correlation form equals,

$$\begin{bmatrix} 1 & .999 & .999 \\ .999 & 1 & .999 \\ .999 & .999 & 1 \end{bmatrix}$$

and the inverse of this matrix is,

$$\begin{bmatrix} 666.7779 & -333.2221 & -333.2221 \\ -333.2221 & 666.7779 & -333.2221 \\ -333.2221 & -333.2221 & 666.7779 \end{bmatrix}$$

In this case, the variance inflation factors of more than 600 suggest that the estimators of the components of β have high variance or are very unstable (for fixed σ^2). In the limit as a tends to zero the correlation matrix equals J_{pxp} which is singular and the parameter vector β is non-identifiable.

Other characterizations of multicollinearity based on measures of

conditioning of matrices (the value of the determinant for example) are possible. However, multicollinearity and its effect on the least squares estimates of parameters is probably illustrated best by transforming the model in the first condition of the Gauss-Markov Theorem to a canonical form,

$$y = Z\alpha + \epsilon \tag{3.11}$$

in which $Z = XP'$ and $\alpha = P\beta$, and P' is the orthogonal matrix whose columns are the normalised eigenvectors of $X'X$. Thus, $X'X = P' \Lambda P$ where Λ is the diagonal matrix of eigenvalues of $X'X$, and $\|\alpha\|^2 = \|\beta\|^2$. The equations which are analogues of (3.8), (3.9) and (3.10) are,

$$\hat{\alpha} = \alpha + \Lambda^{-1}Z'\epsilon \tag{3.12}$$

$$t'\hat{\alpha} = t'\alpha + t'\Lambda^{-1}Z'\epsilon \tag{3.13}$$

and

$$\left.\begin{aligned} \text{var}(\hat{\alpha}) &= \sigma^2\Lambda^{-1} \\ \text{var}(t'\hat{\alpha}) &= \sigma^2 t'\Lambda^{-1}t \end{aligned}\right\} \tag{3.14}$$

When multicollinearity is present, the ill-conditioned matrix $X'X$ has one or more small eigenvalues. Thus, in the canonical form of the model, one or more of the components of α corresponding to the small eigenvalues, or small diagonal elements of Λ are estimated with a high variance or imprecisely (see equation (3.14)). For example, the correlation matrix introduced earlier has eigenvalues,

$$\lambda = \begin{cases} a/b & \text{with multiplicity } p-1 \\ p-(p-1)(a/b) & \end{cases}$$

Thus for the case, $b=10^3$, $a=1$ and $p=3$;

$$\Lambda = \begin{bmatrix} 2.998 & 0 & 0 \\ 0 & .001 & 0 \\ 0 & 0 & .001 \end{bmatrix}$$

so that,

$$\Lambda^{-1} = \begin{bmatrix} 0.3336 & 0 & 0 \\ 0 & 1,000 & 0 \\ 0 & 0 & 1,000 \end{bmatrix} .$$

So, from equation (3.14), the estimation of α_2 and α_3 in this little example is relatively imprecise as $\lambda_2^{-1} = 1,000$ and $\lambda_3^{-1} = 1,000$. The estimation of α_1 is however very precise (more precise than if $X'X$ were orthogonal). More generally, from a perusal of equation (3.14) estimation in the direction of eigenvectors of $X'X$ which correspond to large eigenvalues is relatively precise while estimation in the directions which correspond to small eigenvalues of $X'X$ is relatively imprecise. An estimate of $t'\beta$ where the vector t equals kp_i (k is a scalar, p_i is the normalised eigenvector corresponding to the ith eigenvalue of $X'X$, λ_i) has variance equal to,

$$k^2 \text{var}(p_i' P' \hat{\alpha}) = \sigma^2 \left(\frac{k^2}{\lambda_i} \right) \qquad (3.15)$$

For fixed k and σ^2 this variance is maximised when t is the normalised eigenvector corresponding to the minimum eigenvalue of $X'X$, λ_{min}. Thus the direction in which prediction is the least precise is the direction of the eigenvector associated with λ_{min}. Thus the magnitude of a multicollinearity problem is related to a "set of directions of interest" in which estimation and prediction are required. The quoted phrase in the last sentence is due to Rao (1975) p.112 who stresses that a multicollinearity problem is not merely a conditioning problem, which can be measured by perhaps noting that variance inflation factors are large or that a determinant is small, but is a problem concerning relative precision of estimation and prediction in certain directions of the estimation space. A similar view has been expressed by Silvey (1969).

Various methods have been suggested for dealing with data which exhibit multicollinearity. Since collinearity or high correlation

between "independent" variables has the statistical interpretation that the variables are different labels for the same factor, the conventional statistical strategy in dealing with multicollinearity has been variable selection - of two highly correlated input variables the one which is least highly correlated with the dependent variable is deleted. Hoerl and Kennard (1970a) describe this as dropping factors "to destroy the correlation bonds among the X_i used to form $X'X$" and claim that if the intention is prediction for control and optimization the experimenter is "left with a set of dangling controllables or observables". On a priori grounds variable selection, motivated by a desire to improve conditioning and not by the principle of parsimony, may result in a model which lacks physical credibility - important input variables may be deleted. Kendall (1975) p.101 contains the following example and warning:

> "Criteria of discard by reference to correlations themselves are dangerous. Consider the case of a medical man concerned with conditions of the spine such as displaced vertebrae. Measurements which he might take on the body include the length of the legs. Now the length of one leg is so highly correlated with the length of the other that it might be regarded as a waste of time to measure both. But if we reject one variable as unnecessary we should miss an important contributor to spinal deformation, the difference of leg lengths."

In this quoted passage, in which Kendall's disapproval of such variable selection practices is rather clear, an alternative procedure for dealing with multicollinearity is alluded to, namely, replacing two or more highly correlated input variables with a linear combination of the same variables (in Kendall's example replacing the variables "length of left leg" and "length of right leg" by "the difference in leg lengths"). Such a strategy has been mentioned by Rao (1975) who points out that multicollinearity can be artificially introduced into or removed from a regression analysis by replacing some input variables with hopefully "equally meaningful" linear combinations of those variables.

A limiting case of this heuristic coalescing of highly correlated input variables into one variable, or a set of less highly correlated

variables, is provided by principal component analysis. In principal component analysis a matrix X of n observations on p variables is transformed, as in equation (3.11), to an orthogonal matrix Z of n observations on p linear combinations of the p variables (the original p input variables are replaced by p uncorrelated input variables). Such a transformation, while useful in discovering the structure of, or interdependencies in, a multivariate data set does not, however, overcome the multicollinearity problem itself. This was demonstrated earlier when the linear model form was transformed into a canonical form in equation (3.11). Regressing the dependent variable y on the p principal components does however illuminate the directions in the estimation space in which estimation and prediction are relatively imprecise (see equations (3.14) and (3.15)). Several authors, including Kendall (1957), Massy (1965) and Greenberg (1975) have however suggested principal component regression as a means of "overcoming" the multicollinearity problem. Instead of variable selection on the input variables they suggest variable selection on the principal components by either;

(a) deleting components with the smallest eigenvalues,

or,

(b) deleting components which are not highly correlated with the dependent variable.

Of course, both of these component selection criteria may lead to quite different estimates of the parameter vector β. The component selection procedure (a) ignores the correlations between the dependent variable y and the principal components of the input variables. It is quite possible then, that the principal component or eigenvector associated with the smallest eigenvalue of $X'X$, which is a sure candidate for deletion, might lie in the same direction as, or be a scalar multiple of, the unknown parameter vector β. In this case, the retained principal components are all orthogonal to this deleted eigenvector and the unknown parameter vector β, so that no linear combination of these retained principal

components will give an estimated parameter vector lying in the direction of the true "unknown" parameter vector β. Using this selection criterion and dropping the last p-r "minor" principal components, while it may result in a reduction in the trace of the variance of estimates β_r^+ of β compared with the least squares estimator, that is,

$$\left[\text{tr}(\text{var } \beta_r^+) = \sigma^2 \sum_{i=1}^{r} \frac{1}{\lambda_i}\right] \quad < \quad \left[\text{tr}(\text{var } \hat{\beta}) = \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i}\right] \quad ,$$

results in a biased estimator for β and an increased residual sum of squares. The selection procedure (b), which takes note of the correlation between the dependent variable y and the principal components, also results in a reduction in the trace of the variance of estimates of β but the reduction is not necessarily as great as that achieved using (a) as the p-r principal components dropped are not necessarily those associated with the smallest p-r eigenvalues. However, the increase in lack of fit or residual sum of squares over least squares, in using this criterion, is not as great as that incurred using (a). Both criteria then decrease the trace of the variance matrix of the estimated parameters but the expense incurred in reducing this variance is a non-zero bias and an increase in lack of fit. However criterion (a) attaches more weight to variance reduction than criterion (b). The use of these component selection techniques which play off variance reduction against bias inflation, for overcoming the multicollinearity problem suggests that a consideration of classes of biased estimators may lead to estimation methods which are relatively insensitive to multicollinearity. Biased estimation as a direct consequence of relaxing the fifth condition of the Gauss-Markov Theorem is introduced in section 3.5 and a thorough review of biased estimation procedures for tackling multicollinearity is presented in Chapter 4.

Of course, if it is feasible, the most preferred method of overcoming the multicollinearity problem is to augment the data with an additional collection of observations on the variables. If the

input variables are controllable (this means that the original data set arises from a poorly designed experiment and the whole problem of multicollinearity could have been avoided with a carefully chosen design) then it is possible by judiciously choosing additional values of the input variables to improve estimation and prediction in the directions of the estimation space in which they are least precise. Silvey (1969) has shown that to overcome the imprecise estimation in directions of the estimation space corresponding to small eigenvalues of $X'X$, which is essentially the multicollinearity problem, additional observations of the dependent variable should be taken at values of the input variables which are scalar multiples of the minor principal components of $X'X$. For example, suppose that the matrix $X'X$ has $p-1$ large eigenvalues and one eigenvalue, λ_{min}, close to zero. Then estimation is relatively imprecise in the direction of the normalised eigenvector p_{min} corresponding to λ_{min} or the last principal component of X. From equation (3.15),

$$\begin{aligned} \text{var}(p'_{min}\,\hat{\beta}) &= \text{var}(p'_{min}P'\hat{\alpha}) \\ &= \frac{\sigma^2}{\lambda_{min}} \end{aligned}$$

Suppose an additional observation of the dependent variable, y_{n+1} is taken at $x_{n+1} = cp_{min}$ where c is a scalar. Then the model for the augmented data set is now:

$$\begin{pmatrix} y \\ \cdots \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} X \\ \cdots \\ x'_{n+1} \end{pmatrix}\beta + \begin{pmatrix} \varepsilon \\ \cdots \\ \varepsilon_{n+1} \end{pmatrix}$$

where the error ε_{n+1} associated with the additional observation is assumed to be uncorrelated with errors in the original model and is assumed to have zero expectation and variance σ^2. The least squares estimator of β now becomes:

$$\hat{\beta}_{aug} = \beta + (X'X + x_{n+1}x'_{n+1})^{-1}(X' \vdots x_{n+1}) \begin{pmatrix} \epsilon \\ \dots \\ \epsilon_{n+1} \end{pmatrix}$$

However, since,

$$(X'X + x_{n+1}x'_{n+1})p_{min} = X'Xp_{min} + c^2 p_{min}p'_{min}p_{min}$$

$$= (\lambda_{min} + c^2)p_{min}$$

and, for i=1,2,...,p-1,

$$(X'X + x_{n+1}x'_{n+1})p_i = X'Xp_i + c^2 p_{min}p'_{min}p_i$$

$$= \lambda_i p_i \quad ,$$

the pxp matrices $X'X$ and $X'X + x_{n+1}x'_{n+1}$ have p-1 equal eigenvalues and p equal eigenvectors. The smallest eigenvalue λ_{min} associated with p_{min} in the original model is however increased by an amount c^2 in the augmented model. Thus the variance of an estimate in the direction of the normalised eigenvector p_{min} becomes,

$$\text{var}(p'_{min}\hat{\beta}_{aug}) = \frac{\sigma^2}{\lambda_{min}+c^2}$$

so that for "large enough" c the precision of estimation in the direction for which prediction was most imprecise is greatly improved.

Silvey has also shown that if instead of "improving estimation where it is most imprecise" the aim is to improve the estimation of a specific linear combination of β, $t'\beta$, by taking a further observation y_{n+1} at x_{n+1} with $\|x_{n+1}\|^2$ constrained to equal a constant k, then the optimum direction of x_{n+1} for improving the precision of estimation of $t'\beta$ is that of the vector,

$$k(X'X + kI)^{-1}t \quad . \qquad (3.16)$$

This result does have immediate intuitive appeal. If the matrix

$X=0$ and there is no information about β contained in the original data set, then the best direction for an additional observation to estimate $t'\beta$ is the direction of t. If k is large compared to the components of $X'X$, and is made arbitrarily large, the new observation x_{n+1} dominates the original data set so that it makes sense that the direction of x_{n+1} should tend towards the direction of t. If however k is small then information about $t'\beta$ in the original data set is used to fix a direction of x_{n+1}.

In many circumstances, however, the collection of further data to augment the original multicollinear data set is either not practicable or not possible. Even if further collection is possible it may not be possible to control the collection in the manner suggested by Silvey - values of the input variables may be determined by the system under study and may not be subject to experimental control. Thus when more data is not available and multicollinearity is present in the data the most that can be done classically is an analysis of the canonical structure of the input variables to seek out the directions in which relatively precise estimation and prediction are possible. However a consideration of the classes of estimators alluded to earlier may provide a way forward (see Chapter 4).

Several requirements are contained in the first condition of the Gauss-Markov Theorem. A failure to meet these requirements when the result of the Gauss-Markov Theorem is applied can cause problems. Some of these problems have been referred to in the preceding discussion. There are other aspects of this first condition that have not been discussed. The case when X is stochastic has not been treated but is touched upon in the next section. The unknown vector of parameters β has been assumed to be constant but may, in some circumstances, be regarded as a random variable with unknown mean and variance. Modifying the assumption that β is constant leads to a study of <u>random coefficient regression models</u>. There is clearly much more to the first condition of the Gauss-Markov Theorem than that which has been discussed here.

3.2 Errors in the Variables

Suppose the variable y^* and the p-variate vector x^* have an exact theoretical linear relationship,

$$y^* = (x^*)'\beta$$

where β is an unknown vector of p parameters. Then n realizations of the variable y^* and its corresponding vector variable x^* may be written in the form,

$$y^* = X^*\beta \qquad (3.17)$$

where y^* is an $n\times 1$ vector of realizations of the variable y^* and X^* is an $n\times p$ matrix whose rows are the n realizations of $(x^*)'$. In practice the variables y^* and X^* may not be directly observable or precisely measurable. Those variables which are measured directly and precisely are y and X. Three non-trivial situations may arise:

(i) $y^* \neq y$, $X^* = X$. In this case the p input variables are truly observed without error but the output or dependent variable y^* is observed indirectly. If the error in observation of y^* is defined to be $y-y^* = \epsilon$, say, then, from equation (3.17),

$$y-\epsilon = X^*\beta$$

or,

$$y = X\beta + \epsilon$$

If the error vector ϵ is assumed to be a random variable with zero expectation and variance matrix $\sigma^2 I$ then the usual linear model form is recovered. Clearly, the Gauss-Markov Theorem caters for the case in which the variable y^* is measured with a stochastic error component adjoined.

(ii) $y^* \neq y$, $X^* \neq X$. In this case all variables are measured with errors. If the error in observation of y^* is defined to be $y-y^* = \epsilon$ and the error in observation of X^* is defined to be $X-X^* = \Delta$ then, from equation (3.17),

$$y-\epsilon = (X-\Delta)\beta$$

or,

$$y = X\beta + \epsilon - \Delta\beta$$

If the presence of the errors in the input variables is ignored and the result of the Gauss-Markov Theorem is applied to the observed data the resulting estimator of the parameter vector β may be written,

$$\beta + (X'X)^{-1}X'\epsilon - (X'X)^{-1}X'\Delta\beta \qquad (3.18)$$

where it is assumed the $n\times p$ matrix X has full column rank p. Two situations seem to be of interest to statisticians - the case when ϵ is a random variable but Δ is a constant matrix determined by the value of X and some consistent rounding rule or constant insensitivity in a measuring device, and, the case when both ϵ and Δ are random variables. Both these situations and the merits of equation (3.18) as an estimator of β are discussed in the following subsections.

(iii) $\underline{y^* = y \quad , \quad X^* \neq X}$. In this case the variable y^* is observed directly and precisely but the p input variables are measured indirectly and imperfectly. If the error in observation of X^* is defined to be $X-X^* = \Delta$ equation (3.17) becomes,

$$y = (X-\Delta)\beta$$

If the observed matrix X is of full column rank, the usual least squares estimation procedure gives as an estimator of β,

$$\beta - (X'X)^{-1}X'\Delta\beta \qquad (3.19)$$

If Δ, the unknown error matrix, is constant and non-zero this "estimator" of the unknown parameter vector β is badly biased (in fact the wrong set of equations, $y=X\beta$, has been tackled). However, as long as X is sufficiently close to X^* (i.e. Δ sufficiently small) and as long as $X'X$ is not ill-conditioned, equation (3.19)

shows that the usual least squares procedure may give a good enough approximation to β. If Δ is a matrix of n observations on p random variables such that each observation, or row of Δ, is independently and identically distributed with mean the zero vector and covariance matrix $\Sigma_{p\times p}$, that is,

$$E(\Delta) = 0_{n\times p} \quad , \quad E(\Delta'\Delta) = n\Sigma_{p\times p}$$

and if $\frac{1}{n} X^{*\prime} X^{*} \to C$ as $n\to\infty$ where C is positive definite, then the following limits in probability hold,

$$\operatorname{plim}\left(\frac{1}{n} X'X\right) = \operatorname{plim}\left(\frac{1}{n} X^{*\prime}X^{*} + \frac{1}{n} X^{*\prime}\Delta + \frac{1}{n} \Delta' X^{*} + \frac{1}{n} \Delta'\Delta\right)$$

$$= C + \Sigma$$

$$\operatorname{plim}\left(\frac{1}{n} X'\Delta\beta\right) = \operatorname{plim}\left(\frac{1}{n} X^{*\prime}\Delta + \frac{1}{n} \Delta'\Delta\right)\operatorname{plim}\ \beta$$

$$= \Sigma\beta$$

Thus the limit in probability of the estimator in equation (3.19) is given by,

$$\operatorname{plim}(\beta-(X'X)^{-1}X'\Delta\beta) = \beta - (\operatorname{plim}\ \frac{1}{n} X'X)^{-1}(\operatorname{plim}\ \frac{1}{n} X'\Delta\beta)$$

$$= \beta - (C + \Sigma)^{-1}\Sigma\beta$$

$$= (I - \{C + \Sigma\}^{-1}\Sigma)\beta \ . \qquad (3.20)$$

The equation (3.19) does not provide a <u>consistent</u> estimator of β. As a special case, if Σ is a diagonal matrix and C is diagonal (the design is orthogonal) the components of the resulting estimator tend to be scalar multiples of the corresponding components of β - the dominant effect of the measurement errors in X is a multiplicative scaling of the estimates of the components of β so that the estimate of each component of the parameter vector is somewhat insulated from the others. Further discussion of this situation, in which the input variables are observed with a stochastic error but the "dependent" variable is measured without error is easily

derived from the discussion of the more general case (ii) which takes place in section 3.22.

3.21 Non-stochastic errors in the input variables.

In this instance of case (ii) above, the true model is considered to be the usual linear model form,

$$y = X^*\beta + \epsilon$$

while the candidate model for analysis is,

$$y = X\beta + \epsilon$$

where the matrix $X = X^*+\Delta$. The recorded matrix X, the actual but unrecorded matrix X^*, and the error matrix Δ are considered to be fixed so that Δ may be thought of as a constant, non-stochastic matrix of rounding errors or censored bits of the true matrix of inputs X^*. The usual procedure, when it is suspected that the matrix of recorded or observed input variables X contains a non-stochastic error component, is to proceed with affected innocence with an application of the Gauss-Markov Theorem. The resulting estimator of β, given in equation (3.18), has bias,

$$-(X'X)^{-1}X'\Delta\beta \qquad (3.21)$$

under the usual assumptions on ϵ. Several points immediately arise from an inspection of equation (3.21). If $X'X$ is ill-conditioned the bias is likely to be large, even when the components of Δ are "fairly small". If one of the independent variables is recorded exactly, without any rounding or censoring, the corresponding column of Δ is the zero vector and the corresponding component of the parameter vector β does not contribute to the bias term. This means that bias depends only on those components of β whose associated input variables are measured with error, but, it also means that least squares estimates of the components of the parameter vector β which correspond to precisely known input variables are not insulated from the errors in the imprecisely known input variables. Also, on the negative side, even if the approximate magnitude of Δ is known, in general β - the object of

the estimation exercise - is not known, so that knowledge of the bias is not obtainable. One ray of hope might lie in the fact that a necessary and sufficient condition for the bias to be zero is that Δ should satisfy,

$$X'\Delta = 0_{p\times p} .$$

In general, however, it is not possible to arrange for the experimental design and the rounding or censoring mechanisms in the measuring devices to satisfy this condition. In a sense then, the problem of fixed but unknown measuring errors in the input variables is intractable - in order to ascertain the effects of the unknown errors on estimates of the unknown parameters, the errors and parameters or at least their magnitudes have to be known.

If some knowledge of these magnitudes or knowledge of bounds on the errors and parameter values is available some headway is possible. Swindel and Bower (1972) have produced what they term "useful" bounds for the bias.

A brief summary of their arguments is as follows:

When the linear combination $t'\beta$ is estimated by least squares and non-stochastic errors are present in the input variables the squared bias of the estimator is, from equation (3.21),

$$\|t'(X'X)^{-1}X'\Delta\beta\|^2$$

and the variance of this estimator is, from equation (3.18) and the usual assumptions on ϵ,

$$\sigma^2 t'(X'X)^{-1}t.$$

Swindel and Bower call,

$$\frac{\|t'(X'X)^{-1}X'\Delta\beta\|^2}{\sigma^2 t'(X'X)^{-1}t}$$

the squared relative bias of the estimator. This squared relative

bias can be written,

$$\frac{1}{\sigma^2}\ \frac{T'\Delta\beta\beta'\Delta'T}{T'T} \quad ,$$

where $T = X(X'X)^{-1}t$, from which it can be seen that,

$$0 \leq \text{the relative bias} \leq \frac{\|\Delta\beta\|}{\sigma} \quad , \qquad (3.22)$$

as $\|\Delta\beta\|^2$, which is the maximum eigenvalue, and in fact the only nonzero eigenvalue, of $\Delta\beta\beta'\Delta'$, is the maximum value for all T of the squared relative bias. Note that the inequality (3.22) holds for all X and X^* of full column rank and for all vectors t. Thus from inequality (3.22) if $\|\Delta\beta\|/\sigma$ is known or suspected to be small, the bias in the usual least squares estimator of $t'\beta$ caused by the non-stochastic error matrix Δ is negligibly small.

