University of Cape Town
Department of Mathematical Statistics

COLLINEARITY AND CONSEQUENCES FOR ESTIMATION: A STUDY AND SIMULATION

by

Christien Thiart

A thesis prepared under the supervision of

Dr. T.T. Dunne

in fulfilment of the requirements for the degree of
Master of Science in Mathematical Statistics

CONTENTS

5 FRACTIONAL PRINCIPALS IN ESTIMATION

6 LATENT ROOT REGRESSION ESTIMATION

7 TOTAL LEAST SQUARES

ACKNOVLEDGEIENTS

I would like to thank my supervisor, Dr. Tim Dunne. Vithout his guidance, patience and expertise this thesis would not have been feasible. I record special thanks for correcting my grammar patiently week after week.

Then I would like to thank Prof. Troskie for his guidance in the writing of the simulation program. Derek Chalton has my gratitude for providing the data and verifying my calculations in the simulation study. Shirley Breed shared her expertise in T^3. The Jagger Library, UCT, gave friendliness and help with the literature search.

Finally I would like to thank the FRD for their financial support over the last two years and the staff of the Department of Mathematical Statistics who were always friendly and encouraging.

INTRODUCTION

Collinearity is a phenomenon that occurs in linear statistical models. It cannot be described in simple terms as present or absent but rather in terms of degree and consequences. Consequences include on the one hand the issue of computational accuracy, and on the other hand, the statistically inherent instability of the estimators. The statistical and mathematical theory and background required to understand collinearity is presented in Chapter 1. Chapter 2 discusses collinearity *per se* and in Chapters 3 to 6 we draw from the literature to present various biased statistical estimation procedures and their properties. These procedures include principal components, ridge, generalized ridge, shrunken, fractional and latent root regression methods. In Chapter 7 some attempt is made to incorporate the errors-in-variables model in the discussion of collinearity, in as much as it also admits a perturbation framework. Chapter 8 consists of a summary of all the proposed estimators. In Chapter 9 the question of influence is discussed. Chapter 10 presents the results of a simulation study.

Some original comments on some estimation techniques are presented in the appropriate chapters.

An alphabetical listing of notation is presented in Appendix C for the convenience of the reader interested only in elements of the study.

An extensive bibliography has been completed from a literature search on the MATHSCI database and from additional sources found by the author.

In the published literature there is no study of the relative efficiencies of estimators as comprehensive as the study presented here. Lee (1986), in an unpublished doctoral thesis, apparently examined more biased estimators. This thesis borrows partly from the work of Chalton (1990) and explores the relative efficiencies of the estimators across different parameter-vector orientations, different common error variance sizes, and different collinearity severities within a simple and convenient collinearity framework. Some interesting and anomalous properties of estimators emerge.

Sources for sections of the simulation and estimation programs are acknowledged in Chapter 10, but the main program and subroutines written for this study are listed in Appendix B. Annotations highlight key changes in routines.

Avenues for further research are sketched in Chapters 4, 9 and 10.

Chapter 1

THE LINEAR REGRESSION MODEL

1.1 The Model

The linear regression model is given by

$$Y = X\beta + \epsilon \tag{1.1}$$

where Y is a nx1 observed response vector,

ϵ is a nx1 vector of uncorrelated random error variables with expectation $E(\epsilon) = 0$, and variance matrix $Var(\epsilon) = V(\epsilon) = \sigma^2 I$,

β is a px1 vector of regression coefficients that must be estimated, and

X is a nxp matrix of fixed regressors or independent variables, whose rank is p (we will assume that n>p).

We will not always assume that the X matrix has been standardized. If there is a constant present in the regression model we will assume that it is represented in the X matrix as a column of ones. If we want the X matrix to be scaled so that the product matrix X'X is in correlation form, that will be stated explicity.

By centering we imply that the mean of each regressor column is subtracted from the relevant column. By standardizing X we mean that X has been scaled so that the length of each column of X is unity (eg. $X_i'X_i = 1$, where X_i denotes the i-th column of X). When the columns have been scaled (centered and standardized) the product matrix X'X is in correlation form. In correlation form each of the elements of X'X will lie between -1 and +1. For example let p = 2, then the correlation matrix of X'X is

$$\begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}$$

where r_{12} is the observed coefficient of correlation between the variables represented by the first two columns of X, and

$$r_{12} = \frac{\Sigma(X_{i1} - \bar{x}_1)(X_{i2} - \bar{x}_2)}{[\Sigma(X_{i1} - \bar{x}_1)^2]^{\frac{1}{2}}[\Sigma(X_{i2} - \bar{x}_2)^2]^{\frac{1}{2}}}$$

$$= \frac{\Sigma X_{i1}X_{i2} - n\bar{x}_1\bar{x}_2}{[(\Sigma X_{i1}^2 - n\bar{x}_1^2)(\Sigma X_{i2}^2 - n\bar{x}_2^2)]^{\frac{1}{2}}}$$

where $\bar{x}_j$ is the mean of the j-th column, X_{ij} is the i-th element of the j-th column and the summation is from i = 1(1)n.

Consequences of data centering for collinearity diagnosis are presented in Chapter 2.

1.2 Ordinary Least Square estimation

If $\hat{\beta}$ is the ordinary least square estimator (OLSE) of β in (1.1), minimizing $(Y - X\beta)'(Y - X\beta)$ over all β, then

$$\hat{\beta} = (X'X)^{-1}X'Y \qquad (1.2.1)$$

and the minimum sum of squares of residuals is

$$RSS = (\hat{\epsilon}'\hat{\epsilon}) = (Y - X\hat{\beta})'(Y - X\hat{\beta}) = SSE(\hat{\beta}) \qquad (1.2.2)$$

Properties:

1. $E(\hat{\beta}) = \beta$ (unbiased)

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

3. MSE = $\sigma^2(X'X)^{-1}$

4. TMSE = $\sigma^2\mathrm{tr}[(X'X)^{-1}]$

5. $\hat{\beta}$ is the best linear unbaised estimator (BLUE) of β

6. Let L_1 = Euclidean distance from $\hat{\beta}$ to β then

$$L_1^2 = (\hat{\beta} - \beta)'(\hat{\beta} - \beta)$$

$$E(L_1^2) = \sigma^2\mathrm{tr}(X'X)^{-1}$$

$$E(\hat{\beta}'\hat{\beta}) = \beta'\beta + \sigma^2\mathrm{tr}(X'X)^{-1}$$

7. If ϵ is distributed normally then

$$V(L_1^2) = 2\sigma^4\mathrm{tr}[(X'X)^{-2}] \qquad \text{(from A.1)}$$

8. If the eigenvalues of X'X are denoted by

$\lambda_{max} = \lambda_1 \geq \lambda_2 \geq \ldots\ldots\ldots\ldots \geq \lambda_p = \lambda_{min} > 0$, then

$$E(L_1^2) = \sigma^2\mathrm{tr}[(X'X)^{-1}]$$

$$= \sigma^2\Sigma 1/\lambda_i \quad \text{(from A.2) and}$$

$$V(L_1^2) = 2\sigma^4\Sigma\ 1/\lambda_i^2$$

9. If we assume that the random error terms ϵ_i are Gaussian (normal) in distribution, the maximum likelihood estimators (MLE's) and the BLUE's coincide (Gauss-Markov Theorem, Searle (1971, p87))

1.3 Singular Value Decomposition

The singular value decomposition (SVD) of a matrix is discussed in textbooks such as Stewart (1973, p318), Golub and Van Loan (1983, p16) and Lawson and Hanson (1974, Chapter 4). These discussions can be summarized as follows:

Let X be an nxp matrix of rank k. Then there is an nxn orthogonal matrix U, an pxp orthogonal matrix V, and an nxp matrix Δ such that

$$U'XV = \Delta, \qquad X = U\Delta V'$$

where $\Delta:nxp = \begin{bmatrix} D\alpha & 0 \\ 0 & 0 \end{bmatrix}$

$$D_\alpha = \mathrm{Diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2} \ldots \sqrt{\lambda_k})$$

and $\sqrt{\lambda_1} \geq \sqrt{\lambda_2} \geq \ldots \sqrt{\lambda_k} > 0$

$\sqrt{\lambda_i}$ is called the i-th singular values of X, and λ_i is the i-th characteristic value (eigenvalue) of X′X. The vector columns of U are left singular vectors of X and the columns of V are right singular vectors of X.

Unless the converse is stated explicitly, we will assume that X has full column rank, i.e. the rank of X is p. Observe that we may also write the SVD (also called the basic structure of X) as

$$\begin{aligned} X &= U\Delta V' \quad (X:nxp,\ U:nxp,\ \Delta:pxp,\ V:pxp) \qquad (1.3.1) \\ \Delta &= D_\alpha \\ V &= [v_1 \ldots v_p], \quad v_i:px1 \\ U &= [u_1 \ldots u_p], \quad u_i:nx1 \\ V'V &= U'U = I_p \end{aligned}$$

Note that $V'V = I_p$ but that $UU' \neq I_n$. Using the SVD of X the following equations will be useful in the subsequent chapters.:

1. $$\begin{aligned} X &= [u_1 \ldots u_p]\ \mathrm{Diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2} \ldots \sqrt{\lambda_k})[v_1 \ldots v_p]' \\ &= \sum_{i=1}^{p} \sqrt{\lambda_i} u_i v_i' \qquad (1.3.2) \end{aligned}$$

2. $$\begin{aligned} X'X &= V\ \Delta\ U'U\ \Delta\ V' \\ &= V\ \Delta^2 V' \\ &= \sum_{i=1}^{p} \lambda_i v_i v_i' \qquad (1.3.3) \end{aligned}$$

3.
$$\begin{aligned}(X'X)^{-1} &= V\Delta^{-2}V' \\ &= \sum_{i=1}^{p} v_i v_i'/\lambda_i \qquad (1.3.4)\end{aligned}$$

4.
$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'Y \\ &= V\ \Delta^{-2}V'V\ \Delta\ U'Y \\ &= V\ \Delta^{-1}U'Y \qquad (1.3.5) \\ &= \sum_{i=1}^{p} v_i u_i'Y/\sqrt{\lambda}_i\end{aligned}$$

Let $c_i \ = \ u_i'Y\sqrt{\lambda}_i$ then

$$\hat{\beta} = \sum_{i=1}^{p} v_i\ c_i/\lambda_i \qquad (1.3.6)$$

5. $V(\hat{\beta}) = \sigma^2 \sum_{i=1}^{p} v_i v_i'/\lambda_i$ (from 1.3.4) $\qquad$ (1.3.7)

The SVD of the augmented matrix $[X\ Y]$ is:

$$[X\ Y] = \tilde{U}\ \tilde{\Delta}\ \tilde{V}' \qquad (1.3.8)$$

with
$$\begin{aligned}\tilde{U} &= [\tilde{u}_1 \ldots \tilde{u}_{p+1}],\ \tilde{u}_i : nx1 \\ \tilde{V} &= [\tilde{v}_1 \ldots \tilde{v}_{p+1}],\ \tilde{v}_i : (p+1)x1 \\ \tilde{\Delta} &= \mathrm{Diag}[\omega_1, \omega_2, \ldots \omega_{p+1}] \\ \omega_1 &\geq \omega_2 \geq \ldots\ldots \geq \omega_{p+1} \\ \tilde{U}'\tilde{U} &= \tilde{V}'\tilde{V} = I_{p+1} \qquad (1.3.9)\end{aligned}$$

The following notation will be used in discussing the augmented matrix:

1. Let $\tilde{v}_{i,j}$ be the j-th component of the i-th right singular vector $\tilde{v}_i$.

2. $\tilde{v}_i^0$ is the p-dimensional vector containing the first p components of the i-th right singular vector $\tilde{v}_i$ of $[X\ Y]$ of dimension p+1, thus

$$\tilde{v}_i = [\tilde{v}_i^{0\prime}\ \tilde{v}_{i,p+1}]' \tag{1.3.10}$$

1.4 Distributions

1.4.1 Univariate normal

When the random variable X has a normal (Gaussian) distribution with mean μ and variance σ^2, we will write $X \sim N(\mu,\sigma^2)$. The density function of X, for $-\infty < x < +\infty$, is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\tfrac{1}{2}(x-\mu)^2/\sigma^2] \tag{1.4.1}$$

1.4.2 Multivariate normal

When the random variables in $X' = [X_1\ X_2 \ldots X_n]$ have a joint multivariate normal distribution with vector of means $\boldsymbol{\mu}$ and positive-definite variance-covariance matrix $\mathbf{V}$ we write $X \sim N(\boldsymbol{\mu},\mathbf{V})$. The density function of X is then

$$f(X_1,X_2,\ldots,X_n) = \frac{\exp[-\frac{1}{2}(X-\boldsymbol{\mu})'\mathbf{V}^{-1}(X-\boldsymbol{\mu})]}{(2\pi)^{\frac{n}{2}}|\mathbf{V}|^{\frac{1}{2}}} \tag{1.4.2}$$

When $E(X_i) = \mu$ for all i then $\boldsymbol{\mu} = \mu 1$ and if the X_i's are mutually independent, all with the same variance σ^2, then $\mathbf{V} = \sigma^2 I$ and we write $X \sim N(\mu 1,\sigma^2 I)$.

1.4.3 Central χ^2

When $X \sim N(0,I)$ then $U = \sum_{i=1}^{n} X_i^2$ has the central χ^2 -distribution with n degrees of freedom. The density function is

$$f(u) = \frac{u^{\frac{n-2}{2}} \exp(-u/2)}{(2)^{\frac{n}{2}} \, \Gamma(n/2)} \quad \text{for } u > 0 \qquad (1.4.3)$$

1.4.4 Central F

Two independent variables each having χ^2-distributions form the basis of the F-distribution. Thus if

$$U_1 \sim \chi^2_{n_1} \quad \text{and} \quad U_2 \sim \chi^2_{n_2} \quad \text{then} \quad V = \frac{U_1/n_1}{U_2/n_2} \sim F(n_1,n_2),$$ the F-distribution with $n_1$and n_2 degrees of freedom. The density function is

$$f(v) = \frac{\Gamma(\frac{1}{2}(n_1+n_2)) \, n_1{}^{\frac{1}{2}n_1} \, n_2{}^{\frac{1}{2}n_2} \, v^{\frac{1}{2}n_1-1}}{\Gamma(\frac{1}{2}n_1) \, \Gamma(\frac{1}{2}n_2) \, (n_2+ n_1 v)^{\frac{1}{2}(n_1+n_2)}} \quad \text{for } v > 0 \qquad (1.4.4)$$

1.4.5 Central t

The ratio of a standard normal variable to the root of an independent variable that has a χ^2-distribution is the basis of Student's t-distribution. Thus when $X \sim N(0,1)$ and U is $U \sim \chi^2_n$, independent of X, then

$$Z = X/(U/n)^{\frac{1}{2}} \sim t_n,$$

the t-distribution with n degrees of freedom. Its density function is

$$f(z) = \frac{\Gamma(\frac{1}{2}(n+1))}{\sqrt{n\pi}\;\Gamma(\frac{1}{2}n)} \left[1+\frac{z^2}{n}\right], \quad \text{for } -\infty < z < \infty. \qquad (1.4.5)$$

1.4.6 Non-central χ^2

When $X \sim N(\mu,I)$ and $U = \sum_{i=1}^{n} X_i^2$, the resulting distribution of u is the non-central χ^2 with n degrees of freedom and non-centrality parameter γ,

$$\gamma = \mu'\mu/2 \qquad (1.4.6)$$

Reference to the distribution is by means of the symbol $\chi^2(n,\gamma)$. The density function of the non-central χ^2-distribution $\chi^2(n,\gamma)$ is

$$f(u) = \exp(-\gamma) \sum_{k=0}^{\infty} \frac{\gamma^k \; u^{\frac{1}{2}(n+2k-1)} \exp(-\frac{1}{2}u)}{k! \; (2)^{\frac{1}{2}n+k} \quad \Gamma(\frac{1}{2}n+k)}, \quad \text{for } u > 0 \qquad (1.4.7)$$

Some texts prefer to regard $\mu'\mu$ as the non-centrality parameter with corresponding adjustments to the form of the density function.

1.4.7 Non-central F

If U_1 and U_2 are independent and

$$U_1 \sim \chi^2(n_1,\gamma) \text{ and } U_2 \sim \chi^2_{n_2} \text{ (or } \chi^2(n_2,0))$$

then $$V = \frac{U_1/n_1}{U_2/n_2} \text{ is distributed as } F(n_1,n_2;\gamma),$$

the non-central F-distribution with n_1 and n_2 degrees of freedom and non-centrality parameter γ. Its density function is

$$f(v) = \sum_{k=0}^{\infty} \exp(-\gamma)\frac{\gamma^k\ \Gamma(\frac{1}{2}(n_1+n_2+2k))\quad n_1^{\frac{1}{2}n_1+k}\ n_2^{\frac{1}{2}n_2}\quad v^{\frac{1}{2}n_1+k-1}}{k!\ \ \Gamma(\frac{1}{2}n_1+k)\ \ \Gamma(\frac{1}{2}n_2)\quad (n_2+n_1v)^{\frac{1}{2}(n_1+n_2)+k}},$$

$$\text{for } v > 0 \qquad (1.4.8)$$

Here too some texts prefer to regard $\mu'\mu$ as the non-centrality parameter.

1.4.8 Non-central t

If $X \sim N(\mu,1)$ and if $U \sim \chi^2_n$, independently of X, then $T = X/(U/n)^{\frac{1}{2}}$ has the non-central t-distribution, $t(n,\mu)$, with n degrees of freedom with the non-centrality parameter μ. The density function is

$$f(t) = \frac{n^{\frac{n}{2}}}{\Gamma(\frac{1}{2}n)}\ \frac{e^{-\frac{1}{2}\mu^2}}{(n+t^2)^{\frac{1}{2}(n+1)}} \sum_{k=0}^{\infty} \frac{\Gamma(\frac{1}{2}(n+k+1))\ \ \mu^k\ 2^{\frac{k}{2}}\ t^k}{k!(n+t^2)^{\frac{k}{2}}} \qquad (1.4.8)$$

1.5 Variance Inflation Factors

The variance inflation factors (VIF's) were first defined by Marquardt (1970) as the diagonal elements in the inverse of the correlation matrix of X'X. Thus the i-th variance inflation factor (VIF_i) is:

$$VIF_i = \frac{(X'X)^{-1}_{ii}}{(X_i'X_i)^{-1}} \qquad (1.5.1)$$

where $(X'X)^{-1}_{ii}$ denotes the i-th diagonal elements of $(X'X)^{-1}$ and X_i is the i-th column of X. Note that the columns of X are not necessarily scaled or centered.

Discussion of VIF's and their use in diagnosing collinearity is presented in Chapter 2.

1.6 Variable Diagnostics

1.6.1 Analysis of variance (ANOVA)

Suppose Y is modelled as a linear function of the regressors, with an intercept term. Then $\hat{\epsilon}_i$ (the residual term for the i-th observation) is

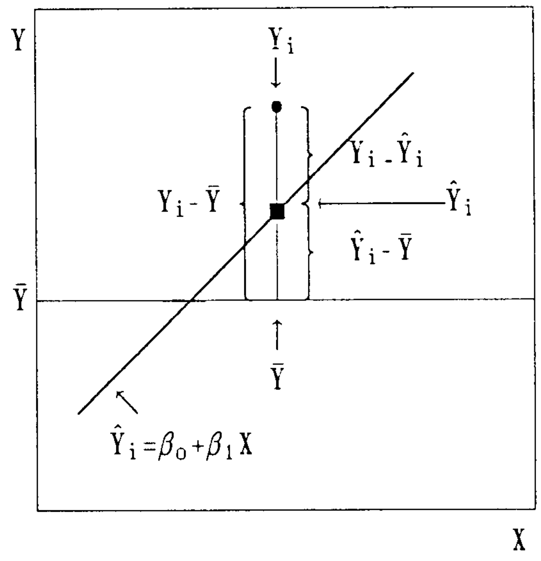
$$\hat{\epsilon}_i = Y_i - \hat{Y}_i$$

and $\hat{\epsilon}$ will be a nx1 column vector of all the n residual values.

The deviation $Y_i - \bar{Y}$ (a quantity measuring the variation of the observations Y_i and around their mean) can be decomposed as follows

$$\underbrace{Y_i - \bar{Y}}_{\text{I}} = \underbrace{\hat{Y}_i - \bar{Y}}_{\text{II}} + \underbrace{Y_i - \hat{Y}_i}_{\text{III}}$$

where I is the total deviation, II is the deviation of fitted OLS regression value around the overall mean and III is the deviation of the observed value from the regression line. The figure below (Neter and Wasserman (1974)) shows this decomposition for one of the observations.



The sums of these squared deviations satisfy the same additive relationship, due to the mixed terms of the expression having zero sum (see, for example, Neter and Wasserman (1974), for a proof).

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

or $\quad SSTO \quad = \quad SSR \quad + \quad SSE$

where SSTO is the total sum of squares (corrected for the mean) with n-1 degrees of freedom (df), SSR is the regression sum of squares with p-1 degrees of freedom (p-1 independent regressor variables) and SSE denotes the error sum of squares with n-p degrees of freedom (p parameters are fitted).

In matrix notation and for any value of p the sums of squares are

$$SSTO = Y'Y - n\bar{Y}^2 \tag{1.6.1}$$

$$SSR = \hat{\beta}'X'Y - n\bar{Y}^2 \tag{1.6.2}$$

$$SSE = Y'Y - \hat{\beta}'X'Y \tag{1.6.3}$$

A sum of squares divided by its degrees of freedom is called a mean square (MS). The breakdown of the total sum of squares and associated degrees of freedom are displayed in the form of an analysis of variance table (ANOVA table).

ANOVA Table

Source	SS	df	MS
Regression	$SSR = \hat{\beta}'X'Y - n\bar{Y}^2$	p-1	MSR = SSR/(p-1)
Error	$SSE = Y'Y - \hat{\beta}'X'Y$	n-p	MSE = SSE/(n-p)
Total	$SSTO = Y'Y - n\bar{Y}^2$	n-1	

Sometimes the random variable SSE will also be specified as SSE $(X_1,X_2,\dots,X_p)$, where the bracket denotes the subset of the independent regressor variables that are included in the model. Where this notation is not explicitly used, the set of regressor variables in the model will be clear in the context.

The use of the term MSE here (for the scalar random variable: mean square error) is not to be confused with the non-stochastic matrix MSE, a matrix of expectations corresponding to the (matrix) sum of the variance and bias matrices of a multivariate parameter estimator.

The coefficient of multiple determination is denoted by R^2 and is defined as

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \tag{1.6.4}$$

$$0 \leq R^2 \leq 1$$

R^2 measures the proportionate reduction in the variation of Y achieved by the introduction of the entire set of X variables considered in the model. Sometimes for clarity R^2 is denoted by $R^2_{y\varphi}$, where φ denotes the set of independent variables that are included in the model (i.e. for $X{:}n{\times}p$ including a column of ones to fit an intercept, the R^2 of the full model is $R^2_{y\varphi} = R^2_{y12...p-1}$).

The coefficient of multiple correlation R is the positive square root of R^2

$$R = \sqrt{R^2} \tag{1.6.5}$$

In the case of simple regression (p=2), R is the absolute value of coefficient of correlation $|r_{ij}|$ where i and j denote the dependent response variable Y and a single regressor variable X. For $p \geq 2$, the value of R is the (simple) correlation coefficient between the observed and estimated Y-values, and is consequently always positive.

The coefficient of partial determination is defined as

$$R^2_{yi.\phi} = \frac{SSE(\{X_\phi\}) - SSE(X_i,\{X_\phi\})}{SSE(\{X_\phi\})} \qquad (1.6.6)$$

where ϕ denotes the set of regressor X variables already in the model prior to fitting X_i. For example when $p = 4$ and we want to find $R^2_{y1.234}$ then $\phi = 234$ and $\{X_\phi\} = X_2,X_3,X_4$. Thus

$$R^2_{y1.234} = \frac{SSE(X_2,X_3,X_4) - SSE(X_1,X_2,X_3,X_4)}{SSE(X_2,X_3,X_4)}$$

The coefficient of partial determination measures the marginal contribution of a regressor variable X_i, given that other specified regressors are already included in the model.

1.6.2 Subset selection of regressor variables

Although p regressor variables are available, not all of them may be necessary for an adequate fit of model to the data. After the functional form of each regressor variable is obtained (i.e. X_i^2, $\log(X_j)$, X_iX_j, and so on), we seek a 'best' subset of regressor variables. This 'best' subset is not necessarily unique but may be one of a unique set of 'best' subsets.

To find a subset there are basically two strategies, all possible regressions and stepwise regression (which we take to include the special cases of forward selection and backward elimination.)

1.6.2.1 All possible regressions

In the all possible regressions search procedure, all possible regression equations are computed and selection of a 'best' equation is performed under some criterion (R^2, Adjusted R^2, MSE and C_p). If there are $(p-1) = k$ independent variables and one intercept term there will be 2^k possible equations. For example if p = 3 (constant, X_1,X_2) the following 2^2 models are possible:

$E(Y) = \beta_0$; $E(Y) = \beta_0 + \beta_1X_1$; $E(Y) = \beta_0 + \beta_2X_2$; $E(Y) = \beta_0 + \beta_1X_1 + \beta_2X_2$; where the meaning and the values of the coefficients β_0, β_1, β_2 is different in each model.

(i) R^2 Criterion

The coefficient of multiple determination R^2 defined in (1.6.4), is computed for each 2^k equations. R^2 will be a maximum when all p regressor variables are included in the equation. We therefore want to find a minimal subset for which R^2 has stabilized close to its maximum (i.e. when including another variable in the model, the increase in R^2 is very small).

(ii) Adjusted R^2

Adding more independent variables to the model can only increase R^2 and never reduce it. A modified measure that recognizes the number of independent variables is introduced. The adjusted coefficient of multiple determination, denoted R_a^2, is defined as:

$$R_a^2 = 1- \frac{n-1}{n-p}\,\frac{SSE}{SSTO} = 1 - \frac{n-1}{n-p}\,(1 - R^2)$$

One then computes R_a^2 for each equation and seeks a set (or more than one set) of independent variables which maximizes R_a^2.

(iii) MSE Criterion

One may compute the MSE for each model equation and seek a set (or more than one set) of independent variables which minimizes MSE. The R^2 criterion does not take into account the number of parameters in the model, whereas the MSE criterion does take that number into account (MSE = SSE/(n-p)).

(iv) C_p Criterion

The C_p criterion, proposed by Mallows (1964), is based on the 'total squared error'. Define the quantity Γ_p

$$\Gamma_p = [\sum_{i=1}^{n} (\nu_i - \eta_i)^2 + \sum_{i=1}^{n} \mathrm{Var}(\hat{Y}_i)]/\sigma^2$$

where $\nu_i = \nu(X_{1i}, X_{2i}, \ldots)$ is the expected value from true equation for the conditional expectation of $(Y_i | X_{1i} \ldots X_{pi})$,

$$\eta_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij}, \text{ is expected value from fitted equation}$$

$$\nu_i - \eta_i = \text{bias at the i-th data point}$$

$$p = k+1 \text{ when } \beta_0 \text{ is present}$$

$$= k \text{ when } \beta_0 \text{ is absent}$$

$$\sum_{i=1}^{n} (\nu_i - \eta_i)^2 = \text{the sum of squared bias } (SSB_p)$$

Now the residual sum of squares (denoted by SSE_p), from a fitted equation involving the p estimated coefficients, has the expectation:

$$E(SSE_p) = SSB_p + (n-p)\sigma^2 \tag{1.6.7}$$

Denote the i-th row of X by x_i', thus

$$X' = [x_1, x_2, \ldots, x_n] \text{ and} \tag{1.6.8}$$

$$X'X = \sum_{i=1}^{n} x_i x_i' \tag{1.6.9}$$

$$\begin{aligned}\mathrm{Var}(\hat{Y}_i) &= \text{variance of the fitted value } \hat{Y}_i \\ &= \mathrm{Var}(x_i'\hat{\beta}) \\ &= \sigma^2 x_i'(X'X)^{-1}x_i \qquad (1.6.10)\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{n} \mathrm{Var}(\hat{Y}_i) &= \sigma^2 \sum_{i=1}^{n} x_i'(X'X)^{-1}x_i \\ &= \sigma^2 \mathrm{tr}\ \{ \sum_{i=1}^{n} x_i'(X'X)^{-1}x_i\} \\ &= \sigma^2 \mathrm{tr}\ \{I_p\} \\ &= \sigma^2 p \qquad (1.6.11)\end{aligned}$$

thus $\Gamma_p = E(SSE_p)/\sigma^2 - (n-p) + p.$

If σ^2 is estimated by $\hat{\sigma}^2$ (after p parameters are fitted), an estimator of Γ_p, denoted by C_p, is:

$$C_p = SSE_p/(\hat{\sigma}^2) - (n - 2p) \qquad (1.6.12)$$

When there is no bias in the regression equation

$$\begin{aligned}E[C_p/\nu_i \equiv \eta_i] &= (n-p)\sigma^2/\sigma^2 - (n - 2p) \\ &= p\end{aligned}$$

Thus, when the C_p values for all possible regressions are plotted against p, those regressions with little bias will tend to cluster near the line $C_p = p$, while those for equations with substantial bias will fall above this line. With this criterion we identify the sets of independent variables that lead to smallest C_p and we would prefer those sets that have little bias (i.e. those near the line).

All possible regressions require much computation (in some contexts this is a disadvantage, but for p large a t-directed search can be performed). For a comprehensive discussion on all possible regression see Daniel and Wood (1980).

1.6.2.2 Stepwise regression

Some practitioners prefer stepwise regression because this technique requires less computation than all-possible subsets regression. This search method computes a sequence of regression equations. At each step an independent variable is added or deleted. The common criterion for adding (or deleting) some regressor variable examines the effect of that variable which produces the greatest reduction (or smallest increase) in the error sums of squares, at each step. Under stepwise regression we can distinguish basically three procedures (i) forward selection, (ii) backward elimination procedure and (iii) forward selection with a view back.

(i) Forward Selection Procedure

In the forward selection procedure, the emphasis is on finding the best single predictor, then the best two predictors (which include the best single predictor) then the best three predictors (which include the best two predictors, and in turn the best single predictor), and so forth. The procedure as outlined by Graybill (1976) is as follows:

1. Compute all correlation coefficients (or R^2_{yi} for i = 1,2,...p) between Y and $X_1, X_2, \ldots, X_p$, that is compute $r^2_{y1}, r^2_{y2}, \ldots, r^2_{yp}$. Choose the largest, suppose it is r^2_{y1}; then X_1 is the best single predictor of Y.

2. Compute all squared multiple correlation coefficients of Y with all pairs of independent variables involving X_1, that is compute R^2_{y12}, R^2_{y13}, $R^2_{y14}, \ldots, R^2_{y1p}$, and select the largest. Suppose it is R^2_{y12} then X_1 and X_2 are the best two predictors of Y which include the best single predictor X_1.

3. Compute all multiple correlation coefficients (or their squares) of Y with all sets of three variables that include (X_1 and X_2, that is compute R^2_{y123}, $R^2_{y124}, \ldots, R^2_{y12p}$ and select the largest.

At every step in the forward selection procedure we want to determine if the addition of one more variable, will 'appreciably' improve the estimation of Y. If we found that a new variable will improve the resulting estimator of Y we include it and continue, but otherwise the forward selection procedure is terminated because a 'best' subset has been found. Estimation is improved if the corresponding estimate of error variance is sufficiently less than the current estimate.

Another way to formulate this strategy is as follows: We ask if regressors 1,2,...,q,q+1 yield a better estimate of Y than do regressors 1,2,...,q (where 1,2,...,q have been determined as above). Thus we examine

$$H_0 : \rho^2_{y12\ldots q+1} = \rho^2_{y12\ldots q} \tag{1.6.13}$$

which is true if and only if

$$H_0 : \rho^2_{yq+1\cdot 12\ldots q} = 0 \tag{1.6.14}$$

Compute the test statistic W, where

$$W = \frac{(n-q-2)R^2_{yq+1\cdot 12\ldots q}}{1-R^2_{yq+1\cdot 12\ldots q}}$$

where R^2 is the estimated sample estimate of the population value ρ^2.

If H_0 is true, W is distributed as F(1,n-q-2). The hypotheses H_0 is rejected (for a size α test) if and only if w the computed value of W satisfies $w \geq F(\alpha\text{:}1,n-q-2)$, the critical value of the F(1,n-q-2) distribution. So the forward selection procedure is terminated at the step where H_0 is rejected. In some contexts α is chosen to be quite large, or almost equivalently, the tabulated F-value criterion is replaced by a suitable constant (eg. $F_{in} = 2.00$ by default in BMDP)

(ii) Backward Elimination Procedure.

This search procedure is the opposite of forward selection. One starts with the full model and then the less important regressors are eliminated one at a time. The basic steps in the procedure are given in Draper and Smith (1981):

1. A regression equation containing all variables is computed.

2. The partial F-test value is calculated for every variable treated as though it were the last variable to enter the regression equation.

3. The lowest partial F-test value, F_L, is compared with a preselected significance level F_0. Then if $F_L < F_0$ we remove the variable X_L from the equation, and recompute the regression equation without X_L, then re-enter step 2 again. If $F_L > F_0$, we adopt the regression equation.

(iii) Forward Selection with a View Back

This method works just like the forward selection with the difference that at each step one looks back at the independent variables already in the model, examines them and decides if one of them should be dropped.

1.7 Case Diagnostics

Outliers are observed values that do not fit the model. Influential cases are observations which can markedly effect the estimation process. Their influence arises from their relationships with the other observations. It is possible for a particular case to be an outlier and to be influential.

1.7.1 Outliers

Generally speaking since the true errors are not observable the analyst has to rely on the estimated error terms. In some situations however the estimated error terms are substantially effected by the influential cases. It is therefore advisable to examine the estimated error terms along with corresponding measures of influence.

1.7.2 Influence

For OLS, the vector of ordinary residuals, $\hat{\epsilon}$ is given by

$$\begin{aligned}\hat{\epsilon} &= Y - X\hat{\beta} \\ &= Y - X(X'X)^{-1}X'Y \\ &= [I - H]Y \qquad (1.7.1)\end{aligned}$$

where $H = X(X'X)^{-1}X'$ and $\hat{Y}$ is the vector of fitted values.

The matrix H is called the Hat matrix, because it maps Y into $\hat{Y} = HY$ (Hoaglin and Welsh (1978)). The matrix H is symmetric ($H' = H$), idempotent ($HH = H$), and a projection matrix (into the column space of X). The diagonal elements of the Hat matrix, whose role as a diagnostic measure will be discussed in Chapter 9 are:

$$h_i = h_{ii} = X_i'(X'X)^{-1}X_i \qquad (1.7.2)$$

where X_i' is the i-th row of X. The diagonal elements are known as the leverage values.

Several transformations of the ordinary residuals have been proposed for use in diagnostic procedures (e.g. see Cook and Weisberg (1982, p17)). The most important are the standardized residuals and the studentized residuals. The standardized residual (also called the studentized residual (Cook and Weisberg (1982)) is defined as

$$r_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}} , \quad i = 1,2,\ldots,n \tag{1.7.3}$$

where $\hat{\sigma}$ is the residual mean square. It does not strictly follow a t-distribution because $\hat{\epsilon}_i$ and $\hat{\sigma}$ are not independent. When σ is estimated by $\hat{\sigma}(i)$, the estimated error variance when the i-th row of X and Y have been deleted, the result is a studentized residual

$$t_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}(i)\sqrt{1-h_{ii}}} , \quad i = 1,2,\ldots,n \tag{1.7.4}$$

which is distributed as Student's t with n-p-1 degrees of freedom. A simple formula for $\hat{\sigma}(i)$ (Belsley *et al.* (1980, p14)) uses

$$(n-p-1)\{\hat{\sigma}(i)\}^2 = (n-p)\hat{\sigma}^2 - \frac{\hat{\epsilon}_i^2}{1-h_{ii}} \tag{1.7.5}$$

The measure DFFITS (Belsley *et al.* (1980, p15)), is the standardized change in the fitted value of a case when it is deleted, is given for the i-th case by

$$\text{DFFITS}_i = \left[\frac{h_{ii}}{1-h_{ii}}\right]^{\frac{1}{2}} \frac{\hat{\epsilon}_i}{\hat{\sigma}(i)\sqrt{1-h_{ii}}} \tag{1.7.6}$$

Cook's (squared) distance (Cook (1977)) of an estimator $\tilde{\beta}$ from the OLS $\hat{\beta}$ is defined as

$$C = D^2 = (\hat{\beta} - \tilde{\beta})'X'X(\hat{\beta} - \tilde{\beta})/(p\hat{\sigma}^2) \tag{1.7.7}$$

The distance is regarded as large when $D^2 > F(1-\alpha,p,n-p)$, where $F(1-\alpha,p,n-p)$ is the $1-\alpha$ probability point of the central F-distribution with p and n-p degrees of freedom. It is widely known that D^2 does not follow an

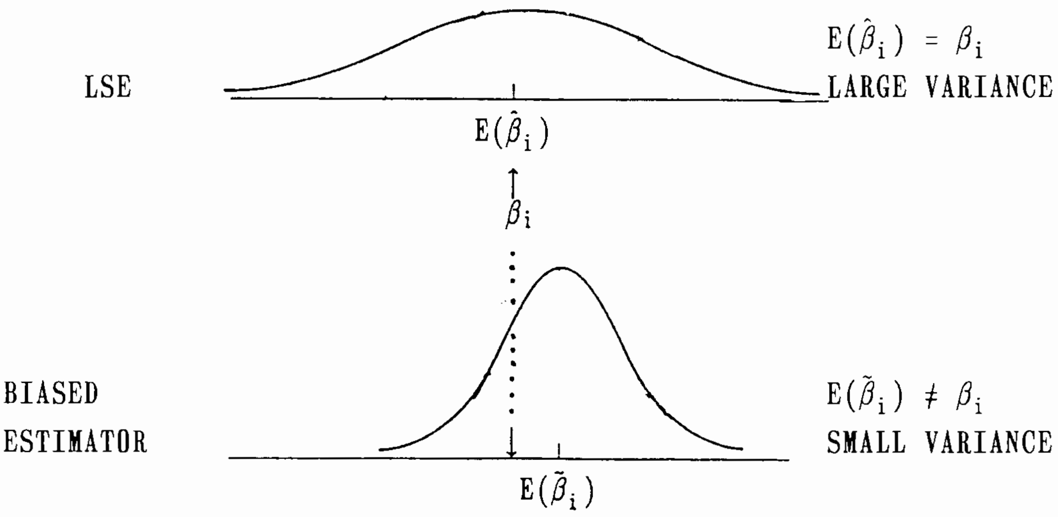
F-distribution, but is an effective measure of relative change in estimated coefficients.

1.8 Bias and Jackknifing

1.8.1 Biased estimation

Least square estimators (LSE's or OLSE's) are the best linear unbiased estimators (BLUE's) of the elements of the parameter vector β. Amongst linear unbiased estimators the LSE's have the smallest variances. In the presence of collinearity one or more of these variances can be inflated to such an extent that the corresponding estimators become unacceptable. The 'fly in the ointment' with the least squares criterion is its requirement of unbiasedness (Marquardt and Snee (1975)). A major reduction in variance can be obtained as a result of allowing a little bias. If one looks beyond the class of unbiased estimators, it is possible to find some biased estimators with smaller variances than the variances of the LSE's. Some of these biased estimators will perform better than LSE's in the presence of collinearity, in the sense of reduced mean square error (MSE).

Variance and bias in an estimator $\hat{\beta}_i$



MSE may be used to assess the performance of regression estimators. In the regression model (1.1)

$$Y = X \beta + \epsilon,$$

if $\tilde{\beta}$ is an estimator of β, the MSE of $\tilde{\beta}$ is defined as

$$\begin{aligned} MSE(\tilde{\beta}) &= E[(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)'] \\ &= V(\tilde{\beta}) + bb' \qquad (1.8.1) \end{aligned}$$

where $b = E(\tilde{\beta}) - \beta$ is the bias vector.

The total mean squared error (TMSE) of $\tilde{\beta}$ is defined as

$$\begin{aligned} TMSE(\tilde{\beta}) &= tr[MSE(\tilde{\beta})] \\ &= \Sigma\ var(\tilde{\beta}_i) + \Sigma\ b_i^2 \qquad (1.8.2) \end{aligned}$$

1.8.2 Jackknifing

The jackknife technique was introduced by Quenouille (1956) and Tukey (1958). The jackknife is a general method for reducing the bias in an estimator and for obtaining a measure of the variance of the resulting estimator by sample reuse.

Let $X = [x_1 \dots . x_n]'$. The subscript -i with any matrix will mean that the i-th row has been deleted, i.e. with X_{-i} we mean the X matrix with its i-th row deleted. In a vector Y the subscript i indicates the i-th element of the vector (i.e. Y_i) but the subscript -i, indicates the subvector of Y remaining after the i-th element has been deleted.

Define

$$\hat{\epsilon}_i = Y_i - x_i'\hat{\beta} \tag{1.8.3}$$

$$\begin{aligned}\hat{\epsilon} &= Y - X\hat{\beta} \\ &= [\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n]'\end{aligned} \tag{1.8.4}$$

$$h_i = x_i'(X'X)^{-1}x_i \tag{1.8.5}$$

The least square estimator obtained by deleting the i-th row (x_i', Y_i) of the data is:

$$\hat{\beta}_{-i} = (X_{-i}'X_{-i})^{-1}X_{-i}'Y_{-i} \tag{1.8.6}$$

$$\begin{aligned}&= [X'X - x_i x_i']^{-1}[X'Y - x_i Y_i] \\ &= [(X'X)^{-1} + (X'X)^{-1}x_i(I - x_i'(X'X)^{-1}x_i)^{-1}x_i'(X'X)^{-1}][X'Y - x_i Y_i] \\ &= \hat{\beta} - (X'X)^{-1}x_i[Y_i - Y_i h_i - x_i'\hat{\beta} + h_i Y_i](1-h_i)^{-1} \\ &= \hat{\beta} - (X'X)^{-1}x_i[Y_i - x_i'\hat{\beta}](1-h_i)^{-1} \\ &= \hat{\beta} - (X'X)^{-1}x_i[\hat{\epsilon}_i](1-h_i)^{-1}\end{aligned} \tag{1.8.7}$$

for i = 1,2,...,n.

This equation illustrates the effect of an influential point (h_i close to 1) on the coefficients. Under the assumption that Y_i can be modelled simultaneously with Y_{-i}, the scalar $\hat{\epsilon}_i/(1-h_i)$ has zero expectation but large variance $\sigma^2/(1-h_i)$, and an outlying x_i in the row space of X_{-i} will tend to have a large influence on the choice of estimates. However if in fact the extrapolation from $Y_{-i} = X_{-i}\beta + \epsilon_{-i}$ to suggest values for $x_i'\beta$ is not justified, using the full data set will lead to $\hat{\beta}$ values that are generally sufficiently different from $\hat{\beta}_{-i}$ as to be misleading, and in particular, biased (for β in the model for the reduced data set).