The usefulness of these bounds on the bias is questionable. However Davies and Hutton (1975) have extended this result of Swindel and Bower. By introducing the following measure of ill-conditioning of the matrix X', called the "distance of the matrix X' from singularity";

$$\rho(X') = \min[\{\text{tr}(D'D)\}^{\frac{1}{2}} : \ (X+DR)' \text{ is singular}] \qquad (3.23)$$

(where D is an $n\chi p$ matrix and the matrix $R = \text{diag}(r_1,\ldots,r_p)$ where the value r_i is the absolute value of the suspected measurement or rounding error in the ith column of X) they have been able to construct a more practical bound on the relative bias. A brief summary of their arguments is as follows:

The distance of X' from singularity can be written,

$$\rho(X') = \|R(X'X)^{-1}R\|^{-\frac{1}{2}} \quad ,$$

where $\|\ldots\|$ denotes the Hilbert norm, $\sup\frac{\|Ax\|}{\|x\|}$ of A, provided $X'X$ is nonsingular. If R is also nonsingular then $\rho(X')$ is equal to

the square root of the minimum eigenvalue of $R^{-1}X'XR^{-1}$. Suppose the elements of R are more precisely defined to be "informed guesses" of the square roots of the diagonal elements of $\frac{1}{n}\Delta'\Delta$, then,

$$\frac{\|\Delta\beta\|}{\sigma} \leq \sqrt{n}\,\frac{|R\beta|}{\sigma}$$

where $|\ldots|$ denotes the sum of the absolute values of the components of a vector. Thus,

$$0 \leq \text{the relative bias} \leq \sqrt{n}\,\frac{|R\beta|}{\sigma} \quad . \tag{3.24}$$

To make this upper bound fully operational the least squares estimator of β, equation (3.18), which is denoted by $\hat{\beta}$, may be substituted for β. However $\hat{\beta}$ is biased so that $|R\hat{\beta}|$ is a biased estimator of $|R\beta|$. But the absolute value of the bias in $|R\hat{\beta}|$ is given by,

$$\left|\, |R\beta| - |RE(\hat{\beta})| \,\right| \leq \frac{\sqrt{np}\;|R\beta|}{\rho(X')}$$

Thus, if $\rho(X')$ is "at least several times the value" of $\sqrt{np}$ the bias incurred in using $|R\hat{\beta}|$ to estimate $|R\beta|$ is negligible. It only remains to estimate σ in inequality (3.24). The usual procedure gives,

$$\hat{\sigma}^2 = y'(I-X(X'X)^{-1}X')y/(n-p). \tag{3.25}$$

The bias in this estimate of σ^2 is clearly,

$$E(\hat{\sigma}^2) - \sigma^2 = \beta'\Delta'(I-X(X'X)^{-1}X')\Delta\beta/(n-p)$$

so that,

$$0 \leq E(\hat{\sigma}^2) - \sigma^2 \leq \|\Delta\beta\|^2/(n-p).$$

Thus if the upper bound of Swindel and Bower on the relative bias term is small then this upper bound on the bias of $\hat{\sigma}^2$ is very small too.

Consequently Davies and Hutton suggest that a safe approximation to an upper bound on the relative bias of the least squares estimator $\hat{\beta}$ in equation (3.18) is,

$$\frac{\sqrt{n}\ |R\hat{\beta}|}{\hat{\sigma}} \tag{3.26}$$

or after further manipulation,

$$\frac{\sqrt{np}\ \|X\hat{\beta}\|}{\hat{\sigma}\ \rho(X')} \tag{3.27}$$

The use of the bound in (3.26) is only recommended when $\rho(X')$ is much greater than $\sqrt{np}$, that is when the distance of X' from singularity is large. The bound in (3.27), which incorporates $\rho(X')$, demonstrates explicitly the joint effect of near singularity and errors on the bias term. Obviously small values of $\rho(X')$ should be avoided if the effect of non-stochastic errors in the input variables on least squares estimates of β is to be minimized.

This work of Swindel and Bower, and Davies and Hutton is a first step in formalising the well known hypersensitivity, in the presence of even moderate multicollinearity, of the least squares estimator of β to small deterministic errors in the input variables. The distance from singularity $\rho(X')$ seems to be a worthwhile index of this hypersensitivity as it enables, in conjunction with the bound (3.27), a researcher to discover whether the bias caused by errors in the input variables is likely to be serious. Other indices of sensitivity or instability have been proposed. They are discussed in the next section.

3.22 Stochastic errors in the input variables.

In this instance of case (ii) both ϵ and the rows of Δ are assumed to be random variables. Much work has been published on the effect of such stochastic errors on the usual least squares estimator but most of it has referred to the bivariate case in which there is only one input variable (see, for example, Cochran (1972)). However,

several alternative estimation procedures for the multivariate case have been suggested in the literature. Foremost among these would be instrumental variable estimation, but in many non-econometric problems there are difficulties in finding suitable jointly observable instrumental variates with which to operate. Thus the usual procedure adopted by researchers is to apply the result of the Gauss-Markov Theorem in spite of the stochastic error component in X and in spite of the existence of possible alternative procedures. Some work has been devoted to an examination of this technique.

The resulting least squares estimator, equation (3.18), can be rewritten as,

$$\beta + \{(X^*+\Delta)'(X^*+\Delta)\}^{-1}(X^*+\Delta)'(\epsilon-\Delta\beta). \qquad (3.28)$$

Two differing approaches have been used to investigate equation (3.28).

Hodges and Moore (1972) proceeded under the assumption that in any experimental situation the error matrix Δ is small so that the series expansion of the inverse of $(X^*+\Delta)'(X^*+\Delta)$ could be conveniently truncated and approximations to the expectation, bias and variance of (3.28) easily evaluated.

The series expansion of the inverse of the aforementioned matrix is,

$$(X^{*\prime}X^*)^{-1} - (X^{*\prime}X^*)^{-1}D(X^{*\prime}X^*)^{-1} + (X^{*\prime}X^*)^{-1}D(X^{*\prime}X^*)^{-1}D(X^{*\prime}X^*)^{-1} \ldots$$

where $D = (\Delta'X^* + X^{*\prime}\Delta + \Delta'\Delta)$ and it is assumed that the $n\times p$ matrix X^* has full column rank p. Hodges and Moore truncated this series by ignoring all terms in $\Delta'\Delta$ and "higher powers" of Δ except the "principal quadratic term",

$$-(X^{*\prime}X^*)^{-1}\Delta'\Delta\beta - (X^{*\prime}X^*)^{-1}\Delta'\Delta(X^{*\prime}X^*)^{-1}X^{*\prime}\epsilon \ .$$

As an approximation to the estimator in expression (3.28) Hodges and Moore gave the following expression,

$$\beta + (X^{*\prime}X^*)^{-1}(X^*+\Delta)'\epsilon - (X^{*\prime}X^*)^{-1}\Delta'X^*(X^{*\prime}X^*)^{-1}X^{*\prime}\epsilon$$
$$- (X^{*\prime}X^*)^{-1}(X^*+\Delta)'\Delta(X^{*\prime}X^*)^{-1}X^{*\prime}\epsilon - (X^{*\prime}X^*)^{-1}(X^*+\Delta)'\Delta\beta \qquad (3.29)$$

If the components of ϵ are assumed to be distributed independently of the rows of Δ in such a way that,

$$E(\epsilon) = 0_{n\times 1} \quad , \quad E(\epsilon\epsilon') = \sigma^2 I_{n\times n}$$

$$E(\Delta) = 0_{n\times p} \quad , \quad E(\Delta'\Delta) = n\Sigma_{p\times p} \quad , \quad E(\Delta'\epsilon) = 0_{p\times 1}$$

then the expected value of (3.29) is,

$$\beta - n(X^{*\prime}X^{*})^{-1}\Sigma\beta.$$

Thus, an approximate expression for the bias in the usual least squares estimator is,

$$-n(X^{*\prime}X^{*})^{-1}\Sigma\beta.$$

Hodges and Moore suggested replacing $(X^{*\prime}X^{*})$ by $(X'X)$, β by expression (3.18), and Σ by "whatever information is available" to arrive at a rough estimate of the bias. Again it can be seen that the conditioning of $X^{*\prime}X^{*}$ and $X'X$ has an important consequence for the bias of the least squares estimator.

If the expression (3.28) is approximated further by ignoring all terms in $\Delta'\Delta$ and higher powers of Δ the result is,

$$\beta + (X^{*\prime}X^{*})^{-1}(X^{*}+\Delta)'\epsilon - (X^{*\prime}X^{*})^{-1}\Delta'X^{*}(X^{*\prime}X^{*})^{-1}X^{*\prime}\epsilon$$
$$- (X^{*\prime}X^{*})^{-1}X^{*\prime}\Delta(X^{*\prime}X^{*})^{-1}X^{*\prime}\epsilon - (X^{*\prime}X^{*})^{-1}X^{*\prime}\Delta\beta \qquad (3.30)$$

This approximation to the estimator in (3.28) has zero bias which is an indication of its crudity as an approximation. However, the variance-covariance matrix of this approximation to the estimator is able to be derived from the assumptions on ϵ and Δ, given earlier, as,

$$\sigma^2(X^{*\prime}X^{*})^{-1} + (X^{*\prime}X^{*})^{-1}X^{*\prime}E(\Delta\beta\beta'\Delta')X^{*}(X^{*\prime}X^{*})^{-1} \ .$$

This expression for the variance-covariance matrix is used by Hodges and Moore (1972, p.189) to claim that the estimated covariance matrix $\hat{\sigma}^2(X'X)^{-1}$ will not be far from $\sigma^2(X^{*\prime}X^{*})^{-1}$, but their notation and reasoning seem confused. They also proposed a "sensitivity analysis" for calculating the sensitivity of each

estimate of a component of β to each observation in the recorded input matrix. However their defined sensitivity depends on the unobserved matrix X^*, the "true" least squares estimator of $\beta_* (X^{*\prime}X^*)^{-1}X^{*\prime}y$, and the "true" residual vector $(I-X^*(X^{*\prime}X^*)^{-1}X^{*\prime})y$.

The second, and more fruitful approach, to an investigation of the properties of expression (3.28) has come from Davies and Hutton (1975). Hodges and Moore assumed that the matrix Δ was small so that with fixed sample size n the expected value of higher powers of Δ could be ignored. Davies and Hutton however examined the asymptotic performance of (3.28) and (3.25) as the sample size, n, tended to infinity. A brief summary of their arguments and results is as follows;

If the components of ϵ and the rows of Δ are assumed to be distributed in the manner described previously and if

$$\lim_{n\to\infty} \frac{1}{n} X^{*\prime}X^* = C \quad , \quad \text{where C is a positive definite}$$

matrix, then in the manner of the arguments leading to equation (3.20) the limit in probability of the estimator in expression (3.28) is,

$$\beta - (C + \Sigma)^{-1}\Sigma\beta \tag{3.31}$$

If Σ is nonsingular (a similar argument holds if Σ is singular) and diagonal, the asymptotic bias can be related to the distance from singularity (equation (3.23)) by,

$$\begin{aligned} \sup \frac{\|(C + \Sigma)^{-1}\Sigma\beta\|}{\|\beta\|} &= \|\Sigma^{\frac{1}{2}}(C + \Sigma)^{-1}\Sigma^{\frac{1}{2}}\| \\ &= [\rho((C + \Sigma)^{\frac{1}{2}})]^{-2} \\ &= \operatorname{plim}\left(\frac{n}{[\rho(X')]^2}\right) \end{aligned}$$

Thus if the distance from singularity $\rho(X')$ is small, large errors may be present in the estimator (3.28). It is also possible to show that, provided the fourth order moments of the elements of Δ exist,

$$\sqrt{n}(\hat{\beta} - (X^{*\prime}X^{*} + n\Sigma)^{-1}X^{*\prime}X^{*}\beta)$$

has an asymptotically normal distribution with mean the zero vector and a complicated covariance matrix. Further to this result if Σ can be written $\frac{1}{\sqrt{n}} S$ then,

$$\sqrt{n}(\hat{\beta}-\beta)$$

has an asymptotically normal distribution with mean $C^{-1}S\beta$ and covariance matrix $\sigma^2 C^{-1}$. Notice that if $C^{-1}S = 0_{p\times p}$, and C^{-1} is small, the bias in the estimator (3.28) is likely to be negligible in large samples. If, however C is ill-conditioned this large sample distributional result indicates that the bias is not likely to be negligible. This suggests that a simple kind of relationship holds among errors, conditioning and probable bias. In fact Davies and Hutton show in the manner of the arguments leading to the bound (3.27) that the bias in expression (3.28) is negligible if,

$$\frac{n\sqrt{p}\ |R\beta|}{\rho(X')\sigma} \tag{3.32}$$

is small compared with one. Analogous with the non-stochastic case the diagonal matrix R is composed of "guestimates" of the square roots of the diagonal elements of Σ. They also show that β and σ in bound (3.32) can be safely replaced by their estimators (3.28) and (3.25) if $\rho(X')$ is large compared with $\sqrt{np}$. They also show that,

$$\text{plim}\ \hat{\sigma}^2 = \sigma^2 + \beta'\Sigma\beta - \beta'\Sigma(C + \Sigma)^{-1}\Sigma\beta$$

so that for Σ close to the null matrix the estimate of the covariance matrix of the least squares estimator is not too far, in the limit, from $\sigma^2(X^{*\prime}X^{*})^{-1}$. This places the claim of

Hodges and Moore (1972, p.189) on a slightly more substantial footing.

Again the distance of the matrix X' from singularity, $\rho(X')$ which is really a measure of how far the recorded matrix X' is away from arising out of a possibly singular unrecorded matrix $X^* = X+\Delta$, where the only information about $\Delta_{n\chi p}$ is contained in $R_{p\chi p}$, seems a good index for measuring the combined effect of multicollinearity and stochastic errors on the estimator in expression (3.18). The estimated value of the bound in (3.32) would seem to be more readily obtainable and hence of more practical value than the approximate bias term derived by Hodges and Moore.

In a slightly different context Beaton, Rubin and Barone (1976) have also studied the effect of stochastic errors in the input variables on the least squares estimator of β. Beaton et al were interested in the effect of non-stochastic rounding errors, errors between -0.5 and +0.4999..., in the last recorded digit of each observation in the input matrix, on the least squares estimator. Using a particular multicollinear test problem they generated 1,000 input matrices X^* by adding rectangularly distributed random numbers between -0.5 and +0.4999... to the last digit of each observation in the recorded input matrix X and then proceeded to compute the 1,000 regressions on these equally plausible "true" input matrices. They found huge variations in the 1,000 values of the estimated coefficients resulting from the regressions on the input matrices $X_k^* = X-\Delta_k$, (k=1,...,1000) but most seemed to be clustered around a mean value that did not necessarily coincide with the least square solution for X. They surmised that these mean values about which the solutions were distributed were estimates of the large-sample limit of the equally likely solutions. The theoretical development they gave is as follows.

Let $X^* = X-\Delta$ where Δ is a matrix of plausible rounding errors. Then the "true" least squares solution vector for a particular Δ is,

$$\hat{\beta} = (X^{*\prime}X^{*})^{-1}X^{*\prime}y$$
$$= (X'X - \Delta'X - X'\Delta + \Delta'\Delta)^{-1}(X-\Delta)'y$$

However the true input matrix X^{*} is unknown, but X is known and fixed, and Δ may vary by the construction of the simulation in such a way that,

$$E(\Delta) = 0 \quad , \quad \frac{1}{n}E(\Delta'\Delta) = D = \text{diag}(d_1,\ldots,d_p)$$

Suppose $\frac{1}{n}X'X = C_{xx}$ for all sampled nxp input matrices X, and $\frac{1}{n}X'y = C_{xy}$. Then, since plim $\frac{1}{n}\Delta'X = 0$ and plim $\frac{1}{n}\Delta'y = 0$,

$$\text{plim } \hat{\beta} = (C_{xx} + D)^{-1}C_{xy} \quad . \tag{3.33}$$

The usual least squares estimator (3.18) based on the matrix X is,

$$\hat{\beta}_x = C_{xx}^{-1}C_{xy} \quad . \tag{3.34}$$

It is easy to show that,

$$\hat{\beta}_x - \text{plim } \hat{\beta} = [I - (I + C_{xx}^{-1}D)^{-1}]\hat{\beta}_x \ .$$

If the matrix, $C_{xx}^{-1}D = 0_{pxp}$ then $\hat{\beta}_x = \text{plim } \hat{\beta}$. If $\text{tr}(C_{xx}^{-1}D)$ is close to zero then $\hat{\beta}_x$ is approximately equal to plim $\hat{\beta}$. So it is that Beaton et al call,

$$\text{tr}(C_{xx}^{-1}D) = n \sum_{i=1}^{p} (X'X)_{ii}^{-1} d_i = \text{P.I.} \tag{3.35}$$

a <u>perturbation index</u>. If the perturbation index is close to zero the usual least squares estimator (3.34) is close to plim $\hat{\beta}$ and is relatively stable, under the influence of the unknown Δ, for large n. The perturbation index is likely to be large, and consequently the usual least squares estimator $\hat{\beta}_x$ unstable, when $X'X$ is ill-conditioned. Again the sensitivity of the usual least squares estimator (3.18) to stochastic errors in the input variables is greatest when multicollinearity is present in the input variables.

The plim solution (3.33) which is the large sample limit of the "true" least squares solution to the regression problem may be rewritten as,

$$(X'X + nD)^{-1}X'y \ . \qquad (3.36)$$

In this form the plim solution has the appearance of a generalised ridge estimator (see section 4.222). Perhaps the usual least squares estimator should be replaced by this estimation procedure when the matrix X is known to contain rounding errors and the rounding errors are assumed to be independently and uniformly distributed. Other alternatives to the least squares method of estimation when stochastic errors are present in the input variables have been suggested. The form of these estimators is similar to (3.36). Theil (1971, p.614) has suggested estimating the parameter vector β by,

$$(X'X - n\hat{\Sigma})^{-1}X'y \ . \qquad (3.37)$$

Here Theil has assumed that $\frac{1}{n} X^{*'}X^{*}$ converges to a positive definite matrix C, and $E(\Delta) = 0_{n\times p}$, $E(\Delta'\Delta) = n\Sigma$, $E(\Delta'\epsilon) = 0_{p\times 1}$. The matrix $\hat{\Sigma}$ is a "guestimate" of Σ and is stochastically independent of ϵ and Δ.

Under these conditions, the probability limit of the usual least squares estimator is, from (3.31),

$$\beta - (C+\Sigma)^{-1}\Sigma\beta = (C+\Sigma)^{-1}C\beta$$

whereas the probability limit of Theil's estimator in expression (3.37) is,

$$(C+\Sigma-\hat{\Sigma})^{-1}C\beta \qquad (3.38)$$

Thus, if the "guestimate" $\hat{\Sigma}$ is close to Σ the inconsistency or asymptotic bias of Theil's estimator is small.

In a similar vein Warren, White and Fuller (1974) have proposed and demonstrated the use of the estimator,

$$(X'X - (n-\alpha)\hat{\Sigma})^{-1}X'y \qquad (3.39)$$

where $\hat{\Sigma}$ is some available estimator of Σ and α is some "constant", usually less than n in value, introduced to reduce the mean square error of the estimator. Warren et al have also proposed a small refinement to the estimator (3.39). If the smallest root $\hat{\gamma}$ of,

$$\left|\frac{1}{n} X'X - \gamma\hat{\Sigma}\right| = 0$$

is less than $\frac{n+1}{n}$ then the coefficient of $\hat{\Sigma}$ should be replaced by $[n\hat{\gamma} - (\alpha+1)]$ as a safeguard that the matrix to be inverted is nonsingular. As an aside, if $\hat{\gamma} < \frac{\alpha+1}{n}$ then the estimator is a generalized ridge estimator.

At first sight there seems to be some conflict in appearance between the plim solution (3.36) of Beaton et al and the alternative estimators (3.37) and (3.39) of Theil and Warren et al. Whereas the "estimator" (3.36) modifies the least squares procedure by adding a diagonal non-negative definite matrix to $X'X$, Theil's estimator and the usual form of (3.39) subtract a matrix from $X'X$. The source of this difference is easily detected. The plim solution of Beaton et al results from assuming that the observed matrix of input variables X is fixed and somehow primordial and that the "true" matrix X^* of input variables is unknown but has many equally likely forms. That is, Beaton et al assume X is fixed and X^* is stochastic. However, Theil and Warren et al assume that the unknown "true" matrix of input variables X^* is fixed and primordial and the observed matrix $X=X^*+\Delta$ is stochastic. The plim estimator of Beaton et al seems to be a response to the question, "What if there are errors in the input variables?" whereas the estimators of Theil and Warren et al are responses to the statement "There are errors in the input variables". If $X'X$ is close to singularity it would seem to be safer to respond to the former statement than to admit to the latter, as the form of (3.36) improves the conditioning of the matrix to be inverted while (3.37) is likely to worsen the conditioning problem.

3.23 Some remarks about Outliers.

Unusual or outlying values of the dependent variable (values which have large residuals when a least squares fit is carried out) may arise as "blunders" or mistakes in recording the values of observations, or may be due to unusual combinations of circumstances not adequately allowed for in the formulation of the linear model. If it is certain that an outlier is the result of a recording error then it is desirable to either delete the observation from the data or correct it (smooth it) in some way. If an outlying value is not a blunder then rejection or correction of the observation means that valuable information not contained in the other observations is being discarded or distorted merely because the assumed model cannot handle it. Establishing the source of an outlier, before any action is taken, is therefore very important.

The sensitivity of the method of least squares to such outlying values is well known. It is possible for one or more outliers to dominate a regression producing a fit determined by only one or a few observations, see for example Andrews (1975). However, detecting outliers after a least squares regression has been carried out can be difficult. Various plots of the residuals may reveal outliers but very often the least squares fit, which responds to the presence of the outliers and correspondingly downgrades the fit to the remaining observations, masks these outliers. Thus an observation which might be termed an outlier prior to the regression (on the basis of some prescreening exercise, inspection of plots of y versus x_i for example) may not appear to be an outlier from an examination of the residuals after a regression. However, the impact of an outlier on the resulting least squares estimator may not be as great as the impact of some other observation. Often the structure of X (its principal components and conditioning) conspires to make the least squares procedure relatively insensitive to the outliers and more sensitive to some other observations. Thus, while least squares may frequently be sensitive to an outlier there is no guarantee that the outlier is having a greater influence on

the least squares estimate than any other observation. Fortunately, Cook (1977) has suggested a measure which detects the most influential observations in a least squares analysis. Cook has proposed that the influence or importance of the ith data point in determining the least squares estimate $\hat{\beta}$ of β be measured by;

$$D_i = \frac{\|X(\hat{\beta}_{(-i)} - \hat{\beta})\|^2}{ps^2}$$

where $\hat{\beta}_{(-i)}$ is the least squares estimate of β with the ith observation deleted and s^2 is the usual estimator of σ^2 from equation (3.25), with i ranging from 1 to n. Under the usual distribution theory,

$$\frac{\|X(B-\hat{\beta})\|^2}{ps^2} \leq F_{1-\alpha}(p,n-p)$$

defines a $(1-\alpha)$x100% confidence hyperellipsoid for β so that if $D_i = F_{.50}(p,n-p)$ the removal of the ith data point shifts the least squares estimate to the edge of the 50% confidence region for β based on $\hat{\beta}$ and $\hat{\beta}$ is consequently greatly influenced by the presence of the ith observation. Cook has suggested that the D_i's should be less than $F_{.10}(p,n-p)$ for an "uncomplicated" regression analysis.

If an outlier is detected by one or more of the more usual methods - examination of the studentized residuals for example - and it also has the highest D value of the n observations then the least squares estimate has certainly been overinfluenced by the combination of this observation and the structure of the observations on the input variables. If the observation is certainly a blunder it must be deleted or smoothed. If the observation is not clearly a blunder its presence poses a problem and further collection of data points and a reformulation of the model may be necessary. If an outlier is detected but its D value is relatively small it is having little influence on the least squares estimate. In this case it may be of lesser

importance whether the observation is a blunder or not as the retention or deletion of the observation has little effect on the least squares estimate. However, the presence of the outlier may still raise doubts about the model if the fit is inadequate.

Previously it was mentioned that least squares regression often masks the presence of outliers by (nonteleologically) attaching more weight to them, in the fit, than the other observations. Some work has been directed to finding estimation and fitting techniques which are less sensitive to the outliers. These techniques attach lesser weights to outliers, preserving the large residuals usually associated with them and consequently allowing the clearer detection of blunders or inadequacies in the model. These techniques are mentioned in section 3.4.

3.3 Generalized Least Squares

The third condition of the Gauss-Markov Theorem, as stated in Chapter 2, makes very specific assumptions about the distribution of the error or disturbance vector ϵ. Fortunately these assumptions can be generalized somewhat to cover possible non-standard situations.

Suppose the variance-covariance matrix for ϵ is given by,

$$E(\epsilon\epsilon') = \Sigma_{n\times n}$$

then several situations of increasing complexity can arise:

(i) Σ may be completely known
(ii) Σ may be known up to a constant scale factor σ^2
(iii) Σ may be unknown but may have some known special pattern
(iv) Σ may be completely unknown.

In the first situation, if Σ is completely known and of full rank then the transformation matrix $\Sigma^{-\frac{1}{2}}$ can be applied to the data set so that,

$$z = \Sigma^{-\frac{1}{2}}y \quad , \qquad \xi = \Sigma^{-\frac{1}{2}}e$$
$$E(z) = \Sigma^{-\frac{1}{2}}X\beta = W\beta$$
$$E(zz') = I \ .$$

The model is now $z = W\beta + \xi$ where $E(\xi) = 0$ and $E(\xi\xi') = I$ which is just a special case of the set up defined in the first three conditions of the Gauss-Markov Theorem.

In the second situation the variance-covariance matrix Σ may take one of three forms,

$\sigma^2 I$
$\sigma^2 G$ in which G has full rank n
$\sigma^2 G$ in which G is singular.

In all three forms the scale factor σ^2 is assumed to be unknown. The first form is just that postulated in the third condition of the Gauss-Markov Theorem. The second form $\sigma^2 G$, in which the known matrix G has full rank, can also be catered for under the third condition as it stands with the application of the transformation matrix $G^{-\frac{1}{2}}$ in the manner of the transformation suggested for situation (i). If the covariance matrix has the form $\sigma^2 G$ where G is singular, then no such transformation exists and the Gauss-Markov Theorem cannot be applied. However, Rao (1971, 1973) has developed an analogue of the Gauss-Markov Theorem for this situation:

Given an input matrix X (which may or may not be of full rank) and the (possibly) singular matrix G, then the <u>best linear unbiased estimator</u> of $t'\beta$ (if it is estimable) is $t'\hat{\beta}$, where $\hat{\beta}$ minimizes

$$(y-XB)'T^{-}(y-XB)$$

with $T = G + k^2XX'$, $k \neq 0$ and T^{-} any generalized inverse of T.

Thus the second situation, Σ known up to a constant factor σ^2, is more or less covered by the third condition of the Gauss-Markov Theorem as it stands.

The third and fourth situations require the estimation of other parameters besides β and the possibly unknown scale factor σ^2. The third situation can be treated as a special case of the fourth situation. Concentrating on the fourth situation where Σ is completely unknown there are $\frac{n(n+1)}{2}$ elements of Σ to be estimated as well as the elements of β. With only n observations, the simultaneous estimation of Σ and β is not possible. If, however, m independent replications of y are available, and if the distribution of the error vector ϵ is assumed to be normal, then maximum likelihood estimators of β and Σ are obtainable using standard results from multivariate analysis. Thus the estimation of β in the most complex case requires extra information in the form of replications of the observations and an extra assumption.

Given these various covariance matrix formulations an immediate problem presents itself to the practioner. What happens if an assumed covariance matrix formulation is less complex than the actual covariance matrix? Or more generally, what is the effect on the least squares estimation procedure of misweightings in the assumed covariance matrix? A related problem concerns heteroscedastic variances - what happens to the least squares estimator if case (ii) is assumed when in reality case (iii) or (iv) holds and there is no common scale factor σ^2? Some of the work which has been directed to these problems is reviewed in the next two sections.

3.31 Misweighting

Suppose $E(\epsilon\epsilon') = \sigma^2 W$ but the covariance matrix is assumed to be $\sigma^2 G$, where W and G are both nonsingular. Then the obtained least squares estimator of $t'\beta$ is,

$$t'(X'G^{-1}X)^{-1}X'G^{-1}y$$

whereas the proper least squares estimator is,

$$t'(X'W^{-1}X)^{-1}X'W^{-1}y \quad .$$

Both estimators are unbiased but are not necessarily equal. If the two covariance matrices differ only by a scalar multiple, that is the ratios of assumed weight to proper weight are nearly equal for every data point then the obtained least squares estimator is equal to or not too far from the "true" least squares estimator. The obtained least squares estimator has variance,

$$\sigma^2 t'(X'G^{-1}X)^{-1}X'G^{-1}WG^{-1}X(X'G^{-1}X)^{-1}t$$

whereas the proper least squares estimator has variance,

$$\sigma^2 t'(X'W^{-1}X)^{-1}t \ .$$

The ratio of the variance of the obtained estimator to the variance of the proper estimator clearly has an upper bound given by the maximum eigenvalue of,

$$(X'W^{-1}X)(X'G^{-1}X)^{-1}X'G^{-1}WG^{-1}X(X'G^{-1}X)^{-1} \qquad (3.40)$$

Without too much loss of generality it can be assumed that G=I and $W=\Sigma=\text{diag}(\mu_1,\ldots,\mu_n)$. The problem now becomes that of finding the maximum eigenvalue of,

$$(X'\Sigma^{-1}X)(X'X)^{-1}X'\Sigma X(X'X)^{-1} \ .$$

If it is also assumed that $X'X = I$ then the problem reduces to that of finding the maximum eigenvalue of,

$$(X'\Sigma^{-1}X)(X'\Sigma X). \qquad (3.41)$$

If X' is partitioned by its first column as $[r_1 \vdots R_2]$ and Σ is correspondingly partitioned as $\text{diag}(\mu_1, U_2)$ then expression (3.41) can be rewritten as,

$$(\mu_1^{-1}r_1r_1' + R_2U_2^{-1}R_2')(\mu_1 r_1r_1' + R_2U_2R_2')$$

$$= (r_1r_1')^2 + \mu_1^{-1}r_1r_1'R_2U_2R_2' + \mu_1 R_2U_2^{-1}R_2'r_1r_1' + R_2U_2^{-1}R_2'R_2U_2R_2'$$

so that the maximum eigenvalue of expression (3.41) is less than,

$$\|A^2\| + \|\mu_{min}^{-1}\mu_{max}AB\| + \|\mu_{max}^{-1}\mu_{min}BA\| + \|B^2\|$$

where $A=r_1r_1'$ and $B = R_2R_2'$ with A+B=I. Thus, the ratio of the variance of the obtained estimator to the variance of the proper estimator is less than,

$$\frac{(\mu_{max} + \mu_{min})^2}{4\,\mu_{max}\mu_{min}} \tag{3.42}$$

This upper bound is quoted by Bloomfield and Watson (1975). The usefulness of the upper bound is open to question. Generally the maximum and minimum eigenvalues of the proper, unused diagonal covariance matrix Σ are unknown. But the use of the bound doesn't necessarily require knowledge of these. If the ratio $\mu_{max}/\mu_{min} = r$ is known, the bound is,

$$\frac{(r+1)^2}{4r}$$

a result quoted by Tukey (1975). Thus if it is suspected that Σ is illconditioned or close to singularity (r very large) then the obtained least squares estimator may be very inefficient.

This bound was first displayed in a rather cryptic paper by Tukey (1948) but only for the case p=1 (the calculation of a weighted average). Much of the work which, temporally but not logically, followed Tukey's paper, by Anderson, Durbin and Watson, and others, was concerned primarily with finding necessary and sufficient conditions for the equality of the least squares estimator,

$$\hat{\beta} = (X'X)^{-1}X'y$$

and the best linear unbiased or Markov estimator,

$$\hat{\beta}_M = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y$$

under the condition that $E(\varepsilon\varepsilon') = \Sigma$. This work was reviewed by Watson (1967). One of the most recent papers concerning the equality of these estimators, for X not necessarily of full column rank, is by Styan (1973). Bloomfield and Watson (1975) and Knott (1975) have presented proofs of the result,

$$1 \leq \frac{|\text{var } \hat{\beta}|}{|\text{var } \hat{\beta}_M|} \leq \left\{\frac{(\mu_{max}+\mu_{min})^2}{4\,\mu_{max}\mu_{min}}\right\}^p \tag{3.43}$$

and proofs of an even tighter upper bound on this ratio. The problem for many practioners is that generally Σ is unknown and the eigenvalues of Σ are unknown, so that the crude upper bounds (3.42), (3.43) are unknown. However, as long as the ratio r is thought not to be too large ($r \leq 3$) the bound (3.42) tells practioners that the choice of covariance matrix is not crucial and a simple weighting scheme is probably sufficient. Such a sentiment has been expressed by Tukey (1975).

3.32 Heteroscedastic Variances

Suppose a special case of situation (ii) is assumed to hold where,

$$\Sigma = \sigma^2 I_{n \times n} \quad \text{with } \sigma^2 \text{ unknown},$$

but in reality a special case of situation (iii) holds where,

$$\Sigma = (\text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2).$$

Then the assumed covariance matrix posits equality of variance for the n disturbances or homoscedasticity whereas, in reality, the errors are heteroscedastic. As pointed out in the previous subsection the obtained least squares estimator of β is still unbiased but is no longer minimum variance or best. However, the loss in efficiency is not too great provided the σ_i^2 do not vary greatly. If the σ_i^2 do vary greatly the obtained least squares estimator rapidly loses efficiency (see inequality (3.43)).

The detection of situations in which markedly different heteroscedastic disturbances are present in a data set and the efficient estimation of β under this heteroscedasticity are clearly of some importance. The presence of nonconstant variances may be detected by an examination of residuals from an initial least squares fit with the assumed homoscedastic error structure. Plots of residuals versus predicted values of y and plots of residuals against time or against the sequence in which the observations were taken may reveal heteroscedastic error structure. A good introduction to the examination and analysis of residual plots is contained in Draper and Smith (1966). Various statistical

tests against heteroscedasticity which utilize the residuals are also available (see, for example, Theil (1971)). If the presence of heteroscedasticity is established, the parameter vector β should be reestimated. One approach to reestimation is to transform the dependent variable y so that the error variances are homogenized. A seminal article on this technique is that of Box and Cox (1964). Another approach is to approximate Σ in some way and perform a weighted least squares analysis and continue in an iterative fashion (monitoring goodness of fit) until residual plots indicate that the heteroscedasticity has been adequately accounted for. The best approach, if it is feasible, is to estimate $\Sigma = \text{diag}(\sigma_1^2,\ldots,\sigma_n^2)$ and if the estimate indicts the homoscedasticity assumption use it to reestimate β. Several techniques for estimating such a diagonal covariance matrix exist in the literature. Most require replication of the observations or in lieu of replication make further restrictive assumptions on the dimensionality or the structure of the input matrix. References to these estimation techniques and comparisons of them can be found in Horn and Horn (1975) and Horn, Horn and Duncan (1975).

3.4 Nonlinear Estimation

The fourth condition of the Gauss-Markov Theorem requires the estimate of the parameter vector β to be a linear function of the vector of observations on the dependent variable y. This requirement seems to originate from two sources:

(i) the need for a computationally feasible procedure,

(ii) the result that the least squares estimator, which is the best linear unbiased estimator (BLUE or MVLUE) of β, is, under the added assumption of a multivariate normal distribution for the error vector ε, the minimum variance unbiased estimator (MVUE) or maximum likelihood estimator (MLE).

The first source of this linearity requirement is no longer as

important as it was in the early 19th century - with the advent of high speed computers and programmable calculators ease of computation and even cost of computation are factors which can no longer limit the choice of an estimator to a class of linear functions of y. The second source of this linearity requirement does not have great impetus. That the linear least squares estimator is the minimum variance unbiased estimator when the error distribution is multivariate normal is a strong theoretical result, but the normality assumption is often difficult to meet in practice. It would seem to be desirable to be able to include nonlinear estimators in the search for an optimal estimator, particularly when the distribution of the error vector is far from normality.

The most common kind of non-normal distribution for the error vector is a "long-tailed" distribution. When the error vector has this kind of distribution the least squares estimator is, of course, no longer minimum variance in the class of all unbiased estimators (it is still MVLUE or BLUE) and may be described as inefficient. Andrews (1974, 1975) and Beaton and Tukey (1974) have suggested robust regression techniques which are insensitive to the outliers thrown up by the long tailed error distributions. These fitting techniques preserve the large residuals usually associated with outlying observations and therefore allow the detection of blunders and model inadequacies. The biweight regression technique of Beaton and Tukey is very similar to an iterative weighted least squares procedure. Whereas the usual least squares estimator is the one-step solution, $\hat{\beta}$, of the equation,

$$X'(y-XB) = 0$$

the biweight solution is the (t+1)th iteration of,

$$X'W_{(t)}\ (y-XB_{(t+1)}) = 0$$

where $W_{(t)}$ is a diagonal nxn weighting matrix whose elements are reset at each step in the manner,

$$W_{(t)_{(i,i)}} = w\left(\frac{y_i-x_i'B_{(t)}}{s(B_{(t)})}\right)$$

where w(u) is some "robustifying" function which attaches small weights to large values of u and relatively large weights to small values of u, with $B_{(t)}$ the solution from the previous iteration and $s(B_{(t)})$ some measure of scale based on the residuals from the previous fit. Beaton and Tukey reported that such an estimation procedure displays resistance or insensitivity to perturbations in the data as well as the planned for insensitivity to outliers and high efficiency over a wide range of error distributions. Thus the nonlinear biweight procedure would, it seems, also be useful in the errors in variables situation discussed in section 3.2.

The robust regression procedures are not meant to replace the usual least squares estimator when the conditions of the Gauss-Markov Theorem hold, although such replacement would not cost much in terms of loss in efficiency, but are more properly to be used as model building tools for detecting possible data irregularities which may prevent the application of the Gauss-Markov Theorem.

Other nonlinear estimators have been proposed in the literature and not necessarily with the robust/resistant motivation described above. Least pth powers and estimation methods based on order statistics, for example, are referred to by Andrews (1974). The James-Stein estimator and the nonlinear operational variants of the minimum mean square error linear estimator (MMSELE) are two more classes of nonlinear estimator which have desirable properties and outperform the least squares estimator with respect to certain optimality criteria. They are however introduced in Chapter 4.

The restriction to estimators which are linear functions of y goes hand-in-hand with the third and sixth conditions of the Gauss-Markov Theorem. The third condition of the Gauss-Markov Theorem only specifies the first and second order moments of the distribution of ϵ and y and the sixth condition of the Gauss-Markov Theorem requires the estimator to be judged by its variance. Using a nonlinear estimator and the variance of the estimator to judge its quality requires knowledge or at least assumptions about higher-than-second-order moments of the distribution of ϵ and y. In some ways the estimation problem begins to lose the little

generality it had when higher order moments of ϵ have to be assumed. Relaxing the linearity requirement may therefore necessitate modifying the sixth condition of the Gauss-Markov Theorem.

3.5 Biased Estimation

Although the word "biased" has unfortunate connotations (connotations which have been described by Lindgren (1968) as "un-American") biased estimators are not necessarily worse than or less preferable to unbiased estimators. An interesting introduction to biased estimation is contained in Efron (1975). Efron points out that to arrange for an estimator of β to be unbiased in mean, that is,

$$E(B) = \beta,$$

does not guarantee that an estimate is close to β. The usual means of illustrating this, for the univariate case, is to present a graph of the distribution of an efficient biased estimator versus a graph of the distribution of a relatively inefficient unbiased estimator. Such graphs are presented in Figure 3.1.

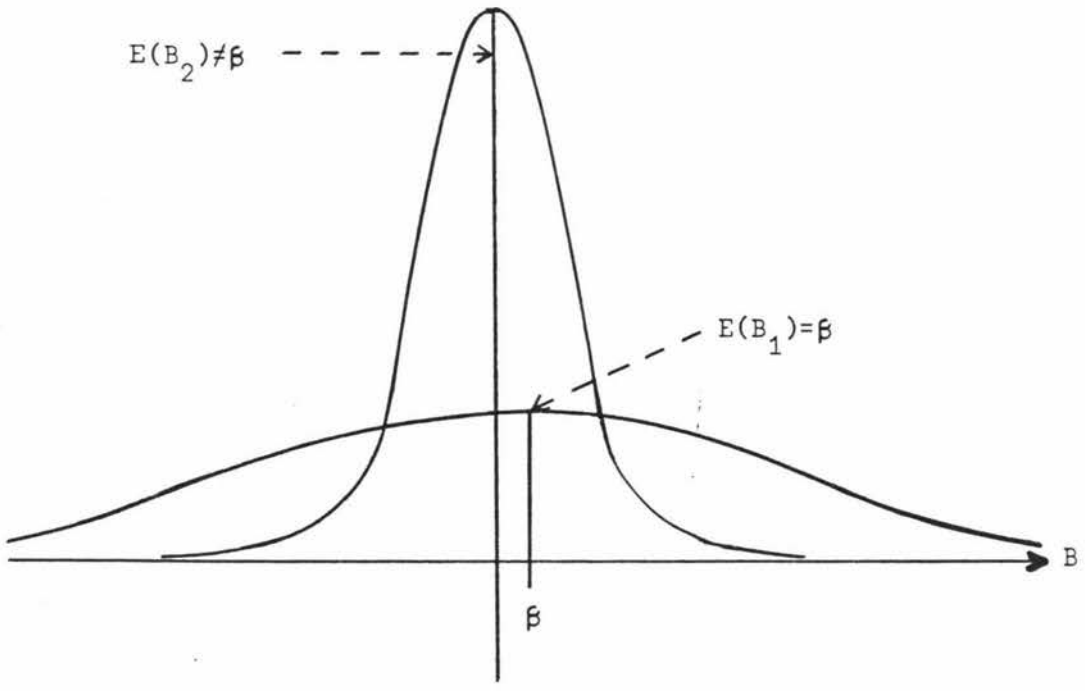

Figure 3.1 Hypothetical distributions of an efficient biased estimator and an inefficient unbiased estimator.

In Figure 3.1 the non-zero bias in the second estimator B_2 is a small fraction of the standard deviation of that estimator and an even smaller fraction of the standard deviation of the unbiased estimator B_1 so that the preferred estimator would have to be the biased estimator B_2. Choosing to use a biased estimator in preference to an unbiased estimator clearly requires knowledge of the value of the bias, the variance of the biased estimator, and the variance of the unbiased estimator. One criterion for choosing between two such competing estimators might involve comparison of the variance plus the squared bias of the biased estimator with the variance of the unbiased estimator (the bias has to be squared so that it has the same dimensionality as the variance). Thus if,

$$\text{Var}(B_2) + (E(B_2)-\beta)^2 < \text{Var}(B_1) \qquad (3.44)$$

where B_1 is unbiased then B_2 might be regarded as a "better" estimator of β than B_1. In the case where β is a vector of parameters the variance terms in (3.44) may be replaced by the traces of the respective covariance matrices and the squared bias term by the square of the Euclidean distance from $E(B_2)$ to β.

The immediate problem with such a criterion and in fact with biased estimators themselves is that β (the object of the estimation exercise) is generally not known so that the value of the bias is unknown. If however some bounds on the value of the bias are available then the criterion may be able to be used.

It is of interest to note that Barnard (1963) has provided a development of the Gauss-Markov Theorem with the unbiasedness requirement replaced by a requirement of unboundedness for the components of β and a requirement of a bounded mean square error (the left hand side of (3.44)) for the estimator. Such requirements lead to the unbiased least squares estimator. If the components of β are bounded then there is a possibility that an alternative biased estimator may have a smaller mean square error. This is demonstrated for specific biased estimators in Chapters 4 and 5.

3.6 Criteria for Estimation

The sixth condition of the Gauss-Markov Theorem requires the quality of linear unbiased estimators of β to be judged by their variance where the variance of such a linear unbiased estimator is given by,

$$\begin{aligned} \text{var}(B) &= E(B-E(B))(B-E(B))' \\ &= E(B-\beta)(B-\beta)' \end{aligned}$$

Thus for two linear unbiased estimators B_1 and B_2, if $\text{var}(B_1) - \text{var}(B_2)$ is non-negative definite then B_2 is judged to outperform B_1 and if $\text{var}(B_1) - \text{var}(B_2)$ is negative definite then B_1 is judged to outperform B_2. The result of the Gauss-Markov Theorem shows that for the least squares estimator $\hat{\beta}$ $\text{var}(B) - \text{var}(\hat{\beta})$ is always non-negative definite for any linear unbiased estimator B. Such an estimation criterion seems reasonable when the second condition of the Gauss-Markov Theorem only specifies the first and second order moments of the distribution of the error vector ε. If other measures of spread were used to characterize the distribution of the error vector then other measures of spread would have to be used to judge the performance of a linear estimator.

If the search for an estimator is widened and can take place outside the class of unbiased estimators then, as alluded to in section 3.5, some estimation criterion that combines variance of the estimator and bias, and allows them to be traded-off, is required. Several such criteria are possible:

$$\begin{aligned} E\|B-\beta\|^2 &= E(B-\beta)'(B-\beta) \\ &= \text{tr var}(B) + (E(B)-\beta)'(E(B)-\beta) \end{aligned} \tag{3.45}$$

$$\begin{aligned} E(B-\beta)'H(B-\beta) &= \text{tr}(HE(B-\beta)(B-\beta)') \\ &= \text{tr}(H\ \text{var}(B)) + \text{tr}(H(E(B)-\beta)(E(B)-\beta)') \end{aligned} \tag{3.46}$$

where H is any pxp non-negative definite matrix.

The first criterion which is the expected value of the square of the Euclidean distance from B to β is known as the mean square error of B. The second criterion is a weighted sum of component mean

square errors and can be described as a generalized mean square error of B. Both criteria require knowledge of β before they can be minimized with respect to B. In some situations knowledge of β may not be strictly necessary as it may be possible to show that for all values of β one estimator outperforms or dominates another with respect to one of these criteria, for example, the domination of least squares by the James-Stein estimator with respect to criterion (i) when β is the unknown mean of a p-variate normal distribution with $p \geq 3$. However the usual procedure in developing alternative biased estimators has been to find a biased estimator B, evaluate the chosen criterion for B and the unbiased least squares estimator $\hat{\beta}$ and find conditions on the values of the unknown parameters β and σ^2 which guarantee the outperformance of $\hat{\beta}$ by B.

The use of the mean square error criterion (3.45) can be objected to. Suppose,

$$E\|B_1-\beta\|^2 < E\|B_2-\beta\|^2 \quad \forall\beta. \tag{3.47}$$

Then the estimator B_1 is judged to outperform B_2, but the criterion used is an ensemble property which may not be true for a particular component of β. The Euclidean norm has been used in (3.47) to measure the distance between the vector B_1 and β. The distances between corresponding components of B_1 and β have not been considered singly but have been coalesced in an unweighted sum. It is possible for (3.47) to be true and for the inequality,

$$E(B_{1_{(i)}} - \beta_{(i)})^2 > E(B_{2_{(i)}} - \beta_{(i)})^2 \quad \text{for some } i$$

to be true simultaneously. Use of the generalized mean square error criterion (3.46) allows attention to be focussed on particular components of β for various choices of the weighting matrix H.

3.7 The Utility of the Theorem.

The first three conditions of the Gauss-Markov Theorem establish

a model for the observations and make assumptions about the structure and error content of the observations. The other three conditions define a class of estimators to be considered in the search for an estimator and a criterion for judging the quality of these estimators. The optimal estimator under all these conditions is the least squares estimator. The literature concerning the effect of relaxation of the first three conditions of the Gauss Markov Theorem on the least squares estimator is rich and vast, and much has been advertently and inadvertently omitted from the foregoing discussion. In retrospect the least squares estimator is surprisingly insensitive to lurking variables and errors in the variables except when the input matrix has a multicollinear structure. The estimator loses little efficiency when the wrong covariance matrix is assumed except in situations where the true unknown covariance matrix is composed of elements of different magnitudes.

4. SOME BIASED ESTIMATION PROCEDURES

One class of alternatives to the least squares estimator is the class of linear and nonlinear biased estimators. Within this broad class of alternatives are a class of shrinkage estimators which, in the phraseology of Dempster (1973), "pull-back" or shrink the least squares estimator to a chosen origin. Such "pulled-back" estimators may differ in "degree" and "pattern" of pull-back. The preliminary test estimators and variable selection rules which lie at the heart of model building are "pulled-back" estimators whose "degree" of pull-back is set prior to the data analysis - the regression coefficients which correspond to the variables in the selected subset are estimated by least squares and not pulled back at all while the regression coefficients corresponding to the unselected variables are pulled back to zero - but whose "pattern" of pull-back is generated by the data and the customary tests of significance. The estimators to be considered in this chapter have their pattern of pull-back chosen prior to the data analysis but their degree of pull back is dependent on the structure of the data.