In what follows, we assume that the same β is operative in the full and reduced data sets. Define the pseudovalue for i = 1,2,...,n as

$$\begin{aligned} P_i &= n\hat{\beta} - (n-1)\hat{\beta}_{-i} \qquad (1.8.8) \\ &= \hat{\beta} + (n-1)(X'X)^{-1}x_i[\hat{\epsilon}_i]/(1-h_i) \end{aligned}$$

then the jackknifed estimator is given by

$$\begin{aligned} \hat{\beta}_J &= n^{-1}\ \Sigma\ P_i \\ &= \hat{\beta} + (n-1)n^{-1}(X'X)^{-1}\ \Sigma\ x_i\hat{\epsilon}_i/(1-h_i) \qquad (1.8.9) \end{aligned}$$

$$\begin{aligned} E(\hat{\beta}_J) &= \beta + (n-1)n^{-1}\ (X'X)^{-1}\ \Sigma\ x_i E(\hat{\epsilon}_i)/(1-h_i) \\ &= \beta \qquad (E(\hat{\epsilon}_i) = 0) \qquad (1.8.10) \end{aligned}$$

Since $\text{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1}$ and $\text{Var}(\hat{\epsilon}) = \sigma^2[I- X(X'X)^{-1}X']$, and because $\hat{\beta}$ and $\hat{\epsilon}$ are uncorrelated, we have

$$\begin{aligned} \text{Var}(\hat{\beta}_J) &= \sigma^2(X'X)^{-1} + \sigma^2((n-1)n^{-1})^2(X'X)^{-1} \\ &\quad \times [\ \Sigma\ x_i[I-x_i'(X'X)^{-1}x_i]x_i'/(1-h_i)^2](X'X)^{-1} \end{aligned} \qquad (1.8.11)$$

The jackknife distribution-free estimate of variance for the parameter estimator $(\hat{\beta}_J)$ is defined as

$$\begin{aligned} V_J &= [n(n-1)]^{-1}\Sigma(P_i - \hat{\beta}_J)(P_i - \hat{\beta}_J)' \qquad (1.8.12) \\ &= \hat{\text{Var}}(\hat{\beta}_J) \end{aligned}$$

This sample moment estimator (on the n pseudo-values) may be used to estimate both $\text{Var}(\hat{\beta}_J)$ and $\text{Var}(\hat{\beta})$ (Hinkley (1977)). He also pointed out the following shortcomings of the method, namely

1. $\hat{\beta}_J$ is different from the original estimator $(\hat{\beta})$, is unbiased for β but has in general a larger variance than the LSE (Gauss Markov).
2. V_J is in general, biased for estimating $\mathrm{Var}(\hat{\beta}_J)$ or $\mathrm{Var}(\hat{\beta})$.

These problems stem from the balanced nature of the ordinary jackknife, which neglects the unbalanced nature of the regression data. Hinkley (1977) proposed a weighted modification. The weighted pseudo-value

$$\begin{aligned} Q_i &= \hat{\beta} + n(1-h_i)(\hat{\beta} - \hat{\beta}_{-i}) \\ &= \hat{\beta} + n(1-h_i)(\hat{\beta} - \hat{\beta} - (X'X)^{-1}x_i[\hat{\epsilon}_i](1-h_i)^{-1}) \\ &= \hat{\beta} + n((X'X)^{-1}x_i\hat{\epsilon}_i) \qquad (1.8.13) \end{aligned}$$

The weighted jackknife estimator (denoted by $\hat{\beta}_{JW}$) is

$$\begin{aligned} \hat{\beta}_{JW} &= n^{-1}\sum_{i=1}^{n} Q_i \qquad (1.8.14) \\ &= \hat{\beta} \end{aligned}$$

and the variance estimator

$$\begin{aligned} V_{JW} &= [n(n-p)]^{-1}\Sigma(Q_i - \hat{\beta}_{JW})(Q_i - \hat{\beta}_{JW})' \\ &= [n(n-p)]^{-1}\Sigma[\hat{\beta}+n((X'X)^{-1}x_i\hat{\epsilon}_i)-\hat{\beta}][\hat{\beta}+n((X'X)^{-1}x_i\hat{\epsilon}_i)-\hat{\beta}]' \\ &= n(n-p)^{-1}(X'X)^{-1}(\Sigma\ \hat{\epsilon}_i^2 x_i x_i')(X'X)^{-1} \qquad (1.8.15) \end{aligned}$$

V_{JW} will be biased in unbalanced cases but is robust against error variance heterogeneity. (Lemma 2, Appendix of Hinkley (1977))

The above description of the jackknife only takes into account the deletion of one single row at a time. Therefore it is called the delete-one jackknife method. Wu (1986) proposed a class of weighted modifications allowing for the deletion of an arbitrary number of observations.

1.9 Vector and Matrix Norms, and Decompositions

1.9.1 Vector norms

A vector norm (or simply a norm) on R^n is a function $\nu:R^n \to R$ that satisfies the following conditions (Stewart (1973)):

$$\begin{aligned} &1.\quad x \neq 0 \Rightarrow \nu(x) > 0, \\ &2.\quad \nu(\alpha x) = |\alpha|\nu(x), \\ &3.\quad \nu(x+y) \leq \nu(x) + \nu(y) \end{aligned} \tag{1.9.1}$$

The conditions 1, 2, 3, are also termed definiteness, homogeneity, and triangle inequality conditions.

Three norms on R^n that are frequently used in analyzing matrix processes, are the 1-, 2-, and ∞-norms.

The 1-norm of a vector y, is defined as

$$\|y\|_1 = \sum_{i=1}^{n} |y_i|, \tag{1.9.2}$$

where y_i is the i-th element of the vector y:nx1.

The 2-norm of a vector y, is defined as

$$\|y\|_2 = \sqrt{y'y} \tag{1.9.3}$$

The 2-norm is sometimes called the Euclidean norm of a vector y.

The ∞-norm of a vector y, is defined as

$$\|y\|_\infty = \max\{|y_i|:i = 1,2,\dots,n\} \tag{1.9.4}$$

and is sometimes called the maximum norm (max-norm) or the Chebyshev norm.

The norms defined in (1.9.2), (1.9.3) and (1.9.4) are special cases of the Hölder norms or vector p-norms defined by

$$\|y\|_p = \sqrt[p]{\sum_{i=1}^{p} |y_i|^p}, \; 1 \leq p < \infty \tag{1.9.5}$$

($\|y\|_\infty$ is $\lim(\|y\|_p)$ as $p \to \infty$)

1.9.2 Matrix norms

A function $\nu : R^{m\times n} \to R$ is a matrix norm on $R^{m\times n}$ if

$$\begin{array}{lllll} 1. & A \neq 0 & \Rightarrow & \nu(A) > 0, & A \in R^{m\times n}, \\ 2. & \nu(\alpha A) & = & |\alpha|\nu(A), & A \in R^{m\times n},\ \alpha \in R, \\ 3. & \nu(A+B) & \leq & \nu(A) + \nu(B), & A,B \in R^{m\times n} \\ 4. & \nu(AB) & \leq & \nu(A)\nu(B). & \end{array} \tag{1.9.6}$$

Condition (4) is known as the submultiplicative or consistency condition. If a function statisfies (1)-(3) and not necessarily (4), it is called a generalized matrix norm.

The Frobenius norm of a matrix A is defined as

$$\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2} \tag{1.9.7}$$

The Frobenius norm can also be shown to satisfy

$$\|A\|_F^2 = \mathrm{tr}[A'A] \tag{1.9.8}$$

This norm (1.9.8) is sometimes called the Euclidean matrix norm, l_2 norm, the Schur norm, or the Hilbert-Schmidt norm.

A unitarily invariant matrix norm is a norm that satisfies

$$\|U'XV\| = \|X\| \tag{1.9.10}$$

for all unitary matrices U and V. (Although the symbols U and V are used in the SVD to indicate unique matrices, here we wish the identity to hold for all other conformable unitary matrices, as well as those U and V of the SVD).

The matrix p-norm of a matrix is defined from vector p-norms as

$$\|A\|_p = \sup_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_p} \tag{1.9.11}$$

where $p \in (1,2,\infty)$

$$\text{e.g.} \qquad \|A\|_2 = \sup_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} \tag{1.9.12}$$

The maximum column-sum matrix norm $\|\cdot\|_1$ of A is defined as

$$\|A\|_1 = \max \sum_{i=1}^{n} |a_{ij}| \ , \ 1 \leq j \leq n \tag{1.9.13}$$

The maximum row-sum matrix norm $\|\cdot\|_\infty$ of A is defined as

$$\|A\|_\infty = \max \sum_{j=1}^{n} |a_{ij}| \ , \ 1 \leq i \leq n \tag{1.9.14}$$

The spectral norm $\|\cdot\|_2$ of A is defined as

$$\|A\|_2 = \max\{\sqrt{\lambda}: \lambda \text{ is an eigenvalue of } A'A\} \qquad (1.9.15)$$

1.9.3 Decomposition

1.9.3.1 SVD

If the SVD of X is given by (1.3.1) and

$$\sqrt{\lambda}_1 \geq \sqrt{\lambda}_2 \geq \ldots \geq \sqrt{\lambda}_r > \sqrt{\lambda}_{r+1} = \ldots \sqrt{\lambda}_p = 0,$$

then

$$r(X) = r \leq p \qquad (1.9.16)$$
$$N(X) = \text{span}\{v_{r+1}, \ldots, v_p\} \qquad (1.9.17)$$
$$R(X) = \text{span}\{u_1, \ldots, u_r\} \qquad (1.9.18)$$

where r(X) is the rank of X, N(X) is the null space of X, and R(X) is the range of X.

Then the Frobenius norm of X can be written as

$$\|X\|_F^2 = \lambda_1 + \lambda_2 + \cdots + \lambda_r \qquad (1.9.19)$$

and the matrix 2-norm of X is

$$\|X\|_2 = \sqrt{\lambda}_1 \qquad (1.9.20)$$

Some authors call (1.9.20) the spectral norm and define it as

$$\|A\|_2 = \max \|Ax\|_2, \text{ for } \|x\| = 1$$

Stewart (1987), omits the subscript 2. Proofs of these properties can be found Golub and Van Loan (1983, Chapter 2) or Horn and Johnson (1987).

1.9.3.2 QR decomposition

The following decomposition of a matrix is known as the QR decomposition:

Let X:nxp and Y:nx1 be given and suppose that an orthogonal matrix Q:nxn exists and is computable, with the property that

$$Q'X = R = \begin{bmatrix} R_1 : pxp \\ \quad 0:(n-p)xp \end{bmatrix} \tag{1.9.21}$$

is upper triangular.

If
$$Q'Y = \begin{bmatrix} c:px1 \\ d:(n-p)x1 \end{bmatrix} \tag{1.9.22}$$

then $\|X\beta - Y\|_2^2 = \|Q'X\beta - Q'Y\|_2^2$ (from 1.9.10)

$$= \left\| \begin{bmatrix} R_1 \\ 0 \end{bmatrix} \beta - \begin{bmatrix} c \\ d \end{bmatrix} \right\|_2^2$$

$$= \|R_1\beta - c\|_2^2 + \|d\|_2^2 \tag{1.9.23}$$

for any $\beta \in R^p$.

If r(X) = p (i.e. X has full rank), then the OLSE $\hat{\beta}$ may be obtained from the upper triangular system $R_1\hat{\beta} = c$, and the minimum sum of squares satisfies $\|X\hat{\beta} - Y\|_2^2 = \|d\|_2^2$.

If X is rank deficient (r(X) < p) then at least one diagonal entry in R is zero and the QR factorization does not necessarily produce an orthonormal basis for R(A). Therefore the QR factorization must be modified to produce an orthonormal basis for the X range. This modified algorithm is known as QR with Column Pivoting:

Let Π be a suitable (pxp) permutation matrix used to interchange the columns of X so that the independent columns are moved to initial column positions. Then

$$X\Pi = QR \quad \text{where } \underset{n\text{x}p}{R} = \begin{array}{c c} \left[\begin{array}{cc} R_{11} & R_{12} \\ 0 & 0 \end{array}\right] & \begin{array}{l} r \\ n-r \end{array} \\ \begin{array}{cc} r & p-r \end{array} & \end{array} \tag{1.9.24}$$

where rank(X) = $r \leq p$, R_{11} is upper triangular and non-singular. Thus

$$\|X\beta - Y\|_2^2 = \|(Q'X\Pi)(\Pi'\beta) - Q'Y\|_2^2$$

$$= \left\| \begin{bmatrix} R_{11} & R_{12} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} - \begin{bmatrix} c \\ d \end{bmatrix} \right\|_2^2$$

$$= \|R_{11}Z_1 - (c - R_{12}Z_2)\|_2^2 + \|d\|_2^2 \tag{1.9.25}$$

where
$$\Pi'\beta = \begin{array}{cc} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} & \begin{array}{l} r \\ n-r \end{array} \end{array} \tag{1.9.26}$$

and $Q'Y$ is defined in (1.9.22). Thus, if $\|X\beta - Y\|_2^2$ is minimized then

$$\Pi'\beta = \begin{array}{cc} \begin{bmatrix} R_{11}^{-1}(c - R_{12}Z_2) \\ Z_2 \end{bmatrix} & \begin{array}{l} r \\ n-r \end{array} \end{array} \tag{1.9.27}$$

If Z_2 is set to zero then we obtain the unique solution of smallest norm

$$\Pi'\beta_{basic} = \begin{array}{cc} \begin{bmatrix} R_{11}^{-1}c \\ 0 \end{bmatrix} & \begin{array}{l} r \\ n-r \end{array} \end{array} \tag{1.9.28}$$

An algorithm for QR with column pivoting can be found on p165 of Golub and Van Loan (1983). Lawson and Hanson (1974) describe the above method on pp78-82 and refered to it as QR with column interchange strategy.

Golub and Van Loan (1983) show by examples that QR with column pivoting is not entirely reliable for detecting rank deficiency but that it works well in practice.

1.10 Subset Selection using Singular Value Decomposition.

In OLS $\hat{\beta}$ is the minimum norm solution, i.e. $\|X\beta - Y\|_2^2$ will be a minimum when β is estimated by $\hat{\beta} = \sum_{i=1}^{p} v_i u_i' Y/\sqrt{\lambda_i}$ (from 1.3.4). In the case of rank deficiency $\hat{\beta}$ can be approximated by

$$\hat{\beta}_{\hat{r}} = \sum_{i=1}^{\hat{r}} v_i u_i' Y/\sqrt{\lambda_i} \qquad (1.10.1)$$

where $\hat{r}$ is an estimate of the rank(X) = r, thus $\hat{\beta}_{\hat{r}}$ minimizes $\|X_{\hat{r}}\beta - Y\|_2$ and

$$X_{\hat{r}} = \sum_{i=1}^{\hat{r}} \sqrt{\lambda_i}\, u_i v_i' \text{ (by using (1.3.2))} \qquad (1.10.2)$$

$$= U_{\hat{r}} \Delta_{\hat{r}} V_{\hat{r}}' \qquad (1.10.3)$$

where $U_{\hat{r}} = [u_1, \ldots, u_{\hat{r}}]$, $V_{\hat{r}} = [v_1, \ldots, v_{\hat{r}}]$ and $\Delta_{\hat{r}} = \text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_{\hat{r}}}\,)$. ($X_{\hat{r}}$ is the matrix that is the closest to X that has rank r). Furthermore the residuals due to this method will be denoted by $\hat{\epsilon}_{\hat{r}}$, and

$$\begin{aligned} \hat{\epsilon}_{\hat{r}} &= Y - X\hat{\beta}_{\hat{r}} \\ &= Y - U_{\hat{r}} U_{\hat{r}}' Y \\ &= (I - U_{\hat{r}} U_{\hat{r}}')Y \end{aligned} \qquad (1.10.4)$$

A subset selection procedure that is based on the SVD has been proposed by Golub, Klema and Stewart (1976). It is also described in Golub and Van Loan (1983, pp414-419). Their method proceeds as follows:

1. Compute the SVD of X (1.3.1) and use it to determine the rank of X. Denote this rank by $\hat{r}$, $\hat{r} \leq p \leq n$.

2. Calculate a permutation matrix P such that the columns of the matrix X_1:n×$\hat{r}$ in $XP = [X_1 \ X_2]$ are 'sufficiently independent'.

3. Predict Y with Xb where $b = P[z' \ 0]'$ and z:$\hat{r}$x1 minimizes $\|X_1 z - Y\|_2^2$. Denote the residuals due to this method by

$$\begin{aligned} \hat{\epsilon}_z &= Y - Xb \\ &= Y - X_1 z \\ &= Y - X_1(X_1'X_1)^{-1}X_1Y \\ &= (I - B_1B_1')Y \end{aligned} \qquad (1.10.5)$$

where $B_1 = X_1(X_1'X_1)^{-\frac{1}{2}}$.

The key step is 2, since

$$\min \|X_1z - Y\|_2 = \|Xb - Y\|_2 \geq \min\|X\beta - Y\|_2$$

We want some bounds on the smallest singular value $\sqrt{\lambda_{\hat{r}}(X_1)}$ of X_1. For clarity the singular values of any matrix, say A, will be written here as $\sqrt{\lambda_i(A)}$. Golub and Van Loan (1983, pp416-418) state and prove the following theorems:

Theorem 1.10.1

Let the SVD of X be given by (1.3.1) and define the matrix X_1:n×$\hat{r}$, for $\hat{r} \leq \text{rank}(X)$, by

$$XP = [X_1 \ X_2]$$

where P is a pxp permutation matrix. If

$$P'V = \begin{array}{cc} \begin{bmatrix} \ddot{V}_{11} & \ddot{V}_{12} \\ \ddot{V}_{21} & \ddot{V}_{22} \end{bmatrix} & \begin{array}{c} \hat{r} \\ p-\hat{r} \end{array} \\ \begin{array}{cc} \hat{r} & p-\hat{r} \end{array} & \end{array} \tag{1.10.6}$$

and $\ddot{V}_{11}$ is nonsingular, then

$$\frac{\sqrt{\lambda_{\hat{r}}(X)}}{\|\ddot{V}_{11}^{-1}\|_2} \leq \sqrt{\lambda_{\hat{r}}(X_1)} \leq \sqrt{\lambda_{\hat{r}}(X)} \tag{1.10.7}$$

Thus the permutation P must be chosen in such away that $\ddot{V}_{11}$ is as well conditioned as possible. A solution to this problem would be QR with column-pivoting factorization of the r̂xp matrix $[V'_{11} \;\; V'_{21}]$ where V is partitioned as follows:

$$V = \begin{array}{cc} \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} & \begin{array}{c} \hat{r} \\ p-\hat{r} \end{array} \\ \begin{array}{cc} \hat{r} & p-\hat{r} \end{array} & \end{array} \tag{1.10.8}$$

Thus

$$Q'[\ddot{V}'_{11} \;\; \ddot{V}'_{21}] = Q'[V'_{11} \; V'_{21}]P = [R_{11} \; R_{12}] \tag{1.10.9}$$

where Q is orthogonal, P is an pxp permutation matrix and R_{11} is upper triangular. Then from (1.10.9)

$$[V'_{11} \; V'_{21}]P = Q[R_{11} \; R_{12}]$$

Thus

$$P'\begin{bmatrix} V_{11} \\ V_{21} \end{bmatrix} = \begin{bmatrix} R'_{11} \\ R'_{21} \end{bmatrix} Q'$$

$$\begin{bmatrix} \ddot{V}_{11} \\ \ddot{V}_{21} \end{bmatrix} = \begin{bmatrix} R'_{11}Q' \\ R'_{21}Q' \end{bmatrix} \quad \text{(from (1.10.6))} \tag{1.10.10}$$

and $$\|\ddot{V}_{11}^{-1}\| = \|(R'_{11}Q')^{-1}\| = \|R_{11}^{-1}\|$$

With column pivoting R_{11} will be well conditioned, which in turn will produce a well conditioned $\ddot{V}_{11}$.

To assess the above method of subset selection, compare the residuals from this subset procedure ($\hat{\epsilon}_z$, defined in (1.10.5)) with those residuals from the nearest rank-$\hat{r}$ LS ($\hat{\epsilon}_{\hat{r}}$, defined in (1.10.4)). Define $\ddot{V}_{11}$ as the leading $\hat{r}$-by-$\hat{r}$ submatrix of $P'V$, then

$$\|\hat{\epsilon}_{\hat{r}} - \hat{\epsilon}_z\|_2 \leq \frac{(\lambda_{\hat{r}+1}(X))^{\frac{1}{2}}}{(\lambda_{\hat{r}}(X))^{\frac{1}{2}}} \|\ddot{V}_{11}^{-1}\|_2 \|Y\|_2 \qquad (1.10.11)$$

The norm $\|\hat{\epsilon}_{\hat{r}} - \hat{\epsilon}_z\|_2$ can also be written as

$$\|\hat{\epsilon}_{\hat{r}} - \hat{\epsilon}_z\|_2 = \|Y - X\hat{\beta}_{\hat{r}} - (Y - X_1Z)\|_2 \quad \text{(from 1.10.4 and 1.10.5)}$$

$$= \|X_1Z - \sum_{i=1}^{\hat{r}} u_i(u_i'Y)\|_2 \quad \text{(from (1.10.1))} \qquad (1.10.12)$$

This sheds light on how well X_1Z can predict the stable component of Y, i.e. $\sum_{i=1}^{r} (u_i'Y)$. Any attempt to approximate $\sum_{i=s}^{n} (u_i'Y)$, where $s = r+1$, can lead to a large norm solution. Moreover (1.10.11) says that if $\lambda_{\hat{r}+1} << \lambda_{\hat{r}}$ then any reasonably independent subset of columns produces essentially the same-sized residual. On the other hand, if there is no well-defined gap in the singular values then the determination of $\hat{r}$ becomes difficult and the entire subset selection problem more complicated.

These observations together with theorem 1.10.1 form the basis of an algorithm proposed by Golub, Klema and Stewart (1976):

Algorithm SX-OLS (SX-OLS comes from subset selection on X where the method of estimation is OLS).

Given $X \epsilon R^{nxp}$, $Y \epsilon R^n$ and a method for computing an integer $\hat{r}$ that approximates rank X, then the following algorithm computes a permutation P and a vector $z \epsilon R^{\hat{r}}$ such that the first $\hat{r}$ columns of XP are independent and such that $\|XP[z' \;\, 0]' - Y\|_2$ is minimized.

Compute the SVD of X (1.3.1) and use it to determine the rank of X. Denote this rank by $\hat{r}$, where $\hat{r} \leq r(X) \leq p \leq n$ and partition V as in (1.10.8).

Step one:

$$V = \begin{array}{cc} \left[\begin{array}{cc} V_{11} & V_{12} \\ V_{21} & V_{22} \end{array}\right] & \begin{array}{c} \hat{r} \\ p-\hat{r} \end{array} \\ \begin{array}{cc} \hat{r} & p-\hat{r} \end{array} & \end{array}$$

Use QR with column pivoting (as described in §1.9) to compute

Step two:

$$Q'[V'_{11} \;\, V'_{21}]P = [R_{11} \;\, R_{12}]$$

and set $$XP = [X_1 \;\, X_2]$$

Step three: Determine $z \epsilon R^{\hat{r}}$ such that $\|Y - X_1 z\|_2 = \min.$

When $\hat{r} > p/2$ it would be more economical to compute P by applying the QR with column pivoting algorithm to $[V'_{22} \;\, V'_{12}]$, because P'V is orthogonal, and $\|\ddot{V}_{11}^{-1}\|_2 = \|\ddot{V}_{22}^{-1}\|_2$ (for a proof see Golub and Van Loan (1983, Theorem 2.4.1). Again it should be noted that these methods are intended to be in the realm of numerical mathematics and do not appeal to any statistical properties *per se*.

1.11 Probability limit

T is the probability limit (plim) of the statistic t_n, derived from a random sample of n observations, if, for any $\epsilon > 0$, the probability of $|t_n - T| < \epsilon$, approaches the limit probability 1 as $n \to \infty$.

Chapter 2

COLLINEARITY

One of the assumptions of the linear regression model (1.1), is that the fixed matrix X of independent variables is a full column-rank matrix. Violation of this assumption leads to problems referred to as collinearity. This phenomenon of collinearity and near-collinearity was first described by Ragnar Frisch (1934) and he warned that in ignoring this structure within the independent regressor variables, one runs the risk of determining a regression equation that is absurd. Frisch believed that 'a substantial part of the regression and correlation analyses which have been made on economic data in recent years is nonsense'.

Some authors refer to collinearity and its analysis as multicollinearity (Gunst (1983)), conditioning (Belsley and Oldford (1986)), confluence analysis (Frisch (1934)), ill-conditioning (Belsley and Oldford (1986)), singularity (Stewart (1987)) and non-orthogonality (Farrar and Glauber (1967)). We favour 'collinearity' as collinearity will always involve two or more vectors, and the prefix 'multi' is unnecessary (Kalman (1984)). To some degree all the foregoing terms can be regarded as synonymous. Collinearity can not be described in simple terms as being present or absent. Rather, what is important is the degree of collinearity and what effect this degree can have on the regression model. For the statistician, near-collinearities inflate the variances of regression coefficients and magnify the effects of error in the regression response variable. For the numerical analyst, collinearities combine with rounding errors to introduce inaccuracies in computations. A great deal of time and effort have been devoted to issues related to collinearity (see attached bibliography) and still the subject has abounds with paradoxes, ambiguities and open questions.

In this chapter some light is shed on this problem (although in the opinion of the writer collinearity is not so much a problem as an inherent part of the data set and model). We discuss definitions of collinearity in §2.1, ways of detecting it in §2.2, the effect of collinearity in §2.3, the issue of centering in §2.4, and the theory of perturbation in §2.5. A summary of approaches to collinearity is presented in §2.6.

2.1 Definitions

Several definitions of collinearity appear in the literature, some more descriptive of symptoms than rigorous.

Collinearity is often viewed as a situation in which two or more predictor variables are highly correlated. This statement is inadequate because correlation is a statistical property of random variables, and the regressor variables need not be stochastic, since they could represent preselected variable values in a designed experiment. When there are no preselected variable values the 'correlation' may be merely a characteristic of a particular data set in a particular time period, say, and might not be expected to occur in other data sets of the same type at other times.

Another view is that collinearities are due to 'weak' or 'deficient' data (Farrar and Glauber (1967)). 'Deficient' implies an aberration in the data-collection process; however, collinearities among predictor variables are sometimes an inherent property of the phenomenon under study, in which case 'deficient' data would actually be a misnomer.

In the foregoing statements, there are connotations which imply more about the predictor variables than simply the existence of a collinearity. A formal definition of collinearity should imply nothing more about the predictor variables than the existence of the collinearity. The following definition is due to Johnston (1963), Silvey (1969), Mason *et al.* (1975) and others, and is in terms of the linear dependence of a set of column vectors, X_j, of the matrix X.

Definition 2.1.1

Vectors $X_1, X_2, \ldots, X_p$ are linearly dependent if there exist non-zero constants $c_1, c_2, \ldots, c_p$ such that

$$\sum_{j=1}^{p} c_j X_j = 0 \qquad (2.1.1)$$

When the relationship in (2.1.1) is exact for a subset of the columns of X, exact collinearity exists. When (2.1.1) is only approximately true, near-collinearity is said to exist.

Gunst (1983) refined definition (2.1.1) as follows:

Definition 2.1.2

Let a multiple linear regression model be defined as in (1.1). If for some specified $\eta \geq 0$ there exists a vector $c' = [c_1 \ c_2 \ldots c_p]$, not all of whose elements are zero, such that

$$\sum_{j=1}^{p} c_j X_j = \mathit{0} \quad \text{with } \|\mathit{0}\| \leq \eta \ \|c\| \qquad (2.1.2)$$

then a collinearity is said to exist among the predictor variables in X.

For definition (2.1.2) to have a practical meaning η must be chosen suitably small. No fixed value of η is suitable for all regression analyses since the predictor variables are not scale invariant and η can be expected to depend on both n and p (the selection of η would be discussed in §2.2.12).

A very loose definition of collinearity in (1.1) is:

Definition 2.1.3

Collinearity exists among the predictor variables if the columns of X are not mutually orthogonal (the matrix X is said to have orthogonal regressors when it is such that $X'X$ is diagonal.). The strength of the collinearity is then determined by some measure of how close the columns of predictor variables are to being linearly dependent.

In definition 2.1.2 the strength of the collinearity can be gauged by the magnitude of $\|\theta\|/\|c\|$. In addition, the loose notion of a collinearity being simply equivalent to the non-orthogonality of X (i.e. collinearity exists when the correlation matrix $X'X$ is not diagonal) is replaced by equation (2.1.2). Clearly collinearity under (2.1.2) implies non-orthogonality, but not conversely.

Smith and Campbell (1980) maintain that high correlations and other measures of collinearity such as in definitions 2.1.2 and 2.1.3 are inadequate since the columns of predictor variables in X can always be transformed so they are mutually orthogonal; i.e., there exists a non-singular matrix T such that $A = XT$ and $A'A = I$. In doing so, they imply that model (1.1) is equivalent to the following model:

$$\begin{aligned} Y &= (XT)(T^{-1}\beta) + \epsilon \\ &= A\alpha + \epsilon \end{aligned} \qquad (2.1.3)$$

They claim that 'the parameters β and α are uniquely related by $\beta = T\alpha$ and their estimates should be also. It should make no diffference whether β is estimated explicitly or implicitly from $\beta = A\alpha$.'

Discussants of the Smith and Campbell paper point out that this type of reasoning ignores an important purpose of the regression analysis. One is usually interested in assessing the influence of the original predictor variables, not the transformed ones, on the response. Their arguments are summarized by Gunst (1983) as follows:

(i) the transformed predictor variables are linear combinations of variables which are not unit- or scale-free,
(ii) in a Bayesian or a sampling-theory framework it is usually easier to obtain prior information on the original model parameters than on arbitrary linear combinations of 'apples and oranges' (see Thisted (1980) section 2.3)

(iii) if one's interest is in specific parameters or parametric functions one must eventually transform back to the original predictor variables, and (iv) if the collinearity is severe enough to affect the computational accuracy of inversion of X it will also affect the computational accuracy of calculations to find and to invert T, in which case the Λ and α are not as well-observed as a superficial reading of (2.1.3) might imply.

The assertion that the collinearity problem is equivalent to a problem of high variances on transformed predictor variables can lead to uncritical, 'black box' mathematical manipulations which obscure the very purpose of a regression analysis: accurate and precise parameter estimation within a chosen and convenient model. Collinearity is an inherent aspect of the analysis and can not be transformed away.

Belsley and Oldford (1986, p104) give the following definition of ill-conditioning:

Suppose we have a system of continuous equations

$$\gamma = f(\varphi) \qquad (2.1.4)$$

where γ,φ,and f(.) are vectors and/or matrices. Then (2.1.4) can describe an estimator, a stochastic model, or, in general, any system of interest in which elements γ are assumed to be dependent upon elements φ. Suppose an additive perturbation $\delta\varphi$ in φ results in a perturbation in γ equal to $\delta\gamma = f(\varphi + \delta\varphi) - f(\varphi)$. A function $g(\delta\varphi) \equiv f(\varphi + \delta\varphi) - f(\varphi)$ may be defined which maps the elements of $\delta\gamma$ of a given domain Ω to elements $\delta\gamma$ in the corresponding range set Λ. That is,

$$g:\delta\varphi\rightarrow\delta\gamma$$

or

$$g:\Omega\rightarrow\Lambda$$

Let Λ^* consists of all those perturbations $\delta\gamma$ which are considered, *a priori*, to be reasonable given the set Ω. Concern arises when, corresponding to some $\delta\varphi$ in Ω, there exists $\delta\gamma$ in Λ which is not in Λ^*; i.e. small perturbations in $\delta\varphi$ result in perturbations $\delta\gamma$ which are not 'reasonably small' e.g. small changes in X give rise to large changes in $(X'X)^{-1}X'Y$. These considerations lead to the following terms and definitions.

Conditioning analysis: The specification of the conditioning triple $K = \{f,\Omega,\Lambda^*\}$ followed by a determination of whether $\gamma = f(\varphi)$ is ill-conditioned.

Ill-conditioning: Given K and its implied Λ, if $\Lambda \not\subseteq \Lambda^*$ then γ is said to be ill-conditioned with respect to φ (or Ω). Equivalently, one can call the system f ill-conditioned. (The concept of a condition number, denoted by K(X), as a measure of ill-conditioning will be defined and discussed in §2.2.6.)

The condition triple K completely specifies the conditioning analysis. Therefore clear and explicit specification of K is essential. The elements of K must be 'contextually or structurally interpretable', meaning that values assigned to the triplet must be interpretable through one's *a priori* knowledge of that situation (e.g. in economics it is usually *a priori* knowledge that allows one to assume some regressor variables may have an error margin of, say, one percent). The values assigned to the triple will depend on the kind of conditioning analysis. The authors distinguish between three kinds of conditioning for the linear model of (1.1):

(i) Data conditioning:

Let γ and φ be defined by

$$\gamma = f(\varphi) \equiv Z\varphi \tag{2.1.5}$$

(Here we use γ, $Z\varphi$ for generality, in the case of OLS their equivalent would be E(Y) or Y and $X\beta$).

Consider perturbations

$$\delta\gamma = g(\delta\varphi) \equiv Z\delta\varphi \tag{2.1.6}$$

with domain set

$$\Omega = \{\delta\varphi : \|\delta\varphi\|/\|\varphi\| = m_1\} \tag{2.1.7}$$

and acceptable response set

$$\Lambda^* = \{\delta\gamma : \|\delta\gamma\|/\|\gamma\| > m_2\} \tag{2.1.8}$$

That is, perturbations $\delta\varphi$ of fixed relative size m_1 are required to result in perturbations $\delta\gamma$ whose length is not less than m_2. (In the linear model context, these conditions imply that changes in β should give rise to changes in the data Y of E(Y), of some minimal size.) If this cannot be the case, then the data of Z are ill-conditioned with respect to Ω (2.1.7). Thus, Z is ill-conditioned if $\|\delta\varphi\|/\|\varphi\| = m_1$ and $\|\delta\gamma\|/\|\gamma\| < m_2$ which implies

$$\|\delta\varphi\| = m_1\|\varphi\| \text{ and } \|\delta\gamma\| < m_2\|\gamma\|$$

thus,

$$\|\delta\gamma\|/\|\delta\varphi\| < (m_2\|\gamma\|)/(m_1\|\varphi\|) \tag{2.1.9}$$

This inequality can be directly associated to the definition of collinearity given by Gunst in (2.1.2). From (2.1.2) the data set Z is ill-conditioned if

$$\|\theta\|/\|c\| \leq \eta$$

If we take

$$\eta = (m_2\|\gamma\|)/(m_1\|\varphi\|) \tag{2.1.10}$$

and replace $Zc = \theta$ by $Z\delta c = \delta\theta$ then we find the data set collinear according to Gunst's definition.

(ii) Estimator conditioning

In OLS the estimator to be examined is

$$\hat{\beta} = (X'X)^{-1}X'Y = X\dagger Y, \quad (X\dagger = (X'X)^{-1}X') \qquad (2.1.11)$$

In the notation of (2.1.4) $\gamma = \hat{\beta}$, and $f(\varphi) = f(X,Y)$. We consider three types of perturbation under this conditioning:

Firstly: If only X is perturbed, the triplet of the condition analysis is:

$$\Omega = \{\delta\varphi \equiv [\delta X, \delta Y] : \|\delta X\|/\|X\| \leq m_1, \ \|\delta Y\| = 0\} \qquad (2.1.12)$$

$$\Lambda^* = \{\delta\hat{\beta} : \|\delta\hat{\beta}\|/\|\hat{\beta}\| \leq m_2\} \qquad (2.1.13)$$

and

$$\delta\hat{\beta} = g(\delta X) = (X + \delta X)\dagger Y - X\dagger Y \qquad (2.1.14)$$

Belsley and Oldford (1986, p111) suggest choices of 0.01 for m_1 and m_2 approximately $20m_1$. If the range Λ of (2.1.14) based on the domain Ω of (2.1.12) contains any element not in Λ^*, the OLS estimate is ill-conditioned with respect to Ω. Thus if small relative changes in the X matrix can produce large relative changes in the estimate, the estimate is said to be ill-conditioned.

To determine if an OLS estimate is ill-conditioned we must calcualate

$$\sup \|\delta\hat{\beta}\|/\|\hat{\beta}\| \leq m_2 \text{ for } \delta X \in \Omega \qquad (2.1.15)$$

To evaluate (2.1.15) is difficult but Belsley and Oldford (1986) show that (2.1.15) is bounded from above by

$$m_1 K(X) R^{-1}[2 + (1 - R^2)^{1/2} K(X)] \tag{2.1.16}$$

where $K(X)$ is the condition number defined in (2.2.5), and R is the uncentered multiple correlation coefficient defined in Chapter one. As a rough guide the quantity $2m_1K(X)R^{-1}$ is compared to m_2, and if it is much larger than m_2, then $\hat{\beta}$ is declared ill-conditioned. Note that $K(X)$ is an important multiplicative factor: for example, 1 percent relative change in X produces approximately $K(X)$ percent change in $\|\delta\hat{\beta}\|/\|\hat{\beta}\|$.

Secondly : If only Y is perturbed, the triplet of the condition analysis is:

$$\Omega = \{\delta\varphi \equiv [\delta X, \delta Y] : \|\delta Y\|/\|Y\| \leq m_1,\ \|\delta X\| = 0\} \tag{2.1.17}$$

$$\Lambda^* = \{\delta\hat{\beta} : \|\delta\hat{\beta}\|/\|\hat{\beta}\| \leq m_2\} \tag{2.1.18}$$

$$\delta\hat{\beta} = g(\delta Y) = X^{\dagger}\delta Y \tag{2.1.19}$$

The relative bound is again the quantity $2m_1K(X)R^{-1}$.

Thirdly: Instead of perturbing Y around its observed value as above we perturb it against its theoretical or expected value, $X\beta$, thus $Y + \delta Y = X\beta + \epsilon$. The conditioning triplet would then be:

$$\Omega = \{\delta Y : \|\delta Y\|/\|X\beta\| \leq m_1\} \tag{2.1.20}$$

$$\Lambda^* = \{\delta\hat{\beta} : \|\delta\hat{\beta}\|/\|\beta\| \leq m_2\} \tag{2.1.21}$$

and

$$\delta\hat{\beta} = X^{\dagger}(Y + \delta Y) - X^{\dagger}Y = \hat{\beta} - \beta \tag{2.1.22}$$

The relative bound for m_2 is $2K(X)\|\epsilon\|/\|X\beta\| \leq 2m_1K(X)$.

Comment:

1. To distinguish this basis for a conditioning analysis from the others, it is called stochastically-based conditioning analysis.
2. Since $\|\delta\hat{\beta}\|^2 = (\hat{\beta} - \beta)'(\hat{\beta} - \beta)$, in determining whether $\hat{\beta}$ is ill-conditioned one also determines whether the maximal squared error of the resulting OLSE is less than some number $m_2\|\beta\|$.
3. m_1^{-1} is chosen to be the minimum 'signal-to-noise' ratio expected to be encountered in model (1.1).

(iii) Criterion conditioning

Parameters β are often estimated by minimizing some criterion function of the data and parameters. Let the criterion function be denoted by $Q(X,Y,\beta)$. In OLS, the criterion to be minimized is

$$Q(X,Y,\beta) = (Y - X\beta)'(Y - X\beta) \tag{2.1.23}$$

It is desirable that large changes $\delta\hat{\beta} \equiv \hat{\beta} - \beta$ from β should be detectable by the selected criterion function. The condition triplet is given by

$$\Omega = \{\delta\beta:\|\delta\hat{\beta}\| = m_1 > 0\} \tag{2.1.24}$$

$$\Lambda^* = \{\delta Q:\|\delta Q\|/(\inf\|\delta Q\|) \leq m_2\} \tag{2.1.25}$$

and

$$\delta Q = g(\delta\hat{\beta}) = 2\delta\hat{\beta}'(X'X)\delta\hat{\beta} \tag{2.1.26}$$

If $\|\hat{\beta}\| \neq 0$, then $\|\delta\hat{\beta}\|$ in (2.1.24) can be replaced by $\|\delta\hat{\beta}\|/\|\hat{\beta}\| = m_1$. The criterion (2.1.23) is ill-conditioned with respect to Ω if $K^2(X)$ is greater then m_2.

In (2.1.25) the worst possible effect is used in defining Λ^*, that is the worst effect denotes the boundary of acceptability. However, should perturbations $\delta\hat{\beta}$ in Ω produce δQ's which are not in Λ^*, then there exist $\hat{\beta}^*$ that differ substantially from $\hat{\beta}$ but which are relatively indistinguishable from $\hat{\beta}$ by criterion $Q(.)$. If $\inf\|\delta Q\| = 0$, for $\delta\hat{\beta} \epsilon \Omega$, then β is said to be inestimable with respect to the criterion. The concept of inestimability will be discussed further in §2.3 under the effects of collinearity.

In conclusion we note how all these conditioning criteria are connected to the condition number. The upper bounds given in the above discussion will be clearer after reading §2.5 on perturbation theory.

2.2 Detecting Collinearity

The distinction between defining and detecting collinearity is very thin. Some authors use ways of detecting collinearity as an implicit method of defining collinearity. For example, (i) if some or one singular value of X is small, then X is collinear; (ii) if the determinant of $X'X$ approaches zero, then X is near-singular; (iii) large VIF's or large condition numbers are taken as indicators of collinearity. The main issue is that the user of OLS must be aware of what is happening in the space of the regressor variables.

Many collinearity measures have been described and we discuss both historical and recent examples.

2.2.1 Nature and sensitivity of estimates

An unexpected sign in estimated coefficient of a model, or a low t-statistic, corresponding to a variable which for other than statistical reasons is viewed as an important regressor, or sensitivity of results to deletion of one or more rows or columns, will generally alert an informed reader to data or model inadequacy. These phenomena are sometimes cited as evidence of collinearity, or their occurrence is ascribed to collinearity.

However none of the phenomena are necessary for collinearity to exist, and they cannot therefore become definitive criteria for detecting collinearity, but rather should be grouped under its possible effects.

2.2.2 Correlation matrix of scaled regressors

Examining the correlation matrix, $X'X$, of the scaled regressors, is a commonly employed procedure because $X'X$ is a standard output of most regression packages. High values of the off-diagonal elements of $X'X$, i.e. $(X'X)_{ij}$, can be an indication of collinearity, between the two regressors X_i and X_j. This method can detect only pairwise collinearity and it is possible for three or more variates to be collinear while no two of the variates taken alone are highly correlated. Thus, the method may be helpful but not conclusive.

2.2.3 Determinant of X'X

When the X matrix is standardized, $0 \leq \det(X'X) \leq 1$, and for X with some columns exactly collinear, $X'X$ is singular and $\det(X'X) = 0$. On the other hand when $\det(X'X) = 1$ the columns of X are mutually orthogonal. These facts lead to the notion that in all other cases some degree of collinearity exists and becomes most severe as the determinant approaches zero.

Stewart (1987) disapproves of the use of the determinant to detect collinearities and describes its use as dependent upon the 'unhappy notion that det(X'X) bears a close relation to near collinearity'. The determinant is excessively sensitive to scaling, for example, $\det(kX'X) = k^p\det(X'X)$, A small determinant, may imply little or nothing about the invertibilty of a matrix. For example, the matrix kI_n, whose determinant is k^n and can be made arbitrarily small, has a simple inverse $[kI_n]^{-1} = k^{-1}I_n$, for $k \neq 0$.