The motivation for considering such estimators originates from the multicollinearity problem discussed in subsection 3.131. When the structure of the input matrix is multicollinear, least squares estimation and prediction are very imprecise in the directions of the minor principal components of X (see equation (3.15)). Another way of characterizing this inflation of the variance of least squares estimators in the presence of multicollinearity is through the mean square error criterion (3.45). The mean square error of the unbiased least squares estimator is,

$$\begin{aligned} E(\hat{\beta}-\beta)'(\hat{\beta}-\beta) &= E(\hat{\beta}'\hat{\beta})-\beta'\beta \\ &= \mathrm{tr}[\mathrm{var}(\hat{\beta})] \\ &= \sigma^2 \mathrm{tr}[(X'X)^{-1}] \\ &= \sigma^2 \sum_{i=1}^{p} 1/\lambda_i \\ &> \sigma^2/\lambda_{min} \end{aligned} \tag{4.1}$$

Thus, if X is highly collinear then $X'X$ is ill-conditioned and $\lambda_{min} \ll 1$ so that, from expression (4.1), the expected value of the square of the Euclidean distance from $\hat{\beta}$ to β is much too long on average and $\hat{\beta}$ is much too long on average,

$$E(\hat{\beta}'\hat{\beta}) > \beta'\beta + \sigma^2/\lambda_{min}$$

Thus when multicollinearity is present in the input matrix, estimation procedures which pull back the too-long-on-average least squares estimator $\hat{\beta}$ and which correspondingly reduce the mean square error by inducing a little bias and causing a reduction in the variance, may be of some worth.

4.1 Best Linear Estimation - a unifying approach to biased Estimation.

Rao (1971, 1973) has shown that the best linear estimator (BLE) of any linear combination $t'\beta$ is, for a particular choice of the symmetric matrix W, $t'\tilde{\beta}$ where,

$$\tilde{\beta} = WX'(I+XWX')^{-1}y \qquad (4.2)$$

This best linear estimator $T'y$ where $T' = t'WX'(I+XWX')^{-1}$ results from attempting to minimize with respect to T the mean square error of $T'y$ or the loss function,

$$E(T'y-t'\beta)^2 = \sigma^2T'T + (X'T-t)'\beta\beta'(X'T-t). \qquad (4.3)$$

Minimization of (4.3) with respect to T is conceptually not possible as the parameters β and σ^2 are unknown. Rao suggested minimizing,

$$S = T'T + (X'T-t)'W(X'T-t) \qquad (4.4)$$

where W is some approximation to $\beta\beta'/\sigma^2$. Three stategies were suggested by Rao for arriving at the estimator (4.2):

(i) Discover an a priori value of $\sigma^{-1}\beta$ and use it to form a W which may be placed in (4.4) and (4.2). Such an approach is probably beyond the capabilities of most data analysts.

(ii) Adopt a Bayesian approach and assume β is a random variable with a prior distribution. The only information that needs to be specified about the prior distribution of β is,

$$E(\beta\beta') = \sigma^2 W.$$

A further expectation of (4.3) with respect to β produces a risk function in W which is a suitable candidate for minimization. Rao (1976) has called the estimator (4.2) which results from this strategy the "Bayes Homogeneous Linear Estimator (BHLE) of β with respect to W".

(iii) Choose W simply on the grounds of how much weight should be attached to either term in (4.4). The first term in (4.4) is a variance term and the second term is a squared bias term so that choosing a W with, for example, large elements attaches more weight to the minimization of bias and less weight to the minimization of variance in the minimization of the mean square error.

The matrix W, which results from the first strategy is of rank one. The W matrices which result from the second and third strategies may have rank greater than one. If W is assumed to be of full rank some simple relationships between the best linear estimator $\tilde{\beta}$, the least squares estimator $\hat{\beta}$, and generalized ridge estimation can be displayed. Noting that,

$$(I+XWX')^{-1} = I-X(X'X+W^{-1})^{-1}X'$$

the best linear estimator can be written as,

$$\begin{aligned}\tilde{\beta} &= WX'(I-X(X'X+W^{-1})^{-1}X')y \\ &= W(X'X-X'X(X'X+W^{-1})^{-1}X'X)\hat{\beta} \\ &= W((X'X)^{-1}+W)^{-1}\hat{\beta} \qquad (4.5)\end{aligned}$$

so that the best linear estimator is a linear transformation of the least squares estimator. The equation (4.5) is easily rewritten as,

$$\tilde{\beta} = (X'X + W^{-1})^{-1}X'y \qquad (4.6)$$

so that the best linear estimator $\tilde{\beta}$ also has the form of a generalized ridge estimator (a result that eluded Farebrother (1975) but which has been established by Rao (1976)).

Best linear estimation provides an important initial unifying approach to biased estimation. If the search for an estimator

is confined to linear estimators, equations (4.2) and (4.3) tell searchers that the "best" estimator can only be found if some prior knowledge is available. Thus the usual unbiased linear least squares estimator can only be bettered if some prior knowledge of β can be found and incorporated into the estimation procedure or if a degree of arbitrariness can be tolerated in an explicit form of the alternative estimator. The equation (4.5) tells searchers that the best linear estimator for a particular full rank choice of W is a linear transformation of the least squares estimator (an analogous result holds for W of less than full rank). Thus a search of pxp transformation matrices which are independent of y might produce useful biased estimators. The fact that the best linear estimator for a particular full rank W is a generalized ridge estimator is also a powerful incentive to explore further the class of such generalized ridge estimators.

4.11 The Minimum Mean Square Error Linear Estimator.

The minimum mean square error linear estimator (MMSELE) results from the direct minimization of equation (4.3). Thus the minimum mean square error linear estimator of β has the form,

$$\tilde{\beta}_{MMSE} = \beta\beta' X'(\sigma^2 I + X\beta\beta' X')^{-1} y$$

$$= \frac{\beta' X' y}{\sigma^2 + \beta' X' X \beta}\,\beta \qquad (4.7)$$

Such an estimator has been studied by Theil (1971), Farebrother (1975), Vinod (1976b), and of course Rao. The estimator is quite useless as it depends on the unknown parameters β and σ^2. Rao, as pointed out in the previous section, suggested replacing the unknown matrix $\beta\beta'/\sigma^2$ by some a priori valued matrix or a quite arbitrary matrix. Farebrother (1975), after showing that (4.7) looked a bit like a ridge estimator for the case p=1, that is,

$$\tilde{\beta}_{MMSE} = \left(x'x + \frac{\sigma^2}{\beta^2}\right)^{-1} x' y$$

suggested making (4.7) operational by replacing the unknown

parameters β and σ^2 with consistent estimators. One such operational variant of (4.7) might be,

$$\frac{SSR}{\frac{RSS}{n-p} + SSR} \cdot \hat{\beta}$$

where $\hat{\beta}$ is the least squares estimator and SSR and RSS are respectively the regression sum of squares and the residual sum of squares for the least squares fit. Such an estimator is no longer linear in y, which means extra assumptions have to be made about the higher order moments of the distribution of y to find the variance and mean square error of the estimator.

Vinod (1976b) suggested replacing the unknowns in (4.7) with initial estimates, possibly least squares estimates and iterating until a fixed point is obtained, in the manner,

$$\tilde{\beta}^{(t+1)} = \frac{\tilde{\beta}^{(t)\prime} X'y}{(\hat{\sigma}^{(t)})^2 + \tilde{\beta}^{(t)\prime} X'X\tilde{\beta}^{(t)}} \cdot \tilde{\beta}^{(t)} .$$

Vinod found a closed form solution for the fixed point by minimizing with respect to B,

$$\|y-XB\|^2 + k(B'X'y - (n-p-1)^{-1}\|y-XB\|^2 - B'X'XB).$$

The resulting fixed point has the form,

$$\left(\frac{k(p-n-1) + 2n-2p-2}{2k(p-n) + 2n-2p-2}\right) \hat{\beta} \qquad (4.8)$$

Such an estimator is a scalar shrinking of the least squares estimator and is a member of a class of estimators to be discussed in section 4.21. Vinod recommended replacing the scalar shrinkage factor in expression (4.8) by the expression,

$$s = \frac{1}{2(n-p)} \{(n-p+1) + \sqrt{(n-p+1)^2 - 4(n-p)/R^2}\} \qquad (4.9)$$

where R^2 is the squared multiple correlation coefficient from the least squares fit. Such a procedure has close affinities

with the James-Stein procedure which is discussed in section 4.4

4.2 Linear Transformation of the Least Squares Estimator

The impetus for considering biased estimators which are linear transformations of the least squares estimator comes from the result that the least squares estimator is too long on average for multicollinear X, that is, $E(\hat{\beta}'\hat{\beta}) > \beta'\beta + \sigma^2/\lambda_{min}$, and from equation (4.5) which demonstrates that the best linear estimator of β is a linear transformation, involving an unknown matrix W, of the least squares estimator. In this section various strategies for determining such transformation matrices are reviewed.

Consider the class of estimators of β which are linear transforms of $\hat{\beta}$, that is estimators of the form $C\hat{\beta}$ where C is any pxp matrix. Then such estimators are biased for $C \neq I$ as

$$E(C\hat{\beta}) = C\beta$$

and have variance,

$$\sigma^2 C'(X'X)^{-1}C$$

and mean square error,

$$\sigma^2 tr\{C'(X'X)^{-1}C\} + \beta'(C-I)'(C-I)\beta$$

The residual sum of squares associated with such estimators is given by,

$$\begin{aligned} RSS(C\hat{\beta}) &= \|y-XC\hat{\beta}\|^2 \\ &= \|y-X\hat{\beta}\|^2 + \|X(C-I)\hat{\beta}\|^2 \\ &= RSS(\hat{\beta}) + \hat{\beta}'(C-I)X'X(C-I)\hat{\beta} \end{aligned} \qquad (4.10)$$

The least squares estimator (C=I) minimizes this sum of squares (this is a direct consequence of the Gauss-Markov Theorem) so that any linear transform of $\hat{\beta}$ has an increased residual sum of squares. If the second term in equation (4.10) is held constant at some amount it defines a hyperellipsoid in the p-dimensional parameter space (see section 4.3) so that whole classes of biased estimators are defined for given increases in lack of fit. Mayer and Willke

(1973) suggested that particular biased estimators could be identified in these classes by considering particular norm minimizations of the estimators. This approach is followed in the next two sections.

4.21 Shrunken Estimators.

Mayer and Willke (1973) introduced a sub-class of the class of linear transformations of the least squares estimator indexed by the transformation matrix,

$$C = \frac{1}{1+k} I$$

with $k \in [0,\infty)$. Such estimators they termed deterministically shrunken estimators. These estimators have variance,

$$\frac{\sigma^2}{(1+k)^2} (X'X)^{-1}$$

and mean square error

$$\frac{\sigma^2}{(1+k)^2} \operatorname{tr}(X'X)^{-1} + \left(\frac{k}{1+k}\right)^2 \beta'\beta$$

Thus the shrunken estimator has smaller mean square error than the least squares estimator if,

$$\frac{\sigma^2}{(1+k)^2} \operatorname{tr}(X'X)^{-1} + \left(\frac{k}{1+k}\right)^2 \beta'\beta < \sigma^2 \operatorname{tr}(X'X)^{-1}$$

or if,

$$\frac{\beta'\beta - \sigma^2 \operatorname{tr}(X'X)^{-1}}{\beta'\beta + \sigma^2 \operatorname{tr}(X'X)^{-1}} < \frac{1}{1+k} < 1.$$

Thus if $\beta'\beta < \sigma^2 \operatorname{tr}(X'X)^{-1}$ any $k > 0$ guarantees the outperformance of least squares with respect to the mean square error criterion (3.45). If $\beta'\beta > \sigma^2 \operatorname{tr}(X'X)^{-1}$ then,

$$0 < k < \frac{2\sigma^2 \operatorname{tr}(X'X)^{-1}}{\beta'\beta - \sigma^2 \operatorname{tr}(X'X)^{-1}}$$

guarantees the domination of least squares by the shrunken estimator with respect to (3.45).

It is trivial to show that the shrunken estimators of Mayer and Willke result from minimizing:

$$\|y-XB\|^2 + k\|XB\|^2 \quad , \qquad k \geq 0$$

Thus the shrunken estimators constrain the squared length of XB while minimizing the residual sum of squares. Thus in a class of estimators with a fixed residual sum of squares the shrunken estimator is the unique estimator identified by the minimization of a "design dependent norm", $\|XB\|$. The use of such a norm to pluck out an estimator from an equivalence class is open to criticism. Constraining the length of XB, while it does produce an estimation procedure which directly shrinks the least squares estimator, does not produce the estimator in the equivalence class with the smallest length. Constraining the length of B may be more appropriate. The components of the shrunken estimator have the same relative magnitudes as the components of the least squares estimator. Such a simplistic, blanket shrinking of all the components of $\hat{\beta}$ ignores a fundamental aspect of the multicollinearity problem, namely, imprecise estimation in certain directions of the estimation space. It could be that some components of β are estimated relatively precisely by least squares so that it may not be necessary to shrink them. Stochastically shrunken estimators are reviewed in section 4.4.

4.22 Ridge-type Estimators.

Hoerl and Kennard (1970a,b) proposed a biased estimation procedure based on a technique - relaxed least squares - given by Riley (1955) for solving ill-conditioned systems of linear equations. Instead of solving,

$$X'XB = X'Y$$

Riley suggested solving the perturbed system of linear equations,

$$(X'X+kI)B = X'Y$$

where k is a small positive "constant" added to the diagonal of $X'X$ to improve the conditioning of $X'X$. Riley suggested values of k between 10^{2-s} and 10^{3-s}, where s is the number of decimal places being carried in the calculations, would be able to improve the conditioning problem without moving the result too far away from the minimum of $||Y-XB||^2$. Thus, the ridge estimator proposed by Hoerl and Kennard is of the form,

$$\beta^* = (X'X+kI)^{-1}X'y \qquad (4.11)$$

where $k \in [0,\infty)$ and with $X'X$ in correlation form and $X'y$ the vector of correlations between the input variables and the dependent variable. The ridge estimator is clearly in the class of linear transforms of the least squares estimator as,

$$\beta^* = (I+k(X'X)^{-1})^{-1}\hat{\beta}.$$

Hoerl and Kennard (1970a) and Mayer and Willke (1973) showed that the ridge estimator resulted from minimizing with respect to B,

$$||y-XB||^2 + k||B||^2 \qquad , \qquad k \geq 0.$$

Thus ridge estimators constrain the length of the vector of regression coefficients while minimizing the residual sum of squares, so that in a class of estimators with constant residual sum of squares the ridge estimator is indexed by the minimization of the Euclidean norm of the estimators.

Hoerl and Kennard showed that the mean square error of the ridge estimator could be written as,

$$\sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)^2} + k^2\beta'(X'X+kI)^{-2}\beta$$

where the λ_i are the p eigenvalues of $X'X$. Consequently the ridge estimator has smaller mean square error than the least squares estimator if,

$$\sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)^2} + k^2 \, tr\{(X'X+kI)^{-2}\beta\beta'\} < \sigma^2 \sum_{i=1}^{p} 1/\lambda_i$$

A sufficient condition for this is,

$$0 < k < \frac{2\sigma^2}{\beta'\beta} \qquad (4.12)$$

although other less strict bounds on k are possible.

The ridge estimation technique has generated much interest and spawned a vast number of research papers. Indeed, the review articles by Marquardt and Snee (1975) and Hocking (1976) indicate a plethora of papers in this area. However, not all data analysts and researchers have endorsed the technique. Kendall (1975) p.107 has commented:

> "It is not plain to me that this admitted distortion of the data (the effect of which is to diminish the correlations among the variables) has any theoretical justification."

Conniffe and Stone (1973, 1975) and Newhouse and Oman (1971) have presented critiques of the method. Some of their criticisms and the criticisms of others are listed here:

C(i) The existence of a k which guarantees a smaller mean square error for the ridge estimator than for the least squares estimator has been established. This does not, however, guarantee that a particular choice of k improves on the least squares estimator. Newhouse and Oman (1971) comment that it is not true that ridge estimators "will necessarily not be better than" least squares but it is true that they "will not necessarily be better than" least squares.

C(ii) In deriving variances and mean square errors of ridge estimators, k is assumed to be a constant. But in practice k is "estimated" from the data. The moments of β^* for fixed k are not the moments of the estimator being used in practice. What is the status of k and the reliability (mean square error) of the estimators used in practice?

C(iii) Arguments based on average mean square error of the components, criterion (3.45), can be misleading. Some components of β may be estimated quite precisely by least squares so that shrinkage of them is unnecessary distortion.

C(iv) Choosing to use ridge-type estimators means abandoning the useful techniques of hypothesis testing and confidence interval construction. No measure of reliability of the estimator is available.

C(v) Multicollinearity means that there are either redundant input variables present or insufficient data for prediction. The remedies are, in the former case, subset selection or model respecification and, in the latter case, data augmentation.

C(vi) Ridge regression encourages prediction for manipulated values of the variables in cases where the prediction equation is based on previous passive observation of the variables.

C(vii) "Ridge analysis is an ad hoc procedure", Newhouse and Oman (1971) p.1.

Clearly, Hoerl and Kennard have a lot to answer for. Curiously, most of the direct responses to these criticisms have not come from Hoerl and Kennard but from Theobald (1974), Smith and Goldstein (1975) and others. The usual kinds of responses by proponents of ridge-type estimation to these criticisms are listed point by point below:

R(i) It is certainly true that while ridge estimation outperforms least squares estimation for certain values of k less than some function of the unknown parameters, there is no guarantee that a particular chosen ridge estimator has total mean square error smaller than the variance of the least squares estimator. This however is a property of all biased linear estimators. In section 4.11 it was shown that the minimum mean square error linear estimator depended on the unknowns β and σ^2. Rao, in his BLE formulation of the biased linear estimation problem, showed that "best" estimators were conditional upon some prior knowledge of $\sigma^{-1}\beta$, or some knowledge of the covariance matrix of a prior distribution for β, or the choice of a quite arbitrary weighting matrix. Uniform domination of least squares by any linear estimator requires prior knowledge of the parameters to be estimated. On the positive side inequality (4.12) indicates that in most situations very small values of k, possibly of the order of magnitude originally suggested by Riley, may safely

reduce the mean square error of the estimation procedure.

R(ii) This criticism is probably the most telling of all the criticisms levelled at ridge-type estimators. Consequently much attention is being paid to devising algorithms for k and evaluating the performance of these algorithms in simulation studies (see, for example, Hoerl, Kennard and Baldwin (1975), Hoerl and Kennard (1976) and Lawless and Wang (1976)). Several methods for selecting k which do not depend on the random variable y have been suggested. Marquardt (1970) suggested that as a "rule of thumb" k should be chosen so that the maximum variance inflation factor, the largest diagonal element of $(X'X+kI)^{-1}X'X(X'X+kI)^{-1}$ where $X'X$ is in correlation form, lies between one and ten. Obenchain (1975a) suggested choosing the k for which the Sum of Squares of all p(p-1)/2 Correlations Between the ridge Coefficients (SSCBC) is minimized. Another suggestion is to choose k less than or equal to the smallest eigenvalue of $X'X$. Such rules certainly produce a k which is independent of y so that the moments and mean square derived by Hoerl and Kennard are applicable. The rules do not, however, and can not guarantee mean square error domination of least squares, as pointed out in the response to the first criticism.

Several criteria which depend on y, for choosing k, have been adopted by users of ridge regression. Indeed Hoerl and Kennard suggested choosing k by inspecting a ridge trace - a plot of the estimated coefficients for various values of k. Most of these criteria attempt to monitor goodness of fit or increases in residual sum of squares for values of $k \geq 0$. Such criteria do produce a k which is not constant with respect to expectation over the values of y and it is these criteria which indict the mean square error results of Hoerl and Kennard and others. These stochastic criteria for k have either complicated forms or are serendipitous in nature so that they are difficult to incorporate into expression (4.11). That they cannot be incorporated into (4.11) prevents the expection operator and mean square error criterion being applied to the ridge estimation procedure. This

stochastic choice of k can be partially rationalized by arguing that it is not k which is estimated from the data but an upper bound on k, say $2\sigma^2/\beta'\beta$, which is estimated from the data. The "constant" k is then chosen to be less than or equal to this estimated upper bound. Thus k does have the status of a constant in the situations in which the expectation operator is applied. Such a rationalization is open to criticism but it seems to be at the basis of the algorithms for k which have been proposed and evaluated in the work cited earlier. Further discussion on the choice of k is postponed until Chapter 6.

R(iii) Theobald (1974) has shown that a sufficient condition for the ridge estimator to outperform the least squares estimator with respect to any generalized mean square error criterion, expression (3.46), is,

$$0 < k < \frac{2\sigma^2}{\beta'\beta}$$

Extensions of this result have been given by Farebrother (1976) and Swindel (1976). More general forms of the ridge estimator in equation (4.11) have been proposed. These forms, which shrink only certain components or linear combinations of the components of the least squares estimator, are discussed in following sections.

R(iv) Obenchain (1975b) has suggested that tests of the general linear hypothesis,

$$H: A\beta = \rho \tag{4.13}$$

where A is a known $r\times p$ matrix and ρ is a known $r\times 1$ vector, can be recovered when least squares is forsaken. Under the hypothesis (4.13) the usual restricted least squares estimator for β is,

$$\hat{\beta}^H = A^-\rho + (I-A^-A)\hat{\beta} \tag{4.14}$$

where $A^- = (X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}$.

Ths usual F statistic for testing the hypothesis (4.13) is,

$$F = \frac{(A\hat{\beta}-\rho)'[A(X'X)^{-1}A']^{-1}(A\hat{\beta}-\rho)}{rs^2} \tag{4.15}$$

where $s^2 = RSS(\hat{\beta})/(n-p-1)$. The ridge estimator of β is of the form, $\beta^* = C\hat{\beta}$. If it can be assumed that the ridge estimator of β under the hypothesis (4.13) is,

$$\beta^{*H} = C\hat{\beta}^H \tag{4.16}$$

then, $A\beta^{*H}$ is an unbiased estimator of $E(A\beta^*)$ under the restriction $A\beta = \rho$. The F statistic for the hypothesis (4.13) using β^* is the same as the F statistic in (4.15), provided C is positive definite, as,

$$(A\beta^* - A\beta^{*H}) = ACA^-(A\hat{\beta}-\rho)$$

and

$$Var(A\beta^* - A\beta^{*H}) = \sigma^2 ACA^- A(X'X)^{-1}A'(ACA^-)' \ .$$

So that under the key assumption in (4.16), ridge estimators supply the usual F statistics and confidence regions supplied by the least squares estimator $\hat{\beta}$.

R(v) Ridge-type estimators were designed primarily for use in situations where the model is assumed to be correctly specified but the data is multicollinear and a further collection of data is impossible. Smith and Goldstein (1975) have commented;

> "No one has ever claimed that Ridge Regression is preferable to more data; merely that it is preferable to least squares when more data is not available and $X'X$ is ill-conditioned."

R(vi) Despite the warnings of Box (1966), which were outlined in section 3.11 and which force experimenters to treat a fitted model as a "black box" when it is based on passive, unplanned, historical data, it was certainly one of the intentions of Hoerl and Kennard when they introduced ridge estimation to provide a method which would give suitable estimates, based on such passively observed data, of the partial derivatives of the expected responses so that control and optimization might proceed. Whether or not such an intention was statistically honourable is cause for debate, however, Obenchain (1975a), who may be classed as a proponent of ridge regression, has seen fit to remind users of ridge-type methods of Box's warnings. However, ridge regression drives the least squares estimates together so that if the warnings of Box are ignored and the regression coefficients from the

passively observed system are interpreted as estimates of the partial derivatives of the expected responses with respect to the regressors, such an interpretation is likely to withstand the changed circumstances in the controlled system better than in the case of ordinary least squares regression. Such a view has also been expressed by Tukey (1975).

R(vii) The use of the phrase ad hoc to describe the ridge regression procedure suggests that the ridge technique is a hastily contrived improvisation with little foundation. This may well be so, but until a suitably well-founded method is proposed for estimating the parameters of a linear model in the presence of multicollinearity data analysts will probably continue to use ridge estimation in exploratory work.

4.221 Other approaches to Ridge-type estimation.

In attempts to overcome some of the resistance to ridge regression several authors have pointed out the similarities between ridge-type estimators and techniques which exist in the literature and have already gained acceptance. Marquardt (1970) and Banerjee and Carr (1971) pointed out that ridge estimation is equivalent to least squares estimation with the actual data augmented by a fictitious set of data points in the manner,

$$\begin{bmatrix} y \\ \cdots \\ 0 \end{bmatrix} = \begin{bmatrix} X \\ \cdots \\ \sqrt{k}I \end{bmatrix} \beta + \begin{bmatrix} \epsilon \\ \cdots \\ \xi \end{bmatrix} . \qquad (4.17)$$

The unbiased least squares estimate of β in (4.17) is thus,

$$(X'X+kI)^{-1}X'y$$

when it is assumed that $\text{var}(\xi) = \sigma^2 I_{p\times p}$ and the p components of ξ are independent of the n components of ϵ. Thus ridge estimation, which is usually regarded as a last resort when augmentation of the actual data set is not possible, amounts to augmentation with an orthogonally designed artificial collection of data points. This fictitious augmentation has similarities with the "usual

constraints" method of handling the non-full rank linear model (see, for example, Searle (1971)).

A reexamination of equation (4.6) indicates that ridge estimation is equivalent to best linear estimation with the assumption of a prior distribution for β in which,

$$E(\beta\beta') = \frac{\sigma^2}{k} I.$$

Hoerl and Kennard (1970a) indicated that the ridge estimator could be viewed in a Bayesian context in which each component of the ridge estimator is a posterior mean based on a prior normal distribution for that component with mean zero and variance σ^2_β. This gives an explicit expression for k of σ^2/σ^2_β. Lindley and Smith (1972), Goldstein and Smith (1974) and Goldstein (1976) have also discussed this approach to ridge estimation and Lindley and Smith gave some discussion to the estimation of $k = \sigma^2/\sigma^2_\beta$. Under this Bayesian approach the ridge estimator is unbiased with respect to the combination of the prior information and the information from the data set if the prior information is correct.

4.222 Directed and Generalized Ridge Estimators

A simple generalization of the ridge estimator in expression (4.11) is,

$$\beta^* = (X'X+Q)^{-1}X'y$$

where Q is any non-negative definite matrix. One choice of Q that has received some attention (see, for example, Goldstein and Smith (1974) and Guilkey and Murphy (1975)) but which was originally introduced by Hoerl and Kennard (1970a) is,

$$Q = P'KP$$

where $K = \text{diag}(k_1,\ldots,k_p)$ and P' is the orthogonal matrix whose columns are the normalized eigenvectors of $X'X$, introduced in section 3.131. Such a generalized ridge estimator has been called a directed ridge estimator. The motivation for considering such estimators is discussed further in section 4.23.

4.223 Robustness and Ridge Estimators

Holland (1973) suggested combining the biweight regression procedure mentioned in section 3.4 with the ridge regression procedure of Hoerl and Kennard. The "ridgified robust estimator" proposed by Holland has the form of a weighted ridge estimator,

$$(X'WX + kI)^{-1}X'Wy \qquad (4.18)$$

where the diagonal matrix W comes from a robust biweight least squares fit. Holland claimed such an estimator would combine the benefits of a robust estimator - insensitivity to outliers and to non-normal error distributions - and the benefits of a ridge estimator - insensitivity to multicollinearity. Such a procedure may be wasteful in terms of computational time and the user's time. Finding the most useful weights for the robust fit may take several iterations and much monitoring of the residuals from each successive fit (see Beaton and Tukey (1974)). Having chosen a suitable weighting matrix W the expression (4.18) has then to be evaluated over a range of values for k. Marquardt (1974) suggested tackling both robustness and stability simultaneously. If smoothed predictions are desired Marquardt's suggestion is to use a generalized ridge estimator with,

$$Q = kX'[\Delta_n^{(2)}]'[\Delta_n^{(2)}]X$$

where $\Delta_n^{(2)}$ is an n dimensional second central difference operator matrix. It can be shown that such an estimation procedure results from minimizing the objective function,

$$\sum_{i=1}^{n} (y_i-\hat{y}_i)^2 + k \sum_{i=1}^{n} (\tfrac{1}{4}\hat{y}_{i-1} - \tfrac{1}{2}\hat{y}_i + \tfrac{1}{4}\hat{y}_{i+1})^2$$

or minimizing with respect to B,

$$\|y-XB\|^2 + k\|\Delta_n^{(2)}XB\|^2.$$

If smoothed coefficients are desired Marquardt's suggestion is to use a generalized ridge estimator with,

$$Q = k[\Delta_p^{(1)}]'[\Delta_p^{(1)}].$$

It can be shown that such an estimation procedure results from minimizing the objective function,

$$\sum_{i=1}^{n} (y_i-\hat{y}_i)^2 + k \sum_{j=1}^{p} (\frac{1}{2} B_j - \frac{1}{2} B_{j-1})^2$$

or,

$$\|y-XB\|^2 + k\|\Delta_p^{(1)} B\|^2.$$

Another suggestion for forcing together the estimates of the regression coefficients might be to constrain not the mean square successive differences of the coefficients but constrain,

$$\sum_{\substack{i=1 \\ i<j}}^{p} \sum_{j=1}^{p} (B_i-B_j)^2$$

The estimator which results from such an additional constraint has the form,

$$\beta^* = (X'X + k(pI-J))^{-1}X'y \qquad (4.19)$$

where $J=11'$. A similar estimator to (4.19) has been proposed by Lindley and Smith (1972) in a Bayesian context in which all the components of β are assumed to have prior normal distributions with equal non-zero means. The motivation for smoothing or forcing together estimates of the coefficients has been discussed by Tukey (1975). Briefly, there is a need, particularly in econometrics, for forecasting equations based on passively observed, historical data which provide some protection against possible catastrophic changes in the historical variance-covariance structure of the data. Forcing together the values of estimated coefficients, providing this does not increase the residual sum of squares too much, is one way of buying such protection. Tukey (1975) has suggested a modification of the estimator in (4.19). Instead of minimizing,

$$\|y-XB\|^2 + kB'(pI-J)B$$

Tukey has suggested minimizing with respect to B,

$$\|y-XB\|^2 + kB'(D-U-U')B$$

where $D = \text{diag}(\sum_j c_{1j}, \sum_j c_{2j}, \ldots, \sum_j c_{pj})$

and $$U = \begin{bmatrix} 0 & c_{12} & c_{13} & \cdots & c_{1p} \\ 0 & 0 & c_{23} & \cdots & c_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & c_{p-1,p} \\ 0 & 0 & \cdots & \cdots & 0 \end{bmatrix} .$$

The constraint term, which may be written, $k\sum_{\substack{i \\ i<j}} \sum_j c_{ij}(B_i-B_j)^2$

contains $P(P-1)/2$ weights c_{ij}. Tukey suggested choosing the weights c_{ij} to be,

$$\left\{\frac{1}{2}\log\left(\frac{1+r_{ij}}{1-r_{ij}}\right)\right\}^2$$

where r_{ij} is the correlation between x_i and x_j.

As yet there have been no published reports of comparisons among these robust ridge estimators. A simulation study comparing the performance of this class of alternatives under various degrees of non-normality and multicollinearity and the robust methods of section 3.4 is called for.

4.224 Other Ridge-type Estimators

The simple ridge estimator of Hoerl and Kennard constrains the squared Euclidean length of the estimated parameter vector and in so doing defines a trajectory through the parameter space from $\hat{\beta}$ to 0 for k increasing without bound from zero. This is shown geometrically in section 4.3. Algebraically,

$$\lim_{k\to\infty} \beta^* = \lim_{k\to\infty} (I+k(X'X)^{-1})^{-1}\hat{\beta} = 0$$

as $(I+k(X'X)^{-1})^{-1}$ tends to the null matrix for $k\to\infty$ (the eigenvalues of this matrix converge to zero). The angle between the ridge estimator and the vector $X'y$ tends to zero as $k\to\infty$ (see, for example, Marquardt (1970)). Thus as $k\to\infty$ the ridge estimator of β tends in Euclidean length to zero and tends in direction to that of $X'y$. As $X'y$ is the least squares estimator of β when X is an orthogonal design ($X'X$ is assumed to be in correlation form) then altering the direction of the estimated parameter vector in the direction of $X'y$ seems, in the absence of any other information about the direction in which β lies, to have some intuitive appeal. Allowing the length of the estimated parameter vector to shrink all the way to zero may however be a little naive. One suggestion may be to minimize,

$$\|y-XB\|^2 + k\|X'y-B\|^2 \quad . \tag{4.20}$$

Such an objective function constrains the Euclidean distance from $X'y$ to the estimated vector for a given fixed residual sum of squares. The estimator which results from the minimization with respect to B of expression (4.20) has the form of a "swollen" ridge estimator,

$$(1+k)(X'X+kI)^{-1}X'y \tag{4.21}$$

Clearly as $k\to\infty$, the estimator (4.21) tends to $X'y$ both in length and direction as,

$$\lim_{k\to\infty}\left(\frac{X'X}{(1+k)} + \frac{kI}{(1+k)}\right)^{-1} = I.$$

It is possible to show that this "swollen" ridge estimator outperforms the least squares estimator with respect to any generalized mean square error criterion of the form (3.46) if and only if the matrix,

$$\sigma^2\{\frac{2(1+k)}{k}(I-X'X)^{-1} + (X'X)^{-1}\} - \beta\beta'$$

is positive definite. In the manner of equation (4.17) the "swollen" ridge estimator is equivalent to the least squares estimator of β in the model,

$$\begin{bmatrix} y \\ \cdots \\ \sqrt{k}X'y \end{bmatrix} = \begin{bmatrix} X \\ \cdots \\ \sqrt{k}I \end{bmatrix} \beta + \begin{bmatrix} \epsilon \\ \cdots \\ \xi \end{bmatrix}$$

Swindel (1975, 1976) has suggested that the objective function,

$$\|y-XB\|^2 + k\|B-b\|^2 \qquad (4.22)$$

be minimized in situations where there is some prior information, in the form of a vector b, available on β. The resulting <u>good ridge estimator</u> has the form,

$$(X'X+kI)^{-1}(X'y+kb). \qquad (4.23)$$

In this estimation procedure the choice of k is seen to be a means of reaching a compromise between the prior information b and the least squares estimator based on the data $\hat{\beta}$. The properties of this estimator are presented in Swindel (1976). Swindel argues that, intuitively, shrinking $\hat{\beta}$ towards b makes more sense than shrinking $\hat{\beta}$ towards 0 or for that matter $X'y$. If no prior information on β is available then shrinking $\hat{\beta}$ towards 0 or $X'y$ does seem rather capricious. The main aim is that the length of $\hat{\beta}$ should be reduced when the data is ill-conditioned. This means that almost any shrinkage rule will effect some improvement. Shrinking and altering the direction of $\hat{\beta}$ towards the direction of $X'y$, as achieved by (4.11) and (4.21), is chosen, in the absence of any prior information about β, simply because $X'y$ is the only other known vector in the parameter space. The good ridge estimator of Swindel can also be regarded as an unbiased least squares estimator of β for the contrived model,

$$\begin{bmatrix} y \\ \cdots \\ \sqrt{k}b \end{bmatrix} = \begin{bmatrix} X \\ \cdots \\ \sqrt{k}I \end{bmatrix} \beta + \begin{bmatrix} \epsilon \\ \cdots \\ \xi \end{bmatrix} .$$

Vinod (1976a) has suggested a rescaling of the ridge estimator. The "two-stage" procedure given by Vinod consists of determining the ridge estimator β^* by choosing an appropriate value of k in

expression (4.11), then minimising with respect to μ,

$$\|y - \mu X\beta^*\|^2 \qquad (4.24)$$

The rescaled ridge estimator is $\mu\beta^*$ where μ, which minimizes expression (4.24), is given by,

$$\frac{\beta^{*\prime} X' y}{\beta^{*\prime} X' X \beta^*} \quad .$$

The philosophy behind such an estimator is that ridge regression often supplies reliable estimates of the relative magnitudes of the components of β but may overshrink the estimated parameter vector. Multiplying the estimated parameter vector by a scale factor blows the estimated parameter vector up to a more appropriate length and gives it the smallest residual sum of squares of all estimated parameter vectors in the same direction if criterion (4.24) is used to fix a value for the scale factor. The rescaled ridge estimator of Vinod has the form,

$$\frac{\beta^{*\prime} X' y}{\beta^{*\prime} X' X \beta^*} \cdot \beta^* \qquad (4.25)$$

and is not a linear estimator of β. Thus an investigation of the mean square error properties of this rescaled ridge estimator requires extra information about the distribution of ϵ.

Ridge regression and shrunken estimation were introduced originally for their mean square error improvement upon least squares estimation in the presence of multicollinearity. The fact that these estimators are least-squares-with-constraints estimators and linear transformations of the least squares estimator has stimulated researchers to look for other such estimators, not primarily with the motive of mean square error reduction but with more specific motives, robustness and prior knowledge incorporation for example. Some of the resulting estimators have been mentioned above and are represented in Table 4.1. Choosing to use a particular biased estimation procedure which is a linear transformation of the least squares estimator clearly requires consideration of robustness and the nature and extent of any prior knowledge which is available, as well as the detection of any

Table 4.1. SOME LINEAR TRANSFORMATIONS OF THE LEAST SQUARES ESTIMATOR

Estimator	Transformation Matrix, $C_{p\times p}$	Objective Function	Usual Form of the Estimator
LEAST SQUARES Gauss et alia, 19th century	I	$\|y-XB\|^2$	$(X'X)^{-1}X'y$
SHRUNKEN Mayer and Willke (1973)	$\frac{1}{1+k}\, I, \quad 0 \le k$	$\|y-XB\|^2+k\|XB\|^2$	$\frac{1}{1+k}\,(X'X)^{-1}X'y$
RIDGE Riley (1955) Hoerl and Kennard (1970)	$(I+k(X'X)^{-1})^{-1}$, $0 \le k \le 1$	$\|y-XB\|^2+k\|B\|^2$	$(X'X+kI)^{-1}X'y$
GENERALIZED RIDGE Hoerl and Kennard (1970)	$(I+(X'X)^{-1}Q)^{-1}$, Q n.n.d.	$\|y-XB\|^2+B'QB$	$(X'X+Q)^{-1}X'y$
DIRECTED RIDGE Guilkey and Murphy (1975)	$(I+(X'X)^{-1}P'KP)^{-1}$, $K=\mathrm{diag}(k_1,\ldots,k_p)$	$\|y-XB\|^2+\|K^{\frac{1}{2}}PB\|^2$	$(X'X+P'KP)^{-1}X'y$
—————— ——————	$(1+k)(I+k(X'X)^{-1})^{-1}$ $0 \le k \le 1$	$\|y-XB\|^2+k\|X'y-B\|^2$	$(1+k)(X'X+kI)^{-1}X'y$

Table 4.1 (Continued)

Estimator	Transformation Matrix, $C_{p\times p}$	Objective Function	Usual Form of the Estimator
GOOD RIDGE Swindel (1975)	$C=(I+k(X'X)^{-1})^{-1}$ $d=(I-C)b$	$\|y-XB\|^2+k\|B-b\|^2$	$C\hat{\beta}+d=(X'X+kI)^{-1}(X'y+kb)$
PREDICTION SMOOTHER Marquardt (1974)	$(I+k(X'X)^{-1}X'\Delta_n^{(4)}X)^{-1}$	$\|y-XB\|^2+k\|\Delta_n^{(2)}XB\|^2$	$\{X'(I+k\Delta_n^{(4)})X\}^{-1}X'y$
COEFFICIENT SMOOTHER Marquardt (1974)	$(I+k(X'X)^{-1}\Delta_p^{(2)})^{-1}$	$\|y-XB\|^2+k\|\Delta_p^{(1)}B\|^2$	$(X'X+k\Delta_p^{(2)})^{-1}X'y$
Lindley and Smith (1972)	$\{I+k(X'X)^{-1}(pI-J)\}^{-1}$	$\|y-XB\|^2+kB'(pI-J)B$	$(X'X+k(pI-J))^{-1}X'y$
Tukey (1975)	$\{I+k(X'X)^{-1}(D-U-U')\}^{-1}$	$\|y-XB\|^2+k\sum_{i<j} c_{ij}(B_i-B_j)^2$ or, $\|y-XB\|^2+kB'(D-U-U')B$	$(X'X+k(D-U-U'))^{-1}X'y$

multicollinearity in the data set. As an aside, the complicated form of Tukey's robust generalized ridge estimator suggests that the threshold at which increased complexity means diminished returns is near.

4.23 Shrinkage in the Canonical Form of the Model

Consider the transformed linear model in equation (3.11) with the modification that the eigenvalues of $X'X$ are relabelled in descending order of magnitude,

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$$

so that the first column of $Z=XP'$ now consists of n observations of the first principal component of X and the pth column of Z consists of n observations on the last or most minor principal component of X. From equation (3.12),

$$\begin{aligned}\hat{\alpha}_i &= \alpha_i + z_i'e/\lambda_i \\ &= z_i'y/\lambda_i \end{aligned} \qquad (4.26)$$

where z_i' is the transpose of the ith column of z. Also from equations (3.14),

$$\text{var}(\hat{\alpha}_1) \leq \text{var}(\hat{\alpha}_2) \leq \ldots \leq \text{var}(\hat{\alpha}_p)$$

as,

$$\frac{\sigma^2}{\lambda_1} \leq \frac{\sigma^2}{\lambda_2} \leq \ldots \leq \frac{\sigma^2}{\lambda_p}$$

Thus as pointed out in section 3.131 the linear combination of β, $t'\beta$, subject to the constraint $t't=1$, which is estimated with the least variance is,

$$\alpha_1 = p_1'\beta$$

and the linear combination $t'\beta$ subject to the same constraint which is estimated with the largest variance is,

$$\alpha_p = p_p'\beta$$

If $\lambda_p \ll 1$ the variance of $\hat{\alpha}_p$ is very large and $\hat{\alpha}_p$ may well be some distance away from α_p. This is the familiar multicollinearity problem.

The ridge regression solution proposed by Hoerl and Kennard (1970a,b) adds a small positive constant value k to each of the eigenvalues of $X'X$ so that equation (4.26) becomes,

$$\alpha_i^* = \left(\frac{\lambda_i}{\lambda_i+k}\right)\alpha_i + \frac{z_i'\epsilon}{\lambda_i+k}$$

$$= z_i'y/(\lambda_i+k) \qquad (4.27)$$

A comparison of equations (4.26) and (4.27) shows that a bias of $(-k/\lambda_i+k)\alpha_i$ is the expense incurred by decreasing the weight given to the error term in the ridge estimator. Coresponding to the decrease in weight given to the error term,

$$\text{var}(\alpha_i^*) = \frac{\sigma^2\lambda_i}{(\lambda_i+k)^2}$$

which is less than

$$\text{var}(\hat{\alpha}_i) = \frac{\sigma^2}{\lambda_i} \quad .$$

The reduction in the variance of estimates of coefficients is not great for those coefficients associated with the major principal components (they are estimated precisely enough by least squares anyway) but it is great for the coefficient associated with the smallest eigenvalue λ_p of $X'X$, particularly if $\lambda_p \ll 1$. The bias in the coefficients which are associated with the larger eigenvalues is not great but the bias in the estimated coefficient associated with λ_p may well be of a higher order of magnitude than α_p itself.

The ridge estimation technique is clearly sensitive to the structure of $X'X$. It may however be made more sensitive to the eigenvalue structure of $X'X$ by allowing different values of k to

be associated with different eigenvalues as Hoerl and Kennard (1970a) and Guilkey and Murphy (1975) have pointed out. Such a directed ridge estimator has the form,

$$(\Lambda + K)^{-1}Z'y$$

where $K = \text{diag}(k_1,\ldots,k_p)$ so that the analogue of equation (4.27) is,

$$\alpha_i^* = \left(\frac{\lambda_i}{\lambda_i+k_i}\right)\alpha_i + z_i'\varepsilon/(\lambda_i+k_i)$$

$$= z_i'y/(\lambda_i+k_i) \ . \qquad (4.28)$$

For the larger eigenvalues, say, $\lambda_i \geq 10^{-c}\lambda_1$, where c is some arbitrary constant, the k_i may be set equal to zero as the corresponding α_i are estimated relatively precisely by least squares. For the smaller eigenvalues, $\lambda_i < 10^{-c}\lambda_1$, the k_i may be set equal to a small positive constant k or, as suggested by Hoerl and Kennard (1970a), Goldstein and Smith (1974) and Guilkey and Murphy (1975) each k_i may be set equal to,

$$\hat{\sigma}^2/\hat{\alpha}_i^2$$

and an iterative procedure initiated with $k_{i,t+1} = \hat{\sigma}^2/(\alpha_{i,t}^*)^2$ until stability of the α_i^* is achieved. The justification for this iterative procedure arises from the result that the mean square error of α_i^*, which is of the form,

$$E\left(\frac{z_i'y}{\lambda_i+k_i} - \alpha_i\right)^2 = \frac{k_i^2\alpha_i^2 + \sigma^2\lambda_i}{(\lambda_i+k_i)^2}$$

has a minimum at,

$$k_i = \sigma^2/\alpha_i^2 \ . \qquad (4.29)$$

Thus the k_i may be estimated initially by $\hat{\sigma}^2/\hat{\alpha}_i^2$, but as,

$$E(\hat{\alpha}_i^2) = \alpha_i^2 + \frac{\sigma^2}{\lambda_i}$$

and σ^2/λ_i is large for the smaller eigenvalues, $\hat{\alpha}_i^2$ may overestimate α_i^2 so that reestimation of k_i using α_i^* and consequent reestimation of α_i in an iterative manner may produce a k_i close in value to (4.29). Guilkey and Murphy (1975) have illustrated the use of this technique.

Transforming the model to its canonical form gives a clearer picture of the manner in which ridge and directed ridge estimators tackle the multicollinearity problem. The ridge-type estimators do attempt to overcome the imprecise estimation in certain directions of the estimation space which is the essential characteristic of the multicollinearity problem. The effect of scalar shrinkage of the least squares estimates is also more clearly discernible when the linear model is in canonical form. The deterministically shrunken estimators of Mayer and Willke (1973) can be written as,

$$\tilde{\alpha}_i = \frac{1}{1+k}\alpha_i + z_i'\epsilon/\lambda_i(1+k)$$

$$= z_i'y/\lambda_i(1+k) \qquad (4.30)$$

Comparison of this estimation procedure with (4.26) and (4.27) reveals that the shrunken estimator does indeed ignore the relative precision of estimation in the directions of the various principal components of X and applies a crude blanket shrinkage to all the estimates $\hat{\alpha}_i$.

The "swollen" ridge estimator in equation (4.21) can be written as,

$$\tilde{\tilde{\alpha}}_i = \frac{\lambda_i(1+k)}{(\lambda_i+k)}\alpha_i + \frac{(1+k)}{(\lambda_i+k)}z_i'\epsilon$$

$$= (1+k)z_i'y/(\lambda_i+k).$$

With the estimator expressed in this form it can be seen quite clearly that for moderate to large eigenvalues, say $\lambda_i > 1$, the relatively precise least squares estimates are not shrunk but are swollen and have their variance increased. The least squares estimates of the α_i which correspond to small eigenvalues, say $\lambda_i < 1$ are shrunk and have their variance reduced. Thus this estimation procedure does not have good mean square error properties when X is not highly multicollinear. If many of the λ_i are less than one then this procedure may be useful, otherwise its usefulness is limited only to situations in which prior knowledge suggests shrinkage towards $X'y$.

The ridge-type and shrunken estimators proceed on the assumption that the original linear model is correctly specified and of full rank, and that the multicollinearity or small eigenvalues arise from a poor collection of data. It could be that the model is in fact misspecified and not of full rank. In this case the last p-r small eigenvalues should not be inflated (by adding k in the case of ridge or multiplying by 1+k in the case of scalar shrinkage) but should perhaps be set equal to zero, reflecting the belief that X actually has column rank r, $1 \leq r \leq p$. Such a procedure is equivalent to deleting the p-r minor principal components (selection criterion (a) of section 3.131) in a principal components regression. Thus the matrix $\Lambda = \text{diag}(\lambda_1,\ldots,\lambda_p)$ is transformed to $\Lambda_r = \text{diag}(\lambda_1,\ldots,\lambda_r,0,\ldots,0)$ so that α is now estimated by,

$$\begin{bmatrix} \overset{+}{\alpha}_{(r)} \\ \cdots \\ 0 \end{bmatrix} = \begin{bmatrix} \hat{\alpha}_1 \\ \vdots \\ \hat{\alpha}_r \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \Lambda_r^{-1} Z'y.$$

Deleting principal components in this fashion was first suggested by Kendall (1957). The method does have some appeal in that while components are dropped in the canonical estimation space, variables in the original estimation space are not deleted, that is,

$$\beta_r^+ = P' \begin{bmatrix} \alpha_{(r)}^+ \\ \cdots \\ 0 \end{bmatrix} = P'_{(r)} \alpha_{(r)}^+$$

where $P' = [P'_{(r)} \vdots P'_{(p-r)}]$. There is a kind of "robustness" here in that the original variables are retained in the final prediction equation even though there are dependencies among them, and their estimated coefficients are arrived at by utilizing the directions in the canonical estimation space in which prediction is most precise.