2.2.4 Departure from orthogonality

Farrar and Glauber (1967) defined collinearity in terms of departures from orthogonality, because they claim that 'orthogonality lends itself easily to formulation as a statistical hypothesis'. The authors then present a three-stage test for the presence, location, and pattern of collinearity. The stages they describe (p104) are:

1) Compute det(X′X), and test for the severity and presence of collinearity. The authors assume X′X is in correlation form and transform det(X′X) into an approximate Chi Square statistic (see their equation (3)). They claim that 'a meaningful scale is provided against which departure from orthogonality, and hence the gradient between singularity and orthogonality, can be calibrated.'

2) In the second stage, the value of the statistic

$$((X'X)^{-1}_{ii} - 1)(n-p)/(p-1), \qquad (2.2.1)$$

where $(X'X)^{-1}_{ii}$ is the i-th diagonal element of $(X'X)^{-1}$, is computed and if the underlying distribution of the i-th column elements is normal, the statistic has the F-distribution with p-1 and n-p degrees of freedom. (For a complete discussion of the second stage see p102)

3) In the third stage, the authors define the coefficient of partial correlation (denoted by $R_{ij\bullet}$) between the i-th and j-th columns of X, while all other columns are held constant, as

$$R_{ij\bullet} = - (X'X)^{-1}_{ij}((X'X)^{-1}_{ii}(X'X)^{-1}_{jj})^{-\frac{1}{2}} \qquad (2.2.2)$$

A transformation of $R_{ij\bullet}$ has a known t-distribution, and this property, according to Farrar and Glauber, permits one to assess the significance of the computed $R_{ij\bullet}$.

The test proposed by Farrar and Glauber (1967) is rejected by several authors, e.g. Kumar (1975) and Belsley *et al.* (1980). The rejections are mostly based on the fact that Farrar and Glauber (1967) assume that the X matrix to be stochastic, and that the rows of the X matrix are independently distributed. O'Hagen and McCabe (1975) question the validity of 'statistical' interpretation of a measure of collinearity, and Belsley *et al.* (1980) added that 'there are no distributional implications from the linear regression model for specific null hypotheses (such as orthogonality) on the nature of the data matrix X, against which tests can be made.'

2.2.5 Smallest singular value

The technique of examining the smallest singular value is very popular amongst numerical analysts and complete discussions can be found in Lawson and Hanson (1974), Stewart (1973,1987) and Wilkinson (1965). Amongst statisticians this idea was suggested by Kendall (1957) and Silvey (1969). Basically one computes the SVD of X and examines the singular values, and collinearity is indicated by the presence of one or more 'small' singular values. Equivalently some authors use the eigenvalues of X′X which are just the squares of the singular values of X. We note that some authors mean by 'small' a singular value near zero, whereas others mean one or more singular values are 'small' in relation to others.

Stewart (1987) expresses the smallest singular value of X in terms of the Euclidean norm $\|.\|$ as

$$\inf(X) \overset{\text{def}}{=} \min \|X\nu\|, \quad \text{for } \|\nu\| = 1 \tag{2.2.3}$$

Its square is the smallest eigenvalue of the crossproduct matrix X′X. The justification is the following result due to Eckart and Young (1936), and generalized by Mirsky (1960):

> inf(X) is the spectral norm of the smallest matrix E such that
> X + E is exactly collinear. (2.2.4)

Thus inf(X) measures the absolute distance of X from collinearity. The fact that inf(X) is an absolute measure makes it difficult to interpret in the absence of information about the size of X, or putting it differently, it is difficult to know what is meant by the term 'small' which clearly implies a basis of comparison. This question leads to our next measure of collinearity, the condition number.

2.2.6 Condition number and condition index

The condition number, K(X), of a matrix X is defined in terms of the (matrix) spectral norm $\|.\|$ as

$$K(X) = \|X\| \, \|X^{\dagger}\| \tag{2.2.5}$$

where $X^{\dagger}$ is the pseudoinverse of X, i.e.

$$X^{\dagger} = (X'X)^{-1}X' \tag{2.2.6}$$

From the SVD, $X = U\Delta V'$, the SVD of $X^{\dagger}$ is $V\Delta^{\dagger}U'$, where $\Delta^{\dagger}$ is the pseudoinverse of Δ, specifically a diagonal matrix with non-zero entries taken as the reciprocals of nonzero diagonal elements of Δ. Hence the singular values of $X^{\dagger}$ are the reciprocals of those of X, whence $\inf(X) = \|X^{\dagger}\|^{-1}$ and it follows that

$$K^{-1}(X) = \frac{\inf(X)}{\|X\|} \tag{2.2.7}$$

Thus K^{-1} is just inf(X) scaled by the norm of X. In terms of the condition number, the Eckart-Young-Mirsky theorem (2.2.4) reads as follows:

The smallest matrix E for which X + E is collinear satisfies

$$\|E\|/\|X\| = K^{-1}(X) \tag{2.2.8}$$

$K^{-1}(X)$ gives a lower bound on the relative distance to collinearity.

The degree of ill-conditioning under the measure K(X) depends on how small the minimum singular value is relative to the maximum singular value. From (2.2.7) the condition number is always greater than one. Thus, using the definitions of a spectral norm, K(X) can be written as:

$$K(X) = \sqrt{\lambda_1}/\sqrt{\lambda_p} \geq 1 \qquad (2.2.9)$$

The lower bound of K(X) will be reached when all the columns of X are orthonormal. The condition number provides a measure of the sensitivity of the solution to the normal equations, to small changes in X or Y. Useful perturbation theory which has been cast in terms of the condition number, will be discussed in §2.5.

The condition number is extended to provide a set of the condition indices, and the i-th condition index is defined as

$$\eta_i = \sqrt{\lambda_1}/\sqrt{\lambda_i} \qquad i = 1,\ldots,p \qquad (2.2.10)$$

The maximum singular value ($\sqrt{\lambda_1}$) is used as a 'yardstick' against which smallness can be measured. The largest condition index is the condition number and the indices can be ordered as

$$K(X) = \eta_p \geq \eta_{p-1} \geq \cdots \geq \eta_2 \geq \eta_1 = 1$$

Any 'large' condition indices can indicate the presence of dependencies. The term 'large' in connection with condition indexes can not be fixed at a definite number. Empirically Belsley *et al.* (1980) suggest that condition indices around 5 - 10 indicate weak dependencies that may be starting to affect the regression estimates. Condition indices of 30 to 100 indicate moderate to strong dependencies, and indices larger than 100 indicate serious collinearity problems. Detecting collinearity by condition indices alone is not enough. Some other insight is necessary to find which columns of X are involved in the collinearity.

We may note that the condition number has its defects:

(i) it is too crude for statistical applications, as it uses matrix norms to distil a large amount of information into a single number and other distillations such as VIF's may be better. Belsley (1987), claims the 'full set of condition indices and variance-decomposition proportions (see §2.2.7) must be compared to other measures', and not the condition indices alone.

(ii) The condition number has its own scaling problem which will be discussed in §2.4.

To bring this numerical measure to bear on statistical questions as to what is the effect of the collinearity, we need methods such as the technique of Variance Decomposition due to Belsley *et al.* (1980).

2.2.7 Regression coefficient variance decomposition

Using the SVD, $X = U\Lambda V'$, the variance-covariance matrix of $\hat{\beta}$ is

$$\mathrm{Var}(\hat{\beta}) = \sigma^2 \sum_{i=1}^{p} v_i v_i'/\lambda_i$$

and for the j-th component of $\hat{\beta}$

$$\mathrm{Var}(\hat{\beta}_j) = \sigma^2 \sum_{i=1}^{p} v_{ij}^2/\lambda_i \qquad (2.2.11)$$

where v_{ij} is the j-th element of the i-th eigenvector.

Note that (2.2.11) decomposes $\mathrm{Var}(\hat{\beta}_j)$ into a sum of components. Those components associated with near dependencies (small λ_i), will be large relative to the other components. This fact suggests that an unusually high proportion of the variance associated with estimators of two or more coefficents (at least two columns of X are involved in the dependency), and provides evidence that the corresponding near dependency is causing problems. Not all regression coefficients need be affected. If the j-th variable is not significantly involved in the near-singularity, its coefficient in the i-th eigenvector, v_{ij}, will be near zero and its regression coefficient will remain stable even in the presence of the collinearity.

Define the i,j-th variance-decomposition proportion of the variance of the j-th regression coefficent associated with the i-th component as

$$\phi_{ij} = v_{ij}^2/\lambda_i \tag{2.2.12}$$

and

$$\phi_{TOTAL} = \sum_{i=1}^{p} v_{ij}^2/\lambda_i \tag{2.2.13}$$

Then the variance-decomposition proportions are

$$\pi_{ij} = \phi_{ij}/\phi_{TOTAL} \tag{2.2.14}$$

Setting up a summary table of the π_{ij}, patterns of high variance-decomposition proportions will be clear. Such a table will have the following form:

Associated	Proportions of			
Singular values	Var($\hat{\beta}_1$)	Var($\hat{\beta}_2$)	...	Var($\hat{\beta}_p$)
$\sqrt{\lambda}_1$	π_{11}	π_{12}	...	π_{1p}
$\sqrt{\lambda}_2$	π_{21}	π_{22}	...	π_{2p}
⋮	⋮	⋮		⋮
$\sqrt{\lambda}_p$	π_{p1}	π_{p2}	...	π_{pp}

A suggested diagnostic procedure for collinearity would then:

1. identify high condition indices (say, greater than 30), associated with
2. high variance-decomposition proportions (say, greater than 0.5) for two or more variances of estimated regression coefficients.

Once the variates involved in each near-dependency have been identified by their high variance-decomposition proportions, the near-dependency itself can be examined, e.g. by regressing one of the variates involved on the others.

In the case of multiple near-singularities, the variance decomposition table may be dominated by the linear transformation with the smallest eigenvalue so that the effect of other near-singularities may not be apparent. To overcome this problem, Rawlings (1988) suggested that the variance contribution of the other linear transformations should then be found by rescaling each column of the table so that the proportions add to one after each removal of a dominating linear transformation.

2.2.8 Mixed condition index

Another measure of collinearity suggested by Thisted (1980) involves the ratios of the squares of the eigenvalues and is defined as

$$\mathrm{mci} = \sum_{i=1}^{p} \lambda_p^2/\lambda_i^2 \qquad (2.2.15)$$

Values of mci near unity indicate high collinearity; values greater than 2.0 indicate little or no collinearity. In this collinearity measure it is the relative size of the smallest eigenvalue to all the others, not its absolute size or its size relative to the largest eigenvalue, that is important.

2.2.9 Variance inflation factors

VIF's were defined by Marquardt (1970) as the diagonal elements in the inverse of the correlation matrix of X'X. Thus the i-th variance inflation factor (VIF_i) is:

$$VIF_i = \frac{(X'X)^{-1}_{ii}}{(X_i'X_i)^{-1}} \geq 1 \qquad (2.2.16)$$

where $(X'X)^{-1}_{ii}$ denotes the i-th diagonal elements of $(X'X)^{-1}$ and X_i is the i-th column of X, not necessarily scaled (centered and standardized). When X is scaled (X'X is in correlation form) then the VIF_i can be written as

$$VIF_i = \frac{1}{1 - R_i^2} \qquad (2.2.17)$$

where R_i^2 is the coefficient of determination from the regression of X_i on the other independent variables. If there are near-singularities involving X_i and the other independent variables, $R_i^2 \simeq 1$ and VIF_i will be large. If X_i is orthogonal to the other independent variables, $R_i^2 \simeq 0$ and $VIF_i \simeq 1$. Indications of serious collinearity will be some $VIF_i > 10$ (if one exists, there are likely be two or more VIF's > 10).

Stewart (1987) demonstrated that the VIF's can be used to detect collinearity in the model (1.1) and called the square root of VIF_i the i-th collinearity index, denoted by κ_i:

$$\kappa_i = (VIF_i)^{\frac{1}{2}}$$

or $\qquad (2.2.18)$

$$VIF_i = \kappa_i^2$$

Stewart also derived and proved the following properties of κ_i:

(1) If X has unit column scaling then

$$\max(\kappa_i) \leq [\inf(X)]^{-1} \leq \sqrt{p}\max(\kappa_i)$$

and (2.2.19)

$$\max(\kappa_i) \leq K(X) \leq \sqrt{p}\max(\kappa_i)$$

(2) The smallest pertubation δX_j in X_j that will make X exactly collinear satisfies

$$\frac{\|\delta X_j\|}{\|X_j\|} = \kappa_j^{-1} \qquad (2.2.20)$$

(For a proof see Golub, Hoffman and Stewart (1987)). This statement is simplified to the following rule of thumb: 'if $\kappa_j^{-1} \cong 10^t$ then perturbations in the t-th digits of the components of X_j can make the problem collinear. Another way of saying the same thing is that one should be troubled about a model if the number of digits in κ_j is not less than the number of accurate digits in the components of X_j'

(3) For j = 1,2,...,p and $i \neq j$

$$\max(\kappa_i) \geq \sqrt{1+(\kappa_j^2-1)/(p-1)^2} \qquad (2.2.21)$$

Thus when there is one large coefficient there will be others. Therefore because collinearity is a 'group phenomenon' the naive use of the condition indices in selecting a variable to delete from an unsatisfactory model, is clearly incorrect.

However the condition indices can be used to assess the ill effects of near-collinearity on regression coefficients. Define the importance of the j-th variable

$$\mathrm{IMP}_j = 2\kappa_j\hat{\sigma}/\|Y\| \qquad (2.2.22)$$

where $\hat{\sigma}$ is the usual estimator of σ. Choose levels of importance, ι_j, (between 0 and 1), above which the X_j would be considered important, e.g. if $IMP_j > \iota_j$ reject the model. For example, a model which produce a IMP_j of 0.5 is unsatisfactory, since X_j accounts for 50% of the response and is judged insignificant.

Stewart (1987) also suggests the following regression diagnostics to assess the effects of errors in the variables:

Let the errors in the j-th column of X have mean μ_j and variance σ_j, and set

$$\begin{aligned} e_j &= (\mu_j^2 + \sigma_j^2)^{1/2}, && \text{if there is no constant term,} \\ &= \sigma_j && \text{otherwise} \end{aligned}$$

Compute

$$\tau_j = [(n-p)^{1/2}\kappa_j \ e_j]/\|X_j\| \qquad (2.2.23)$$

If $\tau_j > 1/3$, reject the model. The errors are so influential that the diagnostic procedure cannot be trusted. For more measures to judge the errors see pp75-77 of his paper.

Schall and Dunne (1987b) generalize the VIF's so that they can be associated with more than one parameter, and more generally, with arbitrary sets of linear functions of β. This generalization, and the distinction of marginal and partial VIF's result in a complete set of collinearity diagnostics. Thus, the generalized set of variance inflation factors can be used to perform the three main tasks of collinearity diagnostics: (i) they measure variance inflation in the parameter estimates (ii) they can be used as collinearity indices and (iii) they can detect the nature and the sources of collinearity.

2.2.9.1 VIF as variance ratio

Under model (1.1) the variance of $\hat{\beta}_i$ is given by

$$V(\hat{\beta}_i) = \sigma^2(X'X)^{-1}_{ii}$$

and to avoid confusion $V(\hat{\beta}_i)$ will be written as $V(X)$ meaning variance of $\hat{\beta}_i$ under the full model (1.1). Context will allow us to dispense with the subscript.

Clearly $V(\hat{\beta}_i)$ can also be written as

$$V(\hat{\beta}_i) = \frac{\sigma^2}{(X_i{}'X_i)^{-1}} \text{VIF}_i \qquad (2.2.24)$$

showing that the variance of the i-th regression coefficient is directly proportional the VIF_i: If VIF_i is large the $V(\hat{\beta}_i)$ will be inflated accordingly.

Consider the model

$$Y - X_{(i)}\beta_{(i)} = X_i\beta_i + \epsilon \qquad (2.2.25)$$

where $X_{(i)}$ and $\beta_{(i)}$ are obtained by dropping the i-th column from X and the i-th component from β, and where the parameters $\beta_{(i)}$ are assumed to be known. Thus the left hand side of (2.2.25) is observable. Under (2.2.25) the variance of β_i, denoted by $V(X_i)$ is given by

$$V(X_i) = \sigma^2(X_i'X_i)^{-1} \qquad (2.2.26)$$

The i-th variance inflation factor can be written as the ratio

$$\text{VIF}_i = \frac{(X'X)^{-1}_{ii}}{(X_i'X_i)^{-1}} = \frac{\sigma^2(X'X)^{-1}_{ii}}{\sigma^2(X_i'X_i)^{-1}} = \frac{V(X)}{V(X_i)} \qquad (2.2.27)$$

Thus, the VIF_i can be described as the loss of information on β_i due to having the covariates $X_{(i)}$ and the unknown parameters $\beta_{(i)}$ in the model (1.1) as compared to the model (2.2.25). Expression of VIF's as variance ratios can be further factorized (Schall and Dunne (1987b), p4) as follows:

Let $X_{(i)}$ be partitioned as

$$X_{(i)} = [X_1 \; X_2] \tag{2.2.28}$$

and denote the variance inflation (VIF_i) due to the covariates $X_{(i)}$ as

$$VIF_i = VIF_i(X_{(i)}) = VIF_i([X_1 \; X_2]) \tag{2.2.29}$$

Thus

$$\begin{aligned} VIF_i &= VIF_i([X_1 \; X_2]) \\ &= \frac{V([X_1 \; X_2 \; X_i])}{V(X_i)} \\ &= \frac{V([X_1 \; X_i])}{V(X_i)} \cdot \frac{V([X_1 \; X_2 \; X_i])}{V([X_1 \; X_i])} \\ &= VIF_i(X_1).VIF_i(X_2|X_1) \end{aligned} \tag{2.2.30}$$

$VIF_i(X_2|X_1)$ is a partial variance inflation factor (PVIF), where as $VIF_i(X_1)$ is a marginal variance inflation factor (MVIF). The factorization provides a refinement in the collinearity diagnostics. If (say) $VIF_i(X_1)$ is large but $VIF_i(X_2|X_1)$ is small (close to one), then collinearity exists between $X_{(i)}$ and X_1, and that given X_1 in the model, having X_2 in the model does not further inflate the variance of $\hat{\beta}_i$.

The authors then generalize the above factorization with respect to single components to a factorization for parameter subsets. Let $I \subset \{1,...p\}$

denote the index set specifying the parameter subset β_I, and X be partitioned as $X = [X_{(I)}\ X_I]$. Then model (2.2.25) can be written as

$$Y - X_{(I)}\beta_{(I)} = X_I\beta_I + \epsilon \tag{2.2.31}$$

The generalized variance inflation factor (GVIF) VIF_I is defined as

$$VIF_I = \frac{|(X'X)^{-1}_{II}|}{|(X_I'X_I)^{-1}|} \tag{2.2.32}$$

where $(X'X)^{-1}_{II}$ denotes the submatrix of $(X'X)^{-1}$ formed by the elements which fall into the rows and columns indexed by I.

2.2.9.2 Use of VIF's

Firstly: Identify disconnected subsets of variables

By disconnected subsets we mean that collinearity may exist within a subset but not between subsets. This identification is performed by partitioning of $X = [X_{(I)}\ X_I]$, and the subsets $X_{(I)}$ and X_I are said to be disconnected if the $GVIF_I$ is small. This procedure is repeated to find disconnected subsets of variables in both $X_{(I)}$ and X_I. If $s \leq p$ disconnected subsets are found then the number $c \leq s$ are subsets containing more than one variable is the 'number of collinearities '.

Secondly: Factorize within each of the c subsets

Insight into the nature of the collinearities can be found by looking at the c subsets individually. Each of them is partitioned as described above. For example assume one disconnected collinear subset is $[X_1\ X_2\ X_i]$ then VIF_i can be partitioned as in (2.2.30) as

$$\begin{aligned} VIF_i &= VIF_i(X_1).VIF_i(X_2|X_1) \\ &= VIF_i(X_2).VIF_i(X_1|X_2) \end{aligned}$$

If $VIF_i(X_1)$ is large, then X_i and X_1 are collinear; similarly, if $VIF_i(X_2)$ is large then X_i and X_2 are collinear. If both MVIF's are small, then the PVIF's $VIF_i(X_2|X_1)$ and $VIF_i(X_1|X_2)$ must be large, and the collinearity involves X_i and at least one variable from both X_1 and X_2.

Computations of all the GVIF's can be obtained easily from the correlation matrix of each step in an all subsets regression problem. In the case where p is too large for all subsets regression a subset of all possible GVIF's can be obtained from stepwise regression procedure.

It is recommended by several authors (Stewart (1987), Schall and Dunne (1987b)) that the X matrix should be mean centered before calculating the VIF's (see also §2.4)

2.2.10 Signal-to-noise tests

Belsley (1982) suggested that the presence of harmful collinearity and other forms of weak data could be assessed through a test for signal-to-noise. In this test the size of the parameter variance (noise) is assessed relative to the magnitude of the parameter (signal). This test is then combined with other collinearity diagnostics (condition number, variance-decomposition) to provide a test for the presence of harmful collinearity and/or short data.

The general test is constructed as follow:

Assume model (1.1) and the following partitioning of the model

$$Y = X_a\beta_a + X_b\beta_b + \epsilon \tag{2.2.33}$$

where X_a:nxp_a, β_a:p_ax1, X_b:nxp_b, β_b:p_bx1 and $p_a+p_b= p$. Then the marginal distribution of the estimate of β_b is

$$\hat{\beta}_b \sim N(\beta_b, V(\hat{\beta}_b|X)) \tag{2.2.34}$$

where $V(\hat{\beta}_b|X)$ is the variance-covariance matrix of $\hat{\beta}_b$ conditional on X and is

$$V(\hat{\beta}_b|X) = \sigma^2(X_b'M_aX_b)^{-1} \qquad (2.2.35)$$

where $M_a = I - X_a(X_a'X_a)^{-1}X_a'$ and $(X_a'X_a)^{-1}$ is assumed to exist.

Let $\overset{*}{\beta}_b$ be any p_b vector. Then the signal-to-noise of the OLSE $\hat{\beta}_b$ relative to $\overset{*}{\beta}_b$ is defined as

$$\tau^2 = (\beta_b - \overset{*}{\beta}_b)'[V(\hat{\beta}_b|X)]^{-1}(\beta_b - \overset{*}{\beta}_b) \qquad (2.2.36)$$

We note that (2.2.36) reduces to $\tau = \beta_b/[V(\hat{\beta}_b|X)]^{\frac{1}{2}}$, when $\overset{*}{\beta}_b = 0$, $p_b = 1$ and β_b is simply the b-th element of β. The inverse of τ is often called the coefficient of variation, but τ itself is the non-centrality parameter of the non-central t-distribution. The term signal-to-noise ratio refers to the magnitude of τ, but the authors use the same name for τ^2.

Belsley (1982) shows that

$$\phi^2 = (\hat{\beta}_b - \overset{*}{\beta}_b)'(X_b'M_aX_b)(\hat{\beta}_b - \overset{*}{\beta}_b)/\{p_b\hat{\sigma}^2\} \sim F(p_b,n-p;\tau^2)$$

a non-central F with p_b and n-p degrees of freedom and non-centrality parameter τ^2. Hence, under $H_0 : \tau^2 = \tau_*^2$, we have

$$\phi^2 \sim F(p_b,n-p;\tau_*^2)$$

A practical drawback of this test is that it requires knowledge of β_b and $V(\hat{\beta}_b|X)$ to stipulate directly a value for τ_*^2. The authors propose a practical definition for an 'adequate level' for the signal to noise ratio that does not require β_b and $V(\hat{\beta}_b|X)$. This measure is an increasing function of a single, selectable parameter $\gamma \epsilon [0,1)$, and can be made stringent (γ chosen near unity) or relaxed (γ chosen small).

The (1-α) critical F-values for the test of

$$H_0 : \tau^2 = \tau_*^2 \quad \text{against}$$

$$H_a : \tau^2 > \tau_*^2$$

are tabulated in Belsley (1982, appendix A), where γ is chosen by the experimenter.

The authors then set up a strategy to determine harmful collinearity and short data, using results from the sequence of (1) the collinearity diagnostics (condition indices and varance-decomposition) of Belsley *et al.* (1980) followed by (2) a test for adequate signal-to-noise for the $\hat{\beta}_b$ (usually the signals are calculated for individual $\hat{\beta}_i$'s) . The four possible outcomes of these tests are given below:

		Collinearity present	
		No	Yes
Inadequacy of Signal-to-noise	No	1	2
	Yes	3	4

In situations 1 to 4 we may say:

(1) everything seems acceptable

(2) collinearity is present, but has not resulted in inadequate signal-to-noise. This situation augurs well for the use of the model for prediction purposes, particularly, but not necessarily only, if the collinear relations continue into the prediction period.

(3) 'the length of X_b is short, e.g. short data', data problems exist, but collinearity is not the culprit.

(4) structural estimation may be adversely affected, but prediction is not necessarily affected if the collinear relations extend into the prediction period.

2.2.11 Prior information use

Swamy *et al.* (1985) argue that any method of detecting collinearity which does not use prior information about β may not be successful in diagnosing the presence of real collinearity. The measure of collinearity should take into account the dependence of $\hat{\beta}$ on $X'X$ as well as the prior and sample information about β as a primary characteristic.

They considered two types of prior information. A non-Bayesian form may investigate $W\beta$ lying on or within the ellipsoid $(\beta - \bar{\beta})'W'\Delta_1^{-1}W(\beta - \bar{\beta}) = r^2$ with known W, $\bar{\beta}$, Δ_1 and r. To incorporate the non-Bayesian ellipsoid of prior information the method of constrained least squares is proposed. The constrained LSE of β (denoted by $\hat{\beta}_c$) subject to $(\beta - \bar{\beta})'W'\Delta_1^{-1}W(\beta - \bar{\beta}) = r^2$ is

$$\hat{\beta}_c = (X'X + \sigma^2\mu W'\Delta_1^{-1}W)^{-1}(X'Y + \sigma^2\mu W'\Delta_1^{-1}W\bar{\beta}) \qquad (2.2.37)$$

where μ is chosen such that $(\hat{\beta}_c - \bar{\beta})'W'\Delta_1^{-1}W(\hat{\beta}_c - \bar{\beta}) = r^2$ (see Swamy *et al.* (1985), p406) for remarks on μ).

Swamy *et al.* (1985) then propose a specific biased version of the estimator in (2.2.37)

$$\hat{\beta}_s = (X'X + s^2\hat{\mu}W'\Delta_1^{-1}W)^{-1}X'Y \qquad (2.2.38)$$

where $s^2 = (Y - X\hat{\beta})'(Y - X\hat{\beta})/(n-p)$ and $\hat{\mu}$ is the value of μ selected according to an empirical rule of Thurman *et al.* (1984). Usually, $\hat{\mu}$ determined by this rule will be a finite, positive quantity (see Swamy *et al.*(1985, p407)). (Although the authors do not state it explicity, they imply that in (2.2.38) the $\bar{\beta}$ is 0).

The modified coefficient of multiple determination is defined as

$$R_s^2 = (\hat{\beta}_s{}'X'X\hat{\beta}_s + 2\hat{\beta}_s{}'X'\hat{\epsilon}_s)/(Y'Y) \qquad (2.2.39)$$

where $\hat{\epsilon}_s = Y - X\hat{\beta}_s$. The 'truncated model' is

$$Y = X_{(h)}\beta_{(h)} + \epsilon_h \qquad (2.2.40)$$

in which the h-th independent variable is not included but the other p-1 variables are included. Then the biased estimates of the truncated model are

$$(\hat{\beta}_s)_{-h} = [X'_{(h)}X_{(h)} + s^2\hat{\mu}(W'\Delta_1^{-1}W)_{-h},_{-h}]^{-1}X'_{(h)}Y \qquad (2.2.41)$$

where $X_{(h)}$is obtained by deleting the h-th column from X, and $(W'\Delta_1^{-1}W)_{-h},_{-h}$ is obtained by deleting the h-th row and the h-th column from $(W'\Delta_1^{-1}W)$. The corresponding coefficient of determination is obtained from

$$[1 - (R_s)^2_{-h}] = (\hat{\epsilon}_s)_{-h}{}'(\hat{\epsilon}_s)_{-h} \qquad (2.2.42)$$

The modified incremental contribution of the h-th variable is then

$$R_s^2 - (R_s)^2_{-h}$$

Based on this increment their measure of collinearity (denoted by $\tilde{m}$) is defined as

$$\tilde{m} = R_s^2 - \sum_{h=1}^{p} (R_s^2 - (R_s)^2_{-h}) \qquad (2.2.43)$$

The authors give a lengthy discussion of the bounds of $\tilde{m}$ (see p410-411) and we only give the computationally simpler bounds for $\tilde{m}$ here, namely

$$m_L = R_s^2 - p[\max(R_s^2 - (R_s)_{-h}^2)], \text{ for } 1 \leq h \leq p \qquad (2.2.44)$$

$$m_U = R_s^2 - p[\min(R_s^2 - (R_s)_{-h}^2)], \text{ for } 1 \leq h \leq p \qquad (2.2.45)$$

The value of $\tilde{m}$ is interpreted as follows:

(1) If $\tilde{m} \cong 0$ then collinearity is absent.

(2) If $X'X$ is non-diagonal $|\tilde{m}|$ is closer to the value of m_L or m_U than to zero and then serious collinearity exists.

A Bayesian view may assume $W\beta$ distributed *a priori* with mean $W\bar{\beta}$ and covariance matrix $W\Delta W'$, where W is a known rectangular matrix with full row rank. Here $\bar{\beta}$ and Δ may be unknown. When the Bayesian prior distribution of $W\beta$ is available the authors derive the posterior mean of β as

$$\hat{\beta}_G = E\beta + \text{cov}(\beta)[\sigma^2(X'X)^{-1} + \text{cov}(\beta)]^{-1}(\hat{\beta} - E\beta) \qquad (2.2.46)$$

where $E\beta$ and $\text{cov}(\beta)$ are respectively the prior mean and the prior covariance matrix of β. For a derivation of (2.2.46) see Thurman *et al.* (1984). Kashyap *et al.* (1984) give an operational version of (2.2.46) which is nearly minimax.

2.2.12 Detection methods and Mason's definition

In §2.1, Mason's definition of collinearity is

$$\sum_{j=1}^{p} c_j X_j = \theta \quad \text{with } \|\theta\| \leq \eta \, \|c\| \qquad (2.1.2)$$

In this whole section we will assume X is scaled so that $X'X$ is in correlation form. Then the above definition (2.1.2) is related to the measures of detection.

Correlation coefficients: Let r_{ij} denote the correlation coefficient between X_i and X_j, as $|r_{ij}| \to 1$ the vectors X_i and X_j approach linear dependence:

$$X_i - \text{sign}(r_{ij})X_j = \theta \quad \text{with } \|\theta\| = \sqrt{2}\ (1 - |r_{ij}|)^{\frac{1}{2}} \qquad (2.2.47)$$

From definition (2.1.2), if $|r_{ij}| \geq 1 - \eta^2$ then a pairwise collinearity exists between X_i and X_j. The choice of η will depend on the goals of the research project.

Variance inflation factors: Consider the definition of VIF_i

$$\text{VIF}_i = \frac{1}{1 - R_i^2} \qquad (2.2.17)$$

where R_i^2 is the coefficient of determination from the regression of X_i on the other independent variables. Denote by $X_{(i)}$ the X matrix with the i-th column deleted and let $b_i = (X_{(i)}'X_{(i)})^{-1}X_{(i)}'X_i$ be the OLS regression coefficient when X_i is regressed on the remaining p-1 columns of X (i.e. on $X_{(i)}$). The corresponding 'residual vector' is then

$$X_i - X_{(i)}b_i = \theta \quad \text{with } \|\theta\| = (1 - R_i^2)^{\frac{1}{2}} \qquad (2.2.48)$$

Thus if $(1 - R_i^2)^{\frac{1}{2}}$ ($= (\text{VIF}_i)^{-\frac{1}{2}}$) $\leq \eta\ (1 + b_i'b_i)^{\frac{1}{2}}$, equation (2.2.48) defines a collinearity among the predictor variable.

The smallest singular values: From (1.3.2) the SVD of X is

$$X = \sum_{i=1}^{p} \sqrt{\lambda}_i u_i v_i'$$

so that

$$Xv_j = \sqrt{\lambda}_j u_j \qquad j = 1,2,\ldots,p. \tag{2.2.49}$$

and

$$Xv_j = \sum_{i=1}^{p} X_i v_{j\,i}$$

where $v_{j\,i}$ is the i-th component of the j-th singular vector v_j. If $\sqrt{\lambda}_j$ is suitably small, equation (2.2.49) defines a collinearity since

$$\sum_{i=1}^{p} X_i v_{j\,i} = \theta \;\text{ with } \|\theta\| = \sqrt{\lambda}_j \tag{2.2.50}$$

satisfies definition (2.1.2) when $\sqrt{\lambda}_j \leq \eta$. The authors suggested a cutoff value for η is about 0.3 (i.e. $\lambda_j \leq \eta^2 = 0.1$).

The condition number: If a condition number of 30 is selected as a cutoff value for collinearity, then $\eta = \sqrt{\lambda}_1/30$ is the appropriate cutoff for $\sqrt{\lambda}_p$. Belsley and Oldford's results in the section on conditioning analysis, §2.1, can be related to Mason's approach.

2.3 Collinearity in Practice

In practice collinearity becomes harmful when estimation or hypothesis testing is influenced more by the relationship between the regressor variables than by the relationship between the response and the regressor variables. Such an influence can result in poor parameter estimates and restrictions on the applicability and generality of the model in use.

2.3.1 Sources and origins

Collinearity or near-singularity may arise in several ways (for detailed discussions see Mason, Gunst and Webster (1975) and Rawlings (1988):

1. An over-defined model is one in which there are more regressor variables than observations. This type of model arises frequently in medical research where many elements of information are recorded on each individual in a study.

2. An in-built mathematical constraint in variables that forces them to add to a constant will generate a collinearity. Generating new variables as transformation of other variables can produce a collinearity among the set of variables involved e.g. ratios or powers of variables frequently may be nearly collinear with the original variables.
3. Component variables of a system may show near linear dependencies because of biological or physical constraints of the system (e.g. various measures of size of an organism will show dependencies). Such correlation structures are properties of the system and can be expected to be present in all observations obtained from the system. Gunst (1983) refered to this type of collinearity as 'population-inherent collinearities'
4. Inadequate sampling occurs when the experimenter unknowingly samples only from a subspace of the space of the regressor variables. Collinearities due to sampling deficiencies are a property of the particular data set which has been collected and would not be expected to occur in data sets arising from alternative sampling.
5. Poor experimental design may give rise to collinearities. If possible the levels of the experimental factors are generally chosen in such a way so that the different treatment factors are statistically orthogonal to each other.
6. Outliers can induce artificial collinearites among the predictor variables, and will be discussed in Chapter 8.

Identifying the origin of collinearity is not always possible but it is important to illustrate likely sources in each instance.

2.3.2 Effects of collinearity

The impact of collinearity on least squares methodology is very serious if primary interest is in the regression coefficients or if the purpose is to identify 'important' variables in the estimation process. The solution is very unstable, i.e. small changes (random noise or rounding effects) in the Y or X, can cause drastic changes in the estimates of the regression coefficients (e.g. change in sign), and the variances of the regression coefficients for the regressor variables involved in the near-singularity, become very large.

In discussing effects of collinearities, the notation of Chapter 1 will be used; for convenience we repeat the following:

The variance covariance matrix of the OLS estimator is given by

$$V(\hat{\beta}) = \sigma^2 \sum_{i=1}^{p} v_i v_i'/\lambda_i . \tag{2.3.1}$$

Because the bias is zero the

$$MSE(\hat{\beta}) = \sigma^2 \sum_{i=1}^{p} v_i v_i'/\lambda_i \tag{2.3.2}$$

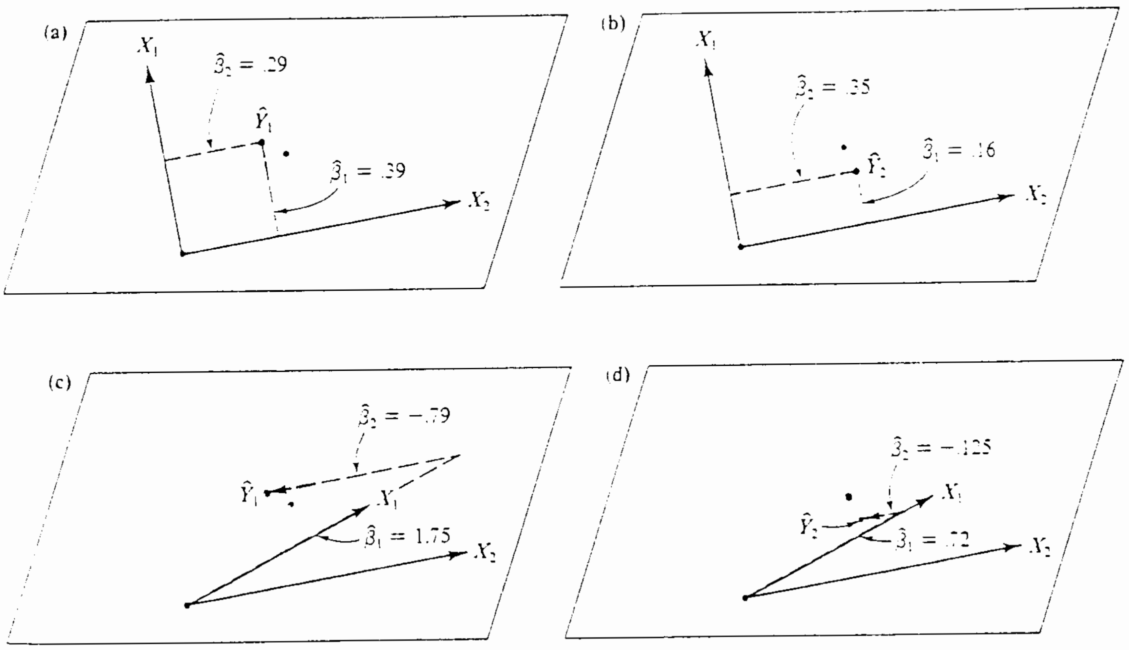
and the TMSE (which is the same as $E(L_1^2)$ in section 1.2) is then

$$TMSE(\hat{\beta}) = \sigma^2 \sum_{i=1}^{p} 1/\lambda_i > \sigma^2/\lambda_p \tag{2.3.3}$$

2.3.2.1 Geometric interpretation of collinearity

The impact of collinearity can be illustrated geometrically. The figure on p2-37 and the interpretation below are from Rawlings (1988, p163).

The plane is the X-space and the heavy dot, with the shaded area around it, represents E(Y). The shaded area represents the distribution of projections $\hat{Y}$ of Y, onto the X-space which one might obtain from repeated samplings of the dependent variable. Panels (a) and (b) represent the case where X_1 and X_2 are orthogonal. Panels (c) and (d) represent the case where X_1 and X_2 are nearly collinear: the angle between the vectors is small. The position of X_2 relative to E(Y) remains the same in all cases; the position of X_1 has been shifted to create the collinearity.



2.3.2.2. Collinearity of two regressor variables

Assume X′X is in correlation form, then

$$(X'X)^{-1} = \begin{bmatrix} 1/(1-r_{12}^2) & -r_{12}/(1-r_{12}^2) \\ -r_{12}/(1-r_{12}^2) & 1/(1-r_{12}^2) \end{bmatrix} \qquad (2.3.4)$$

where r_{12} is the sample correlation coefficient between X_1 and X_2. When the correlation between X_1 and X_2 increases ($r_{12}^2 \rightarrow 1$) and applying L'Hospital's rule

$$\mathrm{Var}(\hat{\beta}_j) \rightarrow \infty \quad \text{as } r_{12}^2 \rightarrow 1$$

and (2.3.5)

$$\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) \rightarrow \pm\, \infty, \text{ depending on whether } r_{12} \rightarrow \pm\, 1$$

A strong pairwise linear relationship between X_1 and X_2 results in very large variances and covariances for the estimators of the regression coefficients. Sastry (1970) was the first to show these limits. His equation (7) is missing a square on the denominator term, $(1-r_{12}^2)$ and thus the limit he derived in his equation (8) is incorrect.

For the two regressor variables the estimator of the regression coefficients are

$$\hat{\beta} = (X'X)^{-1}X'Y = \begin{bmatrix} 1/(1-r_{12}^2)X_1'Y - r_{12}/(1-r_{12}^2)X_2'Y \\ -r_{12}/(1-r_{12}^2)X_1'Y + 1/(1-r_{12}^2)X_2'Y \end{bmatrix}$$

$$= \begin{bmatrix} (X_1'Y - r_{12}X_2'Y)/(1-r_{12}^2) \\ (-r_{12}X_1'Y+X_2'Y)/(1-r_{12}^2) \end{bmatrix} \qquad (2.3.6)$$

Mason *et al.* (1975) (by using the results in Sastry (1970)) show that as $r_{12} \to 1$, and assuming that $X_1'Y$ and $X_2'Y$ becomes identical (the equality is justified since if X_1 and X_2 have perfect correlation between them, they would each have the same correlation with Y). If $r_{12} \to -1$, the correlation between X_1 and Y and between X_2 and Y would be identical but opposite in sign, $\hat{\beta}_1 \to X_1'Y/2$ and $\hat{\beta}_2 \to -X_1'Y/2$. The limit operation forces $\hat{\beta}_1$ and $\hat{\beta}_2$ to become equal but opposite in sign regardless of the true parameter values β_1 and β_2.

The t-ratio for the partial regression coefficient of the first prediction limit for this model (given the second predictor) is

$$t^2 = \frac{r_{1y}^2(1-r_{12})^2(n-3)}{1+2r_{1y}^2r_{12}-2r_{1y}^2-r_{12}^2}$$

where r_{1y} is the sample correlation coefficient between X_1 and Y, r_{2y} is the sample correlation coefficient between X_2 and Y, and we let $r_{1y} = r_{2y}$, when r_{12} approaches 1. The limit of the t-ratio for $|r_{1y}| \neq 0$ is (Crocker (1971)):

$$\lim_{r_{12}\to 1} t^2 = \lim_{r_{12}\to 1} \frac{r_{1y}^2(1-r_{12})^2(n-3)}{(1-r_{12})(1+r_{12}-2r_{1y}^2)}$$

$$= \lim_{r_{12}\to 1} \frac{r_{1y}^2(1-r_{12})(n-3)}{(1+r_{12}-2r_{1y}^2)}$$

$$= \frac{0}{2(1-r_{1y}^2)} = 0$$

2.3.2.3 Inflation of variance

In the presence of near collinearity $\lambda_p \to 0$ so that the var($\hat{\beta}$) is inflated and TMSE($\hat{\beta}$) $\to \infty$. From (2.2.24) the individual variance of the i-th element of $\hat{\beta}$ is

$$V(\hat{\beta}_i) = \frac{\sigma^2}{(X_i{}'X_i)^{-1}} \text{VIF}_i \qquad \text{for } i = 1,2,\ldots p \qquad (2.3.7)$$

Thus the variance of the estimator of the i-th regression coefficient is directly proportional to VIF_i's. If VIF_i is large (indicating collinearity) then the variance will be inflated as well. In the case of a near orthogonal design VIF $\approx$ 1, and there is no effect on the variance.