In fact it is shown geometrically in section 4.3, for the case of two predictors, that such a procedure is equivalent to the coefficient smoothing procedure in equation (4.19). The estimated coefficients are, of course, biased but the sum of their variances is less than the corresponding sum for least squares,

$$\sigma^2 \sum_{i=1}^{r} 1/\lambda_i < \sigma^2 \sum_{i=1}^{p} 1/\lambda_i \quad , \qquad \forall r < p.$$

The warning in section 3.131 should however be noted.

Marquardt (1970) proposed a modification to this principal component selection technique. Instead of assuming X has column rank r when the last p-r eigenvalues of $X'X$ are close to zero, Marquardt suggested the assumption of a fractional rank for X of say s where $s \in (r,r+1)$. This allows for some margin of safety in the divination of the rank of X and in the assessment of which of the ordered eigenvalues of $X'X$ are practically zero. Marquardt called this procedure <u>generalized inverse regression</u>. Properties of this estimator and its relationship to the principal components estimator and other component selection techniques are contained in Marquardt (1970, Goldstein and Smith (1974) and Hocking, Speed and Lynn (1976). These two component selection strategies (principal components regression and generalized inverse regression) are linear transformations of the least squares estimator of α. To allow comparison with the ridge-type and shrunken techniques the transformation matrices of these shrinkage-in-the-canonical-form techniques are presented in Table 4.2.

Table 4.2 SOME LINEAR TRANSFORMATIONS OF THE LEAST SQUARES ESTIMATOR OF α IN THE CANONICAL FORM OF THE LINEAR MODEL.

Estimator	Transformation Matrix, C	Usual form of the Estimator
Least Squares	$I_{p\times p}$	$\Lambda^{-1}Z'y$
Shrunken	$\frac{1}{1+k} I_{p\times p}$	$\frac{1}{1+k} \Lambda^{-1}Z'y$
Ridge	$(I+k\Lambda^{-1})^{-1}_{p\times p}$	$(\Lambda+kI)^{-1}Z'y$
Directed Ridge	$(I+\Lambda^{-1}K)^{-1}_{p\times p}$	$(\Lambda+K)^{-1}Z'y$
Principal Components	$\begin{bmatrix} I_{r\times r} & 0_{r\times(p-r)} \\ 0_{(p-r)\times r} & 0_{(p-r)\times(p-r)} \end{bmatrix}$	$\Lambda_r^{-1}Z'y$
Generalized Inverse or Fractional Rank	$\begin{bmatrix} I_{r\times r} & 0_{r\times 1} & 0 \\ 0_{1\times r} & s-r & 0 \\ 0 & 0 & 0_{(p-r-1)\times(p-r-1)} \end{bmatrix}$	$\begin{bmatrix} \Lambda_{(r)}^{-1} & 0 & 0 \\ 0 & \frac{s-r}{\lambda_{r+1}} & 0 \\ 0 & 0 & 0 \end{bmatrix} Z'y$

From an inspection of Table 4.2 the ridge and shrunken estimators whose transformation matrices are always positive definite are quite different to the two component selection estimators whose transformation matrices are non-negative definite. This difference reflects the difference in assumed rank upon which the two types of estimator proceed. Both component selection procedures define unique generalized inverses of $X'X$. The principal components procedure defines the Moore-Penrose generalized inverse, so that,

$$\beta_r^+ = P'_{(r)} \Lambda^{-1}_{(r)} P_{(r)} X'y = (X'X)^+_{(r)} X'y.$$

The generalized inverse or fractional rank estimator defines the unique generalized inverse,

$$P'_{(r)} \Lambda^{-1}_{(r)} P_{(r)} + \left(\frac{s-r}{\lambda_{r+1}}\right) P_{(r+1)} P'_{(r+1)} .$$

The canonical approach to regression problems and biased estimation exemplified in Goldstein and Smith (1974), Greenberg (1975), Obenchain (1975a) and Hocking, Speed and Lynn (1976) does seem to provide a deeper insight into the manner in which multicollinearity affects estimation techniques. It would seem, therefore, to be a useful strategy in any regression analysis to commence with a principal components analysis of the input matrix to determine whether significant multicollinearity is present and to enable an appropriate estimation technique to be selected. This has been suggested by Kendall (1957) but it seems that his suggestion has been overlooked by many data analysts (see, for example, Longley (1967)).

4.24 General Observations

All the linear transformations of the least squares estimator studied in section 4.2 involve symmetric non-negative definite transformation matrices. These matrices C, listed in Tables 4.1 and 4.2 satisfy the general conditions,

(i) $C(X'X)^{-1}$ is symmetric

(ii) $C(X'X)^{-1} - C(X'X)^{-1}C'$ is non-negative definite. Rao (1976) has shown that conditions (i) and (ii) are necessary and sufficient conditions (when X has full column rank) for the estimator $C\hat{\beta}$ to be linear admissable with respect to criterion (3.46). Rao posits that a study of biased linear admissable estimators should start with the general class $C\hat{\beta}$ where C satisfies conditions (i) and (ii) and then focus on subclasses defined by a particular choice of C. Such an approach has been taken here as the class of Bayes Linear estimators, equation (4.5) under assumption (ii) has been shown by Rao to be precisely the class of admissable linear estimators.

The various biased linear estimators produce their possibly ethereal mean square error properties by distorting the data in some fashion. Such procedures, which have been anathematized by many statisticians are quite common in the physical and engineering sciences especially in situations without a stochastic error formulation. Lawson and Hanson (1974) for example state that since the data defining a least squares problem are uncertain the data can be changed within the bounds of that uncertainty to suit the needs of the data analyst. Such changes are motivated by the desire to achieve stability so that further small changes in the data do not produce large changes in the solution. It is not coincidental then that many of the biased estimation procedures have first appeared, in their statistical setting, in Technometrics. The impression gained is that the biased estimation procedures have been proposed not only with the multicollinearity problem in mind but also with the desire to place on a firmer statistical foundation some of the widely used perturbation methods for solving systems of linear equations.

4.3 Geometric Representation of Some Biased Estimators.

The geometrical interpretation of least squares estimation theory is contained in many textbooks on regression methods, e.g. Scheffé (1959), Draper and Smith (1966) and Theil (1971). The various

biased estimators introduced so far are modifications of the least squares estimator and thus have relatively simple but illuminating geometrical representations.

The vector of observations of the dependent variable y and the p columns of the matrix of predictor variables X in the linear model, equation (3.1), may be represented by vectors in an n-dimensional vector space - the sample space. The p vectors of the independent variables define a hyperplane in the n-dimensional sample space. This hyperplane is a p-dimensional subspace of the sample space and may be thought of as the estimation space. The method of least squares finds the vector (a linear combination of the p predictor-variable vectors) in the hyperplane which is closest to the vector y, i.e.,

$$\min_{B} \|y-XB\|^2 \quad .$$

The resulting vector,

$$\hat{y} = X\hat{\beta} \quad ,$$

where $\hat{\beta}$ is the pX1 vector satisfying the least squares criterion, is the projection of y on the hyperplane defined by the columns of X. The situation is illustrated in Figure 4.1 for the case in which the dimension of the sample space is 3 and the dimension of the estimation space is 2.

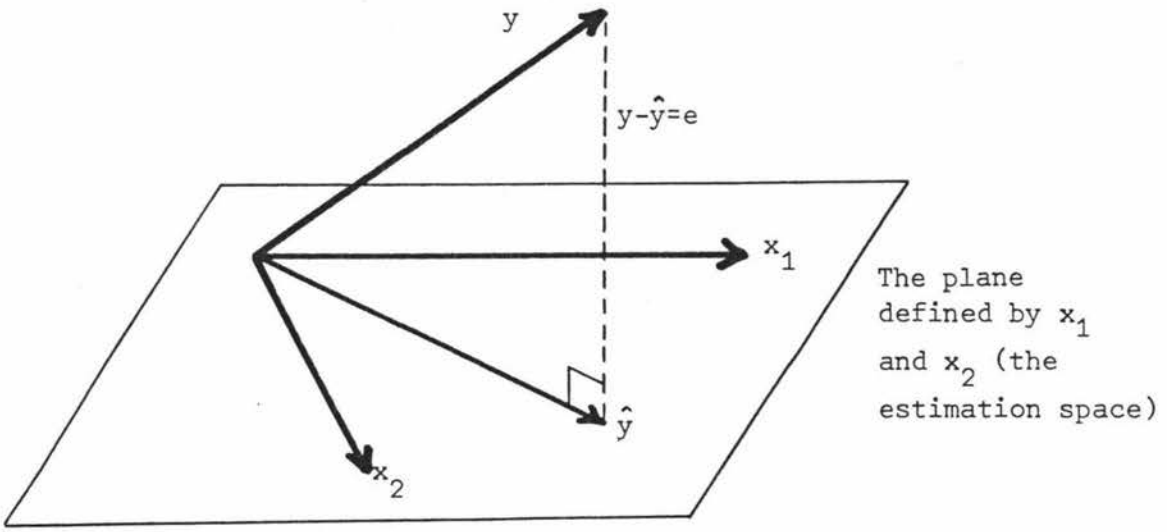

Figure 4.1 Least squares estimation when p=2 and n=3.

From Figure 4.1 it can be seen that multiple regression using the method of least squares consists of splitting the vector y in the sample space into two orthogonal component vectors. One component vector, the predicted value of the dependent variable, $\hat{y}$, lies entirely in the p-dimensional estimation space or hyperplane and the other component vector, the vector of residuals, e, is the projection of y on the direction orthogonal to the estimation space hyperplane.

If an estimation procedure other than least squares is used, the resulting predicted vector $\hat{y}$ in the estimation space will be further from the observed vector y in the sample space. This is a direct consequence of the least squares criterion which minimises the residual sum of squares or the squared length of the vector e. Algebraically the residual sum of squares for any estimator B of β can be written as,

$$\begin{aligned} RSS(B) &= \|y-XB\|^2 \\ &= \|y-X\hat{\beta}\|^2 + \|X(B-\hat{\beta})\|^2 \\ &= RSS(\hat{\beta}) + \|c\|^2 \quad . \end{aligned} \tag{4.31}$$

Thus a predetermined, tolerable increase in residual sum of squares, of say $\|c\|^2$ defines a whole family of predicted vectors in the estimation space and a corresponding family of estimators of β which are alternatives to the least squares estimator $\hat{\beta}$. The alternative predicted vectors XB which arise from a given increase in lack of fit form a p-dimensional hypersphere centred on the least squares prediction vector $\hat{y}$ in the p-dimensional estimation space. This is illustrated in Figure 4.2 for the case in which the sample space has 3 dimensions and the estimation space has 2 dimensions.

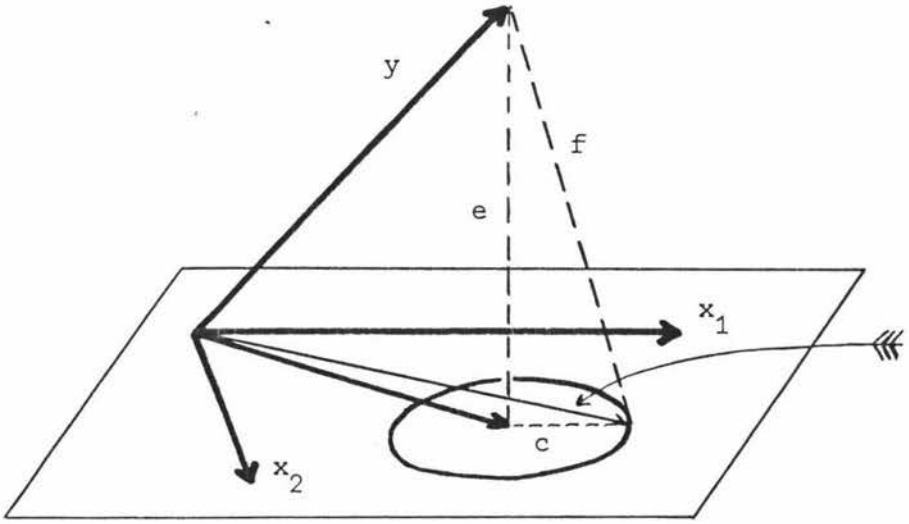

Figure 4.2 The family of predicted vectors in the estimation space with constant residual sum of squares $RSS(\hat{\beta}) + \|c\|^2$. Note that $\|f\|^2 = \|e\|^2 + \|c\|^2$.

Further geometric characterization of the various alternative estimators with fixed residual sum of squares requires a consideration of the p-dimensional parameter space. The parameter space consists of p orthogonal directions along which estimates of each of the p components of β are measured. The parameter space when p=2 is shown in Figure 4.3.

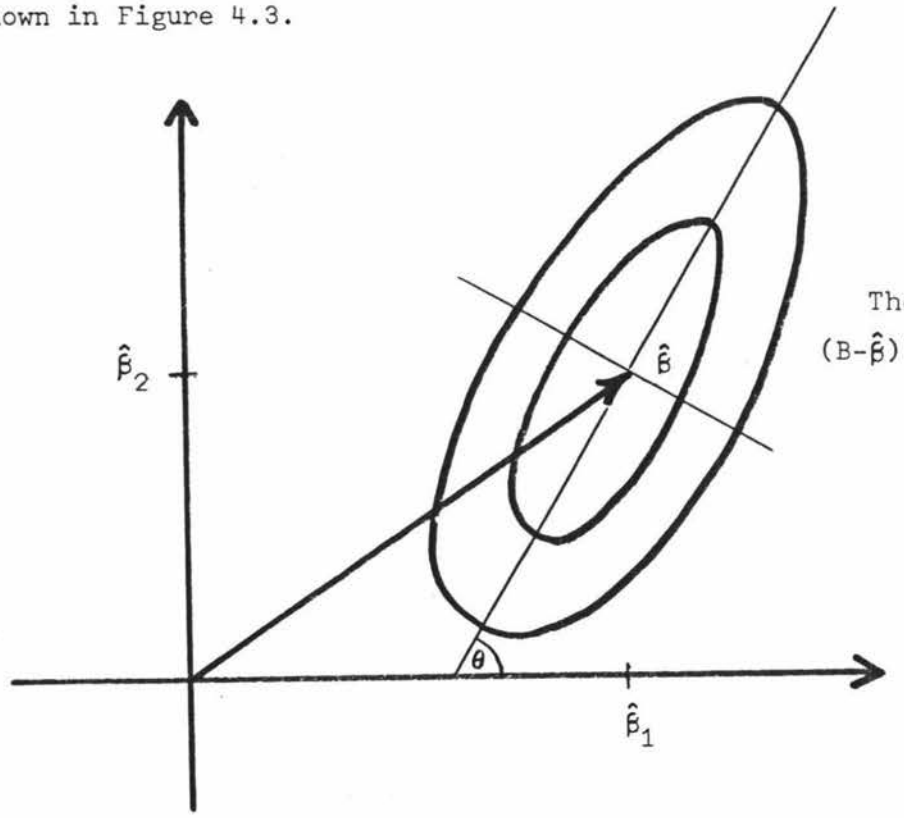

Figure 4.3 The parameter space when p=2.

The least squares estimator $\hat{\beta}$ has minimum residual sum of squares. All other estimators B have residual sums of squares greater than the residual sum of squares for the least squares estimator by an amount, from equation (4.31), $(B-\hat{\beta})'X'X(B-\hat{\beta})$. A fixed increase in residual sum of squares of K gives a family of estimators defined by the hyperellipsoid,

$$(B-\hat{\beta})'X'X(B-\hat{\beta}) = K$$

with centre point at $\hat{\beta}$. Thus the concentric ellipses centred on $\hat{\beta}$ in Figure 4.3 are contours of lack of fit.

The lack of fit hyperellipsoids in the p-dimensional parameter space have three important features:

(i) Under the usual error distribution assumptions, $\varepsilon \sim N(0,\sigma^2 I)$, the hyperellipsoids are $100(1-\alpha)\%$ confidence regions for β, when the constant K is set equal to

$$RSS(\hat{\beta}).\frac{p}{n-p} \; F_{1-\alpha}(p,n-p)$$

(ii) The orientation of the hyperellipsoids is related to the orthogonality of the matrix of predictor variables X. If the p axes of the hyperellipsoids are parallel to the p orthogonal axes defining the parameter space then the least squares estimates of the p components of the vector β are uncorrelated and the matrix X is orthogonal. In Figure 4.3 the axes of the ellipses are not parallel to the axes defining the parameter space so that the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ are correlated and the vectors x_1 and x_2 in the estimation space are not orthogonal.

(iii) The "shape" of the hyperellipsoids is an indication of the precision of the least squares estimates of each of the components of the parameter vector β. In the orthogonal case, the lengths of the p axes of a particular lack of fit hyperellipsoid are proportional to the variances of the least squares estimates of the components of β. In the non-orthogonal case the lengths of the p axes are proportional to the variances of specific linear combinations of the least squares estimates. Thus, in Figure 4.3,

the variance of

$$\hat{\beta}_1 \cos\theta + \hat{\beta}_2 \sin\theta$$

is greater than the variance of

$$-\hat{\beta}_1 \sin\theta + \hat{\beta}_2 \cos\theta \ .$$

The various biased estimation procedures provide rules for plucking a particular estimated vector of parameters from a given family of estimators with common, fixed residual sum of squares $RSS(\hat{\beta})+K$. The simplest biased estimation procedure is provided by Mayer and Willke (1973). Their deterministically shrunken estimators result from minimizing

$$\|y-XB\|^2 + k\|XB\|^2 \ .$$

Thus for a given fixed residual sum of squares the deterministically shrunken estimator is the estimator which minimises the squared length of the predicted vector XB in the estimation space. Correspondingly, in the parameter space, the deterministically shrunken estimator is a scalar shrinking of the least squares estimator (see section 4.21). This is shown for the case n=3, p=2 in Figure 4.4.

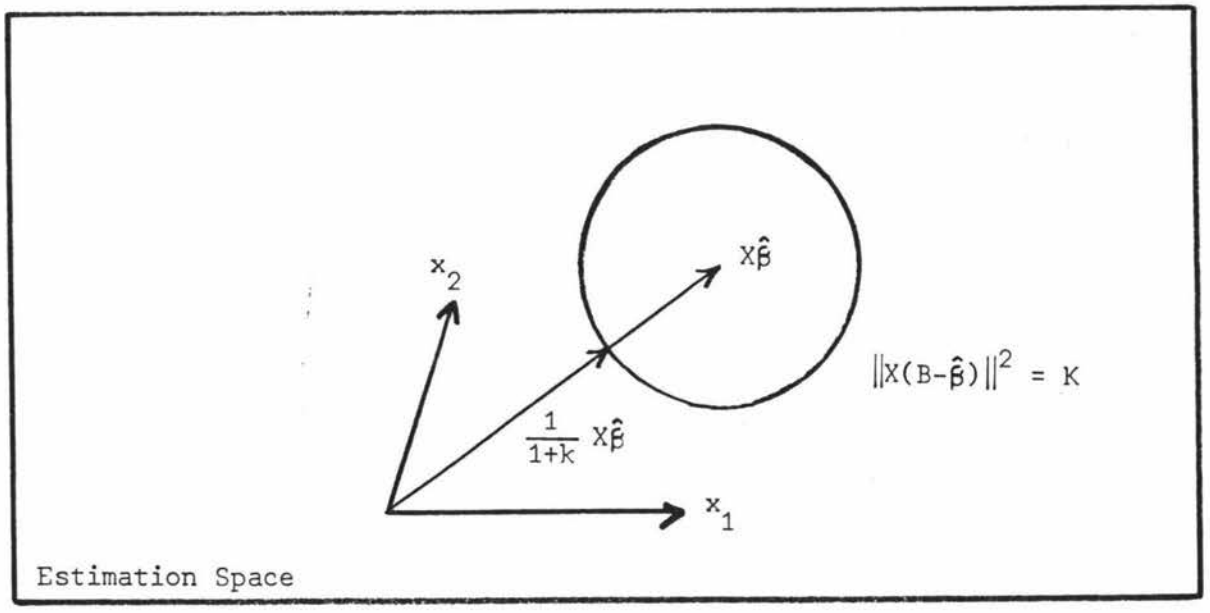

Figure 4.4(a) The estimation space hyperplane when p=2 and the deterministically shrunken estimator with given residual sum of squares.

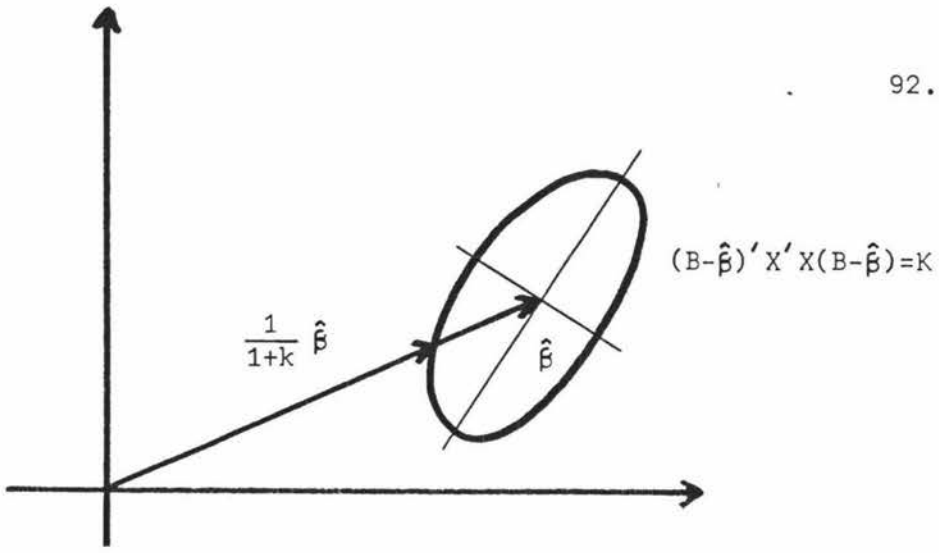

Figure 4.4(b) The deterministically shrunken estimator $\frac{1}{1+k}\hat{\beta}$.

The ridge estimators of Hoerl and Kennard (1970a,b) result from minimising

$$\|y-XB\|^2 + k\|B\|^2 \quad .$$

Thus for a given fixed residual sum of squares the ridge estimator is the estimator with the smallest squared length. Thus the ridge estimator is the shortest vector from the origin of the parameter space to a particular lack of fit hyperellipsoid in the parameter space. Figure 4.5 shows a 2 dimensional parameter space and the uniquely determined ridge estimator with given residual sum of squares.

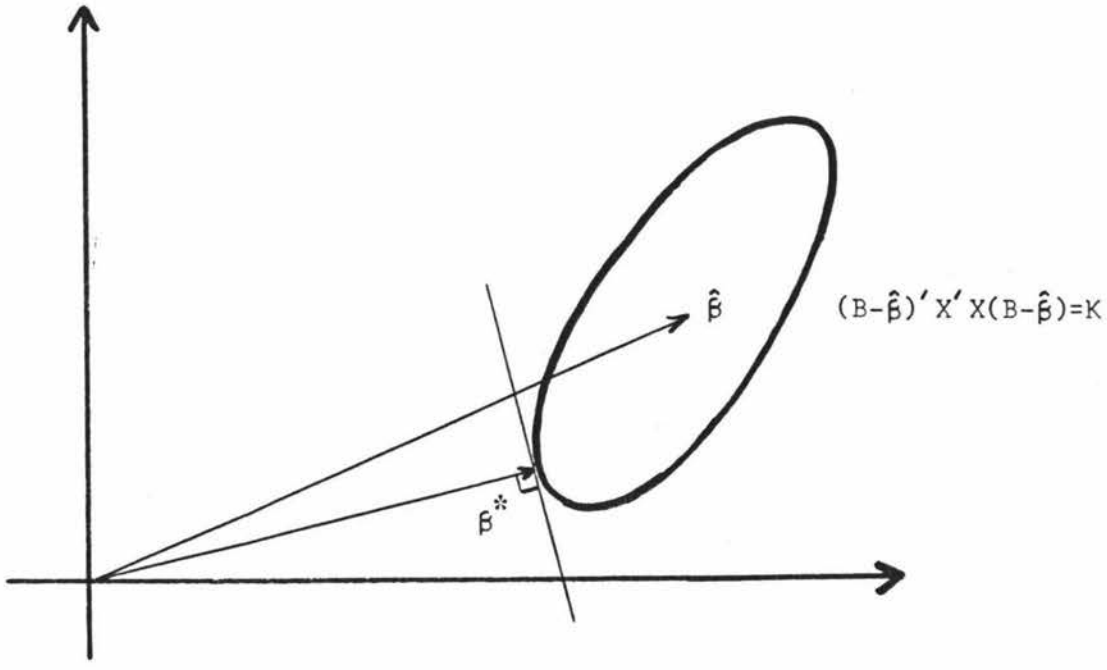

Figure 4.5 The ridge estimator β^*.