Inflation of the variance will also mean that the null hypothesis $H_0:\beta_i=0$ will be more likely to be accepted. For a detail discussion on parametric inference see Gunst (1983).

2.3.2.4 Unexpected coefficient values and signs

Collinearities can result in $\hat{\beta}_i$ to 'have the wrong sign' (Farrar and Glauber (1967)), and magnitudes of values that disagree with well-established theory of previous empirical studies. The notion 'wrong sign' can only be well-defined in a Bayesian framework where the 'correct' sign can be assumed to be known from a prior distribution. Mullet (1976) pointed out that the 'wrong sign' need not be the result of collinearity. Other possible explanations for an incorrect sign are: (i) limited range of regressor variable values, (ii) model misspecification, and (iii) computational error. To these we may add (iv) outliers in the response variable and (v) influential cases.

We illustrate the effect of collinearity by considering the OLSE which, from (1.3.5), can be written as

$$\hat{\beta} = \sum_{i=1}^{p} v_i u_i{}'Y/\sqrt{\lambda}_i$$

$$= \sum_{i=1}^{p} v_i c_i/\lambda_i \quad \text{where } c_i = u_i{}'Y\sqrt{\lambda}_i \qquad (2.3.8)$$

Assume in the SVD of X that the eigenvalues are ordered, e.g.

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{p-1} \geq \lambda_p$$

Suppose that λ_p is much smaller than λ_{p-1} (implying a single very strong collinearity), so much smaller that the summation in (2.3.8) is completely dominated by it, e.g.

$$\hat{\beta} \simeq v_p c_p/\lambda_p.$$

Then we may infer

$$\hat{\beta}_j \simeq v_{pj} c_p/\lambda_p \qquad (2.3.9)$$

Gunst and Mason (1980) claim two characteristics of expression (2.3.9) when a strong collinearity occurs in X and $v_{pj} \neq 0$:

(i) the estimates tend to be large in magnitude due to the multiplier λ_p^{-1}, unless λ_p^{-1} is complemented by a small value of c_p or v_{pj}

(ii) the signs of the estimates tend to be determined more by the collinearity associated with v_p than by relationship of the predictor variables with the response: i.e. if $c_1 > 0$, the sign of $\hat{\beta}_j$ is the same as that of v_{pj}; if $c_1 < 0$, the sign of $\hat{\beta}_j$ is opposite that of v_{pj} .

The second of these claims may be something of an overstatement, because the SVD admits $(-u_i, -v_i)$ in place of (u_i, v_i), and the relationship of the predicted values and the predictor variables with the response is certainly implicit in $c_i = u_i'Y$, for each value of i, even for i = p. It may be better to say that the inherent impression associated with λ_p being as small as suggested, implies that the notion of the sign of the regression coefficients is also imprecise, in that small stochastic variation in Y may result in substantial changes in coefficient estimates, including some sufficiently large as to give rise to apparent sign changes.

2.3.2.5 Unstable regression coefficients

In the presence of collinearity a small perturbation in X or Y can result in a relatively unstable regression coefficient $\hat{\beta}_i$. In §2.2, we have given some insight into situations when perturbation in X will cause harmful collinearity and further results on perturbation will be given in §2.5

2.3.2.6 Linear combinations of regression variables

Poor precision in the estimation of individual parameters does not imply that the estimated model is a poor predictor. Although some individual parameters may be estimated poorly, the Y value may be predicted adequately as the whole vector of $\hat{\beta}_i$'s is used. When collinearity persists into the prediction area the collinearity is not harmful. Use of the model outside the defined field (extrapolation) will result in poor prediction.

2.4. Centering and Standardization of the X matrix

In the literature several conflicting views appear on the question of whether data in the X-matrix should be mean-centered before collinearity is assessed. Belsley (1984) contrasts with authors like Stewart (1987), Schall and Dunne (1987b), Gunst (1983), Marquardt (1980) and Marquardt and Snee (1975) who advocate mean centering. There is less argument on the question on whether X should be standardized although the question of 'how the standardizing must be done could be vague'. Stewart (1987) pointed out that any combination of three elements could be standardized: the matrix X, the vector β (its elements should be close together), or the matrix E (defined in (2.2.4).

The standardizing of X is accomplished by dividing the elements of each column vector by the square root of the sum of squares of the elements, so that the length of each vector, (the root sum of squares of each column) is unity. Standardizing ensures that the measurement of the X variables is uniform (e.g. some columns of the regressor variables may be measured in inches while others could be measured in centimeters) and in fact unit free. Marquardt and Snee (1975) recommended that in some contexts estimates from standardized variates could provide readier parameter interpretability.

Standardizing is essential before eigenanalysis is used for purposes of detecting collinearity, to prevent the eigenanalysis from being dominated by one or two of the independent variables. Independent variables in their original units of measure would contribute unequally to the total sum of squares and, hence, to the eigenvalues.

In §2.2.5 we pointed out that the condition number has its own scaling problem. Stewart (1987) shows that if we partition

$$X = [X_{(p)}\ X_p], \tag{2.4.1}$$

where $X_{(p)}$ is all the columns of X except the p-th column, and write

$$X_\alpha = [X_{(p)}\ \alpha X_p].$$

When a approaches zero then

$$\lim \|X_\alpha\| = \|X_{(p)}\|$$

and

$$\|X_\alpha^\dagger\| \cong a^{-1}\|X_p^\dagger\|$$

where $X_p^\dagger$ is the p-th row of $X^\dagger$. The rest of the matrix becomes so small in comparison to the last row that the norm is only a function of the last row. It follows then that

$$K(X_\alpha) = a^{-1}\|X_{(p)}\| \; \|X_p^\dagger\| \rightarrow \infty \tag{2.4.2}$$

Thus by scaling down any column of X, the condition number can be made arbitrary large and cause 'artificial ill-conditioning'. Therefore it is recommended that before computing the condition number, the columns should be standardized to have unit column length (Belsley *et al.* (1980, appendix 3B and §3.3))

Centering makes all independent variables orthogonal to the intercept column and hence removes any collinearity that involves the intercept (see the discussion later in this section on collinearity involving the intercept term). 'Nonessential collinearity' (Marquardt and Snee (1975)) is thus removed. Centering is recommended in order to eliminate collinearities which are due to the origins of the predictor variables and it can often provide computational benefits when small storage or low precision prevail.

The effect of centering on VIF's is discussed by Schall and Dunne (1987b). Let X from a model that includes a constant term be partitioned $X = [1\ X_{(i)}\ X_i\ ]$, as in (2.2.28), and $\beta = [\beta_0\ \beta'_{(i)}\ \beta_i]'$ where β_0 is the intercept term. Then similarly to (2.2.30)

$$\mathrm{VIF}_i = \mathrm{VIF}_i(1).\mathrm{VIF}_i(X_{(i)}|1) \tag{2.4.3}$$

The partial variance inflation factor $\mathrm{VIF}_i(X_{(i)}|1)$ is the variance inflation factor for the parameter β_i which would have been obtained from the mean

2.5 Perturbation

The numerical data that constitute X and Y have only a limited number of accurate digits and after those digits the data are completely uncertain and hence arbitrary. Situations should be avoided where a 'small' change in the data produces 'large' changes in the solution. Results relating to perturbation of the pseudoinverse ($X^\dagger = (X'X)^{-1}X'$) or the solution of the OLS problem ($\hat{\beta}$) have been given by a number of authors. For treatments of this perturbation problem or special cases of the problem, see Stewart (1977 and 1973), Lawson and Hanson (1974), and Wedin (1969, 1973). The theory is so well described in textbooks (e.g. Lawson and Hanson (1974)) that only a few results will be given here.

2.5.1 Perturbation bounds

The notation for this section is as follows:

Perturbations in X, Y and $\hat{\beta}$ will be denoted by δX, δY, and $\delta\hat{\beta}$ respectively. (In previous sections we sometimes use the notation E for δX, (adopted from Stewart (1987)). The condition number will as previously be denoted by $K(X)$, $\hat{\epsilon}$ will be the residual vector. Let $\hat{\beta} = X^\dagger Y$ and $\tilde{b} = \tilde{X}^\dagger Y$, where $\tilde{X}$ is the perturbed regression matrix $\tilde{X} = X + E$ or $X + \delta X$.

An important inequality, used in §2.1 to derive upper bounds for the different conditioning analysis, is from Golub and Van Loan (1983, 6.1-10, p141):

$$\|\delta\hat{\beta}\|/\|\hat{\beta}\| \le K(X)R^{-1}[2 + (1-R^2)^{1/2}K(X)]\nu + O(\nu^2), \qquad (2.5.1)$$

where $\nu = \max(\|\delta Y\|/\|Y\|, \|\delta X\|/\|X\| < [K(X)]^{-1})$, and R is the uncentered multiple correlation coefficient of Y regressed on X. Both X and $X + \delta X$ are assumed to be of full rank. We note that the condition number plays a vital role in the inequality. A bound for (2.1.15) was established by choosing

$\nu = \|\delta X\|/\|X\| = m_1$. In Chapter 9 of Lawson and Hanson (1974) (2.5.1) is generalized to different ranks of X (e.g rank(X) = k = p < n, n = p = k = rank(X), p > n = k =rank(X)).

A second inequality is due to Stewart (1987, equation 3.4)

$$\frac{\|\tilde{b}-\hat{\beta}\|}{\|\hat{\beta}\|} \leq K(X)\frac{\|E\|}{\|X\|} + [K(X)]^2\frac{\|E\|}{\|X\|}\ \frac{\|\hat{\epsilon}\|}{\|X\|\ \|\hat{\beta}\|} + O(\|E\|^2) \qquad (2.5.2)$$

The derivation of (2.5.2) can be achieved by using equation (3.24) of Stewart (1977) and then applying the triangular and submultiplicative inequalities for matrix norms (appearing in §1.9).

A bound like (2.5.2) is unnecessarily pessimistic, due to the repeated applications of the triangular and submultiplicative inequalities, and the fact that each application represents another backing off from sharpness. Whereas numerical analysts are not concerned about the bound (2.5.2) because their errors originate from rounding on a digital computer and can be made very small by using different routines (for a discussion of various routines to calculate $\hat{\beta}$, see Lawson and Hanson (1974)), the statistician on the other hand must deal with measurement errors in recording data, and here the lack of sharpness hurts.

2.5.2 Computational accuracy

There exists concern about the numerical accuracy of common computer programs calculating OLS solutions. One of the first to study this phenomenon was Longley (1967). He found that different regression programs resulted in very different solutions, including differences in sign and first significant digit. More recently this phenomenon was studied by several authors including Beaton, Rubin and Barone (1976), Lesage and Simon (1988), Simon and Lesage (1988), and comprehensively by Randall and Rayner (1987).

In their study Beaton, Rubin and Barone (1976) took the Longley data set and added small perturbations to it, before running it on various different regression programs. The results exhibited unstable regression coefficients which could have been as a result of the errors in the input data itself (the data was analyzed by different investigators) or problems with the model (collinearity). They suggested that if one chooses a regression program it should be one that uses the Gram-Schmidt or Givens method. If possible matrix inversion should be avoided. The disadvantage with Gram-Schmidt is it is too expensive for a general program as it requires a pass over the data for each independent variable.

In their study on numerical accuracy Randall and Rayner (1987) rate the Cholesky algorithm (described in Chapter 7 of Graybill (1976)) as the best method of performing the least square calculations. It has the following desirable properties: it is compact, whether in recording calculator results or in computer storage requirements, is computationally economical, and is computationally stable. The basis of this method is the factorization of

$$X'X = LL'$$

where L is a unique positive lower triangular matrix of order p. Randall and Rayner (1987) also show how the Cholesky algorithm is related to the QR factorisation. They concentrate their study on the Cholesky rather than the QR as 'its efficiency on the score of economy of computing operation; Chambers (1977) ranks the Cholesky ahead of any QR method in this respect.'

A measure of the accuracy of a calculated coefficient b, is Wampler's (1970) 'count of the number of correct significant digits' in b, and termed the Wampler accuracy (WA) of b. Let $\tilde{b}$ denote the value of the coefficient as computed free from round-off errors, then

(i) $WA = -\log_{10}(|b-\tilde{b}|/|\tilde{b}|)$, if $|b-\tilde{b}| \neq 0$ and $\tilde{b} \neq 0$,

(ii) $WA = -\log_{10}(|b-\tilde{b}|)$, if $|b-\tilde{b}| \neq 0$ and $\tilde{b} = 0$,

(iii) WA = the approximate number of decimal digits (d) with which the machine computes, if $b-\tilde{b} = 0$.

By studying different data sets and implementing the Cholesky method Randall and Rayner (1987) found:

1. Standardizing X (scaling the columns to unity), has no effect on accuracy. They report 'scaling per se can have no appreciable effect on accuracy with the Cholesky method', and can be associated with serious losses of accuracy (see their discussion on pp21-27).

2. Provided double precision is used, the Cholesky method will satisfy all normal requirements of accuracy without the need for centering (unless centering is appropriate for other reasons), even in cases of extreme ill-conditioning. (The average gain in accuracy due to centering may be between 2-3 digits)

3. If at all possible double precision arithmetic should be used at all times.

4. The accuracy of the Cholesky method can be improved further by the double Cholesky (discussed on their pp36-42).

Lesage and Simon (1988) find that QR decomposition provides greater accuracy than the Cholesky decomposition for the Wampler benchmark data set. Furthermore in their data set standardizing of X does not have a substantial impact on accuracy, and they show that although centering does improve the accuracy of Cholesky, this improvement is not due to centering itself. Rather the improvement occurs because Cholesky with centering is a hybrid algorithm mixing Cholesky with the more accurate QR decomposition. In summary, they do not recommend standardizing (it appears to cause a slight decline in accuracy), but support that centering appears useful for the Cholesky algorithm, whereas in the QR algorithm, centering is irrelevant.

Simon and Lesage (1988) study the impact of ill-conditioning on numerical accuracy arising from two types of collinearity: (i) a collinear relation involving the intercept column and each of the independent variables. The 'slope' columns become simultaneously and individually collinear with the intercept and with each other, (ii) a near linear relationship exists among each of the independent variables but does not involve the intercept term.

They investigated the numerical accuracy of different regression software packages by using an artificial data set called 'the modified Wampler Benchmark (the original Wampler benchmark was described by Wampler (1980), modified by Lesage and Simon (1988) and again modified in this study to take into account collinearity involving the intercept.)

The Modified Wampler benchmark data matrix X (nx(n-1)) and dependent variable vector Y (nx1) are shown in expression (2.5.3)

$$X = \begin{bmatrix} 1 & 1+\gamma & 1+\gamma & \ldots 1+\gamma \\ 1 & \gamma+\epsilon & \gamma & \ldots & \gamma \\ 1 & \gamma & \gamma+\epsilon & \ldots & \gamma \\ : & : & : & : & : \\ 1 & \gamma & \gamma & \ldots \gamma+\epsilon \\ 1 & \gamma & \gamma & \ldots & \gamma \end{bmatrix} \qquad Y = \begin{bmatrix} (n-1) + (n-2)\gamma + \epsilon \\ (n-2)\gamma + \epsilon \\ (n-2)\gamma + \epsilon \\ : \\ (n-2)\gamma + \epsilon \\ (n-1) + (n-2)\gamma - \epsilon \end{bmatrix} \qquad (2.5.3)$$

where the parameter labelled γ controls the severity of ill-conditioning of the first type and ϵ controls the severity of the second type. γ and ϵ control the severity of both types independently. As the parameter ϵ is decreased towards zero, the collinearities among all of the last (n-2) columns increases. If γ increases the last (n-2) columns of the matrix X become more nearly collinear with the intercept column.

Their results can be summarized as follows: Both types of collinearity can adversely affect the numerical accuracy of regression estimator calculation. Integer-valued data were found to react quite differently from non-integer data during the computational process of centering, and the computational benefits associated with centering were found to be much greater for integer-valued data than for non-integer data. Centering the data matrix mitigates but does not prevent accuracy problems. Accuracy problems are not confined to the intercept estimate, but extend to all coefficients. The numerical accuracy of the regression algorithm coefficient estimates are measured in the manner of Vampler (1980) as:

$$\mathrm{acc}_j = -\log_{10}(|\beta_j - \hat{\beta}_j|/\beta_j) \qquad (2.5.4)$$

where β_j are the true value and $\hat{\beta}_j$ the estimated value.

If one looks critically at this study there are various important limitations: the Vampler benchmark data matrix X, has approximately rank 1, thus very severe collinearity exists, almost the worst case possible. For the artificial data in (2.5.3) the fitted Y values and the residuals are

$$\hat{Y} = \begin{bmatrix} (n-1) + (n-2)\gamma \\ 1 + (n-2)\gamma + \epsilon \\ 1 + (n-2)\gamma + \epsilon \\ \vdots \\ 1 + (n-2)\gamma + \epsilon \\ 1 + (n-2)\gamma \end{bmatrix} \qquad \hat{\epsilon} = \begin{bmatrix} \epsilon \\ -1 \\ -1 \\ \vdots \\ -1 \\ (n-2)-\epsilon \end{bmatrix} \qquad (2.5.5)$$

which indicate that case 1 in influential, case n is an outlier and the errors in Y are highly correlated.

Furthermore the standard errors of the regression coefficients are very high (calculated for various values of γ and ϵ to be between 78 to 7936, with the possibility of even larger values). If one is working with such high values of SE's of estimators, one may wonder what is the point of centering at all, and of reporting the accuracy of the estimators.

2.6 Detecting and handling collinearity

The first step in successfully coping with collinearity is an understanding of the nature and effects of collinearities and an ability to determine when they occur in a data set (Gunst (1983)).

In §2.2 various ways of detecting collinearities have been discussed. It is of utmost importance that collinearity should be detected. Any method can be used, but it may even be advisable to use several of them. If one were to use for instance VIF's, it is also good practise to look at other methods (e.g condition number, condition indices, variance-decomposition, and small eigenvalues), to get a multi-faceted insight into the problem. What is important here is the identification of collinearity and not which particular method one uses to detect it. The user may have to calculate some of the measures as most regression computer programs are not designed to warn automatically of the presence of near-collinearities.

Once collinearity is identified no easy remedy is at hand. Any remedy will depend on the objective of the model fitting exercise. If the objective of the study is prediction, collinearity will cause no harmful effects if the collinearity proceeds into the prediction area and no serious extrapolation is made within or outside the row-space of X. When primary interest is in estimation of the regression coefficients, other alternatives should be considered. One is augmentation of the data in the directions of the collinearities, e.g. obtain new data or additional data such that the row-space is expanded to remove the near-singularity. Unfortunately this is frequently impractical or impossible.

Subset selection of variables to remove the collinearity should be applied with great care, as this approach may result in removing some of the important regression variables. Hoerl *et al.* (1986) recommend against subset selection as a general strategy to combat collinearity. In the face of severe collinearity one of the best alternatives is to use those biased estimators that are not so severely effected by collinearity. An array of biased estimators will be described elsewhere in this dissertation. Choosing one of them will depend on the circumstances of the problem, and estimates may perform differently in different situations.

2.7 Summary

In this chapter we defined collinearity. We discussed ways of detecting collinearity and the effect of collinearity on regression estimates. The issue of centering and the concepts of perturbation were also introduced. Finally a summary of approaches to collinearity was presented.

Chapter 3

PRINCIPAL COMPONENTS REGRESSION MODEL

In this chapter we discuss Principal Components Regression (PCR). We present in §3.1 a general overview of PCR, in §3.2 the estimator, in §3.3 some properties of PCR estimators, in §3.4 methods of elimination of PC's and in §3.5 comments and critique of PCR.

3.1 Introduction

Let $X'X$ be in correlation form and consider model (1.1), and the SVD of X (1.3):

Then $\quad Y = X\beta + \epsilon \quad$ and $\quad X = U\Delta V'$.

$$Y = XVV'\beta + \epsilon \quad \text{where } (VV' = I_p)$$

Let $\quad Z = XV, \quad \delta = V'\beta \quad$ then

$$Y = Z\delta + \epsilon \tag{3.1.1}$$

$Z = [z_1\ z_2\ \ldots z_p]$ is the nxp matrix of principal components. z_i is the i-th principal component.

$$\begin{aligned} Z'Z &= V'X'XV \\ &= V'V\Delta^2V'V \qquad \text{(from (1.3.3))} \\ &= \Delta^2 \end{aligned} \tag{3.1.2}$$

Using (3.1.1) the LS estimator of δ is:

$$\begin{aligned} \hat{\delta} &= (Z'Z)^{-1}Z'Y \\ &= \Delta^{-2}Z'Y \qquad \text{(from (3.1.2))} \\ &= \Delta^{-2}V'X'Y \\ &= \Delta^{-2}V'X'X\,\hat{\beta} \qquad (X'Y = X'X\,\hat{\beta}) \end{aligned}$$

$$= \Delta^{-2}V'V\Delta^{2}V'\hat{\beta} \qquad \text{(from 1.3.3)}$$

$$= V'\hat{\beta} \qquad (3.1.3)$$

A principal components estimator of β is obtained by deleting one or more of the principal components (z_j) and corresponding parameter (δ_j) and then making a transformation back into the original parameter space.

Equivalently one constructs new parameters (δ_i) and chooses a subset of them. No variables in X are removed, but some PC's are dropped as if the corresponding δ_i were zero. The final model equation is as complex as the one formed using LS (a criticism of PC regression by Wetherill (1986)). This approach may seem feasible rather than meaningful, but in the presence of collinearity it may remove collinearities and is claimed to give an insight into the original objective, namely coefficient estimation.

Let $Z = [Z_1 : Z_2]$ where Z_2 contains the PC's z_i that must be retained and Z_2 the z_j to be deleted. Let us assume that r PC's are deleted; Z_2 will then contain r columns and Z_1 p-r columns. Then (3.1.1) may be rewritten as

$$Y = [Z_1 \ Z_2] \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} + \epsilon \qquad (Z_1 : nx(p-r);\ Z_2 : nxr)$$

$$= Z_1\delta_1 + Z_2\delta_2 + \epsilon$$

Let $Z_1 = XV_1$ (the PC's to be retained and $V = [V_1 \ V_2]$)

Then to remove the z_j's (all those contained in Z_2, which is orthogonal to Z_1) one sets $\delta_2 = 0$, which implies that $Z_2\delta_2 = 0$. Hence we obtain the principal component estimators (PCE) of δ_1:

$$\hat{\delta}_1 = (Z_1'Z_1)^{-1}Z_1'Y$$

$$E[\hat{\delta}_1] = \Delta^{-2}Z_1'Y = \delta_1 \qquad \text{(unbiased for } \delta_1\text{)}$$

$$V[\hat{\delta}_1] = \sigma^2(Z_1'Z_1)^{-1}$$

In (3.1.1) we use a transformation to reparameterize the model:

$$\delta = V'\beta$$

$$\begin{bmatrix}\delta_1\\ \delta_2\end{bmatrix} = \begin{bmatrix}V_1'\\ V_2'\end{bmatrix}\beta$$

PC regression can be viewed as the use of a restricted least squares estimator (Hill *et al.*(1977) and Johnson *et al.*(1973)). The restrictions arise from the assumption that $\delta_2 = 0$, which implies that $V_2'\beta = 0$.

Three advantages of seeing the PCE's as equivalent to Restricted Least Squares (RLS) (in which the restrictions are applied to the noisy collinearities), are claimed by Hill *et al.* (1977) namely:

(i) Data reduction techniques generally impose restrictions upon the associated parameter space.

(ii) Explicit recognition of the restrictions permits an evaluation of their theoretical implications.

(iii) RLS formulation is a convenient vehicle for the statistical investimation of the partitioning of Z.

We comment in §3.5 on these claims.

3.2 The PC estimator

Let $\hat{\beta}_{pc}$ denote the PCE of β. Using (1.3.5) with $\sum_{i=p-r+1}^{p} v_i c_i/\lambda_i$ set to 0 by the r restrictions $V_2'\beta = 0$, the PCE ($\hat{\beta}_{pc}$) of β can be written as

$$\hat{\beta}_{pc} = \sum_{i=1}^{p-r} v_i c_i/\lambda_i \qquad (3.2.1)$$

3.3 Properties of $\hat{\beta}_{pc}$

Some of the properties of PCR discussed in the introduction are restated here for a convenient comparison with ridge regression (Chapter 4).

1. Relationship to OLSE

$$\hat{\beta}_{pc} = \hat{\beta} - \sum_{i=p-r+1}^{p} v_i c_i/\lambda_i \qquad (3.3.1)$$

$$= \hat{\beta} - V_2 V_2{}'\hat{\beta} \qquad (3.3.2)$$

2. Expectation

$$E(\hat{\beta}_{pc}) = [I_p - V_2 V_2{}'] \beta$$
$$= W\beta$$

where $W = [I_p - V_2 V_2{}'] = W' = V_1 V_1'$ (from (1.3.1) and $I_p = V_1 V_1' + V_2 V_2'$).

Thus the bias vector is

$$b = W\beta - \beta$$
$$= (W-I)\beta$$
$$= -V_2 V_2{}'\beta \qquad (3.3.3)$$

3. Variance

$$V(\hat{\beta}_{pc}) = V(W\hat{\beta})$$
$$= W\ V(\hat{\beta})\ W \qquad (W' = W)$$
$$= \sigma^2 W\ (X'X)^{-1} W \qquad (\text{from } \S 1.2)$$
$$= \sigma^2 V_1 V_1{}'(V\ \Delta^{-2} V')V_1 V_1{}'$$
$$= \sigma^2 V_1 V_1{}'(V_1 \Delta_1^{-2} V_1' + V_2 \Delta^{-2} V_2')V_1 V_1'$$
$$= \sigma^2 [V_1 \Delta_1^{-2} V_1{}'] \qquad (V_1 \text{ and } V_2 \text{ are orthogonal})$$
$$= \sigma^2 \sum_{i=1}^{p-r} v_i v_i{}'/\lambda_i \qquad (3.3.4)$$

where $$\Delta = \begin{bmatrix} \Delta_1 0 \\ 0\ \Delta_2 \end{bmatrix} \quad \begin{matrix} \Delta_1 : (p-r)x(p-r) \\ \Delta_2 : rxr \end{matrix} \qquad (3.3.5)$$

4. $$\begin{aligned} MSE(\hat{\beta}_{pc}) &= E[(\hat{\beta}_{pc}-\beta)(\hat{\beta}_{pc}-\beta)'] \\ &= Var(\hat{\beta}_{pc}) + bb' \\ &= \sigma^2 \sum_{i=1}^{p-r} v_i v_i'/\lambda_i + V_2V_2'\beta\beta'V_2V_2' \qquad (3.3.6) \end{aligned}$$

(from (3.3.4) and (3.3.3))

5. $$\begin{aligned} TMSE(\hat{\beta}_{pc}) &= tr(WV(\hat{\beta})W + [I - W]\beta\beta'[I - W]) \\ &= tr\ (W'W\ \sigma^2(X'X)^{-1}) + tr(\beta'(-V_2V_2')(-V_2V_2')\beta) \\ &= \sigma^2\ tr\ ([V_1\Delta_1^{-2}V_1'] \quad + \beta'(V_2V_2')\beta \\ &= \sigma^2 \sum_{i=1}^{p-r} 1/\lambda_i + \beta'(V_2V_2')\beta \quad \text{(from A.2)} \qquad (3.3.7) \end{aligned}$$

3.4 Eliminating PC's

To address the collinearity problem a logical choice might be to include in Z_2 those components corresponding to 'small' eigenvalues of X'X (Massy (1965)). Inclusion will ensure that the variance of $\hat{\beta}_{pc}$ will be small. Simultaneously the bias of PCE will increase so that the MSE and TMSE can be large and predictions of the response poor. Components with small eigenvalues may be highly correlated with the dependent variable. Massy (1965) therefore recommended that the selection rule should be one of balancing the preservation of the sample variation in X'X against the correlation of the PC's with the response.

The choice of which PC's to eliminate will depend on the purpose of the regression. We will distinguish between two purposes. The linear regression model may be estimated to test some theoretical or structural hypothesis (§3.4.1), or its sole use may be that of a prediction equation (§3.4.2).

3.4.1 Theoretical or Structural Norms

PCE's are RLS estimators and can be expressed in terms of structural norms and tests of implicit or explicit restrictions. The term norm here is not exactly the same as the term norm defined in §1.9 although that idea contributes to this section: most of the test statistics defined in this section involve eigenvalues and traces and the test statistics are ratios of norms. Four plausible tests are (Hill *et al.* (1977)):

3.4.1.1 Classical F Test

To exclude r PC's the following restriction matrix (R) is appropriate:

$$\underset{r\mathrm{x}p}{R} = [\underset{r\mathrm{x}p-r}{0} : I_r]$$

To evaluate the hypothesis that $R\delta = 0$ (setting $Z_2\delta_2 = 0$), the F statistic can be used. Under the null hypothesis the statistic V is:

$$V = \left[\frac{(n-p+r)\hat{\sigma}^2_R - (n-p)\hat{\sigma}^2}{(n-p)\hat{\sigma}^2}\right]\left[\frac{n-p}{r}\right] \qquad (3.4.1)$$

where $\hat{\sigma}^2_R$ is the maximum likelihood estimator (MLE) of σ^2 in the restricted model, and $\hat{\sigma}^2$ is the MLE of σ^2 in the full model. Now V has a central F-distribution with r and n-p degrees of freedom, when $R\delta = 0$, and when $R\delta \neq 0$ we have

$$V \sim F(r,n-p;\gamma)$$

where

$$\gamma = \frac{(R\delta)'[R(Z'Z)^{-1}R']^{-1}R\delta}{2\sigma^2} \qquad (3.4.2)$$

$$= \frac{\delta_2{}'\left[[0\ \ I_r]\begin{bmatrix}\Delta_1 & 0\\ 0 & \Delta_2\end{bmatrix}\begin{bmatrix}0\\ I_r\end{bmatrix}\right]^{-1}\delta_2}{2\sigma^2} \qquad \begin{Bmatrix}\text{from } (3.1.2)\\ \text{and } (3.3.5)\end{Bmatrix}$$

$$= \frac{\sum_{i=p-r+1}^{p} \delta_i^2 \lambda_i}{2\sigma^2}$$

To evaluate the hypothesis that $R\delta = 0$ is to test that $\gamma = 0$.

H_0 will be accepted ($\gamma \doteq 0$) when:

(i) σ^2 is very large.

(ii) $\delta_i^2 \lambda_i$ is small for every i

(iii) a combination of (i) and (ii) applies

In (ii) when λ_i is small but δ_i^2 is large enough, the F test will tend to reject the hypothesis.

3.4.1.2 Strong MSE criterion

We say PCE is better in the strong MSE sense if

$$\text{MSE}(\hat{\beta}_{PC}) \leq \text{MSE}(\hat{\beta})$$

$$\text{i.e.} \quad \text{MSE}(h'\hat{\beta}_{PC}) \leq \text{MSE}(h'\hat{\beta}) \qquad (\text{for every } h \neq 0,\ h{:}p\text{x}1)$$

This inequality holds if and only if for γ in (3.4.2) we have

$$\gamma \leq \tfrac{1}{2}$$

Under the null hypothesis that PCE is better than LSE in the strong MSE sense

$$V \sim F(r, n-p; \gamma), \quad \gamma \leq \tfrac{1}{2}$$

This norm is less stringent than the classical F test ($\gamma = 0$).

3.4.1.3 First weak MSE criterion

A weaker criterion than the above is to say PCE is better than the OLSE if

$$\mathrm{tr}(\mathrm{MSE}(\hat{\beta}_{PC})) \leq \mathrm{tr}(\mathrm{MSE}(\hat{\beta}))$$

or $$\mathrm{TMSE}(\hat{\beta}_{PC}) \leq \mathrm{TMSE}(\hat{\beta})$$

where

$$\mathrm{tr}(\mathrm{MSE}(\hat{\beta}_{PC}) = (\sigma^2 \sum_{i=1}^{p-r} 1/\lambda_i + (\beta'(V_2V_2')\beta)) \qquad \text{(from (3.3.5))}$$

$$= \sigma^2 \sum_{i=1}^{p-r} 1/\lambda_i + \sum_{i=p-r+1}^{p} \delta_i^2$$

$$\mathrm{tr}(\mathrm{MSE}(\hat{\beta})) = \sigma^2 \sum_{i=1}^{p} 1/\lambda_i$$

and

$$\mathrm{tr}(\mathrm{MSE}(\hat{\beta}_{PC})) - \mathrm{tr}(\mathrm{MSE}(\hat{\beta})) = \sum_{i=p-r+1}^{p} \delta_i^2 - \sigma^2 \sum_{i=p-r+1}^{p} 1/\lambda_i$$

$$= \sum_{i=p-r+1}^{p} (\delta_i^2\lambda_i - \sigma^2)/\lambda_i \qquad (3.4.3)$$

In this case V ~ $F(r,n-p;\gamma)$ where the non-centrality parameter γ is given by

$$\gamma \leq \tfrac{1}{2}d_L \ \mathrm{tr}[(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}] \qquad (3.4.4)$$

and d_L is the largest eigenvalue of the expression under the trace operator.

This test was proposed by Wallace (1972) and refined by Yancey, Judge and Bock (1973).

For $\hat{\beta}_{PC}$ to be 'better' than LSE (3.4.3) must be negative. We note that (3.4.3) would be very sensitive to collinearity, due to the division by λ_i: the smaller any λ_i the greater the potential advantage of PCE over OLSE under this criterion.

3.4.1.4 Confidence interval norm

A norm proposed by Cheng and Iglarch (1976) generates a highly specialized test, and will only be reported here:

Assume β_i is of primary interest (β_i is any element of β, $(\hat{\beta}_{PC})_i$ is the corresponding element of β_{PC}). If

$$\Pr(|\hat{\beta}_i - \beta_i| < \eta\ ) < \Pr(|(\hat{\beta}_{PC})_i\ -\ \beta_i|< \eta\ )$$

then $\hat{\beta}_{PC}$ is judged superior to $\hat{\beta}$. The value of η is determined by the investigator (η is the measure of how 'close' the investigator wants the estimated β to its true value (see Hill *et al.* (1977) eq. 13)).

3.4.2 Predictive norms

When the purpose of the regression is one of forecasting, a norm involving $\hat{Y}$ ($\hat{Y} = X\hat{\beta}$) is implied. The total mean square error of prediction is (Allen (1974) eq. 6):

$$\begin{aligned} \text{TMSE}(Y_f) &= E(\hat{Y} - Y_f)'(\hat{Y} - Y_f) \\ &= n\sigma^2 + \sum_{i=1}^{n} V(\hat{Y}_i) + \sum_{i=1}^{n} [E(Y_i - \hat{Y}_i)]^2 \\ &= n\sigma^2 + E\{(\hat{Y} - X\beta)'(\hat{Y} - X\beta)\} \\ &= n\sigma^2 + \text{TMSE}(\hat{Y}) \end{aligned} \qquad (3.4.5)$$

Thus (3.4.5) depends upon the unknown population parameters β and σ^2. Allen (1974) suggested using $\hat{\beta}$ and $\hat{\sigma}^2$. The disadvantage of this approach is that this measure can not discriminate among various prospective estimators. Two predictive norms that do not have this disadvantage are discussed in §3.4.2.1 and §3.4.2.2.

3.4.2.1 Weak predictive MSE criterion

Consider two alternative estimators, $X\hat{\beta}_{PC}$ and $X\hat{\beta}$, of $\hat{Y}$. The PCE is a better estimator if

$$E[(X\beta - X\hat{\beta}_{PC})'(X\beta - X\hat{\beta}_{PC})] \leq E[(X\beta - X\hat{\beta})'(X\beta - X\hat{\beta})]$$

The F test can again be applied using V as in (3.3.1). $\hat{Y}_{PC}$ will be better than $\hat{Y}$ if for γ as in (3.4.2) we have

$$\gamma \leq \frac{r}{2}$$

$$V \sim F(r,p-r;\gamma)$$

Note that the degrees of freedom change from n-p to p-r. Critical values for this test are in Goodnight and Wallace (1972).

3.4.2.2 Squared bias of prediction

The term $\sum_{i=1}^{n} [E(Y_i - \hat{Y}_i)]^2$ is the total squared bias of the predictions, and appears in (3.4.5). It is also called the squared bias of prediction. If a norm based on this criterion is considered then it is possible to construct a multiple comparison test for all general hypotheses for the model in (1.1). This norm does not involve σ^2 and it only considers the squared bias. This procedure and other references to it can be found in Hill *et al.*(1977).

3.5 Comments and critique

Some of the above criteria are quite overwhelming. One simple way to eliminate PC's is to use restricted least squares. For more details on the theory and tests of hypotheses involving such restrictions, any textbook on linear models can be consulted. (eg. Searle (1971)).

Rawlings (1988) states that a good working rule is to eliminate those PC's that satisfy two conditions:

(i) they cause serious variance inflation because of small eigenvalues

(ii) the corresponding estimated regression coefficients $(\hat{\beta}_{PC})_i$ are not significantly different from zero.

Lott (1973) described the Strong MSE criterion (§3.4.1.2) and shows in a Monte Carlo study that under this criterion PCE performed better than OLSE.

For more examples on the PC method see Hill *et al.* (1977), Wetherill (1986) and Jolliffe (1972,1973,1982). In Jolliffe (1982) four examples are given to demonstrate that it is not always the PC's with small eigenvalues that are eliminated.

The justification that a MSE criterion can underpin the PCR approach is technically correct, but is only adequate when the actual $V_2'\beta$ parameter values are sufficiently small. Specifically we require

$$\begin{aligned} \text{MSE }(\hat{\beta}_{PC}) &= \sigma^2 \sum_{i=1}^{p-r} v_i v_i'/\lambda_i + V_2 V_2' \beta\beta' V_2 V_2' \\ &\leq \text{var }(\hat{\beta}) \end{aligned}$$

$$\text{Var}(\hat{\beta}) = \sigma^2 \sum_{i=1}^{p-r} v_i v_i'/\lambda_i + \sigma^2 \sum_{p-r+1}^{p} v_i v_i'/\lambda_i$$

and the data may not admit belief that $V_2'\beta$ is in fact small enough. These comments underpin suggestions of Rawlings (1988) previously mentioned.

PC Regression initially amounts to choosing that part of the parameter space spanned by $V_2'\beta$, about which one has virtually no intrinsic information in the data since $\mathrm{Var}(V_2'\hat{\beta}) = \mathrm{diag}(\lambda_{p-r+1}^{-1},\ldots,\lambda_p^{-1})$ is very large. One then claims to have perfect information that $V_2'\beta = V_2'\beta_{PC} = 0$ in the sense that $\mathrm{Var}(V_2'\hat{\beta}_{PC}) = 0$. This additional perfect information does not affect that part of the parameter space spanned by $V_1'\beta$, which is intrinsically independent of any information whatsoever on $V_2'\beta$. In consequence looking at $V_1'\hat{\beta}_{PC}$ will tell us nothing different from what is already known from $V_1'\hat{\beta}$. Any alleged advantages in estimation from PCR must result from the apparently reduced variances of estimators of $AX\beta$ where

$$AX = AXVV' = BV' = [B_1 : B_2]\begin{bmatrix} V_1' \\ V_2' \end{bmatrix},$$

since it follows directly from (3.3.5) that

$$\begin{aligned} \mathrm{Var}(AX\hat{\beta}) &= \mathrm{Var}(B_1V_1'\hat{\beta}) + \mathrm{Var}(B_2V_2'\hat{\beta}) \\ &= \mathrm{Var}(B_1V_1'\hat{\beta}_{PC}) + \mathrm{Var}(B_2V_2'\hat{\beta}) \\ &> \mathrm{Var}(B_1V_1'\hat{\beta}_{PC}) + \mathrm{Var}(B_2V_2'\hat{\beta}_{PC}) \qquad (B_2 \neq 0) \end{aligned}$$

Nonetheless such advantages are artificial, and if the restrictions are not justifiable *a priori*, it makes more sense to simply isolate the region in the parameter space about which further information data could be usefully applied. Equivalently: it should suffice to specify that $V_2'\beta$ is not well approximated by the $X\beta$ of the model.

The point is that a restriction such as $V_2'\beta = V_2'\hat{\beta}$ imposed on the model will admit $V'\hat{\beta}_{PC} = V'\hat{\beta}$ and will reduce the variance of the estimator from $\sigma^2 \sum_{i=1}^{r} 1/\lambda_i$ to $\sigma^2 \sum_{i=1}^{p-r} 1/\lambda_i$ without introducing bias. Such an approach is no more artificial than the ordinary PC restriction $V_2'\beta = 0$. It is also rank-reducing in whatever sense PC estimation is rank reducing.

3.6 Summary

In this chapter we presented a general overview of PCR. We defined the estimator and its properties. We discussed methods of eliminating PC's and gave some critique of PCR.

Chapter 4

RIDGE REGRESSION

4.1 Introduction

Ridge regression (RR) was first proposed by Hoerl and Kennard (1970a and 1970b). Ridge is an estimation procedure for the model (1.1) based on adding a small positive constant to the diagonal elements of $X'X$ (which we assume in this section will be in correlation form). Instead of inverting $X'X$ which is ill-conditioned in the presense of collinearity, one inverts $X'X + kI$, where k is chosen in such a way that the estimators of β become stable.

4.2 The RR estimator

The ridge estimator is defined as the solution to

$$(X'X + kI)^{-1}\hat{\beta}_R = X'Y, \quad k \geq 0 \tag{4.2.1}$$

4.3 Properties of $\hat{\beta}_R$

1. Relationship to OLSE

$$\begin{aligned}\hat{\beta}_R &= WX'Y \qquad \text{where } W = (X'X + kI)^{-1} = W' \\ &= WX'X\hat{\beta} \\ &= Z\,\hat{\beta} \qquad \text{where } Z = W\,X'X = I - kW = Z'\end{aligned} \tag{4.3.1}$$

Note 1. If $k=0$, then $\hat{\beta}_R = \hat{\beta}$.

2. If $k \to \infty$, then Z approaches 0, and $\hat{\beta}_R \to 0$.

2. Expectation

$$\begin{aligned} E(\hat{\beta}_R) &= E(WX'X\hat{\beta}) \\ &= WX'X\ E(\hat{\beta}) \\ &= WX'X\ \beta \\ &= Z\ \beta \\ &= \beta - kW\beta \qquad (4.3.2) \end{aligned}$$

Thus, $\hat{\beta}_R$ is biased for β and we denote the bias of $\hat{\beta}_R$ by $b = -kW\beta$

3 Variance

$$\begin{aligned} V(\hat{\beta}_R) &= V(WX'X\hat{\beta}) \\ &= WX'XV(\hat{\beta})X'XW \\ &= \sigma^2 WX'XW \\ &= \sigma^2 Z(X'X)^{-1}Z' \qquad (4.3.3) \end{aligned}$$

4. The mean squared error (MSE) of $\hat{\beta}_R$ satisfies

$$\begin{aligned} MSE(\hat{\beta}_R) &= E[(\hat{\beta}_R - \beta)(\hat{\beta}_R - \beta)'] \\ &= V(\hat{\beta}_R) + bb' \\ &= \sigma^2 Z(X'X)^{-1}Z' + k^2 W\beta\beta'W' \quad \text{(from (4.3.2) and (4.3.3))} \\ &= \sigma^2 Z(X'X)^{-1}Z' + [Z - I]\beta\beta'[Z - I]' \qquad (4.3.4) \end{aligned}$$

5. The total mean squared error (TMSE) of $\hat{\beta}_R$ satisfies.

$$\begin{aligned} TMSE(\hat{\beta}_R) &= tr(MSE(\hat{\beta}_R)) \\ &= tr(V(\hat{\beta}_R)) + \beta'(Z - I)'(Z - I)\beta \end{aligned}$$

Let

$$\begin{aligned}\gamma_1(k) &= tr[V(\hat{\beta}_R)]\\ &= \sigma^2 tr[Z'Z(X'X)^{-1}]\\ &= \sigma^2 tr[V\Delta^2(\Delta^2+kI)^{-1}V'V(\Delta^2+kI)^{-1}\Delta^2V'V\Delta^{-2}V'] \quad \text{(from (1.3.1))}\\ &= \sigma^2 tr[\Delta^2(\Delta^2+kI)^{-2}]\\ &= \sigma^2 \sum_i \lambda_i(\lambda_i+k)^{-2}\end{aligned}$$

Also let

$$\begin{aligned}\gamma_2(k) &= \beta'(Z-I)'(Z-I)\beta\\ &= k^2\beta'W'W\beta\end{aligned}$$

thus

$$TMSE(\hat{\beta}_R) = \sigma^2 \sum_i \lambda_i(\lambda_i+k)^{-2} + k^2\beta'W'W\beta \qquad (4.3.5)$$

Comments:

(i) $\gamma_1(k)$ is the total variance of the ridge estimators.

(ii) $\gamma_2(k)$ will be considered as the measure of total squared bias that is introduced when $\hat{\beta}_R$ is used rather than $\hat{\beta}$

(iii) $\gamma_1(k)$ is a continuous monotonically decreasing function of k.

(iv) $\gamma_2(k)$ is a continuous monotonically increasing function of k, as can be easily seen from

$$\begin{aligned}\gamma_2(k) &= k^2\beta'[X'X+kI]^{-2}\beta\\ &= k^2\beta'V(\Delta^2+kI)^{-2}V'\beta\\ &= k^2\Sigma\, \delta_i^2/(\lambda_i+k)^2 \quad \text{for } \delta=V'\beta\end{aligned}$$

(v) The squared bias $\gamma_2(k)$ approaches $\beta'\beta$ as an upper limit, as can be easily seen from

$$\begin{aligned}\gamma_2(k) &= k^2\Sigma\, \delta_i^2/(\lambda_i+k)^2\\ &= \Sigma\, \delta_i^2/(\lambda_i/k+1)^2 \text{ , thus}\end{aligned}$$

$$\begin{aligned}\lim_{k\to\infty} \gamma_2(k) &= \delta'\delta \\ &= \beta'\beta\end{aligned}$$

(vi) Hoerl and Kennard (1970b) show that there always exists a value of k such that $\mathrm{TMSE}(\hat{\beta}_R) < \mathrm{TMSE}(\hat{\beta})$. But this value of k is in general bounded by functions of the unknown parameter values β_i (i.e $(k < \sigma^2/\delta^2_{max})$). An optimal choice of $k = p\sigma^2/\beta'\beta$ is operationalized and discussed in §4.6.3.

(vii) The so-called Admissibility Condition of Mayer and Willke (1973) states: A class of estimators E will be called (mean square) admissible if for every problem there is an estimator e in E such that $G(e) < G(\hat{\beta}) = \mathrm{tr}(V(\hat{\beta}))$ (where the symbol G of Mayer and Willke is the TMSE in our terms).

6. The residual sum of squares for the ridge estimator is

$$\begin{aligned}\mathrm{SSE}(\hat{\beta}_R) &= (Y - X\hat{\beta}_R)'(Y - X\hat{\beta}_R) \\ &= Y'Y - Y'X\hat{\beta}_R - \hat{\beta}_R'X'Y + \hat{\beta}_R'X'X\hat{\beta}_R \\ &= Y'Y - \hat{\beta}_R'X'Y - [Y'X - \hat{\beta}_R X'X]\hat{\beta}_R \\ &= Y'Y - \hat{\beta}_R'X'Y - [\hat{\beta}_R'W^{-1} - \hat{\beta}_R X'X]\hat{\beta}_R \\ &= Y'Y - \hat{\beta}_R'X'Y - \hat{\beta}_R'[W^{-1} - X'X]\hat{\beta}_R \\ &= Y'Y - \hat{\beta}_R'X'Y - k\hat{\beta}_R'\hat{\beta}_R \\ &> \mathrm{SSE}(\hat{\beta}) \qquad (4.3.6)\end{aligned}$$

7. The squared length of $\hat{\beta}_R$ is less than the squared length of $\hat{\beta}$ for all $k > 0$ (Hoerl and Kennard (1970a, equation 2.8).

Proof
$$\begin{aligned}\hat{\beta}_R'\hat{\beta}_R &= \hat{\beta}'Z'Z\hat{\beta} \qquad \text{(from (4.3.1))} \\ &= \hat{\beta}'V\Delta^2V'V(\Delta^2 + kI)^{-1}V'V(\Delta^2 + kI)^{-1}V'V\Delta^2V'\hat{\beta} \\ &\qquad\qquad \text{(from (4.3.1) and (1.3.1))} \\ &= \hat{\beta}'\Delta^4(\Delta^2 + kI)^{-2}\hat{\beta} \qquad (4.3.7) \\ &< \hat{\beta}'\hat{\beta} \quad \text{for } k > 0,\end{aligned}$$

Since $\lambda_i^2(\lambda_i + k)^{-2} < 1$ for all i.

8. Marquardt (1970) and Allen (1974) stated that the ridge estimator is formally equivalent to an LSE for the extended data that is obtained when the actual data are supplemented by a fictitious set of data points taken from an orthogonal experiment H_k (H_k is orthogonal up to a scale factor $\sqrt{k}$). The response is taken to be zero for each of these data points.

Let the augmented model be

$$\begin{bmatrix} Y \\ 0 \end{bmatrix} = \begin{bmatrix} X \\ H_k \end{bmatrix}\beta + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix} \tag{4.3.8}$$

If X^* is the augmented matrix, we obtain

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

since H_k is orthogonal up to a scale factor, i.e. $H_k'H_k = kI_p$. It can be seen that the OLSE of the augmented matrix is equivalent to a ridge estimator. We claim that the notion of pseudo-observations at zero for the response variable implicity suggests shrinkage on the usual estimates of β (towards zero).

4.4 Estimation of the ridge parameter k

One way of choosing k involves the use of the 'ridge trace'. The ridge trace is formed by plotting $\hat{\beta}_R$ against k, as k varies through the interval [0,1]. Marquardt (1970) stated that values of k typically are $10^{-4} < k < 1$. A value of k is chosen at the point where the estimates of all β_i have stabilized. If the estimates of some of the regression coefficients do not stabilize, the corresponding variables could be good candidates for deleting (dropping) from the regression equation.

Wetherill *et al.* (1986) suggested that the VIF can also be used by choosing k such that the largest VIF's lie between one and ten. This criterion is Marquardt's (1970) rule of thumb.

Usually there is a range of values of k which give equivalent results from a practical point of view. One problem of selecting an estimate of k (viewed as a parameter), *via* the ridge trace, is that the estimator of k (as a function of y) is technically a random variable. This fact complicates the theory of confidence limits and hypothesis tests, due to introduction of bias and dependence. Because of the introduced bias, the MSE and the TMSE of $\hat{\beta}_R$ are dependent upon the true unknown coefficient vector β (Marquardt and Snee (1975)). In §4.6.2 the fact that k is dependent on the data is investigated further.

For more comments on the use of the ridge trace and the choice of k, the reader is referred to Hoerl and Kennard (1970b, p65), Galpin (1978, p38-58) and section §4.6.3 of this chapter. Examples in the articles to which references have been made, illustrate the effect of varying k. The ridge regression procedure is summarized in recipe form by Rawlings (1988, p340). An alternative method to the ridge trace for estimating k is also discussed by Rawlings.

4.5. General comments

4.5.1 Bayesian view

The ridge estimator can also be derived using Bayesian theory. From a Bayesian viewpoint a prior distribution is defined for the parameters (β,σ^2) and this 'prior knowledge' is then incorporated with the model (1.1) to find the conditional and posterior distributions for the data. (For a general discussion on Bayes estimation in the linear model see Lindley and Smith (1972)).

In the linear regression model (1.1) if we assume that

$$\text{(i)}\quad \epsilon \sim N(0,\sigma^2 I)$$

and we further assume *a priori* knowledge on β:

$$\text{(ii)}\quad \beta \sim N(a,Z) \qquad (a,Z \text{ are known})$$

then the posterior distribution of β given Y, i.e. of the conditional random variable $(\beta|Y)$, can be found using Bayes Theorem (Hsiang (1976)):

$$(\beta|Y) \sim N((Z^{-1} + 1/\sigma^2\ X'X)^{-1}(Z^{-1}a + 1/\sigma^2)X'Y;(Z^{-1} + 1/\sigma^2\ X'X)^{-1}) \tag{4.5.1}$$

If $a = 0$ and $Z = (\sigma^2/k)I$ then the posterior mean of β becomes

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

the ridge estimator and with posterior variance $\sigma^2 W$. Thus a ridge estimator can be considered as the posterior mean based on a normal prior parameter distribution with mean 0 and variance matrix $(\sigma^2/k)I$. Observe that this variance matrix implies a connection between the scale of the residuals (through σ^2) and the degree of the collinearity (through k). This posterior mean estimator is equivalent to the LSE after augmenting the data $Y = X\beta + \epsilon$ with the statistically independent pseudo-observations $a = k^{\frac{1}{2}}\beta + \epsilon_*$, which incorporates the belief that $k^{\frac{1}{2}}\beta = a - \epsilon_*$. In effect the estimator $\hat{\beta}$ is shrunk towards the value $k^{\frac{1}{2}}a$. The Bayesian argument does not depend on X being in a scaled form.

4.5.2 Critique of ridge regression

Smith and Campbell (1980) present the following objections to the use of ridge estimators

1. The standardising and scaling of the $X'X$ matrix to correlation form, is generally presented before the assumption of a unique parameter k.

2. The prior information that is implicit in ridge regression is inadequately justified because of the assumptions that

   (i) parameters have zero means ($\alpha = 0$)

   (ii) parameters have zero covariances

   (iii) parameters have identical variances

   'Ridge regression's weakness is the use of a loose representation of *a priori* beliefs and the relevance at times of ad hoc pseudo-information'.

3. Ridge users may obtain estimates that have preassigned values. 'If one blithely manipulates the data with no regard for the appropriateness of the implicit ridge prior distributions, the ridge estimates may literally be anything.'

4.6 Generalized Ridge Regression Estimators

In this section the generalized ridge regression estimator (GRRE) using a diagonal matrix K in place of kI is discussed and we distinguish between the non-stochastic and the stochastic ridge parameter. We discuss in §4.6.1 the GRRE when K is non-stochastic and in §4.6.2 the GRRE when K is estimated and therefore stochastic. In the latter case the choice of K usually depends on the data and this complicates the sampling distributions of GRRE

4.6.1 K non-stochastic

4.6.1.1 The GRR estimator

By using the model (3.1.1) to obtain the PC reparameterization we have, $Y = Z\delta + \epsilon$. Hoerl and Kennard (1970b) defined the general ridge regression estimator (GRRE) of δ as

$$\hat{\delta}_K = [Z'Z + K]^{-1}Z'Y \tag{4.6.1}$$

where $K = \text{diag}[k_1, k_2, \cdots, k_p]$, $k_i \geq 0 \;\; \forall \; i$, and K is known.

4.6.1.2 Properties of $\hat{\delta}_K$

1. Relationship to OLSE

By manipulation, the GRRE of δ (denoted by $\hat{\delta}_K$) can also be written as:

$$\begin{aligned}\hat{\delta}_K &= [\Delta^2 + K]^{-1}\Delta^2\hat{\delta} \qquad \text{(from (3.1.2) and (3.1.3))}\\ &= [I + \Delta^{-2}K]^{-1}\hat{\delta}\\ &= [I - [\Delta^2 + K]^{-1}K]\hat{\delta} \end{aligned} \tag{4.6.2}$$

and the GRRE of $\beta = V\delta$ is

$$\begin{aligned}\hat{\beta}_K &= V\hat{\delta}_{GR}\\ &= V[I - [\Delta^2 + K]^{-1}K]\hat{\delta}\\ &= V[I - [\Delta^2 + K]^{-1}K]V'\hat{\beta} \end{aligned} \tag{4.6.3}$$

When K = kI the GRRE of β is the ordinary ridge estimator ($\hat{\beta}_R$).

2. Expectation

$$\begin{aligned}E(\hat{\delta}_K) &= [I - [\Delta^2 + K]^{-1}K]E(\hat{\delta})\\ &= [I - [\Delta^2 + K]^{-1}K]\delta \end{aligned} \tag{4.6.4}$$

Thus the bias of $\hat{\delta}_K$ is

$$b = -[\Delta^2 + K]^{-1}K\delta \tag{4.6.5}$$

3. Variance

$$\begin{aligned}\operatorname{Var}(\hat{\delta}_K) &= \operatorname{Var}([\Delta^2 + K]^{-1}\Delta^2\hat{\delta}) \\ &= [\Delta^2 + K]^{-1}\Delta^2\operatorname{Var}(\hat{\delta})\Delta^2[\Delta^2 + K]^{-1} \\ &= \sigma^2[\Delta^2 + K]^{-1}\Delta^2\Delta^{-2}\Delta^2[\Delta^2 + K]^{-1} \\ &= \sigma^2[\Delta^2 + K]^{-1}\Delta^2[\Delta^2 + K]^{-1} \qquad (4.6.6)\end{aligned}$$

4. Then by using (4.6.6) and (4.6.5) the MSE of $\hat{\delta}_K$ is:

$$\begin{aligned}\operatorname{MSE}(\hat{\delta}_K) &= E(\hat{\delta}_K - \delta)(\hat{\delta}_K - \delta)' \\ &= \operatorname{Var}(\hat{\delta}_K) + bb' \\ &= \sigma^2[\Delta^2 + K]^{-2}\Delta^2 + [\Delta^2 + K]^{-1}K\delta\delta'K[\Delta^2 + K]^{-1}\end{aligned}$$
$$(4.6.7)$$

5. The total mean square error (TMSE) of $\hat{\delta}_K$ is

$$\begin{aligned}\operatorname{TMSE}(\hat{\delta}_K) &= \operatorname{tr}(\operatorname{MSE}(\hat{\delta}_K)) \\ &= \sum_{i=1}^{p} \{\sigma^2(\lambda_i + k_i)^{-2}\lambda_i + [\lambda_i + k_i]^{-2}k_i^2\delta_i^2\} \\ &= \sum_{i=1}^{p} \frac{\sigma^2\lambda_i + k_i^2\delta_i^2}{(\lambda_i + k_i)^2} \qquad (4.6.8)\end{aligned}$$

4.6.2 K stochastic

4.6.2.1 The GRR estimator

In the case of K a stochastic variable, the GRRE of δ is

$$\begin{aligned}\hat{\delta}_{\hat{K}} &= [\Delta^2 + \hat{K}]^{-1}\Delta^2\hat{\delta} \\ &= [I - [\Delta^2 + \hat{K}]^{-1}\hat{K}]\hat{\delta} \qquad (4.6.9)\end{aligned}$$

using (4.6.2), where $\hat{K}$ is an estimator of K, i.e $\hat{K} = \text{diag}[\hat{k}_1,\ldots,\hat{k}_p]$

Then the i-th element of $\hat{\delta}_{\hat{K}}$ is

$$[\hat{\delta}_{\hat{K}}]_i = \frac{\lambda_i}{\lambda_i + \hat{k}_i}\,\hat{\delta}_i \tag{4.6.10}$$

which shows that $[\hat{\delta}_{\hat{K}}]_i$ is a shrinkage estimator of δ_i where the shrinkage factor (∂_i) is:

$$\partial_i = \frac{\lambda_i}{\lambda_i + \hat{k}_i} \tag{4.6.11}$$

Minimizing (4.6.6) term-by-term to find an optimum value for k_i yields

$$k_i(\text{opt}) = \frac{\sigma^2}{\delta_i^2} \quad (i = 1,2,\ldots,p) \tag{4.6.12}$$

Hoerl and Kennard (1970b) then estimate k_i by using the LS estimates of σ^2 and δ_i, thus

$$\hat{k}_i = \frac{\hat{\sigma}^2}{\hat{\delta}_i^2} \tag{4.6.13}$$

Inserting these values into (4.6.11) gives

$$\begin{aligned}\partial_i &= \frac{\lambda_i}{\lambda_i + \hat{\sigma}^2/\hat{\delta}_i^2} \\ &= \frac{F_i}{F_i + 1}\end{aligned}$$

where $F_i = \lambda_i\hat{\delta}_i^2/\hat{\sigma}^2$, which is the same as the F-ratio for testing the hypothesis H_0: $\delta_i = 0$. Highly significant values of parameters estimates $\hat{\delta}_i$ will be scarcely shrunk at all, but relatively low values of the estimates $\hat{\delta}_i$ will be shrunken substantially.

4.6.2.2 Properties of $[\hat{\delta}_{\hat{K}}]_i$

Let $\tau_i^2 = \lambda_i \delta_i^2/\sigma^2$ (i.e. a non-centrality parameter of the F-distributions associated with δ_i) and $\nu = (n - p)$, then the first and second moments of $[\hat{\delta}_{\hat{K}}]_i$ are given by (Dwivedi, Srivastava and Hall (1980)):

$$E([\hat{\delta}_{\hat{K}}]_i) = \delta_i \exp(-\tau_i^2/2) \sum_{a=0}^{\infty} \sum_{j=0}^{\infty} \left(\frac{\nu - 1}{\nu}\right)^a \times \frac{(\tau_i^2/2)^j}{\Gamma(j+1)}$$

$$\times \frac{\Gamma(a+(\nu/2))\Gamma(j+(\nu+3)/2)\Gamma(j+(5/2))}{\Gamma(\nu/2)\Gamma(a+j+(\nu+5)/2)\Gamma(j+(3/2))} \tag{4.6.14}$$

$$E([\hat{\delta}_{\hat{K}}]_i^2) = \delta_i^2 \exp(-\tau_i^2/2) \sum_{a=0}^{\infty} \sum_{j=0}^{\infty} (a+1) \text{ x } \left(\frac{\nu - 1}{\nu}\right)^a$$

$$\times \frac{(\tau_i^2/2)^{j-1}}{\Gamma(j+1)} \times \frac{\Gamma(a+(\nu/2))\Gamma(j+(\nu+3)/2)\Gamma(j+(7/2))}{\Gamma(\nu/2)\Gamma(a+j+(\nu+7)/2)\Gamma(j+(1/2))} \tag{4.6.15}$$

Using (4.6.14) and (4.6.15) the total squared bias and total mean square error of $[\hat{\delta}_{\hat{K}}]_i$ can be computed. These authors then define the relative bias (RB) and relative mean square error (RMSE) of $[\hat{\delta}_{\hat{K}}]_i$ as:

$$RB([\hat{\delta}_{\hat{K}}]_i) = E[([\hat{\delta}_{\hat{K}}]_i - [\delta]_i)/[\delta]_i] \tag{4.6.16}$$

$$RMSE([\hat{\delta}_{\hat{K}}]_i) = E[([\hat{\delta}_{\hat{K}}]_i - [\delta]_i)/[\delta]_i]^2 \tag{4.6.17}$$

which are functions of ν and τ_i only. The efficiency of the LSE of δ_i

relative to this ridge estimator is

$$\begin{aligned}\eta_i &= (\mathrm{MSE}([\hat{\delta}_{\hat{K}}]_i)/\mathrm{Var}[\hat{\delta}]_i) \times 100 \\ &= 100\ \tau_i^2\ \mathrm{RMSE}([\hat{\delta}_{\hat{K}}]_i)\end{aligned} \tag{4.6.18}$$

By selecting different values for τ_i^2 (between 0.02 and 50) and ν (between 1 and 100) the authors concluded:

1. The ridge estimator $([\hat{\delta}_{\hat{K}}]_i)$ is biased in a direction opposite to the sign of the coefficient (i.e. shrunk toward zero).

2. The RB is a decreasing function of τ_i^2 and and increasing function of ν (the degrees of freedom for error).

3. RMSE decreases as τ_i^2 increases.

4. Changes in RMSE are more rapid for small values of τ_i^2 and ν.

5. The ridge estimator is more efficient than the LSE as long as $\tau_i^2 < 2$. (We suggest that to operationalise this criterion one estimates τ_i^2 by substitution of the sample estimators. Further research is planned to investigate behaviour of estimators chosen on this basis.)

4.6.3 Choice of k in K = kI

Hoerl, Kennard and Baldwin (1975) give an algorithm for the selection of k (in ordinary ridge estimation). They distinguish between two cases: First if $X'X = I$ then a minimum TMSE is obtained if $k = p\sigma^2/\beta'\beta$. Secondly, in the general form (as discussed in §4.6.1 and §4.6.2), a minimum MSE is obtained when $k_i = \sigma^2/\delta_i^2$ as in (4.6.12). These individual k_i are combined to form a single value for k. Large values of δ_i will blow up the ordinary mean of k_i, therefore they (HKB) suggested the harmonic mean (k_h) of the

k_i's. The two methods give the same value of k since

$$
\begin{aligned}
1/(k_h) &= \frac{1}{p}\sum_{i=1}^{p}(1/k_i) \\
&= \frac{1}{p}\sum_{i=1}^{p}(\delta_i^2/\sigma^2) \\
&= (1/p\sigma^2)\delta'\delta \\
&= (1/p\sigma^2)\beta'VV'\beta \\
\therefore\ k_h &= p\sigma^2/\beta'\beta
\end{aligned}
$$

and $$\hat{k}_h = p\hat{\sigma}^2/\hat{\beta}'\hat{\beta} \qquad (4.6.19)$$

Substituting $\hat{k}$ from (4.6.19) in the TMSE (4.6.8) give the estimated minimum TMSE as

$$
\begin{aligned}
TMSE(\hat{\delta}_K) &= \sum_{i=1}^{p}\frac{\hat{\sigma}^2\lambda_i+(p\hat{\sigma}^2/\hat{\delta}'\hat{\delta})^2\hat{\delta}_i^2}{(\lambda_i + (p\hat{\sigma}^2/\hat{\delta}'\hat{\delta}))^2} \\
&= \sum_{i=1}^{p}\frac{\hat{\sigma}^2\lambda_i(\hat{\delta}'\hat{\delta})^2+(p\hat{\sigma}^2)^2\hat{\delta}_i^2}{(\hat{\delta}'\hat{\delta}\lambda_i + p\hat{\sigma}^2)^2}
\end{aligned}
$$

By using simulation on three data sets they show that the algorithm (discussed in the previous paragraph for selecting k) has the following properties (p111):

1. The use of the ridge estimator with biasing parameter $\hat{k}_h = p\hat{\sigma}^2/\hat{\beta}'\hat{\beta}$ has a probability greater than 0.5 of producing estimates with a smaller MSE than that of OLSE.

2. The probability of a smaller TMSE (than that of LSE) using $\hat{k}_h$ increases as p (the number of independent variables) increases.

3. The probability of a smaller TMSE using $\hat{k}_h$ increases as X′X becomes less well-conditioned.

4. The probability of a smaller TMSE using $\hat{k}_h$ increases as the noise (measured by $\hat{\sigma}^2$) increases.

They concluded that every ridge trace should have the point corresponding to $k = \hat{k}_h$ computed explicitly.

4.7 Jackknifed Ridge Estimator

One disadvantage of the ridge estimator is that it may have a serious bias. To reduce this bias Singh, Chaubey and Dwivedi (1986) use the jackknife procedure (as described in §1.8.2 for OLSE) on the generalized ridge estimator to construct a jackknifed ridge estimator (JRE). Some authors (e.g. Nomura (1988)) call this JRE the almost unbiased generalized ridge regression (AUGRR) estimator. This estimator is similar in form to the ridge estimator and has the same asymptotic properties. For the moment K is fixed (non-stochastic).

4.7.1 The JR estimator

Let Y_{-i} and Z'_{-i} denote respectively the vector Y with its i-th coordinate deleted and the matrix Z′ with its i-th column deleted, and let $(\hat{\delta}_K)_{-i}$ be given by (4.6.1) with Z′ and Y replaced by Z'_{-i} and Y_{-i} respectively. Clearly Z_{-i} is the matrix Z with its i-th row deleted.

$$(\hat{\delta}_K)_{-i} = [Z'_{-i}Z_{-i} + K]^{-1}Z'_{-i}Y_{-i} \qquad (4.7.1)$$

$$\text{Note that } Z'Z = [Z'_{-i} \ \ z_i]\begin{bmatrix} Z_{-i} \\ z'_i \end{bmatrix}$$

$$= [Z'_{-i}Z_{-i}+z_i z'_i] \qquad (4.7.2)$$

$$\text{and}\quad Z'Y = [Z'_{-i}\ \ z_i]\begin{bmatrix} Y_{-i} \\ y_i \end{bmatrix}$$
$$= [Z'_{-i}Y_{-i} + z_i y_i] \qquad (4.7.3)$$

Define $A = Z'Z + K$, $w_i = z_i'A^{-1}z_i$ (corresponding to h_i in the OLS estimation) and $u_i = y_i - z_i'\hat{\delta}_K$ (corresponding to the OLS $\hat{\epsilon}_i$). Then (4.7.1) can be be written as

$$\begin{aligned}
(\hat{\delta}_K)_{-i} &= [Z'Z - z_i z_i' + K]^{-1}[Z'Y - z_i y_i] \\
&= [A - z_i z_i']^{-1}[Z'Y - z_i y_i] \\
&= [A^{-1} + A^{-1}z_i(1- z_i'A^{-1}z_i)^{-1}z_i'A^{-1}][Z'Y - z_i y_i] \\
&= A^{-1}Z'Y - A^{-1}z_i y_i + A^{-1}z_i(1-w_i)^{-1}z_i'A^{-1}[Z'Y - z_i y_i] \\
&= (\hat{\delta}_K) - A^{-1}z_i(1-w_i)^{-1}[y_i(1-w_i) - z_i'A^{-1}[Z'Y - z_i y_i]] \\
&\qquad \text{(from (4.6.1))} \\
&= (\hat{\delta}_K) - A^{-1}z_i(1-w_i)^{-1}[y_i - y_i w_i - z_i'A^{-1}Z'Y + w_i y_i] \\
&= (\hat{\delta}_K) - A^{-1}z_i(1-w_i)^{-1}[y_i - z_i'(\hat{\delta}_K)] \\
&= \hat{\delta}_K - \frac{A^{-1}z_i}{(1-w_i)}\Big[u_i\Big] \qquad (4.7.4)
\end{aligned}$$

In the same manner as (1.8.8) the pseudovalues are defined as

$$P_i = n\hat{\delta}_K - (n-1)(\hat{\delta}_K)_{-i} \qquad (4.7.5)$$

and the jackknife ridge estimator (JRE) is

$$\begin{aligned}
\hat{\delta}_J &= \frac{1}{n}\Sigma P_i \\
&= \frac{1}{n}\Sigma(n\hat{\delta}_K - (n-1)(\hat{\delta}_K)_{-i}) \\
&= \hat{\delta}_K + \frac{n-1}{n} A^{-1} \sum_{i=1}^{n} \frac{z_i u_i}{(1-w_i)} \qquad (4.7.6)
\end{aligned}$$

Singh, Chaubey and Dwivedi (1986) reject these pseudo-values (P_i), as they are defined symmetrically with respect to the observations, while the regression models are often unbalanced with the lack of balance reflecting through the 'squared distance' w_i. Further, since the variance of the differences $(\hat{\delta}_K - (\hat{\delta}_K)_{-i}) = A^{-1}\frac{z_i u_i}{(1-w_i)}$ is an increasing function of w_i, they define the pseudo-values in the manner of (1.8.13) as

$$Q_i = \hat{\delta}_K + n(1-w_i)(\hat{\delta}_K - (\hat{\delta}_K)_{-i}) \qquad (4.7.7)$$

Then the weighted jackknife estimator (denoted by $\hat{\delta}_{JW}$) is

$$\begin{aligned}\hat{\delta}_{JW} &= \frac{1}{n}\Sigma Q_i \\ &= \hat{\delta}_K + A^{-1}\Sigma z_i u_i \quad \text{(using (4.7.4)} \\ &= \hat{\delta}_K + A^{-1}\Sigma z_i(y_i - z_i'\hat{\delta}_K) \\ &= \hat{\delta}_K + A^{-1}[\Sigma z_i y_i - \Sigma z_i z_i'\hat{\delta}_K] \\ &= \hat{\delta}_K + A^{-1}[Z'Y - Z'Z\hat{\delta}_K] \\ &= \hat{\delta}_K + A^{-1}[A\hat{\delta}_K - Z'Z\hat{\delta}_K] \quad (\hat{\delta}_K = A^{-1}Z'Y) \\ &= \hat{\delta}_K + A^{-1}K\hat{\delta}_K \quad (A = Z'Z + K) \qquad (4.7.8)\end{aligned}$$

When $k = k_1 = \dots = k_p$ the AUGRR estimator is called the almost unbiased ordinary ridge regression (AUORR) estimator (see (4.7.21) for an estimate of k).

4.7.2 Properties of $\hat{\delta}_{JW}$ (non-stochastic K)

1. Relationship to OLSE

The weighted JRE can also be written as

$$\begin{aligned}\hat{\delta}_{JW} &= [I + A^{-1}K][I - A^{-1}K]\hat{\delta} \quad \text{(from (4.6.2))} \\ &= [I - [A^{-1}K]^2]\hat{\delta} \qquad (4.7.9)\end{aligned}$$

and the i-th element of $\hat{\delta}_{JW}$ is:

$$[\hat{\delta}_{JW}]_i = \left[1 - \frac{k_i^2}{(\lambda_i + k_i)^2}\right]\hat{\delta}_i \qquad (4.7.10)$$

2. Expectation

$$E(\hat{\delta}_{JW}) = [I - [A^{-1}K]^2]\delta$$

$$\text{Bias}(\hat{\delta}_{JW}) = - [A^{-1}K]^2\delta \qquad (4.7.11)$$

and the bias of the i-th element is:

$$\text{Bias}[(\hat{\delta}_{JW})_i] = -\left[\frac{k_i^2}{(\lambda_i+k_i)^2}\right]\delta_i \qquad (4.7.12)$$

Note that the JRE ($\hat{\delta}_{JW}$) has smaller bias than that of the GRRE, where bias of the i-th element $[(\hat{\delta}_K)]_i$ is $-\left[\frac{k_i}{(\lambda_i+k_i)}\right]\delta_i$, thus

$$|\text{Bias}[(\hat{\delta}_K)]_i| - |\text{Bias}[(\hat{\delta}_{JW})_i]| = \left[\frac{\lambda_i k_i}{(\lambda_i+k_i)^2}\right]|\delta_i| > 0 \qquad (4.7.13)$$

3. Variance

$$\begin{aligned}\text{Var}([(\hat{\delta}_{JW})] &= [I - [A^{-1}K]^2]\text{Var}(\hat{\delta})[I - [A^{-1}K]^2] \\ &= [I - [A^{-1}K]^2]\sigma^2\Delta^{-2}[I - [A^{-1}K]^2])\end{aligned}$$

and the variance of the i-th component is:

$$\begin{aligned}\text{Var}[(\hat{\delta}_{JW})_i] &= \left[1- \frac{k_i^2}{(\lambda_i+k_i)^2}\right]^2 \frac{\sigma^2}{\lambda_i} \\ &= \sigma^2\left[\frac{\lambda_i(\lambda_i+2k_i)^2}{(\lambda_i+k_i)^4}\right] \qquad (4.7.14)\end{aligned}$$

4. By using (4.7.14) and (4.2.12) the MSE of the components of $\hat{\delta}_{JW}$ is

$$\begin{aligned}\text{MSE}[(\hat{\delta}_{JW})_i] &= \text{Var}[(\hat{\delta}_{JW})_i] + \text{bias}^2 \\ &= \sigma^2\left[\frac{\lambda_i(\lambda_i+2k_i)^2}{(\lambda_i+k_i)^4}\right] + \left[\frac{k_i^4}{(\lambda_i+k_i)^4}\right]\delta_i^2 \qquad (4.7.15)\end{aligned}$$

5. The total mean square error (TMSE) of $\hat{\delta}_{JW}$ is obtained from (4.7.15) and the definition of TMSE (1.8.2):

$$\text{TMSE}(\hat{\delta}_{JW}) = \sum_{i=1}^{p}\left\{\sigma^2\frac{\lambda_i(\lambda_i+2k_i)^2}{(\lambda_i+k_i)^4} + \frac{k_i^4}{(\lambda_i+k_i)^4}\delta_i^2\right\} \tag{4.7.16}$$

4.7.3 Additional comments on $(\hat{\delta}_{JW})$ (non-stochastic K)

In comparing component-wise the JRE (or AUGRR estimator) with the GRRE and the OLSE in terms of the MSE, the following theorems were developed by Nomura (1988). The comparison of the various estimators will be discussed further in Chapter 8.

From (4.6.7) and (4.6.15) we have

$$\begin{aligned}
\text{MSE}(\hat{\delta}_i) &= \sigma^2/\lambda_i, \qquad i = 1,\ldots,p \\
\text{MSE}[(\hat{\delta}_K)_i] &= [\sigma^2\lambda_i+\delta_i^2 k_i{}^2]/(\lambda_i+k_i)^2 \\
\text{MSE}[(\hat{\delta}_{JW})_i] &= \{\sigma^2\lambda_i^3+4\sigma^2\lambda_i^2 k_i+4\sigma^2\lambda_i k_i^2+\delta_i^2 k_i^4\}/(\lambda_i+k_i)^4
\end{aligned}$$

Theorem 4.7.3.1

1. $\text{MSE}[(\hat{\delta}_{JW})_i] > \text{MSE}[(\hat{\delta}_K)_i]$ for $0 < k_i < K1_i$,
2. $\text{MSE}[(\hat{\delta}_{JW})_i] < \text{MSE}[(\hat{\delta}_K)_i]$ for $K1_i < k_i < \infty$,

where $K1_i = [3\sigma^2 - \lambda_i\delta_i^2 + \{(3\sigma^2+ \lambda_i\delta_i^2)^2 + 4\sigma^2\lambda_i\delta_i^2\}^{1/2}]/4\delta_i^2 > 0$.
Proof is given in Nomura (1988, p732).

Theorem 4.7.3.2

1. If $\lambda_i\delta_i^2 - \sigma^2 \leq 0$, then

   $\text{MSE}[(\hat{\delta}_{JW})_i] < \text{MSE}[(\hat{\delta})_i]$ for $0 < k_i < \infty$,

2. If $\lambda_i \delta_i^2 - \sigma^2 \leq 0$, then there exists a strictly positive constant $K2_i = [2\sigma^2\lambda_i + (2\sigma^4\lambda_i{}^2 + 2\sigma^2\lambda_i^3\delta_i^2)^{1/2}]/(\lambda_i\delta_i^2 - \sigma^2 > 0$ such that $MSE[(\hat{\delta}_{JW})_i] < MSE[(\hat{\delta})_i]$ for $0 < k_i < K2_i$ and

   $MSE[(\hat{\delta}_{JW})_i] > MSE[(\hat{\delta})_i]$ for $K2_i < k_i < \infty$.

Proof is given in Nomura (1988, p733)

Furthermore, differentiating the $MSE[(\hat{\delta}_{JW})_i]$ with respect to k_i yields:

$$\frac{dMSE[(\hat{\delta}_{JW})_i]}{dk_i} = \frac{4\lambda_i k_i(\delta_i^2 k_i^2 - 2\sigma^2 k_i - \sigma^2\lambda_i)}{(\lambda_i + k_i)^5} \qquad (4.7.17)$$

thus the optimal value of k_i for which $MSE[(\hat{\delta}_{JW})_i]$ is a minimum is:

$$k_i(opt) = \frac{\sigma^2 + \sqrt{\sigma^4 + \sigma^2\lambda_i\delta_i^2}}{\delta_i^2} \qquad (4.7.18)$$

If $\lambda_i\delta_i^2 - \sigma^2 > 0$ then

$$K1_i < k_i(opt) < K2_i \qquad (4.7.19)$$

We observe that this is a condition that depends on the data source not the data and is therefore not operationable.

In the case of AUORR defined below (4.7.8), a single value of k can be formed by combining all the k_i through the harmonic mean (compare this with §4.6.3). Thus

$$k_h = p\sigma^2/(\sum_{i=1}^{p}[\delta_i^2/\{1+\sqrt{1+\lambda_i(\delta_i^2/\sigma^2)}\}]) \qquad (4.7.20)$$

4.7.4 Additional comments on ($\hat{\delta}_{JW}$) (stochastic K)

In equation (4.7.20) an estimate of k_h can be obtained from estimating δ and σ by their LSE's

$$\hat{k}_h = p\hat{\sigma}^2/(\sum_{i=1}^{p} [\hat{\delta}_i^2/\{1+\sqrt{1+\lambda_i(\hat{\delta}_i^2/\hat{\sigma}^2)}\}]) \qquad (4.7.21)$$

where $\hat{\sigma}^2 = (Y - Z\hat{\delta})'(Y - Z\hat{\delta})/(n-p)$.

Nomura (1988) performed a simulation study on six estimators namely:

1. OLSE
2. AUORR estimator using the operational $\hat{k}_h$ of (4.7.21)
3. Ordinary RR estimator using the Hoerl, Kennard and Baldwin ridge parameter (4.6.19)
4. AUORR estimator using the operational $\hat{k}_h$ of (4.6.19)
5. Ordinary RR estimator using the $\hat{k}$ as suggested by Lawless and Wang (1976) $\hat{k}_{LW} = p\hat{\sigma}^2/\sum_{i=1}^{p} \lambda_i \hat{\delta}_i^2$
6. AUORR estimator using the $\hat{k}_{LW}$.

For a description of the simulation, the data and the results see pp735-742 of his article. He reported the following findings for different estimators referenced 1 to 6 as above.

(1) The absolute value of the bias for 2 and 4 is smaller than that of 3 and 5. This result coincides with (4.7.13) for k non-stochastic.

(2) The estimators 2 and 4 are more efficient in terms of the MSE and predictive MSE (defined as the $E(\tilde{\delta} - \delta)'X'X(\tilde{\delta} - \delta))$ than 3 and 4 when the signal to noise ratio is relatively large, and more efficient than 1 (OLSE) when the signal to noise ratio is relatively small.

(3) The estimator 2 performs very well against 4.

We note that these findings are applicable to the data set and models used by Nomura and are thus only suggestive of what one might expect.

The AUGRR discussed so far was due to Singh, Chaubey and Dwivedi (1986) and was not operationable. Ohtani (1986) considered the operational AUGRR estimator based on the idea of Kadiyala (1984). Following Kadiyala (1984) the almost unbiased generalized ridge estimator (non-operationable) for the i-th element of δ is:

$$[\delta_0]_i = [\delta_K]_i + k_i[\delta_K]_i/(\lambda_i + k_i) \qquad (4.7.22)$$

In the above expression if k_i and $[\delta_K]_i$ are replaced by $\hat{k}_i$ and $[\hat{\delta}_K]_i$ we obtain the following operational AUGRR estimator:

$$\begin{aligned}[\hat{\delta}_{\hat{0}}]_i &= (1 + \hat{k}_i/(\lambda_i + \hat{k}_i))[\hat{\delta}_{\hat{K}}]_i \\ &= ((\lambda_i + 2\hat{k}_i)/(\lambda_i + \hat{k}_i))[\lambda_i\hat{\delta}_i/(\lambda_i + \hat{k}_i)] \quad \text{(from (4.6.10))} \\ &= \frac{(\lambda_i\hat{\delta}_i^2 + 2\hat{\sigma}^2)\lambda_i\hat{\delta}_i^3}{(\lambda_i\hat{\delta}_i^2 + \hat{\sigma}^2)^2} \quad \text{(from (4.6.13))} \qquad (4.7.23)\end{aligned}$$

Let $\tau_i^2 = \lambda_i\delta_i^2/\sigma^2$ and $\nu = n - p$, then the first and second moments of $[\hat{\delta}_{\hat{0}}]_i$ are

$$\begin{aligned}E([\hat{\delta}_{\hat{0}}]_i) &= [\delta_i \exp(-\tau_i^2/2)/\{\sqrt{\pi}\,\Gamma(\nu/2) \\ &\quad \times \sum_{g=0}^{\infty} [2^{g+1}(\tau_i^2)^g \Gamma(g+(\nu+3)/2)/(2g+1)! \\ &\quad \times \int_0^1 [(t+2(1-t)/\nu)/(t+(1-t)/\nu)^2] \\ &\quad \times t^{g+3/2}(1-t)^{\nu/2-1}dt \qquad (4.7.24)\end{aligned}$$

$$\begin{aligned} E([\hat{\delta}_{\hat{0}}]_i^2) = {} & [\delta_i^2 \exp(-\tau_i^2/2)/\{\sqrt{\pi}\,\Gamma(\nu/2) \\ & \times \sum_{g=0}^{\infty} [2^{g+1}(\tau_i^2)^{g-1}\Gamma(g+(\nu+3)/2)/(2g)!] \\ & \times \int_0^1 [(t+2(1-t)/\nu)^2/(t+(1-t)/\nu)^4] \\ & \times t^{g+5/2}(1-t)^{\nu/2-1}dt \end{aligned} \qquad (4.7.25)$$

Then the relative bias (RB) and relative mean square error (RMSE) of $[\hat{\delta}_{\hat{0}}]_i^2$ as defined in (4.6.16) and (4.6.17) are

$$RB([\hat{\delta}_0]_i^2) = E[([\hat{\delta}_0]_i - [\hat{\delta}]_i)/[\hat{\delta}]_i] \qquad (4.7.26)$$

$$RMSE([\hat{\delta}_{\hat{0}}]_i^2) = E[([\hat{\delta}_0]_i - [\hat{\delta}]_i)^2/[\hat{\delta}]_i^2] \qquad (4.7.27)$$

which are functions of τ_i^2 and ν only. Ohtani (1986) then defined the relative efficiencies of the AUGRR estimator and the GRRE to the OLSE as:

$$REFF([\hat{\delta}_{\hat{0}}]_i) = MSE([\hat{\delta}]_i)/MSE([\hat{\delta}_{\hat{0}}]_i) \qquad (4.7.28)$$

$$REFF([\hat{\delta}_{\hat{K}}]_i) = MSE([\hat{\delta}]_i)/MSE([\hat{\delta}_{\hat{K}}]_i) \qquad (4.7.29)$$

In (4.6.18) efficiency of OLSE to GRRE was defined as the inverse of (4.7.29) and presented as a percentage.

By selecting different values for τ_i^2 (between 0.25 and 0.40) and ν (between 10 and 60) the author concluded:

1. The reduction in bias is greater for AUGRR estimator than for GRRE but the relative efficiency of the AUGRR estimator is less than the GRRE.