The "swollen" rescaled ridge estimator proposed by Vinod (1976a) is a scalar multiple of a predetermined ridge estimator, $\mu\beta^*$, where μ is chosen to minimise RSS($\mu\beta^*$). Clearly $\mu\beta^*$ is the vector lying in the same direction as β^* with the smallest residual sum of squares. The "swollen" ridge estimator therefore meets or cuts one of the axes of the concentric family of hyperellipsoids in the parameter space in the manner of Figure 4.6(a). Depending on the general shape of the hyperellipsoids, or the structure of $X'X$, the rescaled ridge estimator may or may not pass through the axis concerned (see Figure 4.6(b)).

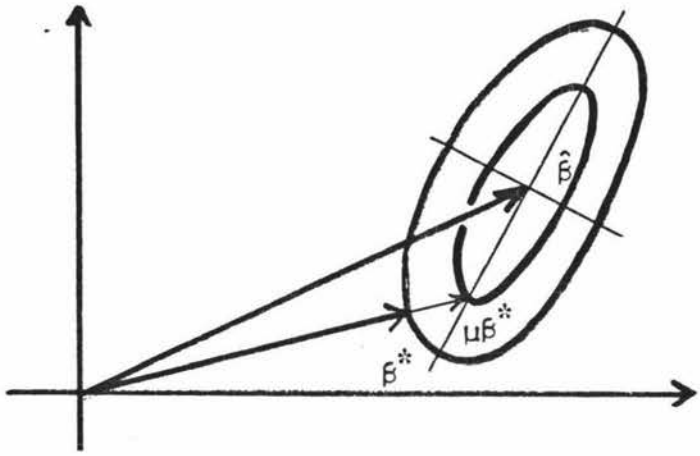

Figure 4.6(a) The rescaled ridge estimator $\mu\beta^*$ where $\mu = (\beta^{*\prime}X'y)/(\beta^{*\prime}X'X\beta^*)$

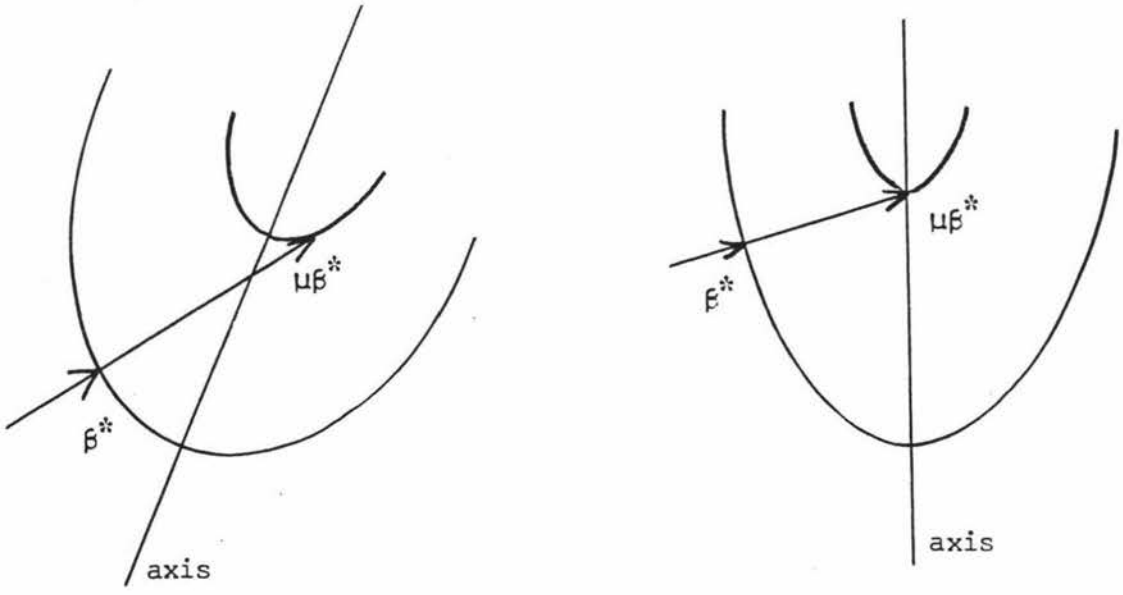

Figure 4.6(b) Two possible intersections of the rescaled ridge estimator with a principal axis of the lack of fit ellipses.

The estimators in equations (4.21), (4.23) and (4.19) are presented in Figures 4.7, 4.8 and 4.9 respectively. In each figure the same lack of fit ellipsoid is shown, the same least squares parameter vector and the same ridge estimation vector are shown too. The estimator (4.21) which is represented in Figure 4.7 identifies the estimate closest to the vector $X'y$ for a given family of estimators with the same residual sum of squares. The estimator (4.23) which is represented in Figure 4.8 plucks out the estimate, in the family of estimators with the same residual sum of squares, which is closest to the a priori vector b. The estimation rule (4.19) which is equivalent to Marquardt's coefficient smoother for the case p=2 and which is represented in Figure 4.9 identifies the estimate on the ellipse which is closest to the line $\beta_1=\beta_2$. At first sight the three different estimates of β may seem alarming (in this illustration all three estimates lie in different directions and have different lengths). However, the three estimates result from three different sets of a priori circumstances. The estimates in Figures 4.7 and 4.8 would be identical if $b=X'y$, that is if the prior knowledge in both cases was identical. The estimate in Figure 4.9 arises from a desire to achieve robustness for a given value of the residual sum of squares. Prior knowledge and the purpose for which an estimator is to be used are clearly important determinants in the choice of an estimation procedure.

The generalized inverse estimator of Marquardt and the principal component estimator are identical when r=s. In the case when p=2 and λ_2 is near zero the assumption that $\lambda_2=0$ or that the rank of $X'X$ is essentially one is equivalent to the constraint $\beta_1=\beta_2$ (see, for example, Marquardt (1970) and Hocking (1976)). Thus Figure 4.9 also shows a generalized inverse estimator with assigned rank 1 for the case p=2.

As a final visual aid the trajectories defined by the ridge estimator (4.11) and the good ridge estimator (4.23) for values of k increasing without bound from zero are shown in Figure 4.10. The intuitive appeal of Swindel's good ridge estimator, when the relevant prior information is available, is illustrated in this figure.

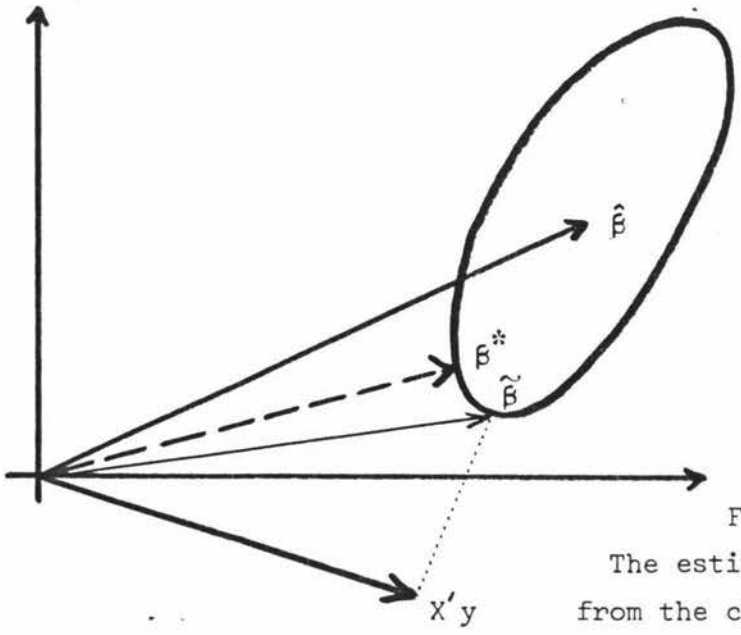

Figure 4.7
The estimator $\tilde{\beta}$ resulting from the constraint $\|X'y-B\|^2$

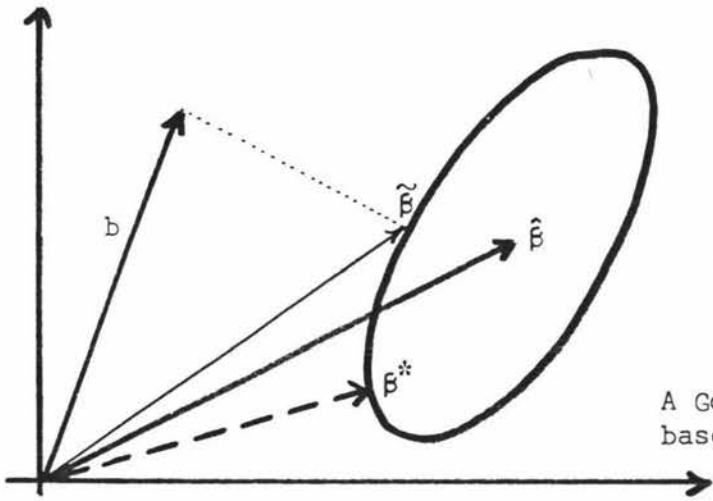

Figure 4.8
A Good Ridge Estimator $\tilde{\beta}$ based on prior information b.

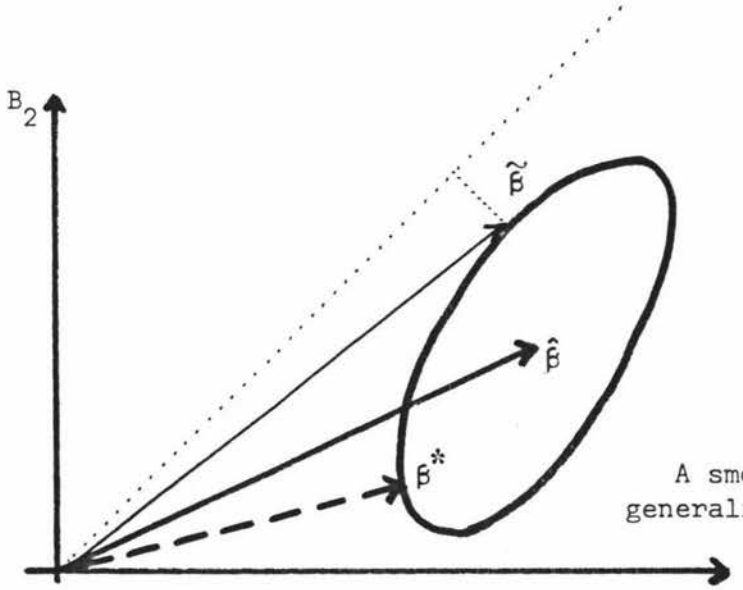

Figure 4.9
A smoothed coefficient, generalised ridge estimator, $\tilde{\beta}$.

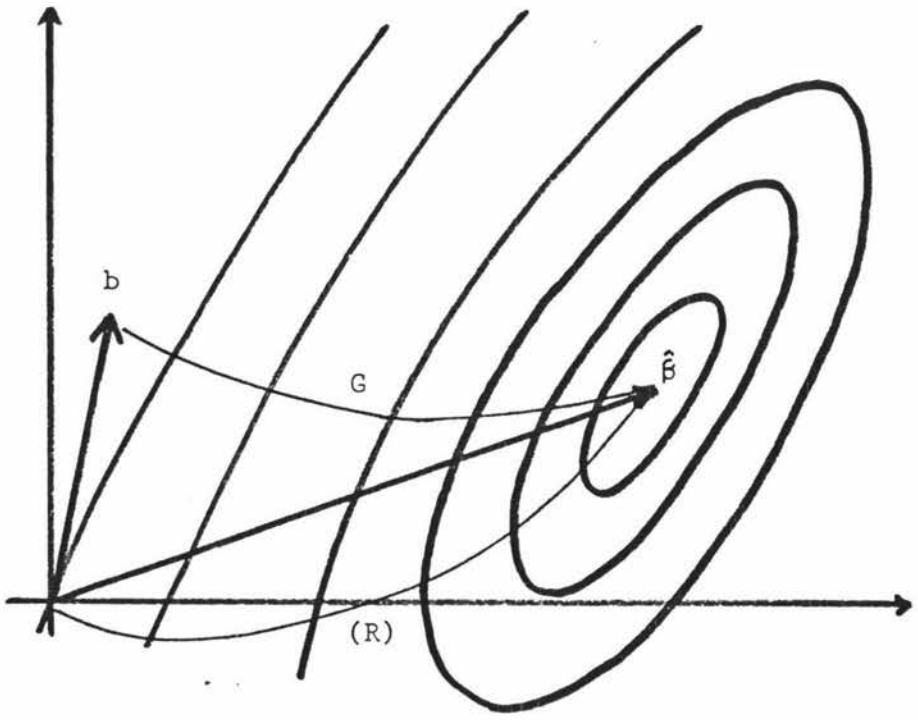

Figure 4.10 The ridge locus (R) and the good ridge locus (G).

4.4 Other Biased Estimation Procedures.

The discussion in this chapter has been limited to biased estimators which are linear in y. The main reason for this limitation is, as outlined in section 3.4, that the higher order moments of the error distribution and the form of the error distribution are generally unknown. If the form of the error distribution is known or its higher order moments are known then various biased nonlinear estimators may be shown to outperform the least squares estimator and the various biased linear estimators outlined in the preceding sections. One such biased, nonlinear, stochastically shrunken, estimator is the James-Stein estimator. In, for example, Stein (1966) it was shown that in the following situation,

$$\begin{aligned} x_{ij} &\sim N(\theta_i, \sigma^2) && i=1,\ldots,p \\ & && j=1,\ldots,n \\ y_i &\sim N(\theta_i, \frac{\sigma^2}{n}) && i=1,\ldots,p \end{aligned}$$

where $y_i = \sum_{j=1}^{n} x_{ij}/n$ or $y \sim N(\theta, \frac{\sigma^2}{n} I)$ so that y is the usual least squares estimator of θ;

$$E\|(1 - \frac{p-2}{n+2} \frac{S}{\|y\|^2})y - \theta\|^2 < E\|y-\theta\|^2 , \quad \forall\theta$$

when $p \geq 3$ and where S is distributed independently of y as $(\sigma^2/n)\chi^2_n$.

Thus with respect to the mean square error criterion (3.45) the estimator

$$\left(1 - \frac{p-2}{n+2} \frac{S}{\|y\|^2}\right) y$$

dominates the usual least squares estimator y for all values of the unknown θ when $p \geq 3$. This result has been generalized to the usual regression situation so that Mayer and Willke (1973), for example, quote a result of Sclove's that the estimator,

$$\left[1 + \frac{p-2}{(n-p+2)} \frac{RSS(\hat{\beta})}{\|\hat{\beta}\|^2}\right] \hat{\beta} \quad , \qquad p \geq 3$$

dominates the least squares estimator with respect to the criterion,

$$\min_B E\|X(B-\beta)\|^2 \quad .$$

There are many variants of the James-Stein estimator of β in the linear model situation. That there are many variants is probably one reason why the estimator has not gained the popularity of the ridge estimator. Also, very little work has been published on the robustness of these estimators. It would be useful to know how rapidly the various James-Stein estimators lose their optimality, if indeed they do when the error distribution is far from being normal.

A class of biased estimation procedures similar to the principal components regression estimator have been proposed by Gunst, Webster and Mason (1976). These latent root regression estimators result from an eigenvalue analysis of the matrix $A'A$ where,

$$A = [y \mid X]$$

The interested reader is referred to the article cited above and Hawkins (1975).

5. A DOUBLY RIDGED ESTIMATOR

5.1 Two-Parameter Ridge Estimators

Goldstein and Smith (1974) and Obenchain (1975a) have suggested that an examination of two-parameter generalised ridge estimators might produce worthwhile estimation procedures. The two-parameter ridge estimators have the form,

$$\alpha_i^*(k,q) = \left(\frac{\lambda_i}{\lambda_i+k\lambda_i^q}\right)\alpha_i + \frac{z_i'\epsilon}{\lambda_i+k\lambda_i^q}$$

$$= z_i'y \Big/ (\lambda_i+k\lambda_i^q) \qquad (5.1)$$

when the linear model is in canonical form, equation (3.11). Two familiar special cases of the estimator in equation (5.1) are the estimators obtained when q=+1 and q=0, the deterministically shrunken estimators of Mayer and Willke (1973), equation (4.30), and the simple ridge estimators of Hoerl and Kennard (1970a,b), equation (4.27), respectively. Goldstein and Smith suggested 1-q should be an integer $m \geq 0$ so that their formulation of equation (5.1) looked like,

$$\alpha_i^*(k,m) = \left(\frac{\lambda_i^m}{\lambda_i^m+k}\right)\alpha_i + \frac{\lambda_i^{m-1}z_i'\epsilon}{\lambda_i^m+k}$$

$$= \lambda_i^{m-1}z_i'y \Big/ (\lambda_i^m+k) \qquad (5.2)$$

where $0 \leq k \leq 1$ and $m \in \{0,1,2,\ldots\}$. Rewriting the estimator (5.2) in terms of the original model gives,

$$\beta^*(k,m) = [(X'X)^m + kI]^{-1}(X'X)^{m-1}X'y$$

$$= [X'X + k\{(X'X)^{-1}\}^{m-1}]^{-1}X'y \qquad (5.3)$$

Goldstein and Smith claimed that such an estimator is "more sensitive to variation in the eigenvalue spectrum" of $X'X$ for $m \geq 2$ (a comparison of equation (5.2) with equation (4.27) should satisfy the reader that this is so). One large disadvantage with this estimator is arriving at a choice of m. Opponents of ridge-type estimation who justifiably baulk at the task of choosing k would probably justifiably wilt when faced with the task of divining m. A suggestion may be to fix m quite arbitrarily at the value two since the estimators for m=0 and m=1 are already well established. Then fixing m=2 gives,

$$\tilde{\beta} = (X'X + k(X'X)^{-1})^{-1}X'y \ . \qquad (5.4)$$

Using the estimator (5.4) requires the evaluation of $(X'X)^{-1}$ which may pose a numerical problem if $X'X$ is ill-conditioned (which it almost certainly is as multicollinearity provides the motivation for considering these estimators). Evaluation of $(X'X+kI)^{-1}$ for some $k \in (0,1]$ may be less of a problem numerically so that possible alternative estimators to (5.4) could be,

$$^{*}\beta = [X'X + k(X'X+kI)^{-1}]^{-1}X'y \qquad (5.5)$$

$$\beta^{**} = [X'X + kI + k(X'X+kI)^{-1}]^{-1}X'y \ . \qquad (5.6)$$

The estimator in equation (5.6) may be described as a <u>doubly ridged estimator</u>. Equation (5.6) has been proposed by Rutishauser (1968) in a slightly different context. Rutishauser has called equation (5.6) the doubly relaxed least square solution of,

$$XB = Y.$$

Recall that the ridge estimator, equation (4.11) arises as the relaxed least squares solution of the same equation, and was proposed by Riley (1955) for solving ill-conditioned systems of linear equations. Similarly, the doubly relaxed least squares solution has been proposed by Rutishauser as a method for solving ill-conditioned systems of linear equations. Rutishauser favoured (5.6) doubly relaxed least squares over (4.11) relaxed least squares as a method for keeping cancellation under control in the evaluation of XB. The equation (5.6) may therefore be

worthy of investigation and application in the linear statistical models area.

5.2 Properties of the Doubly Ridged Estimator

Some properties of the doubly ridged estimator are listed below. The close similarity between the ridge estimator and this estimator is readily perceived (the doubly ridged estimator is a generalized ridge estimator).

(i) The doubly ridged estimator is a linear transformation of the least squares estimator,

$$\beta^{**} = [I+k(X'X)^{-1}(I+(X'X+kI)^{-1})]^{-1}\hat{\beta}$$

(ii) The doubly ridged estimator can be regarded as a constrained least squares estimator resulting from the minimization with respect to B of the objective function,

$$\|y-XB\|^2 + k\|B\|^2 + kB'(X'X+kI)^{-1}B$$

(iii) The doubly ridged estimator is "shorter in length" than the least squares estimator, that is,

$$\|\beta^{**}\| < \|\hat{\beta}\| \qquad \text{for all } k > 0.$$

The proof of this is readily obtained in the manner of Riley (1955).

(iv) For the same choice of $k > 0$,

$$\|\beta^{**}\| < \|\beta^{*}\|$$

(v) In the canonical form of the usual linear model, the doubly ridged estimator of α_i has the form,

$$\alpha_i^{**} = \left(\frac{\lambda_i}{\lambda_i+k+k/(\lambda_i+k)}\right)\alpha_i + \frac{z_i'e}{\lambda_i+k+k/(\lambda_i+k)}$$

Comparing this expression with equation (4.27) it can be seen that the doubly ridged estimator takes a little more account of the eigenvalue spectrum of $X'X$ than does the ridge estimator, in the reweighting of α_i and the error vector, with the addition of the term $k/(\lambda_i+k)$ in the denominator.

(vi) An important result concerning the generalized mean square error admissability of the doubly ridged estimator over the usual least squares estimator arises out of some work by Theobald (1974) on the admissability of the ridge estimator and extensions of Theobald's work by Farebrother (1976) and Gunst and Mason (1976). Theobald (1974) established the general result that for any two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ of a parameter vector θ the following conditions are equivalent:

(a) M_1-M_2 is non-negative definite

(b) $m_1-m_2 \geq 0$ for all non-negative definite H.

where,

$$M_j = E(\hat{\theta}_j-\theta)(\hat{\theta}_j-\theta)' \quad , \quad j=1,2$$

and where,

$$m_j = E(\hat{\theta}_j-\theta)'H(\hat{\theta}_j-\theta) \quad , \quad j=1,2$$

This means that an estimator $\hat{\theta}_2$ can be considered better in generalized mean square error than an estimator $\hat{\theta}_1$ if and only if the difference of their second order moment matrices M_1-M_2 is positive definite. Theobald applied this result to the ridge and least squares estimators of β in the linear model (3.1) and found that a sufficient condition for ridge estimation to outperform least squares estimation with respect to generalized mean square error was $0 < k \leq 2\sigma^2/\beta'\beta$. A slight generalization of Theobald's result for the ridge estimator which ties together the results of Farebrother (1976), Theobald (1974) and to some extent Lowerre (1974) is as follows:

Let $C\hat{\beta}$ be a linear transformation of the least squares estimator $\hat{\beta}$ such that C is non-negative definite, I-C is positive definite (this ensures that C shrinks all the components of the least squares estimates of the parameter vector in the canonical form of the model), and that C and $X'X$ commute. If it is also assumed that the matrix X has full column rank then a theorem may be proved.