2. The AUGRR estimator is more efficient than the OLSE as long as the value of $\tau_i^2 < 1.5$ (For GRRE the corresponding value was 2).

Ohtani (1986) concluded that the AUGRR estimatior is rather inferior to the GRRE.

4.8 Summary

In this chapter we introduced ridge, generalized ridge and the jackknifed ridge estimators. The estimators were defined, and their properties discussed. Methods of estimating k of K were introduced.

Chapter 5

FRACTIONAL PRINCIPALS IN ESTIMATION

In this chapter the principle of fractional estimation using the augmented matrix and in terms of shrinkage estimators, will be discussed.

5.1 Shrunken Estimators

5.1.1 The SH estimator

Mayer and Willke (1973) discuss the use of several shrunken (SH) estimators of the form:

$$\hat{\beta}_{SH} = d\hat{\beta} \qquad (5.1.1)$$

where $0 \leq d \leq 1$ is a deterministically (fixed) or stochastically defined constant. By using (1.3.6) this estimator can be written as

$$\hat{\beta}_{SH} = d \sum_{i=1}^{p} v_i c_i/\lambda_i \qquad (5.1.2)$$

Thus the effect of d/λ_i is to reduce the magnitude of the estimates.

5.1.2 Properties (d fixed)

1. Relationship to OLSE

From the definition it is clear that

$$\hat{\beta}_{SH} = d\hat{\beta} \qquad (5.1.1)$$

2. Expectation

$$E[\hat{\beta}_{SH}] = d\beta = \beta - (1-d)\beta \qquad (5.1.3)$$

Thus the estimator is biased, and the bias is

$$b = -(1-d)\beta \qquad (5.1.4)$$

3. Variance

$$\mathrm{Var}(\hat{\beta}_{SH}) = d^2\mathrm{Var}(\hat{\beta})$$

$$= d^2\sigma^2(X'X)^{-1} \qquad (5.1.5)$$

$$= \sigma^2 d^2 \sum_{i=1}^{p} v_i v_i'/\lambda_i \qquad (5.1.6)$$

4. $$\mathrm{MSE}[\hat{\beta}_{SH}] = \mathrm{Var}(\hat{\beta}_{SH}) + bb'$$

$$= \sigma^2 d^2(X'X)^{-1} + (1-d)^2\beta\beta' \qquad (5.1.7)$$

5. $$\mathrm{TMSE}[\hat{\beta}_{SH}] = \mathrm{tr}(\mathrm{Var}(\hat{\beta}_{SH}) + bb')$$

$$= \sigma^2 d^2 \sum_{i=1}^{p} 1/\lambda_i + (1-d)^2\mathrm{tr}(\beta\beta')$$

$$= \sigma^2 d^2 \sum_{i=1}^{p} 1/\lambda_i + (1-d)^2\beta'\beta \qquad (5.1.8)$$

The SHE with $0 \leq d \leq 1$ and d non-stochastic guarantees smaller variances than OLS for the parameter estimators. The bias, however could offset the reduction in variances. Furthermore the SHE is (total mean square) admissible (see comment 7 in the ridge section), as proved by Mayer and Willke (1973):

$$\text{TMSE}[\hat{\beta}_{SH}] < \text{tr}[\text{Var}(\hat{\beta})] \text{ if and only if}$$

$$d > \frac{\beta'\beta - \text{tr}(\text{Var}(\hat{\beta}))}{\beta'\beta + \text{tr}(\text{Var}(\hat{\beta}))} \tag{5.1.9}$$

Since the RHS in (5.1.9) is less than 1, there does exist a number d < 1 such that $\text{TMSE}[\hat{\beta}_{SH}] < \text{tr}[\text{Var}(\hat{\beta})]$. In fact if we minimize (5.1.8) with respect to d we get:

$$\frac{\delta \; \text{TMSE}(\hat{\beta}_{SH})}{\delta \; d} = \sigma^2 2d \sum_{i=1}^{p} 1/\lambda_i - 2(1-d)\beta'\beta = 0$$

thus

$$\sigma^2 d \sum_{i=1}^{p} 1/\lambda_i + d\beta'\beta = \beta'\beta$$

$$d = \frac{\beta'\beta}{\beta'\beta + \text{tr}(\text{Var}(\hat{\beta}))} \tag{5.1.10}$$

The d in (5.1.10) minimize (5.1.8), and the minimum TMSE of the SHE is

$$\begin{aligned}
\min\{\text{TMSE}[\hat{\beta}_{SH}]\} &= d^2\text{tr}(\text{Var}(\hat{\beta}) + (1-2d+d^2)\beta'\beta \\
&= d^2(\beta'\beta + \text{tr}(\text{Var}(\hat{\beta})) + \beta'\beta - 2d\beta'\beta \\
&= \frac{(\beta'\beta)^2}{\beta'\beta + \text{tr}(\text{Var}(\hat{\beta}))} + \beta'\beta - 2\frac{(\beta'\beta)^2}{\beta'\beta + \text{tr}(\text{Var}(\hat{\beta}))} \\
&= \frac{\beta'\beta(\text{tr}(\text{Var}(\hat{\beta}))}{\beta'\beta + \text{tr}(\text{Var}(\hat{\beta}))} \qquad (5.1.11)
\end{aligned}$$

5.1.3 Properties (d stochastic)

The only known SHE of the form given in (5.1.1) with d stochastic, and possessing any optimal properties, is the estimator due to James and Stein (1961).

Providing $p \geq 3$ and $X'X = I$ the SHE is given by (5.1.1) with

$$d = \max\{0, (1-cv/\hat{\beta}'\hat{\beta})\}$$

where $0 < c < 2(p-2)/(v+2)$ and v is the degrees of freedom for the residual sums of squares in OLS. The drawback of this SHE is the requirement that $p \geq 3$ and $X'X = I$. Particularly the latter requirement eliminates most practical regression problems and completely eliminates the possible presence of collinearity.

Sclove (1968) modified the James-Stein estimator by suggesting shrinkage of just a subset of the components of $\hat{\beta}$. If the shrinkage is applied to those p-r (where r is the rank of X) components with smallest eigenvalues, his estimator (denoted be $\hat{\beta}_{SSH}$) may be written as,

$$\hat{\beta}_{SSH} = \begin{bmatrix} I_r & 0 \\ 0 & dI_{p-r} \end{bmatrix} \hat{\beta} \qquad (5.1.12)$$

for $0 \leq d \leq 1$. Note that for $d = 0$ this estimator reduces to the PCE.

5.2 Fractional Principal Component (FPC) regression

In this section we firstly show that most of the biased estimators so far discussed, can be considered as FPC estimators. Then we obtain two new biased estimators due to Lee and Birch (1988).

5.2.1 FPC estimators

Consider the model defined in (3.1.1):

$$Y = Z\delta + \epsilon \qquad (5.2.1)$$

with $$Z'Z = \Delta^2 \quad \text{(from (3.1.2))} \qquad (5.2.2)$$

and $$\hat{\delta} = (Z'Z)^{-1}Z'Y \quad \text{(from (3.1.3))}$$

$$= V'\hat{\beta} \qquad (5.2.3)$$

Then a typical member of the class of FPC regression estimators is defined as:

$$\hat{\delta}_{\text{FPC}} = F\hat{\delta} \qquad (5.2.4)$$

where $F = \text{Diag}(f_1, f_2, \ldots, f_p)$ and $0 \leq f_j \leq 1$, $j = 1, \ldots, p$. In addition the FPC estimator of $\hat{\beta}$ is

$$\hat{\beta}_{\text{FPC}} = V\hat{\delta}_{\text{FPC}} = VF\hat{\delta} = VFV'\hat{\beta} \qquad (5.2.5)$$

The diagonal matrix F is termed the fraction matrix and the diagonal elements, f_j, are called the fractions. (This FPC estimator is one of a special class of linear transforms of $\hat{\delta}$ introduced by Mayer and Willke (1973)).

5.2.2 Relationship to other estimators

1. Least squares:

$$\hat{\delta}_{\text{FPC}} = \hat{\delta} \text{ with } F = I.$$

2. Ridge

The ridge estimator of δ is

$$\hat{\delta}_R = (\Delta^2 + kI)^{-1}\Delta^2\hat{\delta} \quad \text{(from (4.3.2)}$$
$$= F\hat{\delta}$$

with $F = (\Delta^2 + kI)^{-1}\Delta^2$, where $k > 0$

3. Generalized ridge

The i-th element of the generalized ridge estimator is

$$[\hat{\delta}_K]_i = \lambda_i(\lambda_i+k_i)^{-1}\hat{\delta}_i$$

Thus, take $f_i = \lambda_i(\lambda_i+k_i)^{-1}$, $k_i \geq 0$.

4. Principal components

The PCE applies 1 to the first p-r components (those components retained in the estimator) and 0 to the rest (those components that the estimator deletes).

5. Shrinkage

The concept of the FPC estimator is closely related to that of SHE's (5.1.1). Both estimators shrink the length of the LSE vector of the parameters toward the origin. For SHE the fractions are constant ($f_i = d$ for all i) but FPC takes the individual PC's and shrinks each of them. Thus some PC's receive greater emphasis in estimating δ than others.

5.2.3 Properties of FPC estimators

1. Expectation

$$E(\hat{\delta}_{FPC}) = F\delta$$
$$= \delta - [I-F]\delta \qquad (5.2.6)$$

Thus, the bias is

$$b = - [I-F]\delta \tag{5.2.7}$$

2. Variance

$$\begin{aligned} Var(\hat{\delta}_{FPC}) &= FVar(\hat{\delta})F \\ &= \sigma^2 F\Delta^{-2}F \end{aligned} \tag{5.2.8}$$

3. $$\begin{aligned} MSE(\hat{\delta}_{FPC}) &= Var(\hat{\delta}_{FPC}) + bb' \\ &= \sigma^2 F\Delta^{-2}F + [I-F]\delta\delta'[I-F] \end{aligned} \tag{5.2.9}$$

Thus FPCE is an improvement over OLS in terms of MSE if $MSE(\hat{\delta})-MSE(\hat{\delta}_{FPC}) = \sigma^2\Delta^{-2}(I-F^2) - (I-F)\delta\delta'(I-F) = S$ (say), is a positive semi-definite matrix. S will be positive semi-definite if $y'Sy \geq 0$ for any vector y:nx1. Thus, where $(I-F^2)^-$ is the generalized inverse of $(I-F^2)$, we have

$$y'\Delta^2 S(I-F^2)^- y = y'[\sigma^2 I - \Delta^2(I-F)\delta\delta'(I-F)(I-F^2)^-]y \quad \text{for all } y$$

Consequently, the necessary and sufficient condition for S to be positive semi-definite is

$$\sigma^2 \geq \delta'\Delta^2(I-F)^2(I-F^2)^-\delta \tag{5.2.10}$$

Thus the positive semi-definitenes of S will depend on the vector δ, the degree of collinearity (Δ), and σ.

4. $$\begin{aligned} TMSE(\hat{\delta}_{FPC}) &= tr[MSE((\hat{\delta}_{FPC})] \\ &= \sigma^2 \sum_{i=1}^{p} f_i^2/\lambda_i + \sum_{i=1}^{p} (1-f_i)^2\delta_i^2 \end{aligned} \tag{5.2.11}$$

We note that:

(i) If $f_i \rightarrow 0$, then the i-th diagonal element of $\text{Var}(\hat{\delta}_{FPC}) \rightarrow 0$ and the $(\text{bias})^2 \rightarrow \delta_i^2$.

(ii) If $f_i \rightarrow 1$, then the i-th diagonal element of $\text{Var}(\hat{\delta}_{FPC}) \rightarrow [\text{Var}(\hat{\delta})]_i$ and the $(\text{bias})^2 \rightarrow 0$.

5.2.4 Optimal values for fractions

One set of optimal values of the fractions is obtained by minimizing $\text{TMSE}(\hat{\delta}_{FPC})$ with respect to the f_j's. Thus

$$\frac{d\text{TMSE}(\hat{\delta}_{FPC})}{df_j} = 2\sigma^2 f_j/\lambda_j - 2\delta_j(1-f_j)$$

Let the j-th optimal value be denoted by f_j^0, thus

$$f_j^0 = \delta_j^2\lambda_j(\sigma^2 + \delta_j^2\lambda_j)^{-1} \tag{5.2.12}$$

and the min TMSE will be

$$\min\{\text{TMSE}(\hat{\delta}_{FPC})\} = \sigma^2 \sum_{i=1}^{p} \delta_j^4\lambda_j^2(\sigma^2+\delta_j^2\lambda_j)^{-2}/\lambda_i + \sum_{i=1}^{p} \sigma^4(\sigma^2+\delta_j^2\lambda_j)^{-2}\delta_j^2$$

$$= \sum_{i=1}^{p} (\sigma^2/\lambda_i)[\delta_j^2\lambda_i](\sigma^2+\delta_j^2\lambda_j)^{-1}$$

$$= \sum_{i=1}^{p} (\sigma^2/\lambda_i)f_i^0 \tag{5.2.13}$$

$$= \sum_{i=1}^{p} (\sigma^2/\lambda_i)[1-(1-f_i^0)]$$

$$= \text{TMSE}(\hat{\delta}) - \sum_{i=1}^{p} (\sigma^2/\lambda_i)(1-f_i^0) \tag{5.2.14}$$

The second term in the above equation (i.e. $\sum_{i=1}^{p} (\sigma^2/\lambda_i)(1-f_i^0))$ is the maximum reduction in TMSE when the optimal FPC is used instead of OLS. This quantity will always be positive and will increase as the degree of collinearity increases (because of the term λ_i^{-1}).

The theoretically optimal fraction matrix cannot be used in practice since the f_j^0, defined in (5.2.12), contain the unknown parameters δ_j and σ^2. Therefore Lee and Birch (1988) suggested two new biased estimators both using the optimal FPC but with different approaches in estimating f_j^0.

5.2.5 Optimal FPC estimators

In §4.6.1 (generalized ridge) it was shown that an optimal value for k_i is obtained when $k_i = \sigma^2/\delta_i^2$. Hoerl and Kennard (1970) estimated k_i by estimating σ^2 and δ_i^2 as their OLS estimates. They used an iterative method to obtain a new value for $\hat{k}_i = \hat{\sigma}^2/\hat{\delta}_K^2$ and continued until the estimate stabilised. The same method is applied to form the iterative (optimal) FPC estimator (abbreviated as FPCI). For the FPCI estimator the iterative scheme of the optimal fraction is:

$$f_{j,GR}(t+1) = \frac{\lambda_j}{\lambda_j + s^2/[\hat{\delta}_K(t)]_j^2}, \quad t=0,1,2,... \tag{5.2.15}$$

where t denotes the iteration number, s^2 is the OLS estimate of σ^2 and $[\hat{\delta}_K(t)]_j$ is the generalized ridge estimate of δ_j at the t-th iteration with $[\hat{\delta}_K(0)]_j = \hat{\delta}_j$. The iteration continues until there is stability achieved in the length of the generalized ridge estimator $([\hat{\delta}_K(t)]_j)$.

In the presence of collinearity the starting values of OLS may be severely perturbed and therefore it may be more beneficial to use a biased estimator as initial value. If one considers a PCE as the initial value, then the iterative scheme (5.2.15) can be used to compute the fractions, where the starting value for $[\hat{\delta}_K(0)]_j$ would be $[\hat{\delta}_{PC}]_j$ and s^2 is replaced with $s^2(t) = (Y-Za(t))'(Y-Za(t))/(n-p)$. Here $a(t)$ is the estimate of δ at the t-th iteration, $a(0) = \hat{\delta}_{PC}$. Fractions obtained in this way will be denoted by $f_{j,PC}(t+1)$.

The final resulting estimator, denoted by $\hat{\delta}_{FPCI}$, with the limiting fraction matrix F_{PCI}, is

$$\hat{\delta}_{FPCI} = F_{PCI}\hat{\delta} \tag{5.2.16}$$

where $F_{PCI} = \mathrm{Diag}(f^*_{1,PC},\ldots,f^*_{p,PC})$, and $f^*_{j,PC} = \lim_t [f_{j,PC}(t+1)]$, for $j = 1,\ldots,p$. The FPCI estimator is thus formed from the combined concepts of the PC estimator and the iterative generalized ridge estimator. It is interesting to note that as early as Marquardt (1970), the possibility is suggested of a combined estimator i.e. $\hat{\delta}_C = (Z_1'Z_1 + kI)^{-1}Z_1'Y$ where k is chosen to deflate the effects of the remaining near-zero eigenvalues left over in Z_1.

The second biased estimator due to Lee and Birch (1988) is based on the iterative ridge estimator concept. In this scheme the fraction in (5.2.15) becomes

$$f_{j,R}(t+1) = \frac{\lambda_j}{\lambda_j + s^2/[\hat{\delta}_R(t)'\hat{\delta}_R(t)/p]}, \quad t = 0,1,2,\ldots \tag{5.2.17}$$

where $\hat{\delta}_R(t)$ is the ridge estimate of δ at the t-th iteration with $\hat{\delta}_R(0) = \hat{\delta}$. Just as in the FPCI estimator, the authors replaced the OLSE by the PCE $(\hat{\delta}_R(0) = \hat{\delta}_{PC})$ and the s^2 by $s^2(t)$, and denoted the fractions by $f_{j,PCV}(t+1)$. The resulting estimator, denoted by $\hat{\delta}_{FPCV}$, is defined as:

$$\hat{\delta}_{FPCV} = F_{PCV}\hat{\delta} \tag{5.2.18}$$

where $F_{PCV} = \mathrm{Diag}(f^*_{1,PCV},\ldots,f^*_{p,PCV})$, and $f^*_{j,PCV} = \lim_t [f_{j,PCV}(t+1)]$, for $j = 1,\ldots,p$.

The authors have observed that a 1-step version of both (5.2.15) and (5.2.17) exhibited already improved estimation properties over other biased estimators in the data sets they studied. More comments on these estimators and their performances against other estimators will appear in Chapter 8 and Chapter 10.

5.3 Summary

In this chapter we introduced the shrinkage estimators as well as various fractional principal component regression estimators. The estimators were defined, and their properties given and discussed.

Chapter 6

LATENT ROOT REGRESSION ESTIMATION

Latent root regression (LRR)is similar to PC regression with the difference that it operates on the augmented matrix of explanatory variables and response in correlation form rather than the matrix of regressors alone. The method was originally proposed by Webster *et al.* (1974) and by Hawkins (1973).

6.1 The LRR estimator

Define the centered and standardized vector Y^* (nx1) as

$$Y^* = (Y - \bar{y}1)/s_y \qquad (6.1.1)$$

where $$\bar{y} = \sum_{i=1}^{n} Y_i/n \quad \text{and} \quad s_y^2 = \sum_{i=1}^{n} (Y_i - \bar{y})^2 \qquad (6.1.2)$$

(Instead of just saying Y is centered and standardized we introduce this notation as it is useful in developing the estimator).

Assume that X is centered and standardized and form the augmented matrix $[X\ Y^*]$, so that $[X\ Y^*]'[X\ Y^*]$ will be in correlation form. Then by using the SVD of an augmented matrix (1.3.8),

$$[X\ Y^*] = \tilde{U}\,\tilde{\Delta}\,\tilde{V}'$$

with

$$\tilde{U} = [\tilde{u}_1 \ldots \tilde{u}_{p+1}],\ \tilde{u}_i\text{:nx1}$$

$$\tilde{V} = [\tilde{v}_1 \ldots \tilde{v}_{p+1}],\ \tilde{v}_i\text{:(p+1)x1}$$

$$\tilde{\Delta} = \text{diag}[\omega_1, \omega_2, \ldots \omega_{p+1}]$$

$$\omega_1 \geq \omega_2 \geq \ldots\ldots \geq \omega_{p+1}$$

$$\tilde{U}'\tilde{U} = \tilde{V}'\tilde{V} = I_{p+1}$$

and the following notation defined in Chapter 1:

1. Let $\tilde{v}_{i,j}$ be the j-th component of the i-th right singular vector $\tilde{v}_i$.

2. $\tilde{v}_i^0$ is the p-dimensional vector containing the first p components of the i-th right singular vector $\tilde{v}_i$ of $[X\ Y^*]$ of dimension p+1, thus

$$\tilde{v}_i = [\tilde{v}_i^{0\prime}\ \tilde{v}_{i,p+1}]'.$$

Then by using the definition of an eigenvalue, one can write

$$[X\ Y^*]\omega_i{}^2 = [X\ Y^*][\tilde{v}_i^{0\prime}\ \tilde{v}_{i,p+1}]'.$$

If $\omega_i{}^2 \approx 0$ (indicating a collinearity) then

$$[X\ Y^*][\tilde{v}_i^{0\prime}\ \tilde{v}_{i,p+1}]' \approx 0 \quad (0\text{:nx1}) \text{ or}$$

$$X\tilde{v}_i^0 + Y^*\tilde{v}_{i,p+1} \approx 0 \tag{6.1.3}$$

If $\tilde{v}_{i,p+1}$ (the weight of Y^*) is not trivially close to zero, the response variable (Y^*) is involved in the collinearity. The key difference between detecting collinearities in X'X and those in the augmented correlation matrix is that Y can affect the collinearities in $[X\ Y^*]'[X\ Y^*]$, and any such collinearity has predictive value since (6.1.3) implies

$$\hat{Y}^* \approx -(\tilde{v}_{i,p+1})^{-1}X\tilde{v}_i^0 \tag{6.1.4}$$

$$\hat{Y} = \bar{y}1 - s_y(\tilde{v}_{i,p+1})^{-1}X\tilde{v}_i^0 \quad \text{(from (6.1.1))} \tag{6.1.5}$$

Definition 6.1: When $\omega_i \approx 0$ and $(\tilde{v}_{i,p+1}) \approx 0$ the multicollinearity is referred to as a non-predictive collinearity, i.e., a collinearity among the predictor variables that is of little value in predicting the response variable.

Gunst and Mason (1980) suggest values of $\omega_i \leq 0.1$ and $(\tilde{v}_{i,p+1}) \leq 0.1$ should be investigated for non-predictive collinearities)

Providing all $(\tilde{v}_{i,p+1}) \neq 0$, one can define p+1 prediction equations like (6.1.5) (one for each eigenvector). Let the i-th prediction equation be denoted by $\hat{\mathbf{Y}}_i$, thus, for i = 1,2,...,p+1

$$\hat{\mathbf{Y}}_i = \bar{y}\mathbf{1} - s_y(\tilde{v}_{i,p+1})^{-1}X\tilde{v}_i^0. \tag{6.1.6}$$

Linear combinations of the predictors in (6.1.6) will be used to obtain estimates of the parameters of the model. Consider the following arbitrary linear combination of the predictors:

$$\hat{\mathbf{Y}} = \sum_{i=1}^{p+1} a_i(\tilde{v}_{i,p+1})\hat{\mathbf{Y}}_i.$$

A value for a_i will be obtained in (6.1.15). Imposing the restriction

$$\sum_{i=1}^{p+1} a_i(\tilde{v}_{i,p+1}) = 1 \tag{6.1.7}$$

yields

$$\begin{aligned}\hat{\mathbf{Y}} &= \sum_{i=1}^{p+1} a_i(\tilde{v}_{i,p+1})\{\bar{y}\mathbf{1} - s_y(\tilde{v}_{i,p+1})^{-1}X\tilde{v}_i^0\} \\ &= \bar{y}\mathbf{1} - s_yX\sum_{i=1}^{p+1} a_i\tilde{v}_i^0 \end{aligned} \tag{6.1.8}$$

The residual sum of squares using this predictor is

$$(Y - \hat{Y})'(Y - \hat{Y}) = (Y-\bar{y}\mathbf{1} + s_y X \sum_{i=1}^{p+1} a_i\tilde{v}_i^{0})'(Y-\bar{y}\mathbf{1} + s_y X \sum_{i=1}^{p+1} a_i\tilde{v}_i^{0})$$

$$= s_y^2[(Y^* + X \sum_{i=1}^{p+1} a_i\tilde{v}_i^{0})'(Y^* + X \sum_{i=1}^{p+1} a_i\tilde{v}_i^{0})$$

$$= s_y^2\{[X\ Y^*][\sum_{i=1}^{p+1} a_i\tilde{v}_i^{0\prime}\ 1]'\}'\{[X\ Y^*][\sum_{i=1}^{p+1} a_i\tilde{v}_i^{0\prime}\ 1]'\}$$

and $[\sum_{i=1}^{p+1} a_i\tilde{v}_i^{0\prime}\ 1]' = [\sum_{i=1}^{p+1} a_i\tilde{v}_i^{0\prime}\ \sum_{i=1}^{p+1} a_i(\tilde{v}_{i,p+1})]'$ (from (6.1.7))

$$= [\sum_{i=1}^{p+1} a_i[\tilde{v}_i^{0\prime}\ \tilde{v}_{i,p+1}]]'$$

$$= [\sum_{i=1}^{p+1} a_i\tilde{v}_i{}']' \qquad \text{(from 1.3.10)}$$

$$= [a'\tilde{V}']'$$

$$= \tilde{V}a \tag{6.1.9}$$

where $a' = [a_1,\ldots,a_{p+1}]$. Thus the RSS is

$$(Y - \hat{Y})'(Y - \hat{Y}) = s_y^2\{[X\ Y^*]\tilde{V}a\}'[X\ Y^*]\tilde{V}a$$

$$= s_y^2 a'\tilde{V}'[X\ Y^*]'[X\ Y^*]\tilde{V}a$$

$$= s_y^2 a'\tilde{\Lambda}^2 a \quad \text{(from (1.3.8))} \tag{6.1.10}$$

$$= s_y^2 \sum_{i=1}^{p+1} a_i^2\omega_i^2 \tag{6.1.11}$$

To find the value of a for which this residual sum of squares will be a minimum, one has to minimize f(a) where

$$f(a) = s_y^2 \sum_{i=1}^{p+1} a_i^2\omega_i^2 - 2u_0(\sum_{i=1}^{p+1} a_i(\tilde{v}_{i,p+1}) - 1)$$

and $2u_0$ is a Lagrangian multiplier. Now, for $j = 1,\ldots,p+1$

$$\frac{df(a)}{da_j} = 2s_y^2 a_j \omega_i^2 - 2u_0(\tilde{v}_{j,p+1}) = 0 \tag{6.1.12}$$

so
$$a_j = u_0(\tilde{v}_{j,p+1})/(s_y^2\omega_j^2) \tag{6.1.13}$$

and from the restriction in (6.1.7)

$$\sum_{j=1}^{p+1} a_j(\tilde{v}_{j,p+1}) = u_0 \sum_{j=1}^{p+1} (\tilde{v}_{j,p+1})^2/(s_y^2\omega_j^2)$$

$$u_0 = 1/\{\sum_{j=1}^{p+1} (\tilde{v}_{j,p+1})^2/(s_y^2\omega_j^2)\} \tag{6.1.14}$$

Substituting this value in (6.1.13) for $j = 1,\ldots,p+1$ gives

$$\begin{aligned} a_j &= [1/\{\sum_{j=1}^{p+1} (\tilde{v}_{j,p+1})^2/(s_y^2\omega_j^2)\}](\tilde{v}_{j,p+1})/(s_y^2\omega_j^2) \\ &= (\tilde{v}_{j,p+1})\omega_j^{-2}\{\sum_{j=1}^{p+1} (\tilde{v}_{j,p+1})^2/\omega_j^2\}^{-1} \end{aligned} \tag{6.1.15}$$

Then if we put this value in (6.1.8):

$$\begin{aligned} \hat{Y} &= \bar{y}1 - s_y X \sum_{i=1}^{p+1} (\tilde{v}_{i,p+1})\omega_i^{-2}\{\sum_{j=1}^{p+1} (\tilde{v}_{j,p+1})^2/\omega_j^2\}^{-1}\tilde{v}_i^0 \\ &= \bar{y}1 + X \sum_{i=1}^{p+1} f_i\tilde{v}_i^0 \end{aligned} \tag{6.1.16}$$

where

$$f_i = -s_y(\tilde{v}_{i,p+1})\omega_i^{-2}\{\sum_{j=1}^{p+1} (\tilde{v}_{j,p+1})^2/\omega_j^2\}^{-1} \tag{6.1.17}$$

Suppose that there are s non-predictive near-singularities, i.e.

$$\omega_{p+1-s} = \omega_{p+1-s-1} = \ldots = \omega_{p+1-1} = \omega_{p+1} \approx 0$$

and
$$\tilde{v}_{p+1-s,p+1} = \ldots = \tilde{v}_{p+1-1,p+1} = \tilde{v}_{p+1,p+1} \approx 0.$$

Then $$\hat{Y} = \bar{y}1 + X \sum_{i=1}^{p+1-s} f_i \tilde{v}_i^0 \qquad (6.1.18)$$

Then the latent root estimator ($\hat{\beta}_{LR}$ of β) is

$$\hat{\beta}_{LR} = \sum_{i=1}^{p+1-s} f_i \tilde{v}_i^0 \qquad (6.1.19)$$

and f_i is as defined in (6.1.17) for $i = 1,2,\ldots,p+1-s$

$$f_i = -s_y(\tilde{v}_{i,p+1})\omega_i^{-2}\{\sum_{j=1}^{p+1} (\tilde{v}_{j,p+1})^2/\omega_j^2\}^{-1}$$

Some authors (i.e. Gunst and Mason (1980) define $\hat{\beta}_{LR} = \sum_{i=1}^{p+1} f_i \tilde{v}_i^0$ where $f_i = 0$ if $\omega_i \approx 0$ and $\tilde{v}_{p+1,p+1} \approx 0$; otherwise f_i will have the value defined in (6.1.17).

If $\tilde{V} = [\tilde{v}_1 \ldots \tilde{v}_{p+1}]$, $\tilde{v}_i:(p+1)x1$ is partitioned as $\tilde{V} = [\tilde{V}_1 \ \tilde{V}_2]$ where $\tilde{V}_1 = [\tilde{v}_1 \ldots \tilde{v}_{p+1-s}]$, $\tilde{V}_2 = [\tilde{v}_{p+2-s} \ldots \tilde{v}_{p+1}]$, then by defining the vector $f = [f_A' \ f_B']'$, and $f_A = [f_1, f_2, \ldots f_{p+1-s}]'$, $f_B = [f_{p+2-s} \ldots f_{p+1}]'$, LRR can be viewed as the use of a restricted least squares estimator where the restriction is $\tilde{V}_2 f_B = 0$

When $s = 0$ we have $\hat{\beta}_{LR} = \hat{\beta}$, since both minimize SSE without imposing any restriction on the estimators and the RSS will be given by (6.1.11) and (6.1.15) as

$$\begin{aligned} RSS &= s_y^2 \sum_{j=1}^{p+1} a_j^2 \omega_j^2 \\ &= s_y^2 \sum_{j=1}^{p+1} (\tilde{v}_{j,p+1})^2 \omega_j^{-4} \{\sum_{j=1}^{p+1} (\tilde{v}_{j,p+1})^2/\omega_j^2\}^{-2} \omega_j^2 \\ &= s_y^2 \{\sum_{j=1}^{p+1} (\tilde{v}_{j,p+1})^2/\omega_j^2\}^{-1} \\ &= u_0 \end{aligned} \qquad (6.1.20)$$

When s ≠ 0 then

$$\text{RSS} = s_y^2 \sum_{j=1}^{p+1-s} a_j^2 \omega_j^2$$

$$= s_y^2 \{ \sum_{j=1}^{p+1-s} (\tilde{v}_{j,p+1})^2/\omega_j^2 \}^{-1} \qquad (6.1.21)$$

Comment:

1. $\sum_{j=1}^{p+1-s} (\tilde{v}_{j,p+1})^2/\omega_j^2 \leq \sum_{j=1}^{p+1} (\tilde{v}_{j,p+1})^2/\omega_j^2$ so that RSS in (6.1.21) will be larger than RSS in (6.1.20) (the full model).

2. If one is removing non-predictive near-singularities (6.1.21) will be approximately equivalent to (6.1.20) as the s non-predictive collinearities that have been removed contribute very little to $\sum_{j=1}^{p+1} (\tilde{v}_{j,p+1})^2/\omega_j^2$. Thus the estimated $\hat{\sigma}^2$ using (6.1.20) and (6.1.21) will be reasonably close. When the number of non-predictive collinearities (s) increases, $\hat{\sigma}^2$ for the full model will be larger then that of (6.1.21) as the degrees of freedom for the latter will be n-p-1+s (pooling the error associated with the s non-predictive collinearities with the error sums of squares of the restricted model (6.1.21)).

6.2 Properties and Problems

1. Expectation

$$E(\hat{\beta}_{LR}) = E[\sum_{i=1}^{p+1} f_i \tilde{v}_i^0 - \sum_{i=t}^{p+1} f_i \tilde{v}_i^0] \quad \text{where } t = (p+1-s)+1$$

$$= \beta - E[\sum_{i=t}^{p+1} f_i \tilde{v}_i^0] \qquad (6.2.1)$$

2. The LRRE has no known MSE function and so can not be directly compared with other estimators. Gunst *et al.* (1976) report that in many cases the LRRE behaves similarly to the PCE so that

$$\mathrm{TMSE}(\hat{\beta}_{LR}) \approx \mathrm{TMSE}(\hat{\beta}_{PC}) \tag{6.2.2}$$

Thus LRRE yields a biased estimator, but one which apparently greatly reduces the TMSE of the estimator. The theoretical conditions necessary for (6.2.2) to hold are unknown at the present time.

3. No criterion of the smallness of the eigenvalues and last element of the corresponding eigenvector that admits deletion, is as yet adequately defined. The decision will depend on the number of independent variables, the sample size (n) and the degree of ill-conditioning. Webster *et al.* (1974) suggested in their example (n = 12 and six regressors) that suitably small would mean $\omega_i^2 \leq 0.05$ and $\tilde{v}_{i,p+1} \leq 0.10$.

4. Though Webster *et al.* (1974) suggested that accuracy of computer algorithms for determining eigenvectors and eigenvalues under various degrees of ill-conditioning should be investigated, this may not be a relevant problem today.

Note that LRRE is in some sense simply a computational device like TLS with a justification but no complete statistical development.

6.3. Elimination of independent variables

Webster *et al.* (1974) as well as Gunst *et al.* (1976) discussed and compared OLS and LRRE on their performances in eliminating independent variables *via* the stepwise technique. Both authors preferred the backward elimination procedure (see Chapter 1) and Webster *et al.* (1974) proposed an F-statistic for LRR similar to the OLS F-statistic. They also suggested that perhaps several regressor variables could be deleted at the first stage of LRR.

Thus in a backward elimination procedure, they applied the following variable selection rule:

OLS: Determine the minimum F-value for each of the p variables in the model (1.1); if the minimum $F < 1$, delete that particular regressor from the predictor.

LRR: Determine each of the F-values for the p variables in the model (6.1.18); for each $F < 1$, delete the corresponding regressor from the predictor.

The advantage of LRR over OLS is that by using LRR one firstly eliminates the non-predictive collinearities (unmasks them) while in OLS they remain masked. The true influences of the independent variables on the dependent variable are more clearly represented. In many cases several independent variables may be eliminated at the first stage when the computations are the easiest.

6.4 Summary

In this chapter we introduced latent root regression estimation. We discussed the estimator, its properties, related problems and methods of eliminating independent variables.

Chapter 7

TOTAL LEAST SQUARES

Total least squares (TLS) is a method of fitting that is appropriate when there are errors in both the observation vector Y (nx1) and in the independent variable matrix X (nxp). Errors in X-variables models contrast with the classical regression model where the data matrix X is assumed to be error free.

7.1 Introduction

The problem of parameter estimation when all the variables involve error, has a long history. For an overview of the historical development see Van Huffel and Vandewalle (1985). Recently the linear errors-in-variables (all observations are coming from some unknown true values plus measurement errors) has been treated by the method of total least squares (TLS) based on the singular value decomposition (SVD) as proposed by Golub and Van Loan (1980) and further developed by Van Huffel, Vandewalle and Staar (1984).

In §7.2 the TLS technique is discussed, the use of TLS in the errors-in-variables model in §7.3 and subset selection based on TLS for prediction, in §7.4. The comparison of TLS estimator with the estimators discussed in Chapters 3 to 6 as well as the OLSE will be discussed in Chapter 8.

7.2 TLS Technique

7.2.1 Assumptions and notation

1. The Total Least Square Estimate (TLSE) of β will be denoted by $\hat{\beta}_{TLS}$.

2. The SVD of the augmented matrix [X Y] is given in (1.3.8):

$$[X\ Y] = \tilde{U}\tilde{\Delta}\tilde{V}'$$

with
$$\tilde{U} = [\tilde{u}_1 \ldots \tilde{u}_{p+1}],\ \tilde{u}_i : nx1$$
$$\tilde{V} = [\tilde{v}_1 \ldots \tilde{v}_{p+1}],\ \tilde{v}_i : (p+1)x1$$
$$\tilde{U}'\tilde{U} = \tilde{V}'\tilde{V} = I_{p+1}$$
$$\tilde{\Delta} = \mathrm{diag}[\omega_1, \omega_2, \ldots, \omega_{p+1}]$$
$$\omega_1 \geq \omega_2 \geq \ldots\ldots \geq \omega_{p+1} \qquad (7.2.1)$$

For simplicity, assume $\omega_1 > \omega_2 > \ldots\ldots > \omega_{p+1}$.

3. $\tilde{v}_{i,j}$ is the j th component of the i th right singular vector $\tilde{v}_i$ the i-th column of $\tilde{V}$

4. $\overset{0}{\tilde{v}_i}$ is the p-dimensional vector containing the first p components of the i-th right singular vector $\tilde{v}_i$ of [X Y], so we have
$$\tilde{v}_i = [\overset{0}{\tilde{v}_i}\ \tilde{v}_{i,p+1}]' \qquad (7.2.2)$$

5. $[X\ \hat{Y}]$ is the LS approximation of [X Y] with $\hat{Y}$ the orthogonal projection of Y onto the column space R(X) of X. $[\tilde{X}\ \tilde{Y}]$ is the TLS approximation of [X Y].

6. The following theorem is due to Van Huffel and Vandewalle (1987) and is useful in the generalization of TLS:

If $1 \leq j \leq p+1$ then

(a) $\tilde{v}_{j,p+1} = 0 \Rightarrow Y \perp \tilde{u}_j$ and $\tilde{Y} \perp \tilde{u}_j$ (7.2.3)

(when $\tilde{v}_{j,p+1} = 0$, then $\tilde{u}_j \omega_j \tilde{v}_j{}' = \omega_j [\tilde{u}_j \overset{0}{\tilde{v}_j}\ 0]$)

(b) The eigenvalues of X′X interlace with those of [X Y]′[X Y] as follows:

$$\omega_1^2 \geq \lambda_1 \geq \omega_2^2 \geq \lambda_2 \geq \cdots \geq \omega_p^2 \geq \lambda_p \geq \omega_{p+1}^2 \qquad \text{(from A.3)}$$

and if we assume that the singular values of X are disjoint. Then

$$\tilde{v}_{j,p+1} = 0 \Rightarrow \omega_j^2 = \lambda_i,\ \tilde{u}_j = \pm u_i,\ \text{and } \tilde{v}_j = [\pm v_i{}'0]' \quad \text{with } i \in \{j-1,j\},\ 1 \leq i \leq p, \qquad (7.2.4)$$

$$\tilde{v}_{j,p+1} = 0 \Rightarrow \hat{Y} \perp \tilde{u}_j \qquad (7.2.5)$$

7. The correction vector $\epsilon = Y - \hat{Y}$ is the OLS approximation error and the correction matrix $[\Delta\tilde{X}\ \Delta\tilde{Y}] = [X - \tilde{X},\ Y - \tilde{Y}]$ the TLS approximation error.

8. By manipulation the pseudo-equation $Y = X\beta$ can be written as (Golub and Van Loan (1980)):

$$X\beta - Y = 0$$

$$[X\ Y]\begin{bmatrix}\beta \\ -1\end{bmatrix} = 0$$

$$[X\ Y][\beta'\ -1]' = 0 \qquad (7.2.6)$$

9. The following result of Eckart and Young (1936) is used in the development of the TLS technique:

Let $r(X) = k \leq p \leq n$ and let $\hat{X}$ be an approximation to X that satisfies

$$r(\hat{X}) \leq k \qquad (7.2.7)$$

$$\|\hat{X} - X\| = \inf_{r(\bar{X}) \leq k} \|(\bar{X} - X)\| \qquad (7.2.8)$$

where $\|\cdot\|$ is a unitarily invariant matrix norm as defined in (1.9.10).

Set

$$\sqrt{\lambda}_{k+1} = \sqrt{\lambda}_{k+2} = \ldots = \sqrt{\lambda}_p = 0$$

then by using the SVD (1.3.1), $\hat{X}$ can be written as

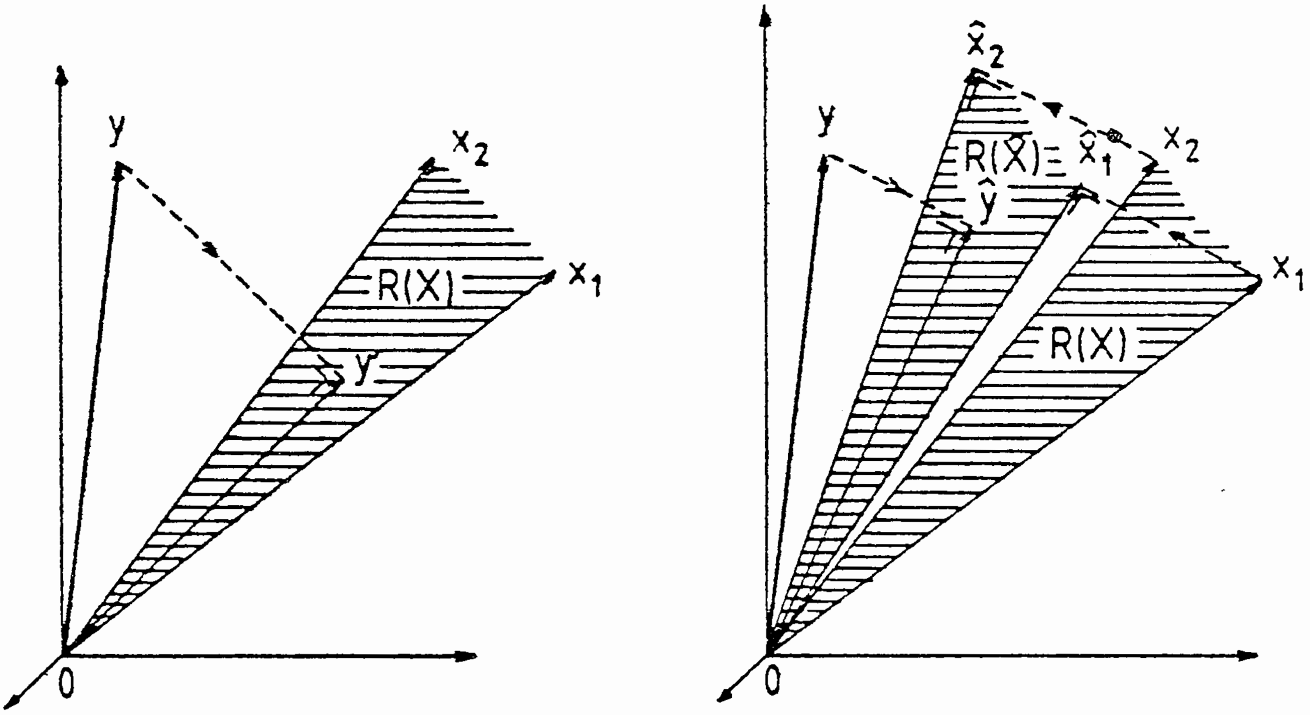
$$\hat{X} = U\hat{D}_a V' \tag{7.2.9}$$

where $$\hat{D}_a = \mathrm{Diag}[\sqrt{\lambda_1},\sqrt{\lambda_2},\ldots\sqrt{\lambda_k},0,\ldots 0]$$

In the case of the Frobenius norm, Eckart and Young (1936) showed that the $\hat{X}$ of (7.2.9) satisfies (7.2.7) and (7.2.8). For a generalization of the Eckart and Young approximation see Golub, Hoffman and Stewart (1987).