Theorem: $M(\hat{\beta}) - M(C\hat{\beta})$ is positive definite if and only if,

$$\beta'\{2(I-C)^{-1}C(X'X)^{-1} + (X'X)^{-1}\}^{-1}\beta \quad < \sigma^2$$

Proof:

$$M(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

$$M(C\hat{\beta}) = E(C\hat{\beta}-\beta)(C\hat{\beta}-\beta)'$$

$$= \sigma^2C(X'X)^{-1}C + (C-I)\beta\beta'(C-I)$$

Thus

$$M(\hat{\beta}) - M(C\hat{\beta}) = (I-C)[\sigma^2\{2(I-C)^{-1}C(X'X)^{-1} + (X'X)^{-1}\} - \beta\beta'](I-C)$$

which is positive definite if and only if,

$$\sigma^2\{2(I-C)^{-1}C(X'X)^{-1} + (X'X)^{-1}\} - \beta\beta' \tag{5.7}$$

is positive definite. Now $\{2(I-C)^{-1}C(X'X)^{-1}+(X'X)^{-1}\}$ is positive definite, so that $M(\hat{\beta})-M(C\hat{\beta})$ is positive definite if and only if,

$$\beta'\{2(I-C)^{-1}C(X'X)^{-1} + (X'X)^{-1}\}^{-1}\beta \; < \; \sigma^2 \; . \tag{5.8}$$

This result means that $C\hat{\beta}$ outperforms $\hat{\beta}$ with respect to any generalized mean square error criterion if and only if equation (5.8) holds. Various sufficient conditions are available. Since $(X'X)^{-1}$ is positive definite a sufficient condition for (5.7) to be positive definite is,

$$\sigma^2\{2(I-C)^{-1}C(X'X)^{-1}\} - \beta\beta' \tag{5.9}$$

is non-negative definite, or,

$$\beta'\{2(I-C)^{-1}C(X'X)^{-1}\}^{-1}\beta \; \leq \; \sigma^2.$$

The matrix $2\sigma^2(I-C)^{-1}C(X'X)^{-1}$ is non-negative definite so that a sufficient condition for (5.7) to be positive definite is,

$$\sigma^2(X'X)^{-1} - \beta\beta' \tag{5.10}$$

is positive definite, or,

$$\beta'X'X\beta < \sigma^2$$

Applying these results to the doubly ridged estimator recall from property (i) that,

$$C = [I + k(X'X)^{-1}(I + (X'X + kI)^{-1})]^{-1}$$

so that C is positive definite, commutes with $X'X$, and $(I-C)$ is positive definite for $k > 0$. Thus by the theorem the doubly ridged estimator β^{**} outperforms the least squares estimator $\hat{\beta}$ with respect to any generalized mean square error criterion if and only if,

$$\beta'\{\frac{2}{k}[I + (X'X+kI)^{-1}]^{-1} + (X'X)^{-1}\}^{-1}\beta < \sigma^2.$$

The sufficient condition (5.9) becomes,

$$\beta'\{\frac{2}{k}[I + (X'X+kI)^{-1}]^{-1}\}^{-1}\beta \leq \sigma^2$$

A further sufficient condition for this is,

$$k\left\{1 + \frac{1}{\lambda_{min}+k}\right\} < \frac{2\sigma^2}{\beta'\beta} \quad .$$

For the simple ridge estimator the sufficient condition (5.9) becomes,

$$k < \frac{2\sigma^2}{\beta'\beta}$$

Comparing these admittedly sufficient conditions it can be seen that the choice of k for the doubly ridged estimator is more closely linked to the conditioning or eigenvalue structure of $X'X$ than for the ridge estimator. Of course, the choice of k in the doubly ridged procedure still depends as in the ridge estimation procedure on the unknowns σ^2 and β. Thus the doubly ridged estimator is open to the same kinds of objections as the ridge estimator (recall the criticisms presented in section 4.22). The problem of choosing a k for the doubly ridged estimation procedure could be approached in the various algorithmic, iterative ways currently being evaluated for ridge estimators (see, for example, Dempster, Schatzoff and Wermuth (1977) and the discussion following their article by Hoerl).

Further comparison of the doubly ridged estimator and the ridge estimator is possible. Let,

$$\beta^{*} = C_1\hat{\beta} \quad , \text{ where } C_1 \text{ depends on the constant}$$

k_1, and let,

$$\beta^{**} = C_2\hat{\beta} \quad , \text{ where } C_2 \text{ depends on the constant } k_2.$$

Then,

$$M(\beta^{*}) - M(\beta^{**}) = \sigma^2(C_1+C_2)(X'X)^{-1}(C_1-C_2)$$
$$+ (C_1-I)\beta\beta'(C_1-I)$$
$$- (C_2-I)\beta\beta'(C_2-I)$$

and a sufficient condition for this matrix to be positive definite is that the matrix (C_1-C_2) is positive definite. A sufficient condition for (C_1-C_2) to be positive definite is,

$$k_1 < k_2 + \frac{k_2}{\lambda_{max}+k_2} \quad .$$

Thus sufficient conditions for the doubly ridged estimator β^{**} to outperform both the ridge estimator β^{*} and the least squares estimator $\hat{\beta}$ with respect to the generalized mean square error criterion are,

$$k_1 < k_2 + \frac{k_2}{\lambda_{max}+k_2} \quad \text{and} \quad k_2 + \frac{k_2}{\lambda_{min}+k_2} < \frac{2\sigma^2}{\beta'\beta} \quad .$$

when the ridge estimator also outperforms the least squares estimator.

The doubly ridged estimator is probably worth investigating further by way of simulation studies. An application of the estimator is presented in Chapter 6.

6. AN APPLICATION

6.1 The Longley Problem

Longley (1967) used a particular set of data to investigate the accuracy of several different multiple regression programs on several different computers. The data consisted of 16 observations on the variables;

- y, Total derived employment in the U.S.A. (in thousands)
- x_1, Gross National Product Implicit Price Deflator (in tenths)
- x_2, Gross National Product (in millions)
- x_3, Unemployment (in thousands)
- x_4, Size of Armed Forces (in thousands)
- x_5, Noninstitutional Population 14 years and over (in thousands)
- x_6, Year.

Longley's data, which can be found in either Longley (1967) or Beaton, Rubin and Barone (1976), features a highly ill-conditioned input matrix and has become quite popular as a test problem for illustrating the application of new techniques. Obenchain (1975b) has used Longley's data to illustrate the use of several solution selection criteria, Beaton, Rubin and Barone (1976) have used the data to illustrate the use of their perturbation index (see section 3.22), and Cook (1977) has used the data to illustrate the use of his detector of influential observations in linear regression (see section 3.23). It was, therefore, decided to use the Longley problem to compare the performance of the ridge and doubly ridged estimators. The OMNITAB II, VERSION 5.00 computer package (see Hogben, Peavy and Varner (1971)), as implemented on the Massey University Burroughs B6700, was used for this purpose. The variables were rescaled so that $X'X$ was in correlation form and so that the vector $X'y$ was composed of the correlations between the dependent variable y and the variables x_i. The resulting matrix X was then transformed so that the model was in canonical form and a principal component analysis was carried out on the matrix of input variables (see Table 6.1). The first principal component accounts for about 77% of the variation in the input

Table 6.1 Principal Component Analysis of the Longley Data in Correlation Form

		Eigenvector Coefficients					
i	Eigenvalues λ_i	x_1	x_2	x_3	x_4	x_5	x_6
1	4.6 0 3 3 8	.461835	.461504	.321317	.201510	.462279	.464940
2	1.1 7 5 3 4	-.0578428	-.0532123	.595514	-.798193	.0455445	-.000618788
3	.2 0 3 4 2 5	-.149120	-.277682	.728306	.561608	-.195985	-.128116
4	.0 1 4 9 2 8 3	-.792874	.121621	-.00764580	.0772550	.589745	.0522866
5	.0 0 2 5 5 2 0 7	.337938	-.149573	.00923196	.0242525	.548578	-.749543
6	.0 0 0 3 7 6 7 0 8	.135187	-.818481	-.107453	-.0179710	.311571	.450409
	6.0 0 0 0 0						

matrix, so that it appears that most of the observations on the six independent variables are arranged in a single direction, or arranged about a straight line in the estimation space. This is one indication of the multicollinearity in X. The first principal component seems to be a general weighted average of the six input variables. The weights are almost in the same proportions as the components of $X'y$ since,

$$X'y = \begin{bmatrix} .9709 \\ .9836 \\ .5025 \\ .4573 \\ .9604 \\ .9713 \end{bmatrix} .$$

The second principal component accounts for 20% of the variation in X and is dominated by the variables x_3 and x_4, unemployment and size of armed forces respectively. This component might be interpreted as a measure of the level of non-paid, non-productive employment (the unemployed do not contribute to the GNP and are virtually unpaid while the armed forces, who do not contribute to the nation's production, are well paid - hence the negative sign). Such interpretations, like many in principal components analysis, are highly speculative, especially when it is remembered that principal components are not invariant to changes in scale. The transformation to principal components does however illuminate the extent of the ill conditioning of $X'X$; λ_4, λ_5 and λ_6 are all substantially less than one. Beaton et al (1976) comment that "most statisticians would advise a client not to fit a model" with such a conditioning problem. This is probably good advice but statisticians could at least point out to clients that least squares prediction in the direction of the first two or three principal components might nevertheless be quite precise (see section 3.131).

In the present case a least squares fit was carried out for the full model in correlation form. The least squares estimates of the parameters are shown in the first row of Table 6.2 and Table 6.3. The value of R^2 was 0.9955 indicating a very close fit. Several ridge regressions and doubly ridged regressions were performed

Table 6.2 Ridge Estimates for $X'X$ in Correlation Form

k	β_1	β_2	β_3	β_4	β_5	β_6	RSS
0.0	.04628	-1.0137	-.5375	-.2047	-.1012	2.4797	.004521
0.0001815	-.002501	- .5035	-.4669	-.1900	-.2375	2.1019	.004712
0.02	.2554	.3390	-.2795	-.1027	.1714	.4314	.01267
0.04	.2651	.3298	-.2474	-.07848	.2177	.3481	.01464
0.06	.2638	.3204	-.2216	-.05989	.2313	.3152	.01648
0.08	.2606	.3119	-.1997	-.04460	.2359	.2965	.01837
0.10	.2571	.3042	-.1808	-.03170	.2371	.2838	.02030
0.20	.2412	.2758	-.1146	.0110	.2303	.2508	.02973
0.30	.2294	.2568	-.07476	.03426	.2211	.2340	.03818
0.40	.2202	.2429	-.04816	.04819	.2131	.2226	.04579
0.50	.2125	.2320	-.02924	.05698	.2061	.2138	.05286

Table 6.3 Doubly Ridged Estimates for $X'X$ in Correlation Form

k	β_1	β_2	β_3	β_4	β_5	β_6	RSS
0.0	.04628	-1.0137	-.5357	-.2047	-.1012	2.4797	.004521
0.001955	.2927	.3501	-.2884	-.1042	.2870	.2789	.01407
0.02	.2628	.3058	-.1729	-.01803	.2635	.2542	.02101
0.04	.2492	.2824	-.1134	.02402	.2475	.2426	.02847
0.06	.2412	.2687	-.07974	.04647	.2379	.2357	.03391
0.08	.2357	.2595	-.05780	.06007	.2315	.2308	.03796
0.10	.2314	.2527	-.04225	.06898	.2267	.2270	.04114
0.20	.2183	.2332	-.002670	.08664	.2126	.2153	.05120
0.30	.2098	.2220	.01478	.09038	.2043	.2075	.05806
0.40	.2031	.2136	.02493	.09058	.1977	.2012	.06424
0.50	.1972	.2066	.03160	.08955	.1921	.1956	.07032

for values of k ranging from 0 to 0.6 in steps of 0.02. All calculations were carried out with the canonical form of the correlation form of the model and the resulting estimates of α transformed to estimates of β in the correlation form of the model by premultiplying the vector of estimates of α by P', the transpose of the matrix in the last six columns of Table 6.1. The ridge estimation results are summarized in Table 6.2 and the doubly ridged estimation results are summarized in Table 6.3.

A brief inspection of Table 6.2 and Table 6.3 indicates that the doubly ridged procedure shrinks the least squares estimates faster than the ridge procedure for the same values of k. Correspondingly the residual sum of squares for the doubly ridged estimator increases faster than the residual sum of squares for the ridge estimation procedure, for the same increasing values of k. There does not seem to be any great difference in pattern of shrinkage for both procedures in this example.

6.2 Solution Selection

Several methods for choosing a value of the biasing parameter k have been proposed in the literature (see, for example, Obenchain (1975a,b) and McDonald (1975)). In the present case visual inspection of ridge and doubly ridged traces suggested values of $0.1 < k < 0.2$ and $0.06 < k < 0.16$ respectively.

An iterative procedure for choosing k based upon,

$$k_{(t+1)} = \frac{2\hat{\sigma}^2}{\beta^{*\prime}_{(t)}\beta^{*}_{(t)}}$$

where $\hat{\sigma}^2$ is the usual estimate of σ^2 from the least squares fit and $\beta^{*}_{(0)} = \hat{\beta}$, converged in eight steps to a four-significant-figures value of k of 0.0001815 for ridge estimation of β (the corresponding estimate of β is shown in Table 6.2). A similar iterative procedure for choosing k in the doubly ridged regression

converged in five steps to a four-significant-figures value of k of 0.001955 (the corresponding estimate of β is shown in Table 6.3). Whether or not these choices of k have produced an estimate of β with smaller mean square error than the least squares estimator $\hat{\beta}$ is essentially unknowable. Simulation studies in the manner of Hoerl and Kennard (1976) and Dempster, Schatzoff and Wermuth (1977) may indicate whether or not the doubly ridged procedure possesses any great advantage over the simple ridge estimator.

7. SUMMARY

The utility of the method of least squares has been subject to review. When the conditions of the Gauss-Markov Theorem obtain the method produces the MVLUE of β. If the noise or disturbance component of the dependent variable is normally distributed the method produces the MVUE of β. When the conditions of the Gauss-Markov Theorem do not obtain the method loses its minimum variance property and may in the case of the misspecified model and the errors in variables situation produce a badly biased estimator of β. This is particularly so when the matrix $X'X$ is ill-conditioned. Ill-conditioning of $X'X$ or multi-collinearity in X therefore provides a strong motive for searching for alternatives to least squares in the wider class of biased, non-linear estimators. One sub-class of this class of alternatives - linear transformations of the least squares estimator - has been reviewed. Several areas worthy of investigation have been proposed in the course of this review, namely,

(i) A comparative study of robust generalized ridge estimators possibly in the form of a simulation study.

(ii) A study of the robustness of the James-Stein estimator and the possibility of constructing robust variants of it.

(iii) An investigation of two parameter ridge-type estimators for fixed values of the second parameter.

(iv) A simulation study to discover the merits, if any, of the doubly ridged estimator over the ordinary ridge estimator.

(v) A wider search for non-linear estimators of β with good mean square error properties. The MMSELE has been used as a starting point so far.

The method of least squares is still a useful estimation procedure but as Tukey (1975) has pointed out the least squares procedure has to be modified when the conditions usually assumed in its application are not met. In these situations "least squares embedded in modification processes" is called for. The robust weighted least squares estimators and ridge and shrunken estimators are only two classes of such modifications.

BIBLIOGRAPHY

Andrews, D.F. (1974) A robust method for multiple linear regression. Technometrics, 16, 523-531.

Andrews, D.F. (1975) Alternative calculations for regression and analysis of variance problems. Proceedings of the Conference at Dalhousie University, Halifax, May 1974, in R.P. Gupta, ed., Applied Statistics, New York, American Elsevier Publishing Company, 1-7.

Banerjee, K.S. and Carr, R.N. (1971) A comment on ridge regression. Biased estimation for non-orthogonal problems. Technometrics, 13, 895-898.

Barnard, G.A. (1963) The logic of least squares. Journal of the Royal Statistical Society, Series B, 25, 124-127.

Beaton, A.E., Rubin, D.R. and Barone, J.L. (1976) The acceptability of regression solutions: Another look at computational accuracy. Journal of the American Statistical Association, 71, 158-168.

Beaton, A.E. and Tukey, J.W. (1974) The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. Technometrics, 16, 147-185.

Bloomfield, P. and Watson, G.S. (1975) The inefficiency of least squares. Biometrika, 62, 121-128.

Box, G.E.P. (1966) Use and abuse of regression. Technometrics, 8, 625-629.

Box, G.E.P. and Cox, D.R. (1964) An analysis of transformations. Journal of the Royal Statistical Society, Series B, 26, 211-243.

Cochran, W.G. (1972) Some effects of errors of measurement on linear regression. Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley and Los Angeles, University of California Press, 1, 527-539.

Conniffe, D. and Stone, J. (1973) A critical view of ridge regression. The Statistician, 22, 181-187.

Conniffe, D. and Stone, J. (1975) A reply to Smith and Goldstein. The Statistician, 24, 67-68.

Cook, R.D. (1977) Detection of influential observation in linear regression. Technometrics, 19, 15-18.

Davies, R.B. and Hutton, B. (1975) The effect of errors in the independent variables in linear regression. Biometrika, 62, 383-391.

Dempster, A.P. (1973) Alternatives to least squares in multiple regression. Proceedings of the Research Seminar at Dalhousie University, Halifax, March 23-25, 1972, in D.G. Kabe and R.P. Gupta, eds., Multivariate Statistical Inference, New York, American Elsevier Publishing Company, 25-40.

Dempster, A.P., Schatzoff, M. and Wermuth, N. (1977) A simulation study of alternatives to ordinary least squares. Journal of the American Statistical Association, 72, 77-106.

Draper, N.R. and Smith, H. (1966) Applied Regression Analysis, New York, John Wiley and Sons.

Efron, B. (1975) Biased versus unbiased estimation. Advances in Mathematics, 16, 259-277.

Farebrother, R.W. (1975) The minimum mean square error linear estimator and ridge regression. Technometrics, 17, 127-128.

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

Goldstein, M. (1976) Bayesian analysis of regression problems. Biometrika, 63, 51-58.

Goldstein, M. and Smith, A.F.M. (1974) Ridge-type estimators for regression analysis. Journal of the Royal Statistical Society, Series B, 36, 284-291.

Greenberg, E. (1975) Minimum variance properties of principal component regression. Journal of the American Statistical Association, 70, 194-197.

Guilkey, D.K. and Murphy, J.L. (1975) Directed ridge regression techniques in cases of multicollinearity. Journal of the American Statistical Association, 70, 769-775.

Gunst, R.F. and Mason, R.L. (1976) Generalized mean squared error properties of regression estimators. Communications in Statistics - Theory and Methods, A5, 1501-1508.

Gunst, R.F., Webster, J.T. and Mason, R.L. (1976) A comparison of least squares and latent root regression estimators. Technometrics, 18, 75-83.

Hawkins, D.M. (1975) Relations between ridge regression and eigenanalysis of the augmented correlation matrix. Technometrics, 17, 477-480.

Hocking, R.R. (1976) The analysis and selection of variables in linear regression. Biometrics, 32, 1-49.

Hocking, R.R., Speed, F.M. and Lynn, M.J. (1976) A class of biased estimators in linear regression. Technometrics, 18, 425-437.

Hodges, S.D. and Moore, P.G. (1972) Data uncertainties and least squares regression. Applied Statistics, 21, 185-195.

Hoerl, A.E. and Kennard, R.W. (1970a) Ridge regression: Biased estimation for non-orthogonal problems. Technometrics, 12, 55-67.

Hoerl, A.E. and Kennard, R.W. (1970b) Ridge regression: Applications to non-orthogonal problems. Technometrics, 12, 69-82.

Hoerl, A.E. and Kennard, R.W. (1976) Ridge regression: Iterative estimation of the biasing parameter. Communications in Statistics - Theory and Methods, A5, 77-88.

Hoerl, A.E., Kennard, R.W. and Baldwin, K.F. (1975) Ridge regression: Some simulations. Communications in Statistics, 4, 105-123.

Hogben, D., Peavy, S.T. and Varner, R.N. (1971) Omnitab II: User's reference manual. National Bureau of Standards Technical Note 552, Washington D.C., U.S. Government Printing Office.

Holland, P.W. (1973) Weighted ridge regression: Combining ridge and robust regression methods. National Bureau of Economic Research Working Paper No. 11, Cambridge, Massachusetts, National Bureau of Economic Research.

Horn, S.D. and Horn, R.A. (1975) Comparison of estimators of heteroscedastic variances in linear models. Journal of the American Statistical Association, 70, 872-879.

Horn, S.D., Horn, R.A. and Duncan, D.B. (1975) Estimating heteroscedastic variances in linear models. Journal of the American Statistical Association, 70, 380-385.

Kendall, M.G. (1957) A Course in Multivariate Analysis, London, Charles Griffin and Company Limited.

Kendall, M.G. (1975) Multivariate Analysis, London, Charles Griffin and Company Limited.

Knott, M. (1975) On the minimum efficiency of least squares. Biometrika, 62, 129-132.

Lawless, J.F. and Wang, P. (1976) A simulation study of ridge and other regression estimators. Communications in Statistics - Theory and Methods, A5, 307-323.

Lawson, C.L. and Hanson, R.J. (1974) Solving Least Squares Problems, Englewood Cliffs, New Jersey, Prentice-Hall.

Lindgren, B.W. (1968) Statistical Theory, 2nd ed., New York, MacMillan Publishing Company.

Lindley, D.V. and Smith, A.F.M. (1972) Bayes estimates for the linear model. Journal of the Royal Statistical Society, Series B, 34, 1-41.

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.

Lowerre, J.M. (1974) On the mean square error of parameter estimates for some biased estimators. Technometrics, 16, 461-464.

Marquardt, D.W. (1970) Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. Technometrics, 12, 591-612.

Marquardt, D.W. (1974) Discussion No. 2. Technometrics, 16, 189-192.

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

Massy, W.F. (1965) Principal components regression in exploratory statistical research. Journal of the American Statistical Association, 60, 234-256.

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

McDonald, G.C. (1975) Discussion of: Ridge analysis following a preliminary test of the shrunken hypothesis. Technometrics, 17, 443-445.

Newhouse, J.P. and Oman, S.D. (1971) An evaluation of ridge estimators. United States Air Force Project Rand Report R-716-PR, Santa Monica, The Rand Corporation.

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

Obenchain, R.L. (1975b) The associated probability of a ridge estimator. Unpublished manuscript.

Plackett, R.L. (1949) A historical note on the method of least squares. Biometrika, 36, 458-460.

Rao, C.R. (1971) Unified theory of linear estimation. Sankhyā, Series A, 33, 371-394.

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

Rao, C.R. (1975) Some thoughts on regression and prediction, Part 1. Sankhyā, Series C, 37, 102-120.

Rao, C.R. (1976) Estimation of parameters in a linear model. The Annals of Statistics, 4, 1023-1037.

Riley, J.D. (1955) Solving systems of linear equations with a positive definite, symmetric, but possibly ill-conditioned matrix. Mathematical Tables and other Aids to Computation, 9, 96-101.

Rutishauser, H. (1968) Once again: The least square problem. Linear Algebra and Its Applications, 1, 479-488.

Scheffé, H. (1959) The Analysis of Variance, New York, John Wiley and Sons.

Searle, S.R. (1971) Linear Models, New York, John Wiley and Sons.

Silvey, S.D. (1969) Multicollinearity and imprecise estimation. Journal of the Royal Statistical Society, Series B, 31, 539-552.

Smith, A.F.M. and Goldstein, M. (1975) Ridge regression: Some comments on a paper of Conniffe and Stone. The Statistician, 24, 61-66.

Snee, R.D. (1973) Some aspects of nonorthogonal data analysis. Part 1. Developing prediction equations. Journal of Quality Technology, 5, 67-79.

Styan, G.P.H. (1973) When does least squares give the best linear unbiased estimate? Proceedings of the Research Seminar at Dalhousie University, Halifax, March 23-25, 1972, in D.G. Kabe and R.P. Gupta, eds., Multivariate Statistical Inference, New York, American Elsevier Publishing Company, 241-246.

Swindel, B.F. (1975) Good ridge estimators based on prior information (Abstract only). Biometrics, 31, 602-603.

Swindel, B.F. (1976) Good ridge estimators based on prior information. Communications in Statistics - Theory and Methods, A5, 1065-1075.

Swindel, B.F. and Bower, D.R. (1972) Rounding errors in the independent variables in a general linear model. Technometrics, 14, 215-218.

Theil, H. (1971) Principles of Econometrics, New York, John Wiley and Sons.

Theobald, C.M. (1974) Generalizations of mean square error applied to ridge regression. Journal of the Royal Statistical Society, Series B, 36, 103-106.

Tukey, J.W. (1948) Approximate weights. Annals of Mathematical Statistics, 19, 91-92.

Tukey, J.W. (1975) Instead of Gauss-Markov least squares, what? Proceedings of the Conference at Dalhousie University, Halifax, May 1974, in R.P. Gupta, ed., Applied Statistics, New York, American Elsevier Publishing Company, 351-372.

Vinod, H.D. (1976a) Application of new ridge regression methods to a study of Bell system scale economies. Journal of the American Statistical Association, 71, 835-841.

Vinod, H.D. (1976b) Simulation and extension of a minimum mean squared error estimator in comparison with Stein's. Technometrics, 18, 491-496.

Warren, R.D., White, J.K. and Fuller, W.A. (1974) An errors-in-variables analysis of managerial role performance. Journal of the American Statistical Association, 69, 886-893.

Watson, G.S. (1967) Linear least squares regression. Annals of Mathematical Statistics, 38, 1679-1699.