7.2.2 Geometric view of TLS

The figures below are from Van Huffel and Vandewalle (1985):



(a) The LS solution is obtained by projecting Y orthogonally onto R(X) and solving $X\hat{\beta} = Y$.

(b) The TLS solution is obtained by approximating the columns X_i of X and Y by X_i and $\hat{Y}$ until $\hat{Y}$ is in the space $R(\hat{X})$ generated by the columns $\hat{X}_i$ and solving $\hat{X}\beta = \hat{Y}$.

7.2.3 Definition

Given an over-determined set of n linear equations $X\beta = Y$ in p unknowns, the total least-squares (TLS) solution is the minimum-norm solution $\hat{\beta}_{TLS}$ of the set of n linear equations

$$\tilde{X}\hat{\beta}_{TLS} = \tilde{Y} \qquad (7.2.10)$$

where $\tilde{X}$ and $\tilde{Y}$ are determined such that

$$\tilde{Y} \epsilon R(\tilde{X}) \qquad (7.2.11)$$

$$\|\Delta X \ \Delta Y\|_F = \|[X \ Y] - [\tilde{X} \ \tilde{Y}]\|_F \text{ is minimal} \qquad (7.2.12)$$

Using the Eckart-Young principle, equations (7.2.11) and (7.2.12) will be satisfied by making ω_{p+1} in (7.2.1) zero.

The TLS approximation of $\tilde{\Delta}$ defined in (7.2.1) can then itself be approximated by

$$\hat{\tilde{\Delta}} = \text{diag}[\omega_1, \omega_2, \ldots, \omega_p, 0]$$

and $$[\tilde{X} \ \tilde{Y}] = \tilde{U}\hat{\tilde{\Delta}}\tilde{V}'$$

The lower rank approximation for (7.2.6) is then

$$[\tilde{X} \ \tilde{Y}][\hat{\beta}'_{TLS} \ -1]' = 0 \quad (7.2.13)$$

The TLS solution is obtained by scaling the last column $\tilde{v}_{p+1}$ of $\tilde{V}$ so that its last component $\tilde{v}_{p+1,p+1} = -1$. It is easy to verify that

$$[\hat{\beta}'_{TLS} \ -1]' = -\tilde{v}_{p+1}/\tilde{v}_{p+1,p+1} \qquad (7.2.14)$$

by showing

$$\begin{aligned}
[\tilde{X}\ \tilde{Y}][-\tilde{v}_{p+1}/\tilde{v}_{p+1,p+1}] &= \tilde{U}\hat{\tilde{\Delta}}\tilde{V}'[-\ \tilde{v}_{p+1}/\tilde{v}_{p+1,p+1}] \\
&= \tilde{U}\hat{\tilde{\Delta}}[\tilde{v}_1\ldots\tilde{v}_{p+1}]'[-\ \tilde{v}_{p+1}/\tilde{v}_{p+1,p+1}] \\
&= \tilde{U}\hat{\tilde{\Delta}}[0,0,\ldots,0,-1/\tilde{v}_{p+1,p+1}]' \\
&\qquad\qquad \text{(the columns of } \tilde{V} \text{ are orthogonal)} \\
&= 0{:}\text{nx}1 \qquad (\hat{\tilde{\Delta}}[0,0,\ldots,0,-1/\tilde{v}_{p+1,p+1}]'=0))
\end{aligned}$$

If we transpose (7.2.14) and multiply each side by $[X\ Y]'[X\ Y]$, we obtain

$$\begin{aligned}
\begin{bmatrix} X'X & X'Y \\ Y'X & Y'Y \end{bmatrix}\begin{bmatrix} \hat{\beta}_{TLS} \\ -1 \end{bmatrix} &= \tilde{V}\tilde{\Delta}^2\tilde{V}'[-\tilde{v}_{p+1}/\tilde{v}_{p+1,p+1}] \quad \text{(by using (7.2.1))} \\
&= \tilde{V}\tilde{\Delta}^2[0,0,\ldots,0,-1/\tilde{v}_{p+1,p+1}]' \\
&= \tilde{V}[0,0,\ldots,0,-\sigma^2_{p+1}/\tilde{v}_{p+1,p+1}]' \\
&= -\omega^2_{p+1}\tilde{v}_{p+1}/\tilde{v}_{p+1,p+1} \\
X'X\hat{\beta}_{TLS} - X'Y &= \omega^2_{p+1}\ \hat{\beta}'_{TLS} \\
X'X\hat{\beta}_{TLS} - \omega^2_{p+1}\hat{\beta}_{TLS} &= X'Y \\
[X'X-\omega^2_{p+1}I]\hat{\beta}_{TLS} &= X'Y \\
\hat{\beta}_{TLS} &= [X'X-\omega^2_{p+1}I]^{-1}X'Y \qquad (7.2.15)
\end{aligned}$$

Using the SVD (1.3.1) of X this estimator can be written as

$$\begin{aligned}
\hat{\beta}_{TLS} &= [V\Delta^2V' - \omega^2_{p+1}I]^{-1}V\Delta U'Y \\
&= V[\Delta^2 - \omega^2_{p+1}I]^{-1}V'V\Delta U'Y \\
&= \sum_{j=1}^{p} v_j(\lambda_j - \omega^2_{p+1})^{-1}\sqrt{\lambda_j}u_j'Y \\
&= \sum_{j=1}^{p}(\lambda_j - \omega^2_{p+1})^{-1}\sqrt{\lambda_j}(u_j'Y)v_j \qquad (7.2.16)
\end{aligned}$$

where $u_j'Y$ is a scalar. Equation (7.2.16) is applicable when $\omega_p > \omega_{p+1}$ and $\tilde{v}_{p+1,p+1} \neq 0$. If $\tilde{v}_{p+1,p+1} = 0$, (7.2.14) become infinite and Golub and Van Loan's (1980) algorithm for TLS fails to compute a finite TLS solution. Consequently such problems can not be solved with TLS. However Van Huffel and Vandewalle (1987) claimed that the TLS approximation [X Y] can still be determined by making the next smallest singular value in (7.2.1) zero.

Note that the authors definition of non-predictive collinearities is: $\tilde{v}_{j,p+1}$ $(j = k,\dots,p+1)$ are exactly zero and ω_j is suitably small; in contrast to the definition 6.1 (below (6.1.5) of the Chapter 6) where $\tilde{v}_{j,p+1}$ $(j = k,\dots,p+1) \approx 0$.

7.2.4 Generalized definition

In generalizing of TLS to cases where ω_{p+1} coincides with other singular values $\omega_{k-1} > \omega_k = \dots \omega_p = \omega_{p+1}$, and some or all of the $\tilde{v}_{j,p+1}$ $(j = k,\dots,p+1)$ are zero, Van Huffel and Vandewalle (1987) give two definitions (their equations 2.8 and 2.9):

Definition one: If ω_{p+1} coincides with other singular values $\omega_{k-1} > \omega_k = \dots \omega_p = \omega_{p+1}$, then $[\hat{\beta}_{TLS}' \ -1]'$ is a linear combination of the corresponding right singular vectors $\tilde{v}_k \dots \tilde{v}_{p+1}$ such that the solution $\hat{\beta}_{TLS}$ has minimum norm.

If not all components $\tilde{v}_{j,p+1}$ $(j = k,\dots,p+1)$ are zero the solution is

$$\hat{\beta}_{TLS} = \sum_{j=k}^{p+1} \varrho_j \tilde{v}_j^0 \qquad \text{with } \varrho_j = -\frac{\tilde{v}_{j,p+1}}{\sum_{j=k}^{p+1} \tilde{v}_{j,p+1}^2} \qquad (7.2.17)$$

When k = p+1, (7.2.17) reduces to

$$\hat{\beta}_{TLS} = \varrho_{p+1}\tilde{v}^{0}_{p+1} \quad \text{with } \varrho_{p+1} = -\frac{\tilde{v}_{p+1,p+1}}{\tilde{v}^{2}_{p+1,p+1}}$$

$$= -\ \tilde{v}^{0}_{p+1}/\tilde{v}_{p+1,p+1}$$

When r([X Y]) = p+1-r (r singular values are near zero) we will assume $\omega_{p+1-r} > \omega_{p+1-r+1} = \dots \omega_{p+1-1} = \omega_{p+1}$, and k = (p+1) - (r-1) = p+2-r.

If, however, all r collinearities which appear in [X Y] are nonpredictive, then TLS approximates [X Y] with $[\tilde{X}\ \tilde{Y}]$ by making the next larger singular value ω_{p+1-r} zero. Then by using equation (7.2.3) to (7.2 5) the TLS estimator is given by

Definition two: If r > 0, $\tilde{v}_{j,p+1}$ (j = p+2-r,...,p+1) are exactly zero (hence by using (7.2.3) to (7.2.5) we know that Y is orthogonal to $\tilde{u}_j$) and with $\omega_{p-r} > \omega_{p+1-r}$, we have:

$$\hat{\beta}_{TLS} = \sum_{j=1}^{p-r} (\lambda_j - \omega^2_{p+1-r})^{-1} \sqrt{\lambda_j}(u_j'Y)v_j \tag{7.2.18}$$

with r the number of r non-predictive collinearities.

From a computational point of view definition 6.1 would be more practical, and in applying (7.2.18) the $\tilde{v}_{j,p+1}$ (j = p+2-r,...,p+1) are exactly zero only when known *a priori* or assumed and set to 0.

7.2.5 Properties of TLS estimator

For simplicity we work with (7.2.18):

$$\hat{\beta}_{TLS} = \sum_{j=1}^{p-r} (\lambda_j - \omega^2_{p+1-r})^{-1} \sqrt{\lambda_j}(u_j'Y)v_j$$

$$= \sum_{j=1}^{p-r} (\lambda_j - \omega^2_{p+1-r})^{-1} c_j v_j, \text{ where } c_j = \sqrt{\lambda_j}(u_j'Y)$$

The following properties are reported by Van Huffel and Vandewalle (1985, p15):

$$E(\hat{\beta}_{TLS}) = \beta + \omega^2_{p+1-r} \sum_{j=1}^{p-r} (\lambda_j - \omega^2_{p+1-r})^{-1} (v_j'\beta) v_j - \sum_{j=t}^{p} (v_j'\beta) v_j \tag{7.2.19}$$

where t = p-r+1

$$\text{Var}(\hat{\beta}_{TLS}) = \sigma^2 \sum_{j=1}^{p-r} \lambda_j (\lambda_j - \omega^2_{p+1-r})^{-2} v_j v_j' \tag{7.2.20}$$

$$\text{MSE}(\hat{\beta}_{TLS}) = \sigma^2 \sum_{j=1}^{p-r} \lambda_j (\lambda_j - \omega^2_{p+1-r})^{-2} v_j v_j' + \omega^4_{p+1-r} \sum_{j=1}^{p-r} (\lambda_j - \omega^2_{p+1-r})^{-2} (v_j'\beta)^2 - \sum_{j=t}^{p} (v_j'\beta)^2 \tag{7.2.21}$$

The above properties ((7.2.19), (7.2.20) and (7.2.21)) were given without a proof. Van Huffel (pers. comm, 1990) confirms that her proofs assume the following:

(i) ω^2_{p+1-r} is fixed and thus the expectation of the TLSE is the same as that of the RRE with $k = -\omega^2_{p+1-r}$.

(ii) Furthermore $\sum_{j=t}^{p} (\lambda_j - \omega^2_{p+1-r})^{-1} \sqrt{\lambda_j} (u_j'Y) v_j = 0$, where t = p+1-r, on the grounds that TLS and PC estimation delete the same collinearities.

Statisticians may reject the expectation of the TLSE (7.2.19) and what follows firstly as ω^2_{p+1-r} is not a constant but a function of Y, and therefore random. Furthermore the quantity $\sum_{j=t}^{p} (\lambda_j - \omega^2_{p+1-r})^{-1} \sqrt{\lambda_j} (u_j'Y) v_j$, is actually infinite (division by 0).

Van Huffel (pers. comm, 1990) comments 'my main motivation for the expressions ((7.2.19), (7.2.20) and (7.2.21)) is to give the reader the feeling that TLS is not appropriate for estimation of parameters in models with error-free predictor variables. TLS is not devised as a biased regression estimator like RR (although close similarities between the expressions of both estimators exist) but is only appropriate for parameter estimation in models with errors in all variables.'

We may further observe that from a computational point of view if $X'X$ is ill-conditioned $[X'X - \omega_{p+1}^2]^{-1}$ will be worse-conditioned. TLS is therefore an alternative estimation procedure to handle the errors-in-variables phenomenon rather than collinearity. As a device it approximates X with $\tilde{X}$ generally of lower rank and imposes restrictions (7.2.10) in place of the normal equations. If there are in fact no errors in the variables then (7.2.15) implies that $\hat{\beta}_{TLS}$ would be biased and have a larger variance than the OLSE $\hat{\beta}$.

7.3 Use of TLS in the errors-in-variables model

7.3.1 The Model

The general errors-in-variables model is (Van Huffel and Vandewalle (1985)):

$$\begin{aligned} Y_0 &= X_0\beta_0 + \epsilon \\ X &= X_0 + X_e \\ Y &= Y_0 + Y_e \end{aligned} \qquad (7.3.1)$$

where

β_0:px1 is the unobservable vector of parameters to be estimated
X: nxp is the observed matrix on the p independent variables
Y: nx1 is the observed response vector
Y_0, X_0 contain the true but unobservable variables, and
X_e and Y_e are their measurement or observation errors (unobserved).

In addition to the errors in the variables (X_e, Y_e) there is an error vector ϵ in the equation. In the classical or pure error-in-variables model $\epsilon = 0$.

It is assumed that the error variances of $[X_e]_{ij}$ and $[Y_e]_i$ and their covariances $E([X_e]_{ij}, [Y_e]_i)$ are known while β_0 and the variance σ_ϵ^2, of the equation, are unknown.

Define the nx(p+1) error matrix as

$$\Delta_e = [X_e \quad Y_e + \epsilon] \tag{7.3.2}$$

Denote the i-th row of Δ_e by $[\Delta_e]_i'$ and partition X_e by its rows and denote each row by a subscript. Thus $X_e = [(X_e)_1, (X_e)_2, .., (X_e)_n]'$ and $(X_e)_i'$ is a 1xp row vector. This implies that $[\Delta_e]_i' = [(X_e)_i' \quad [(Y_e)_i + \epsilon_i]]$, where $(Y_e)_i + \epsilon_i$ is the i-th element of the sum of Y_e and ϵ. Assume the following

(i) Δ_e and X are stochastically independent random matrices

(ii) row vectors of Δ_e are stochastically independent and identically distributed

(iii) $E(\Delta_e) = 0$

Define the covariance matrix of the error variables as

$$\begin{aligned} \Sigma_{\Delta_e} &= E[\ [\Delta_e]_i [\Delta_e]_i'\ ] \quad (:(p+1)x(p+1) \text{ matrix}) \qquad (7.3.3) \\ &= E\begin{bmatrix} (X_e)_i (X_e)_i' & (X_e)_i [(Y_e)_i + \epsilon_i] \\ [(Y_e)_i + \epsilon_i](X_e)_i' & [(Y_e)_i + \epsilon_i][(Y_e)_i + \epsilon_i] \end{bmatrix} \end{aligned}$$

Denote

$E[(X_e)_i (X_e)_i'] = \Sigma_{X_e}(pxp)$, $E[(X_e)_i [(Y_e)_i + \epsilon_i]] = \sigma_{Y_e X_e}(px1)$, $E[\epsilon_i^2] = \sigma_\epsilon^2(1x1)$

and $E[(Y_e)_i^2] = \sigma^2_{Y_e}$ (1x1) and from the assumptions:

$$\Sigma_{\Delta_e} = \begin{bmatrix} \Sigma_{X_e} & \sigma_{Y_eX_e} \\ \sigma'_{Y_eX_e} & \sigma^2_{Y_e} + \sigma^2_\epsilon \end{bmatrix} \qquad (7.3.4)$$

The term $\sigma_{Y_eX_e}$ depends on the error variables only and is assumed to be known, and $[\Sigma_{\Delta_e}]_{n+1,n+1}$ depends on σ^2_ϵ (unknown).

The model (7.3.1) can be manipulated to give

$$\begin{aligned} Y_0 &= X_0\beta_0 + \epsilon \\ Y - Y_e &= (X - X_e)\beta_0 + \epsilon \\ Y &= X\beta_0 - X_e\beta_0 + \epsilon + Y_e \\ &= X\beta_0 + \theta \end{aligned} \qquad (7.3.5)$$

where

$$\begin{aligned} \theta &= (Y_e + \epsilon) - X_e\beta_0 \\ &= -\Delta_e[\beta_0' \ -1]' \quad \text{(from 7.3.2)} \end{aligned} \qquad (7.3.6)$$

This is not a simple regression situation: X_e is a random matrix and it is correlated with the error term θ in contrast to the regression situation defined in (1.1).

7.3.2 Estimation by Least Squares

The OLS estimation of β_0 (denoted by $\hat{\beta}_0$) in (7.3.5) will yield an inconsistent estimator for β_0:

$$\begin{aligned} \hat{\beta}_0 &= (X'X)^{-1}X'Y \\ &= \beta_0 + (X'X)^{-1}X'\theta \\ &= \beta_0 + ((X'X)/n)^{-1}(X'\theta/n) \end{aligned} \qquad (7.3.7)$$

Schneeweiss (1976) shows that the following probability limits (plim) (defined in §1.11) exist and we will denote them by:

$$
\begin{aligned}
\text{plim}(n^{-1}\ X_0{}'X_0) &= \mathbf{M}_{X_0} \\
\text{plim}(n^{-1}\ \Delta_e{}'\Delta_e) &= \mathbf{M}_{\Delta_e} \\
\text{plim}(n^{-1}\ X_0{}'\Delta_e) &= 0 \\
\text{plim}(n^{-1}\ X'X) &= \mathbf{M}_{X_0} + \Sigma_{X_e} = \mathbf{M}_X \qquad (7.3.8) \\
\text{plim}(n^{-1}\ X'\Delta_e) &= [\Sigma_{X_e}\ \ \sigma_{Y_eX_e}]\ :\ \text{px}(\text{p}+1) \\
\text{plim}(n^{-1}\ X'\theta) &= [\Sigma_{X_e}\ \ \sigma_{Y_eX_e}][\beta_0'\ -1]' \\
\text{plim}(n^{-1}\theta'\theta) &= \beta_0'\Sigma_{X_e}\beta_0 - 2\sigma'_{Y_eX_e}\beta_0 + (\sigma_\epsilon^2 + \sigma_{Y_e}^2) = \sigma_\theta^2
\end{aligned}
$$

Then the plim of $\hat{\beta}_0$ is:

$$
\begin{aligned}
\text{plim}(\hat{\beta}_0) &= \beta_0 + \text{plim}\{((X'X)/n)^{-1}(X'\theta/n)\} \\
&= \beta_0 + \mathbf{M}_X^{-1}[\Sigma_{X_e}\beta_0 - \sigma_{Y_eX_e}] \quad (\text{from } (7.3.8)) \qquad (7.3.9)
\end{aligned}
$$

Also the OLS estimator of σ_θ^2 is inconsistent:

$$
\begin{aligned}
(n-p)\hat{\sigma}_\theta^2 &= Y'(I - X(X'X)^{-1}X')Y \\
&= \theta'(I - X(X'X)^{-1}X')\theta
\end{aligned}
$$

Then the plim of $\hat{\sigma}_\theta^2$ is:

$$
\begin{aligned}
\text{plim}(\hat{\sigma}_\theta^2) &= \sigma_\theta^2 - \text{plim}(n(n-p)^{-1}(\theta'X/n)(X'X/n)^{-1}(X'\theta]/n)) \\
&= \sigma_\theta^2 - [\Sigma_{X_e}\beta_0 - \sigma_{Y_eX_e}]'\mathbf{M}_X^{-1}[\Sigma_{X_e}\beta_0 - \sigma_{Y_eX_e}] \qquad (7.3.10)
\end{aligned}
$$

From (7.3.9) and (7.3.10) it is clear that $\hat{\beta}_0$ is asymptotically biased and $\hat{\sigma}^2_\theta$ will always be asymptotically underestimated. When Σ_{X_e} and $\sigma_{Y_eX_e}$ are known, this bias can be removed by the corrected least square (CLS) estimator (Schneeweiss (1976) equation 3.1), defined as:

$$\hat{\beta}_{CLS} = (X'X - n\Sigma_{X_e})^{-1}(X'Y - n\sigma_{Y_eX_e}) \qquad (7.3.11)$$

Then

$$\begin{aligned}
\hat{\beta}_{CLS} - \beta_0 &= (X'X/n - \Sigma_{X_e})^{-1}(X'Y/n - \sigma_{Y_eX_e} - (X'X/n - \Sigma_{X_e})\beta_0) \\
&= (X'X/n - \Sigma_{X_e})^{-1}(X'X\beta_0/n + X'\theta/n - \sigma_{Y_eX_e} - X'X\beta_0/n + \Sigma_{X_e}\beta_0) \\
&\qquad \text{(from (7.3.5))} \\
&= (X'X/n - \Sigma_{X_e})^{-1}(X'(-\Delta_e[\beta_0' \ -1]')/n + \Sigma_{X_e}\beta_0 - \sigma_{Y_eX_e}) \\
&\qquad \text{(from (7.3.6))} \\
&= (X'X/n - \Sigma_{X_e})^{-1}(-X'\Delta_e[\beta_0' \ -1]'/n + [\Sigma_{X_e} \ \ \sigma_{Y_eX_e}][\beta_0' \ -1]') \\
&= (X'X/n - \Sigma_{X_e})^{-1}((-X'\Delta_e/n + [\Sigma_{X_e} \ \ \sigma_{Y_eX_e}])[\beta_0' \ -1]')
\end{aligned}$$

and hence

$$\begin{aligned}
\text{plim}(\hat{\beta}_{CLS} - \beta_0) &= M_{X_0}^{-1}((-[\Sigma_{X_e} \ \ \sigma_{Y_eX_e}] + [\Sigma_{X_e} \ \ \sigma_{Y_eX_e}])[\beta_0' \ -1]') \\
&\qquad \text{(from (7.3.8))} \\
&= 0
\end{aligned}$$

Thus $\hat{\beta}_{CLS}$ is a consistent estimator of β_0 (i.e. is asymptotically unbiased). A consistent estimator of σ^2_θ is

$$\hat{\sigma}^2_{CLS} = \hat{\theta}'\hat{\theta}/n,$$

where
$$\begin{aligned}\hat{\theta} &= Y - X\hat{\beta}_{CLS} \\ &= X\beta_0 + \theta - X\hat{\beta}_{CLS} \quad \text{(by using (7.3.5))} \\ &= X(\beta_0 - \hat{\beta}_{CLS}) + \theta\end{aligned}$$

From
$$\begin{aligned}\hat{\sigma}^2_{CLS} &= \hat{\theta}'\hat{\theta}/n \\ &= [(\beta_0 - \hat{\beta}_{CLS})'X' + \theta'][X(\beta_0 - \hat{\beta}_{CLS}) + \theta]/n \\ &= [(\beta_0-\hat{\beta}_{CLS})'X'X(\beta_0-\hat{\beta}_{CLS})+2\theta'X(\beta_0-\hat{\beta}_{CLS})+\theta'\theta]/n\end{aligned}$$

we have
$$\begin{aligned}\text{plim}(\hat{\sigma}^2_{CLS}) &= \sigma^2_\theta \quad \text{(from (7.3.8))} \\ &= \beta_0'\Sigma_{X_e}\beta_0 - 2\sigma'_{Y_eX_e}\beta_0 + (\sigma^2_\epsilon + \sigma^2_{Y_e}) \qquad (7.3.12)\end{aligned}$$

Hence,
$$(\sigma^2_\epsilon + \sigma^2_{Y_e}) = \sigma^2_\theta - \beta_0'\Sigma_{X_e}\beta_0 + 2\sigma'_{Y_eX_e}\beta_0$$

The CLS estimate of $(\sigma^2_\epsilon + \sigma^2_{Y_e})$ is:

$$(\hat{\sigma}^2_\epsilon + \hat{\sigma}^2_{Y_e})_{CLS} = \hat{\sigma}^2_{CLS} - \hat{\beta}'_{CLS}\Sigma_{X_e}\hat{\beta}_{CLS} + 2\sigma'_{Y_eX_e}\hat{\beta}_{CLS} \qquad (7.3.13)$$

and Schneeweiss (1976) shows that this estimator is consistent. He also shows that

$$(\hat{\sigma}^2_\epsilon)_{CLS} = (\hat{\sigma}^2_{CLS} - \hat{\beta}'_{CLS}\Sigma_{X_e}\hat{\beta}_{CLS} + 2\sigma'_{Y_eX_e}\hat{\beta}_{CLS}) - (\hat{\sigma}^2_{Y_e})_{CLS}$$

estimates (σ^2_ϵ) consistently under the assumptions that $E(Y_e) = E(\epsilon) = 0$, $E(Y_e\epsilon') = 0$, $E(Y_eY_e') = \sigma^2_{Y_e}I$, $E(\epsilon\epsilon') = \sigma^2_\epsilon I$, and $\sigma^2_{Y_e}$ is known.

An alternative formula for (7.3.13) is

$$(\hat{\sigma}^2_\epsilon + \hat{\sigma}^2_{Y_e})_{CLS} = Y'Y/n - (Y'X/n - \sigma'_{Y_eX_e})\hat{\beta}_{CLS} \qquad (7.3.14)$$

When ϵ, Y_e, X_e and X_0 are normally distributed, with the rows of Δ_e and X_0 are identically and independently distributed and under the assumptions already stated, then all these estimators are maximum likelihood estimators. (Johnston (1963)).

7.3.3 Estimation by Total Least Squares

The TLS estimator of β, defined in (7.2.15), is:

$$\hat{\beta}_{TLS} = [X'X-\omega_{p+1}^2 I]^{-1}X'Y$$

and the CLS estimator of β, defined in (7.3.11), is:

$$\hat{\beta}_{CLS} = (X'X - n\Sigma_{X_e})^{-1}(X'Y - n\sigma_{Y_e X_e})$$

Hence, TLS and CLS will yield the same consistent estimator if

(i) $\sigma_{Y_e X_e} = 0$, (zero correlation between the measurement errors), and

(ii) $\Sigma_{X_e} = \sigma_\upsilon^2 I_p$, and ω_{p+1}^2/n is a consistent estimator of σ_υ^2, i.e. all error variables $(X_e)_i$ are stochastically independent and have equal variance σ_υ^2 , consistently estimated by ω_{p+1}^2/n.

Van Huffel and Vandewalle (1985, Theorem 4.3-1) prove that σ_υ^2 is consistently estimated by ω_{p+1}^2/n if $(\sigma_\epsilon^2 + \sigma_{Y_e}^2) = \sigma_\upsilon^2$. They use (7.3.14) to show that $(\sigma_\epsilon^2 + \sigma_{Y_e}^2)$ can be consistently estimated by

$$(\sigma_\epsilon^2 + \sigma_{Y_e}^2)_{TLS} = \omega_{p+1}^2/n \qquad (7.3.15)$$

Their theorem also holds in case of s non-predictive collinearities. The variance of the equation error σ_ϵ^2 is always unknown, so a consistent estimator is only possible is $\sigma_\epsilon^2 = 0$. Hence, TLS estimates consistently the parameters of a pure errors-in-variables model with known covariance matrix of the error variables. This condition implies that (7.3.4) will be known totally (all error variables are stochastically independent with equal variance σ_υ^2):

$$\Sigma_{\Delta_e} = \begin{bmatrix} \sigma_\upsilon^2 I_p & 0 \\ 0 & \sigma_\upsilon^2 \end{bmatrix} \tag{7.3.16}$$

Since Σ_{X_e} is known it is always possible to obtain independence and constancy of variance by transforming the data. This means that TLS always yields a consistent estimate of the parameters of a classical errors-in-variables model provided the appropriate scaling of the data has been performed.

Furthermore, Σ_{X_e} must only be known, up to a factor of proportionality.

The assumption of known error variance means that if the data is not scaled exactly or if the equation error is not zero, the TLS estimate may not be consistent ($E(\sigma_\upsilon^2) \neq \omega_{p+1}^2/n$). Ketellapper (1983) investigated the impact of the violation of this assumption and found that TLS is then also preferable to OLS, except when the error variance is highly overestimated.

For a discussion of the asymptotic properties of TLS in the error-in-variable models see Van Huffel and Vandewalle (1985) and Schneeweiss (1976).

7.4 Subset Selection based on TLS for Prediction

In §1.10 subset selection via QR factorization with column pivoting was discussed, and this technique can be extended to TLS. Van Huffel and Vandewalle (1985) distinguish between three methods:

1. **Algorithm SX-TLS** (subset selection on X where the method of estimation is TLS)

This method is a variant of SX-OLS (§1.10). The third step of Algorithm SX-OLS was to compute a OLS solution for the subset equation $X_1Z = Y$. If some of the variables are perturbed then a higher accuracy of the solution Z and the predicted response Y can be obtained by using TLS. Thus Algorithm SX-TLS is the same as SX-OLS except that the third step is replaced by:

built upe $Z \epsilon R^{\hat{r}}$ such that $\|[\hat{X}_1]_e \ \hat{Y}_e\|_F^2$ is minimized and $\hat{Y}_e \ \epsilon \ R(\hat{X}_1)$.

2. **Algorithm SXY-TLS** (subset selection on [X Y] where the method of estimation is TLS)

Given X:nxp, Y:nx1 and a method of computing an integer r that approximates the rank of [X Y], the following algorithm computes a permutation P and a vector z:$\hat{r}$x1 such that the first $\hat{r}$ columns X_1 of XP are independent and such that $\|[\hat{X}_1]_e \ \hat{Y}_e\|_F^2$ is minimized and $\hat{Y} \ \epsilon \ R(\hat{X}_1)$.

First step: Compute the SVD of [X Y] (1.3.8) and determine $\hat{r} \leq$ rank[X Y], partition $\tilde{V}$

$$\tilde{V} = \begin{array}{cc} \left[\begin{array}{cc} \tilde{V}_{11} & \tilde{V}_{12} \\ \tilde{V}_{21} & \tilde{V}_{22} \\ \tilde{V}_{1y} & \tilde{V}_{2y} \end{array}\right] & \begin{array}{c} \hat{r} \\ p-\hat{r} \\ 1 \end{array} \\ \begin{array}{cc} \hat{r} & \quad p-\hat{r}+1 \end{array} & \end{array} \tag{7.4.1}$$

Second step: use QR with column piving to compute

$$Q'[\tilde{V}'_{11} \ \tilde{V}'_{21}]P = Q'[\ddot{V}'_{11} \ \ddot{V}'_{21}] = [R_{11} \ R_{12}] \text{ and set } XP = [\underset{\hat{r}}{X_1} \ \underset{p-\hat{r}}{X_2}]$$

Third step: determine the TLS solution z:$\hat{r}$x1 of $X_1Z = Y$. R_{11} is non-singular and $\|\ddot{V}_{11}^{-1}\|_2 = \|[R_{11}^{-1}]\|_2$.

The SXY-TLS algorihtm tends to maximize the r-th singular value of $[X_1\ Y]$. Indeed from (1.10.7) we have

$$\frac{\omega_{\hat{r}}}{\|\ddot{V}_{11}^{-1}\|_2} \leq \omega_{\hat{r}}([X_1 Y]) \leq \omega_{\hat{r}} \qquad (7.4.2)$$

The extra information provided by the variable Y in this algorithm might make the stable components of Y predicted by SXY-TLS, i.e. $\sum_{i=1}^{\hat{r}} \tilde{u}_i{}'Y$, superior to the stable components predicted by SX-LS $\sum_{i=1}^{\hat{r}} u_i{}'Y$, when errors are introduced into all variables of the model under consideration.

3. **Algorithm SXY-VTLS** (subset selection on $[X\ Y]$ where the method of estimation is TLS, with a variant)

It follows step one and two of SXY-TLS and then instead of computing step three, z is obtained directly from step two as

$$z = R_{11}^{-1}Q'V_{1y}$$

(Here the authors assumed that the $(n-\hat{r}+1)$ smallest singular values of $[X\ Y]$ are non-zero only due to perturbations.)

7.5 Summary

In this chapter TLS was introduced. The estimator was defined and its properties and use of for errors-in-variables models were discussed, along with subset selection based on TLS. It was observed that TLS is not an appropriate tool for dealing with collinearity.

Chapter 8

COMPARISON OF ESTIMATORS

In the previous chapters regression estimation techniques (OLS, PC, RR, GRR, LRR, FPC, SHE and TLS) have been defined, their properties discussed and some comparisons or remarks have been made. In this chapter we represent some of these remarks and properties in a unified manner in order to compare estimators and give some general direction on finding 'best' estimators. The estimators can be divided into two groups: The first group includes all those that operate on the regressor matrix X, namely OLS, PC, RR, GRR (and its variations), SH and FPC (and its variations): the second group consist of those that operate on the augmented matrix [X Y], namely LRR and TLS.

8.1 Basic Comparisons

In this section we present tabular summaries of the expectation and expected square error properties of the estimators. The tables 8.1 and 8.2 are separated largely for convenience of presentation, as the properties to which they refer are essentially interrelated. Table 8.3 presents some choices (mostly those used in Chapter 10) of k, K, and d.

Table 8.1 Expectation properties of estimators and references

Type	Chapter	Definition	Bias	Variance
OLS	1	$\hat{\beta} = \sum_{i=1}^{p} v_i c_i/\lambda_i$,	0	$\sigma^2 \sum_{i=1}^{p} v_i v_i'/\lambda_i$
	3	$\hat{\delta} = (Z'Z)^{-1}Z'Y$ $= V'\hat{\beta}$	0	$\sigma^2 \sum_{i=1}^{p} 1/\lambda_i$
PC	3	$\hat{\beta}_{PC} = \sum_{i=1}^{p-r} v_i c_i/\lambda_i$,	$-V_2V_2'\beta$	$\sigma^2 \sum_{i=1}^{p-r} v_i v_i'/\lambda_i$
RR	4.2	$\hat{\beta}_R = \sum_{i=1}^{p} (\lambda_i+k)^{-1}c_i v_i$	$-k(X'X+kI)^{-1}\beta$	$\sigma^2 V[\Delta^2+kI]^{-2}\Delta^2 V'$
GRR	4.6	$\hat{\delta}_K = (\Delta^2 + K)^{-1}\Delta^2\hat{\delta}$	$-(\Delta^2 + K)^{-1}K\delta$	$\sigma^2(\Delta^2+K)^{-2}\Delta^2$
AUORR	4.7	$\hat{\delta}_{JW} = [I-[kA^{-1}]^2]\hat{\delta}$ where $A = Z'Z + K$	$-[kA^{-1}]^2\delta$	$\sigma^2[I-[kA^{-1}]^2]^2\Delta^{-2}$
AUGRR	4.7	$[\hat{\delta}_0]_i = \dfrac{(\lambda_i\hat{\delta}_i^2+2\hat{\sigma}^2)\lambda_i\hat{\delta}_i^3}{(\lambda_i\hat{\delta}_i^2+\hat{\sigma}^2)^2}$	see (4.7.24)	and (4.7.25)
SH	5	$\hat{\beta}_{SH} = d \sum_{i=1}^{p} v_i c_i/\lambda_i$	$-(1-d)\beta$	$\sigma^2 d^2 \sum_{i=1}^{p} v_i v_i'/\lambda_i$
FPC	5	$\hat{\delta}_{FPC} = F\hat{\delta}$	$-[I-F]\delta$	$\sigma^2 F\Delta^{-2}F$
	5	$\hat{\beta}_{FPC} = VFV'\hat{\beta}$	$-[I-VFV']\beta$	$\sigma^2 VF\Delta^{-2}FV'$
LRR	6	$\hat{\beta}_{LR} = \sum_{i=1}^{p+1-s} f_i\tilde{v}_i^0$	$-E[\sum_{i=t}^{p+1} f_i\tilde{v}_i^0]$	•
TLS	7	$\hat{\beta}_{TLS} = \sum_{j=1}^{p} (\lambda_j - \omega_{p+1}^2)^{-1} c_j v_j$	•	•

Table 8.2 Mean square errors of estimators

Type	MSE	TMSE
OLS	$\sigma^2 \sum_{i=1}^{p} v_i v_i'/\lambda_i$	$\sigma^2 \sum_{i=1}^{p} 1/\lambda_i$
PC	$\sigma^2 \sum_{i=1}^{p-r} v_i v_i'/\lambda_i + V_2 V_2'\beta\beta' V_2 V_2'$	$\sigma^2 \sum_{i=1}^{p-r} 1/\lambda_i + \beta' V_2 V_2'\beta$
RR	$\sigma^2 V[\Delta^2+kI]^{-2}\Delta^2 V'+k^2(X'X+kI)^{-1}\beta\beta'(X'X+kI)^{-1}$	$\sum_{i=1}^{p} \dfrac{\sigma^2\lambda_i+k^2\delta_i^2}{(\lambda_i + k)^2}$
GRR	$\sigma^2(\Delta^2+K)^{-2}\Delta^2+(\Delta^2+K)^{-1}K\delta\delta' K(\Delta^2+K)^{-1}$	$\sum_{i=1}^{p} \dfrac{\sigma^2\lambda_i+k_i^2\delta_i^2}{(\lambda_i + k_i)^2}$
AUORR	$\sigma^2[I-[kA^{-1}]^2]^2\Delta^{-2}+[kA^{-1}]^2\delta\delta'[kA^{-1}]^2$	$\sum_{i=1}^{p}\left\{\sigma^2\dfrac{\lambda_i(\lambda_i+2k_i)^2}{(\lambda_i+k_i)^4}+\dfrac{k_i^4\delta_i^2}{(\lambda_i+k_i)^4}\right\}$
AUGRR	see (4.7.24) and (4.7.25)	
SH	$\sigma^2 d^2 \sum_{i=1}^{p} v_i v_i'/\lambda_i+(1-d)^2\beta\beta'$	$\sigma^2 d^2 \sum_{i=1}^{p} 1/\lambda_i + (1-d)^2\beta'\beta$
FPC	$\sigma^2 F\Delta^{-2}F + [I-F]\delta\delta'[I-F]$	$\sigma^2 \sum_{i=1}^{p} f_i^2/\lambda_i + \sum_{i=1}^{p} (1-f_i)^2\delta_i^2$
LRR	$\approx \mathrm{MSE}(\hat{\beta}_{PC})$	$\approx \mathrm{TMSE}(\hat{\beta}_{PC})$
TLS	unknown	unknown

Table 8.3 Choices of k, K, and d

type	estimated value	use for	suggested by
k (HK)	$p\hat{\sigma}^2/\hat{\beta}'\hat{\beta}$	RRE	Hoerl, Kennard and Baldwin
k (LW)	$p\hat{\sigma}^2/\sum_{i=1}^{p}\lambda_i\hat{\delta}_i^2$	RRE	Lawless and Wang
K (HKB)	$\hat{k}_i = \hat{\sigma}^2/\delta_i^2$	GRRE	Hoerl, Kennard and Baldwin
K (T)	$\hat{k}_i = \lambda_i/(F_i+1),\ F_i = \lambda_i\hat{\delta}_i^2/\hat{\sigma}^2$	GRRE	Troskie (1990)
k	$p\hat{\sigma}^2/(\sum_{i=1}^{p}[\hat{\delta}_i^2/\{1+\sqrt{1+\lambda_i(\hat{\delta}_i^2/\hat{\sigma}^2)}\}$	AUORR	Nomura
d	$\hat{\beta}'\hat{\beta}/\{\mathrm{tr}([\mathrm{var}(\hat{\beta})] + \hat{\beta}'\hat{\beta}\}$	SHE	*

* an estimated value that minimizes the TLSE of the SHE (5.1.10)

8.2 Judgement of estimators

8.2.1 Unbiasedness

In the class of unbiased estimators, the OLSE is the best linear estimator (BLUE) in the sense of minimum variance. In the presense of collinearity, the variance of OLSE can be inflated (due to small λ_i's) so that other biased estimators will be more suitable under a changed criterion, e.g. mimimum mean square error. If the collinearity between some of the regressors is however consistently continued in the prediction area, the effect of this collinearity on predictions will be less serious. If σ^2 is sufficiently small, β can be estimated by OLS with good accuracy even if strong collinearities exist in X. Thus the choice of whether or not to use the OLSE should be based on the magnitudes of the λ_i and unknown σ^2.

8.2.2. MSE criteria

Consider two competing estimators b_1 and b_2. If the matrix difference

$$S = \text{MSE}(b_2) - \text{MSE}(b_1) \qquad (8.2.1)$$

is positive semi-definite (psd), then b_1 is to be preferred to b_2. S will be psd if $w'Sw \geq 0$ for any non-zero vector w:nx1. The TMSE equivalent is that b_1 is prefered to b_2 whenever

$$\text{TMSE}(w'b_2) \geq \text{TMSE}(w'b_1), \text{ for every vector } w \qquad (8.2.2)$$

This MSE criterion is the so called 'strong MSE criterion', and a weaker criterion for b_1 to be preferred to b_2 is that

$$\text{TMSE}(b_2) \geq \text{TMSE}(b_1) \qquad (8.2.3)$$

Looking at these criteria it is worthwhile to point out:

(i) $\text{TMSE}(b_i)$ is the Euclidian distance between b_i and β. One therefore seeks an estimator that minimizes this norm.

(ii) The criteria (8.2.1) to (8.2.3) were defined in PC estimation as ways to determine which PC's to eliminate.

(iii) Although only some criteria are explicity stated here, there is a whole range available. For instance, all those criteria applied in PC to eliminate PC's can be generalized. A detailed discussion of criteria appears in Vinod and Ullah (1981, Chapter 2).

(iv) Some authors perform comparisons based on the relative efficiency (RE) of each estimator to OLSE (i.e. see section 4.7 where this RE concept was used to compare the various GRRE's). There appears to be no statistical analysis of these efficiency ratios, literally only comparisons of the values (i.e. $2 < 3$, or 2.9 is slightly better than 3, and so on)

(v) Empirical comparisons of estimators reveals that no one estimator is clearly superior to the others. The conditions for superiority depend on the degree of collinearity, the orientation of δ, and the value of σ^2. These factors should always be considered when choosing an estimator. Although some rough guidelines can be given, the 'best' estimator of any problem will be unique to that particular problem and no 'recipe' is available.

(vi) Simulation studies have been performed to compare the different estimators under various degrees of collinearity, changing σ^2 values, and different orientations of the δ's. Some of these results have been reported in the appropriate foregoing chapters, and a study of some FPC estimators will be reported in the next section.

8.3 Comparison of FPC estimators

In Chapter 5 it was shown that most estimators (OLS, RR, GRR, PC) can be considered as FPC estimators. All the FPC estimators give an improvement over OLS in terms of the MSE criteria if

$$\sigma^2 \geq \delta'\Delta^2(I-F)^2(I-F^2)^{-}\delta \qquad (5.2.10)$$

Thus the improvement over OLS will depend on the vector δ, the degree of collinearity (Δ), and σ^2.

In (5.2.12) the optimal values for the fractions were obtained as

$$f_j^0 = \delta_j^2\lambda_j(\sigma^2 + \delta_j^2\lambda_j)^{-1}$$

and the min TMSE will be

$$\min\{TMSE(\hat{\delta}_{FPC})\} = TMSE(\hat{\delta}) - \sum_{i=1}^{p} (\sigma^2/\lambda_i)(1-f_i^0) \quad \text{(from (5.2.14))}$$

By using this optimal value two new estimators (FPCI and FPCV) were defined in (5.2.16) and (5.2.18).

To evaluate different FPC estimators under a variety of conditions Lee and Birch (1988) set up the following artificial data sets:

p = 4: n = 20; eigenvalues : 2.96072, 1.02801, 0.0112, and 0.00012; condition number = 157.07762.

Two orientations of δ were considered:

I: δ = [385.112 91.9509 333.138 -198.061]

II: δ = [-51.4399 -86.844 -28.844 542.225]

Two values of σ were chosen as 5 and 10.

They then compared OLS, RR (where k is obtained by an iterative method starting with OLS), GRR (K is obtained by an iterative method as in (5.2.15) starting with OLS), two PC estimators (called PC(1) and PC(2) to denote the number of deleted eigenvalues and vectors) and the two new estimators (FPCI and FPCV with iteration initial value PC(1)), by means of a Monte Carlo study with the number of repetitions set at 50. The performances of the estimators across the four combinations and the 50 repetitions was summarized by computing the standardized empirical mean square error defined as

$$\mathrm{SEMSE(b)} = \sum_{j=1}^{4} \sum_{i=1}^{50} \frac{(b_{ji} - \beta_j)^2}{50\sigma^2_{j,LS}} \qquad (8.3.1)$$

where b_{ji} is the j-th element of b, an estimator of β, in the i-th Monte Carlo repetition and $\sigma^2_{j,LS}$ is the theoretical variance of $b_{j,LS}$, j =1,...,4. Lee (1986) shows that the SEMSE($\hat{\beta}$) is independent of both the orientation of the parameter vector and the value of σ^2_ϵ but does depend on the structure of the regressor variables. The use of the SEMSE(b) facilitates comparisons of estimators both within a given combination and across the combinations. The results of this Monte Carlo study are displayed in Table II of Lee and Birch (1988), where the evaluations are based on the relative efficiency of each estimator compared to $\hat{\beta}$. Their results and suggestions can be summarized as follows:

1. While PC(1) is far superior to LS in terms of SEMSE, just the opposite is true for PC(2) estimator. This suggests that the user of PC regression should exercise caution in determining which principal components should be deleted. This finding is analogous to the test and ideas described in Chapter 3. It is interesting that the authors find for orientation II and σ = 5, a PCE with only the third component deleted more efficient than LS, PC(1) and PC(2). They do not however give any explanation as to why this phenomenon occurred and it may constitute another reason why one should always be on the alert, when deleting PC's.

2. The relative behaviour of the RR, GRR, FPCI and FPCV estimators is dependent upon the orientation of the parameter vector. The success of the FPCI and FPCV depends on the proper choice of the PC estimator used as the initial values. Once this choice is made the authors suggest that either FPCI or FPCV could be computed, and for their simulation each gives practically the same results.

3. As σ^2 increases, the relative efficiency of estimators increased with respect to OLS.

4. The FPCV estimator is always more efficient than the RR estimator, but not always more efficient than the GRR estimator. The FPCI estimator is not always more efficient than the RR and GRR estimator.

5. In conclusion: among the biased estimators, no single estimator dominates the others, in terms of TMSE, across all conditions.

8.4 Summary

In this chapter the various estimators were summarized. Judgement and comparisons of estimators were discussed and the idea of a class of FPC estimators was introduced.

Chapter 9

INFLUENCE AND COLLINEARITY

In this chapter the influence of observations as well as variables on the collinearity of the model is investigated. We introduce in §9.1 the concept of influential points, ways of determining collinearity-influential points in §9.2, influential variables in §9.3 and in §9.4 the concept of influence in ridge regression estimators and weighted least squares. In this chapter we will assumme that X in (1.1) is a full column rank matrix.

9.1 Influential points

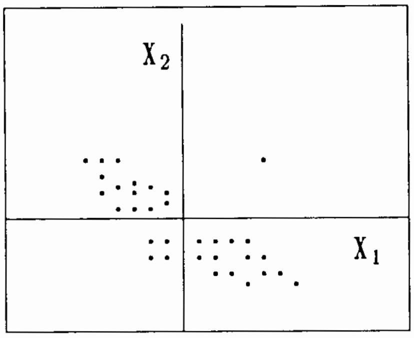
Traditionally, collinearity has been associated with the columns and the column space of X. However, as already pointed out in Chapter 2, the collinearity structure of the data can be strongly affected by a few observations (Belsley, *et al.* (1980), Mason and Gunst (1985), Draper and John (1981)). The term influential is used to describe an observation whose inclusion in a data set substantially changes regression coefficent estimates, predicted responses, or the results of inferential procedures (Mason and Gunst (1985)). Not all outliers are necessarily collinearity-influence points and *vice versa*.

A formal definition for a collinearity-influential point or case is given by Walker (1989):

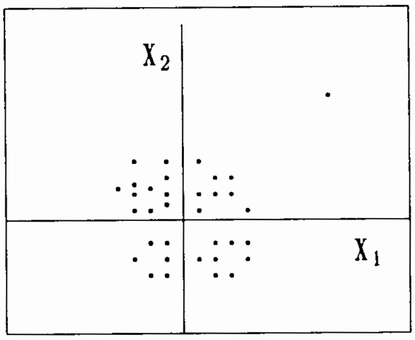
> Let x_i' be the i-th row of the X matrix and let η_k and $\eta_k(-i)$ be the k-th condition indices computed with and without x_i', respectively. The i-th point (x_i') is a collinearity-influential point if, for a predetermined value δ,
>
> $$|\eta_k - \eta_k(-i)| > \delta\eta_k \quad \text{for } k = 2,3,\ldots,p \qquad (9.1.1)$$

This definition focusses upon the potential of the case to influence the estimation and inference procedures. But the actual influence depends on the Y-value observed for that case. Influential points in the sense of

Walker can be classified into two broad categories: those that mask collinearity and those that induce collinearity, illustrated in the following figures from Walker (1989)



Masking collinearity
$(\eta_k << \eta_k(-i))$



Inducing collinearity
$(\eta_k >> \eta_k(-i))$

9.2 Detecting collinearity-influential points

The first step in finding collinearity-influential points is to detect outliers. Various methods of detecting outliers are available in the literature. Common methods include graphical representation of the residuals and standardized residuals versus the individual fitted Y and observed X variables, and normal probability plots (see for instance Cook and Weisberg (1982) and Daniel and Wood (1980)). High leverage, as defined below, is also an indication that a point is potentially influential. Once a observation is potentially influential it is necessary to see what actual influence it has on the estimator and inference. Various measures of this influence will be discussed.

9.2.1 Leverage

From §1.7 the Hat matrix is

$$H = X(X'X)^{-1}X'$$

It maps Y into $\hat{Y}$

$$\hat{Y} = HY$$

and the diagonal elements of H are known as the leverage values:

$$h_{ii} = x_i'(X'X)^{-1}x_i \tag{1.7.2}$$

The following properties of h_{ii} are pertinent here:

(i) $h_{ii} = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2$ (from H idempotent) (9.2.1)

(ii) $0 \leq h_{ii} \leq 1$ (H is a projection matrix) (9.2.2)

(iii) $\sum_{i=1}^{n} h_{ii}^2 = p$ (X is full column rank) (9.2.3)

When h_{ii} is large, the i-th case is called a high-leverage point. Hoaglin and Welsh (1978) suggested that a point has high leverage if $h_{ii} > 2p/n$. Mason and Gunst (1985) use $h_{ii} > 2(p+1)/n$. The influence of the response value Y_i on the fitted value $\hat{Y}_i$ is reflected in the corresponding leverage. For the two extreme cases, (h_{ii} = 0 or 1) we have:

If $h_{ii} = 0$, then $\sum_{j \neq i} h_{ij}^2 = 0$ (from (9.2.1)) and thus $h_{ij} = 0$ and $\hat{Y}_i = 0$. Thus $\hat{Y}_i$ must be fixed at zero by design - it is not effected by any Y_i.

If $h_{ii} = 1$, then $\sum_{j \neq i} h_{ij}^2 = 0$ (from (9.2.1)) and thus $h_{ij} = 0$ and $\hat{Y}_i = Y_i$, implying that $\hat{\epsilon}_i = 0$. The model fits the data value exactly.

9.2.2 Outlier sum of squares

Once outliers or influential observations are detected model (1.1) can be partitioned as

$$\begin{bmatrix} Y_a \\ Y_b \end{bmatrix} = \begin{bmatrix} X_a \\ X_b \end{bmatrix} \beta + \begin{bmatrix} \epsilon_a \\ \epsilon_b \end{bmatrix} \tag{9.2.4}$$

where by some rearrangement of rows the k observations that are influential or outlying are contained in $[Y_b \; X_b]$. Thus Y_a and ϵ_a are (n-k)x1 vectors, Y_b and ϵ_b are kx1 vectors, X_a is a (n-k)xp matrix and X_b is a kxp matrix. The residuals $\hat{\epsilon} = [I-H]Y$ under this model can also be partitioned as

$$\begin{bmatrix} \hat{\epsilon}_a \\ \hat{\epsilon}_b \end{bmatrix} = \begin{bmatrix} I_{n-k} - X_a(X'X)^{-1}X_a' & -X_a(X'X)^{-1}X_b' \\ -X_b(X'X)^{-1}X_a' & I_k - X_b(X'X)^{-1}X_b' \end{bmatrix} \begin{bmatrix} Y_a \\ Y_b \end{bmatrix} \tag{9.2.5}$$

Deleting the suspected outliers gives a model with $E[Y_a] = X_a\beta$. Draper and John (1981) suggested the following alternative model:

$$\begin{bmatrix} Y_a \\ Y_b \end{bmatrix} = \begin{bmatrix} X_a & 0 \\ X_b & I_k \end{bmatrix} \begin{bmatrix} \beta \\ \alpha \end{bmatrix} + \begin{bmatrix} \theta_a \\ \theta_b \end{bmatrix} \tag{9.2.6}$$

where α is a kx1 vector of additional parameters. The estimates of β and α under this model are:

$$Y_a = X_a\tilde{\beta}, \text{ thus } \tilde{\beta} = (X_a'X_a)^{-1}X_a'Y_a \tag{9.2.7}$$

$$Y_b = X_b\tilde{\beta} + \tilde{\alpha}$$

Thus, by using (9.2.5)

$$\begin{aligned} \tilde{\alpha} &= Y_b - X_b\tilde{\beta} \\ &= [I-X_b(X'X)^{-1}X_b']^{-1}[\hat{\epsilon}_b+X_b(X'X)^{-1}X_a'Y_a] - X_b(X_a'X_a)^{-1}X_a'Y_a \\ &= [I-X_b(X'X)^{-1}X_b']^{-1}[\hat{\epsilon}_b+X_b(X'X)^{-1}X_a'Y_a-[I-X_b(X'X)^{-1}X_b']X_b(X_a'X_a)^{-1}X_a'Y_a] \end{aligned}$$

and

$$
\begin{aligned}
&X_b(X'X)^{-1}X_a'Y_a - [I-X_b(X'X)^{-1}X_b']X_b(X_a'X_a)^{-1}X_a'Y_a \\
&\quad = X_b(X'X)^{-1}X_a'Y_a - X_b(X_a'X_a)^{-1}X_a'Y_a + X_b(X'X)^{-1}X_b'X_b(X_a'X_a)^{-1}X_a'Y_a \\
&\quad = X_b(X'X)^{-1}[I-(X'X)(X_a'X_a)^{-1}+X_b'X_b(X_a'X_a)^{-1}]X_a'Y_a \\
&\quad = X_b(X'X)^{-1}[I-\{(X_a'X_a+X_b'X_b)-X_b'X_b\}(X_a'X_a)^{-1}]X_a'Y_a \\
&\quad = X_b(X'X)^{-1}[0]X_a'Y_a \\
&\quad = 0
\end{aligned}
$$

we have

$$
\begin{aligned}
\tilde{\alpha} &= [I-X_b(X'X)^{-1}X_b']^{-1}[\hat{\epsilon}_b+0] \\
&= [I-X_b(X'X)^{-1}X_b']^{-1}\hat{\epsilon}_b \qquad (9.2.8)
\end{aligned}
$$

The adjusted observations Y_b with expectations $X_b\beta$ can be estimated by $Y_b - \tilde{\alpha}$, then model (9.2.4) becomes

$$
\begin{bmatrix} Y_a \\ Y_b-\tilde{\alpha} \end{bmatrix} = \begin{bmatrix} X_a \\ X_b \end{bmatrix} \beta + \begin{bmatrix} \theta_a \\ \theta_b \end{bmatrix} \qquad (9.2.9)
$$

thus,

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

$$
\begin{aligned}
Y_b - \tilde{\alpha} &= Y_b - [I-X_b(X'X)^{-1}X_b']^{-1}\hat{\epsilon}_b \quad \text{(from 9.2.8)} \\
&= [I-X_b(X'X)^{-1}X_b']^{-1}X_b(X'X)^{-1}X_a'Y_a \quad \text{(from 9.2.5)}
\end{aligned}
$$

$$
\begin{aligned}
\hat{\theta}_b &= (Y_b-\tilde{\alpha}) - X_b\hat{\beta} \\
&= [I-X_b(X'X)^{-1}X_b']^{-1}X_b(X'X)^{-1}X_a'Y_a - X_b(X'X)^{-1}X_a'Y_a \\
&\qquad -X_b(X'X)^{-1}X_b'[I-X_b(X'X)^{-1}X_b']^{-1}X_b(X'X)^{-1}X_a'Y_a \\
&= [I-[I-X_b(X'X)^{-1}X_b']-X_b(X'X)^{-1}X_b'][I-X_b(X'X)^{-1}X_b']^{-1}X_b(X'X)^{-1}X_a'Y_a \\
&= 0 \qquad (9.2.10)
\end{aligned}
$$

$$\begin{aligned}\hat{\theta}_a &= Y_a - X_a\hat{\beta}\\ &= Y_a - X_a(X'X)^{-1}X_a'Y_a - X_a(X'X)^{-1}X_b'[I-X_b(X'X)^{-1}X_b']^{-1}X_b(X'X)^{-1}X_a'Y_a\\ &= [I- X_a(X'X)^{-1}X_a' - X_a(X'X)^{-1}X_b'[I-X_b(X'X)^{-1}X_b']^{-1}X_b(X'X)^{-1}X_a']Y_a\end{aligned} \tag{9.2.11}$$

The extra sum of squares due to fitting α in model (9.2.6), as compared with model (9.2.4) is

$$Q_k = \hat{\epsilon}_b'[I-X_b(X'X)^{-1}X_b']^{-1}\hat{\epsilon}_b \tag{9.2.12}$$

Gentleman and Wilk (1975), called Q_k the outlier sum of squares associated with the observations Y_b. Then the F-statistic, associated with the hypothesis H_0: $\alpha = 0$ is:

$$\begin{aligned}F &= \frac{\hat{\epsilon}_b'[I-X_b(X'X)^{-1}X_b']^{-1}\hat{\epsilon}_b}{\hat{\epsilon}'\hat{\epsilon}-\hat{\epsilon}_b'[I-X_b(X'X)^{-1}X_b']^{-1}\hat{\epsilon}_b}\cdot\frac{n-p-k}{k}\\ &= \frac{Q_k}{RSS-Q_k}\cdot\frac{n-p-k}{k}\end{aligned} \tag{9.2.13}$$

9.2.3 Andrews-Pregibon statistic

Andrews and Pregibon (1978) consider the following two augmented matrices:

$X_1^* = [X\ Y]$ (from model (1.1))

$X_2^* = [X\ D\ Y]$ (from model (9.2.6))

where D is the matrix with which the X matrix was augmented in (9.2.6)

$$D = \begin{bmatrix}0\\ I_k\end{bmatrix}$$

The Andrews-Pregibon (AP) statistic is then defined as:

$$AP^{k}_{ij..} = |X_2^{*\prime}X_2^{*}|/|X_1^{*\prime}X_1^{*}| \qquad (9.2.14)$$

where ij.. denote the k subscripts that estimate Y_b (the outliers). For example AP^{2}_{34} means the Andrews-Pregibon statistic for 2 suspected outliers cases 3 and 4. Then from the result of Draper and John (1981, p25) we have

$$|X_2^{*\prime}X_2^{*}| = |X'X|.RSS.|(I-X_b(X'X)^{-1}X_b'|.(1-Q_k/RSS)$$

$$(9.2.15)$$

$$|X_1^{*\prime}X_1^{*}| = |X'X|.RSS$$

where RSS is the residual sums of squares from the full model (1.1). Then (9.2.14) can be written as the dimensionless quantity

$$AP^{k}_{ij..} = |(I-X_b(X'X)^{-1}X_b'|.(1-Q_k/RSS) \qquad (9.2.16)$$

and Andrews and Pregibon (1978) regard the quantity $(1-AP^{k}_{ij..})$ 'as the proportion of the volume generated by X_1^{*} attributable to the k observation (ij...)'. Hence an interpretation of $AP^{k}_{ij..}$ is that small values of $AP^{k}_{ij..}$ are associated with influential observations.

The AP statistic in (9.2.16) consists of two factors:

(i) $|(I-X_b(X'X)^{-1}X_b'|$ only involves the regressor variables, and $X_b(XX)^{-1}X_b'$, is the leverage of Hoaglin and Welsh (1978). For k = 1 the leverage will be $x_i'(X'X)^{-1}x_i = h_{ii}$, and thus $|(I-X_b(X'X)^{-1}X_b'| = 1-h_{ii}$. Small values of $|(I-X_b(X'X)^{-1}X_b'|$ correspond to large h_{ii}, revealing high-leverage points. This factor reflects potential influence of the cases.

(ii) $(1 - Q_k/RSS) = (RSS - Q_k)/RSS$ is a decreasing function of the F-statistic of (9.2.13). Small values of $(RSS - Q_k)/RSS$ indicate high influence (large sum of squares for outliers). This factor reflects actual influence of the observed responses on the fitting.

9.2.4 Cook's statistic

In (1.7.7) Cook's distance was defined as

$$C = D^2 = (\hat{\beta} - \tilde{\beta})'X'X(\hat{\beta} - \tilde{\beta})/p\hat{\sigma}^2 \tag{1.7.7}$$

To determine the degree of influence the i-th data point has on the estimate, $\tilde{\beta}$ will be replaced by $\hat{\beta}_{-i}$, where $\hat{\beta}_{-i}$ is the least square estimate of β with the i-th point deleted. Thus for $i = 1,2,\ldots,n$

$$C_i = (\hat{\beta} - \hat{\beta}_{-i})'X'X(\hat{\beta} - \hat{\beta}_{-i})/p\hat{\sigma}^2 \tag{9.2.17}$$

Computed values of C_i can be compared to the $F(p,n-p)$ distribution. For example if C_i equals the 0.50 value of the corresponding F-distribution, then deletion of the i-th case moves the estimate of β to the edge of a 50% confidence ellipsoid relative to $\hat{\beta}$. Cook (1977) suggests that for an uncomplicated analysis one would like each $\hat{\beta}_{-i}$ to stay well within a 10% confidence region. The comparison of C_i to an F is only used for converting C_i to a familiar scale, and C_i is not distributed as F (Cook and Weisberg (1982)).

The statistic C_i can be simplified as follows:

$$\hat{\beta} - \hat{\beta}_{-i} = \hat{\beta} - \hat{\beta} + (X'X)^{-1}x_i\hat{\epsilon}_i/(1-h_{ii}) \quad \text{(from (1.8.7))}$$

$$= (X'X)^{-1}x_i\hat{\epsilon}_i/(1-h_{ii}) \tag{9.2.18}$$

thus

$$\begin{aligned} C_i &= x_i'(X'X)^{-1}x_i(\hat{\epsilon}_i(1-h_{ii})^{-1})^2/p\hat{\sigma}^2 \\ &= p^{-1}[\hat{\epsilon}_i/\{\hat{\sigma}(1-h_{ii})^{\frac{1}{2}}\}]^2\ [h_{ii}/(1-h_{ii})] \\ &= p^{-1}r_i^2[h_{ii}/(1-h_{ii})] \qquad (9.2.19) \end{aligned}$$

Here $r_i = \hat{\epsilon}_i/\{\hat{\sigma}(1-h_{ii})^{\frac{1}{2}}\}$ is the standardized residual as defined in (1.7.3) and it is a measure of the degree to which the i-th observation can be considered as an outlier from the assumed model. We note that r_i is a monotonic function of Q_1 (9.2.12). The ratio $h_{ii}/(1-h_{ii})$ measures the relative sensitivity of the estimate $\hat{\beta}$, to potential outlying values, so that large values of this ratio indicate the associated point has heavy weight in the determination of $\hat{\beta}$.

The (squared) distance C_i can be extended to contexts in which more than one case is an outlier ($k > 1$, in the outlier model (9.2.6)). Thus $\tilde{\beta}$ in (1.7.7) is computed from model (9.2.9):

$$\begin{aligned} \tilde{\beta} &= (X'X)^{-1}X'Y \\ &= (X'X)^{-1}[X_a'\ X_b'][Y_a'\ (Y_b-\tilde{a})']' \\ &= (X'X)^{-1}[X_a'Y_a\ + X_b'Y_b - X_b'\tilde{a}] \\ &= (X'X)^{-1}[X'Y - X_b'\tilde{a}] \\ &= \hat{\beta} - (X'X)^{-1}X_b'\tilde{a} \qquad (9.2.20) \end{aligned}$$

thus

$$\hat{\beta} - \tilde{\beta} = (X'X)^{-1}X_b'\tilde{a} \qquad (9.2.21)$$

The distance C_i is now indicated by $C_{ij..}$, where the subscripts ij.. denote the cases contained in $[Y_b\ X_b]$, thus

$$C_{ij..} = \frac{\tilde{a}'X_b(X'X)^{-1}(X'X)(X'X)^{-1}X_b'\tilde{a}}{p\hat{\sigma}^2}$$

$$= \frac{Q_k}{p\hat{\sigma}^2}\ \frac{\tilde{a}'X_b(X'X)^{-1}X_b'\tilde{a} + \tilde{a}'\tilde{a} - \tilde{a}'\tilde{a}}{Q_k}$$

$$= \frac{Q_k}{p\hat{\sigma}^2}\left[\frac{\tilde{a}'\tilde{a}}{Q_k} - \frac{\tilde{a}'[I - X_b(X'X)^{-1}X_b']\tilde{a}}{Q_k}\right]$$

$$= \frac{Q_k}{p\hat{\sigma}^2}\left[\frac{\tilde{a}'\tilde{a}}{Q_k} - \frac{\hat{\epsilon}_b'[I - X_b(X'X)^{-1}X_b']^{-1}\hat{\epsilon}_b}{Q_k}\right] \quad \text{(from (9.2.8))}$$

$$= \frac{Q_k}{p\hat{\sigma}^2}\left[\frac{\tilde{a}'\tilde{a}}{Q_k} - 1\right] \quad \text{(from (9.2.12)} \qquad (9.2.22)$$

9.2.5 DFFITS

The change of fit on forecasting when an observation is deleted, is defined (Belsley *et al.* (1980)) as

$$\text{DFFIT}_i = \hat{Y}_i - [\hat{Y}(-i)]_i$$

$$= x_i'[\hat{\beta} - \hat{\beta}_{-i}]$$

$$= h_{ii}\hat{\epsilon}_i/(1-h_{ii}) \quad \text{(from (9.2.18))} \qquad (9.2.23)$$

where $[\hat{Y}(-i)]_i$ is the i-th element of $[\hat{Y}(-i)]$, and $[\hat{Y}(-i)]$ is the estimated Y obtained by using $\hat{\beta}_{-i}$. For scaling purposes (9.2.23) is divided by

$\hat{\sigma}(i)\{h_{ii}\}^{\frac{1}{2}}$, where $\hat{\sigma}(i)$ is the estimated error variance when the i-th row of X and Y have been deleted (see (1.7.5)). The measure DFFITS (Belsley *et al.* (1980, p15)) is the standardized change in the fitted value of a case when it is deleted, and is given for the i-th case by

$$\text{DFFITS}_i = \left[\frac{h_{ii}}{1-h_{ii}}\right]^{\frac{1}{2}} \frac{\hat{\epsilon}_i}{\hat{\sigma}(i)\sqrt{1-h_{ii}}} \qquad (1.7.6)$$

The authors warned against the misuse of deleting high-influence data points solely to effect a desired change in a particular estimated coefficient, or t-value. Once a high-influence point is identified it should only be deleted if shown to be in error, or if it has the effect of inducing collinearity, which, can be an undesirable property of the model.

The authors suggested a size-adjusted cut-off value for DFFITS_i as $2\sqrt{p/n}$, taking into account the sample size (n) as well as the number of variables in the model. If the DFFITS_i are divided into distinct groups and if a noticable gap appears, caution should be exercised in deleting observations.

The statistic DFFITS can be extended to include more than one data point. If the data points to be deleted are indicated by I, where $I \subset \{1,2,\ldots,n\}$ then

$$\text{MDFFITS}_I = [\hat{\beta} - \hat{\beta}_{-I}]'X'_{-I}X_{-I}[\hat{\beta} - \hat{\beta}_{-I}] \qquad (9.2.24)$$

where $\hat{\beta}_{-I}$ is the OLS estimator with the {I} rows deleted and X_{-I} is the X matrix with the {I} rows deleted.

9.2.6 Variance inflation factor

Schall and Dunne (1987b) suggested the following statistic for the detection of collinearity-influential points:

$$R_{ij} = \frac{\mathrm{VIF}_i^{-j}}{\mathrm{VIF}_i} \tag{9.2.25}$$

where VIF_i^{-j} denotes the i-th variance inflation factor obtained after deletion of the j-th observation from model (1.1). When $R_{ij} \approx 1$ for all i and j, there are no collinearity influential points; if $R_{ij} << 1$, x_j is a point that induces collinearity; and if $R_{ij} >> 1$, x_j hides (masks) collinearity.

A straightforward generalization of (9.2.25) can be defined as

$$R_{IJ} = \frac{\mathrm{VIF}_I^{-J}}{\mathrm{VIF}_I} \tag{9.2.26}$$

where $I \subset \{1,\ldots,p\}$, $J \subset \{1,\ldots,n\}$ and VIF_I^{-J} is the variance inflation factor of the set of I variables and J is the subset of data points that are deleted before calculating it.

The augmented model (9.2.6), where I_k reduces to u_j, is

$$Y = [X\ u_j]\begin{bmatrix}\beta\\ \lambda\end{bmatrix} + \epsilon \tag{9.2.27}$$

Since here u_j is the j-th unit vector in the space R^n (not the j-th column of U in the SVD of X), VIF_i^{-j} can be obtained as the partial variance inflation factor $\mathrm{VIF}_i(X_{-i}|u_j)$:

$$R_{ij} = \frac{VIF_i^{-j}}{VIF_i}$$

$$= \frac{VIF_i(X_{-i}|u_j)}{VIF_i(X_{-i})}$$

$$= \frac{VIF_i([X_{-i}\ \ u_j])}{VIF_i(X_{-i})\cdot VIF_i(u_j)} \qquad (9.2.28)$$

Let e_i denote the residual vector of X_i when this variable is regressed on X_{-i}, thus

$$e_i'e_i = X_i'[I - X_{-i}(X'_{-i}X_{-i})^{-1}X'_{-i}]X_i \qquad (9.2.29)$$

and, similarly let e_i^{-j} denote the residual vector of X_i when the chosen regressors are $[X_{-i}\ \ u_j]$, thus

$$(e_i^{-j})'(e_i^{-j}) = X_i'\{I - [X_{-i}\ \ u_j]\begin{bmatrix} X'_{-i}X_{-i} & X'_{-i}u_j \\ u_j'X_{-i} & u_j'u_j \end{bmatrix}^{-1}[X_{-i}\ \ u_j]'\}X_i \qquad (9.2.30)$$

Note that the first diagonal element of $\{[X\ \ u_j]'[X\ \ u_j]\}^{-1}$ is

$$\{[X\ \ u_j]'[X\ \ u_j]\}^{-1}_{11} = \begin{bmatrix} X_i'X_i & X_i'X_{-i} & X_i'u_j \\ X'_{-i}X_i & X'_{-i}X_{-i} & X'_{-i}u_j \\ u_j'X_i & u_j'X_{-i} & u_j'u_j \end{bmatrix}^{-1}_{11}$$

$$= \{X_i'X_i - [X_i'X_{-i}\ \ X_i'u_j]\begin{bmatrix} X'_{-i}X_{-i} & X'_{-i}u_j \\ u_j'X_{-i} & u_j'u_j \end{bmatrix}^{-1}\begin{bmatrix} X'_{-i}X_i \\ u_j'X_i \end{bmatrix}\}^{-1}$$

$$= \{X_i'X_i - X_i'[X_{-i}\ \ u_j]\begin{bmatrix} X'_{-i}X_{-i} & X'_{-i}u_j \\ u_j'X_{-i} & u_j'u_j \end{bmatrix}^{-1}\begin{bmatrix} X'_{-i} \\ u_j' \end{bmatrix}X_i\}^{-1}$$

$$= \{X_i'[I - [X_{-i}\ \ u_j]\begin{bmatrix} X'_{-i}X_{-i} & X'_{-i}u_j \\ u_j'X_{-i} & u_j'u_j \end{bmatrix}^{-1}\begin{bmatrix} X'_{-i} \\ u_j' \end{bmatrix}]X_i\}^{-1}$$

$$= \{(e_i^{-j})'(e_i^{-j})\}^{-1} \qquad (9.2.31)$$

The first diagonal element of $[X'X]^{-1}$, is

$$[X'X]_{11}^{-1} = \begin{bmatrix} X_i'X_i & X_i'X_{-i} \\ X'_{-i}X_i & X'_{-i}X_{-i} \end{bmatrix}_{11}^{-1}$$

$$= \{X_i'X_i - X_i'X_{-i}(X'_{-i}X_{-i})^{-1}X'_{-i}X_i\}^{-1}$$

$$= \{X_i'[I - X_{-i}(X'_{-i}X_{-i})^{-1}X'_{-i}]X_i\}^{-1}$$

$$= \{e_i'e_i\}^{-1} \qquad (9.2.32)$$

and the first diagonal element of $\{[X_i\ \ u_j]'[X_i\ \ u_j]\}^{-1}$ is

$$\{[X_i\ \ u_j]'[X_i\ \ u_j]\}_{11}^{-1} = \begin{bmatrix} X_i'X_i & X_i'u_j \\ u_j'X_i & u_j'u_j \end{bmatrix}_{11}^{-1}$$

$$= \begin{bmatrix} X_i'X_i & x_{ij} \\ x_{ij} & 1 \end{bmatrix}_{11}^{-1}$$

$$= 1/(X_i'X_i - x_{ij}^2)\begin{bmatrix} 1 & -x_{ij} \\ -x_{ij} & X_i'X_i \end{bmatrix}_{11}$$

$$= 1/(X_i'X_i - x_{ij}^2) \qquad (9.2.33)$$

Then

$$\text{VIF}_i(X_{-i}\ \ u_j) = \frac{V(X_i\ \ X_{-i}\ \ u_j)}{V(X_i)}$$

$$= \frac{\{[X\ \ u_j]'[X\ \ u_j]\}_{ii}^{-1}}{(X_i'X_i)^{-1}}$$

$$= \{(e_i^{-j})'(e_i^{-j})\}^{-1}/(X_i'X_i)^{-1} \quad \text{(from (9.2.31))}$$

$$= (X_i'X_i)/\{(e_i^{-j})'(e_i^{-j})\} \qquad (9.2.34)$$

$$\mathrm{VIF}_i(X_{-i}) = \frac{V(X_i \; X_{-i})}{V(X_i)}$$

$$= \frac{(X'X)^{-1}_{ii}}{(X_i'X_i)^{-1}}$$

$$= \{e_i'e_i\}^{-1}/(X_i'X_i)^{-1} \quad \text{(from (9.2.32))}$$

$$= (X_i'X_i)/\{e_i'e_i\} \qquad (9.2.35)$$

and

$$\mathrm{VIF}_i(u_j) = \frac{V(X_i \; u_j)}{V(X_i)}$$

$$= \frac{\{[X_i \; u_j]'[X_i \; u_j]\}^{-1}_{11}}{(X_i'X_i)^{-1}}$$

$$= (X_i'X_i)/(X_i'X_i - x_{ij}^2) \quad \text{(from (9.2.33))} \qquad (9.2.36)$$

where x_{ij} are the j-th element of X_i. Then by inserting (9.2.34), (9.2.35) and (9.2.36), into (9.2.28), R_{ij} can be written as

$$R_{ij} = \frac{\mathrm{VIF}_i([X_{-i} \; u_j])}{\mathrm{VIF}_i(X_{-i})\cdot \mathrm{VIF}_i(u_j)}$$

$$= \frac{(X_i'X_i)}{(e_i^{-j})'(e_i^{-j})} \quad \frac{e_i'e_i}{X_i'X_i} \quad \frac{X_i'X_i - x_{ij}^2}{X_i'X_i}$$

$$= \frac{e_i'e_i}{(e_i^{-j})'(e_i^{-j})} \; \frac{X_i'X_i - x_{ij}^2}{X_i'X_i}$$

$$= \frac{e_i'e_i}{e_i'e_i - e_{ij}^2/(1-[P_{-i}]_{jj})} \; \frac{X_i'X_i - x_{ij}^2}{X_i'X_i} \qquad (9.2.37)$$

$$= R_{ij}^{1} \cdot R_{ij}^{2}$$

where $[P_{-i}]_{jj} = X_{-i}(X'_{-i}X_{-i})^{-1}X'_{-i}$ is the projection matrix onto the column space of the variables X_{-i}, and e_{ij} is the j-th elements of e_i. Schall and Dunne (1987b) claim that the quantity $(R_{ij}^{1} - 1)/(n-p)$ follows an $F(1,n-p;\gamma)$ distribution under the assumption that the x_{ij}, $j = 1,\ldots,n$ are independent and identically distributed normal variates. The F-distribution is central (γ=0) when X_{-i} includes an intercept. The quantity $(R_{ij}^{2} - 1)/(n-1)$ follows an $F(1,n-1;\gamma)$-distribution which in general is non-central when X_i is not centered. Thus the two factors in (9.2.37) can be calibrated against these distributions. A similar factorization of the multiple case (9.2.26) is possible.

9.2.7 Condition index

The condition indices of a matrix are one of the measures proposed in Chapter 2 to detect collinearity. It is important to determine whether an index is high because of collinearity-inducing points or low because of collinearty masking points. The effect on the condition indices when a single row is deleted were studied by Walker (1989) and Hadi (1988), and based on this knowledge Hadi (1988) proposed two measures to detect influential points, one of which was also proposed by Walker (1989).

9.2.7.1. Row deletion and condition indices

Let the condition indices of the matrix X_{-i} be denoted by

$$\eta_j(-i) \tag{9.2.38}$$

where the subscript j denotes the j-th condition index and (-i) is an indicator of the deleted row. If one wants to assess the influence of the i-th row, the condition indices of X_{-i} can be computed and compared to the condition indices of the full matrix. One drawback of this approach is that it requires the computation of the eigenvalues of (n+1) matrices each of order k×k. Hadi (1988) derives an approximation to $\eta_j(-i)$ without actually computing the eigenvalues of $X'_{-i}X_{-i}$, $i = 1,2,\ldots,n$. His development consist of two special cases (case one and two) and then the generalized case (case 3).

Case one: A theorem due to Kempthorne (1985), when X is an n×2 matrix, is summarized by (Hadi (1988), p146):

Theorem one: If X is n×2 and the columns of X are normalized to have length 1, then the square of the condition index of X_{-i} is

$$\eta_p(-i) = \frac{1 + (1-4/S_i)^{\frac{1}{2}}}{1 - (1-4/S_i)^{\frac{1}{2}}} \tag{9.2.39}$$

where

$$S_i = \frac{[(1-x_{i1}^2)+(1-x_{i2}^2)]^2}{(1-x_{i1}^2)(1-x_{i2}^2)-(\xi-x_{i1}x_{i2})^2} \tag{9.2.40}$$

and the i-th row of X is partitioned as

$$x_i' = [x_{i1} \; x_{i2}]$$

and

$$\xi = \sum_{i=1}^{n} x_{i1}x_{i2} \qquad (9.2.41)$$

Second case: The second special case, due to Dorsett (1982), involves the i-th row of X lying in the direction of the j-th eigenvector, and is summarized by (Hadi (1988), p147):

Theorem two: Let $v_1, v_2, \ldots, v_p$ be the set of orthonormal eigenvectors associated with the eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ of $X'X$. If $x_i = av_j$, then the (possibly unordered) eigenvalues of $X'_{-i}X_{-i}$ are

$$\lambda_1, \lambda_2, \ldots, \lambda_{j-1}, \lambda_j - a^2, \lambda_{j+1}, \ldots, \lambda_p \qquad (9.2.42)$$

Thus when x_i lies on the direction of the j-th eigenvector, the deletion of x_i will deflate the j-th eigenvalue. If the conditions of Theorem 2 hold, the following three conclusions can be made:

(i) If $j = 1$, then $[\eta_p(-i)]^2 = \max(\lambda_1 - a^2, \lambda_2)/\lambda_p$, and the deletion of x_i will decrease the condition index,

(ii) If $j = p$, then $[\eta_p(-i)]^2 = \lambda_1/(\lambda_p - a^2)$, and the deletion of x_i will increase the condition index, and

(iii) If $1 < j < p$, then $[\eta_p(-i)]^2 = \lambda_1/\min(\lambda_p; \lambda_j - a^2)$, and the deletion of x_i will have no effect on the condition index as long as $\lambda_j - \lambda_p > a^2$

A similar theorem is also given by Walker (1989) as Corollary 1 on p1681, but apparently has a slight error. The author assumed that $\max(\lambda_1 - a^2, \lambda_2) = (\lambda_1 - a^2)$.

Third case: Suppose there are no restrictions on X: for example, the restrictions of theorem one and two are lifted, X may or may not contain a constant column, or may or may not be normalized or standardized.

Let $v_1, v_2, \ldots, v_p$ be the set of orthonormal eigenvectors associated with the eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ of $X'X$. Let $\gamma_{i1} \geq \gamma_{i2} \geq \ldots \geq \gamma_{ip}$ be the eigenvalues of $X'_{-i}X_{-i}$. Define $z_{ij} = x_i' v_j$ and

$$c_i = \sum_{j=1}^{p} (z_{ij}/\lambda_j)^2$$

then

(i) if j = 1

$$\gamma_{i1} \simeq \frac{\lambda_1^2 - 2\lambda_1 z_{i1}^2 + z_{i1}^2 x_i' x_i}{\lambda_1 - z_{i1}^2} \qquad (9.2.43)$$

(ii) if 1 < j < p

$$\gamma_{ij} \simeq \lambda_j - z_{ij}^2 \qquad (9.2.44)$$

(iii) if j = p

$$\gamma_{ip} \simeq \min(A_j, A_p) \qquad (9.2.45)$$

where

$$A_j \simeq \min(\lambda_j - z_{ij}^2), \text{ for } j \neq p$$

and

$$A_p \simeq \frac{\lambda_p(1-h_{ii})[\lambda_p(1-h_{ii})+z_{ip}^2]}{\lambda_p(1-h_{ii})^2+2z_{ip}^2(1-h_{ii})+z_{ip}^2\lambda_p c_i}$$

(iv) the condition index of X_{-i} can be approximated by

$$[\eta_p(-i)]^2 \simeq \gamma_{i1}/\gamma_{ip} \qquad (9.2.46)$$

The results in (i) and (iii) are proven in Hadi (1988, Theorem five). In simulation studies the author finds that the approximations of the condition number in (iv) are close to the actual computed condition indixes. If the condition of theorem two holds, then (iv) holds exactly. Although the author could not prove (ii) theoretically, he uses an empirical investigation to suggest it is a good approximation.

9.2.7.2 Collinearity-influential points

To assess the influence of the i-th row on the condition index of X, Hadi (1988) suggested the following two measures:

Define

$$\delta_i = \frac{[\eta_p(-i)] - [\eta_p]}{[\eta_p]}, \quad i = 1,2,\dots,n \qquad (9.2.47)$$

where $[\eta_p(-i)]$ is defined in (9.2.46) and η_p is the ordinary condition index, defined in Chapter 2. Thus δ_i is an approximation to the relative distance between the condition index of X_{-i} and X. If $\delta_i >> 1$ then the i-th case is a row/point that masks collinearity, and if $\delta_i << 1$ then the i-th case is a point that induces collinearity.

Although Hadi does not suggest it, it is possible that the approximation of all the collinearity indices could be derived from the result in (9.2.44), and (9.2.46) be generalized from the p-th index to all the collinearity indices. Once δ_i is identified as unequal to one, and the other collinearity indices examined and their δ_i's computed, and say δ_{35} (3-rd case deleted, 5-th condition index) is relatively large, is the x_{35} point the outlier, or can one say anything about the 5-th variable? The above questions could be a field for further investigation.

The second measure proposed by Hadi (1988) to diagnose collinearity-influential points involves diagonal elements of the Hat matrix expressed a little differently. The Z matrix, whose elements are $z_{ij} = x_i'v_j$ can be expressed as Z = XV (this is the same Z as defined in Chapter 3). Thus the Hat matrix can be expressed as

$$\begin{aligned} H &= X(X'X)^{-1}X' \\ &= ZV'(VD^2V')^{-1}VZ \qquad \text{(by using the SVD of } X'X) \\ &= ZD^{-2}Z \\ &= WW' \qquad (9.2.48) \end{aligned}$$

where $W = ZD^{-1}$, then h_{ii} can be expressed as

$$h_{ii} = \sum_{j=1}^{p} z_{ij}^2/\lambda_j = \sum_{j=1}^{p} w_{ij}^2 \qquad (9.2.49)$$

The diagnostic measures δ_i and h_{ii} are supplemented by plots to give the analyst a comprehensive picture of the eigenstructure of the data. Collinearty-influential points can easily be seen on a graphical display of δ_i, such as stem and leaf displays. Pairwise scatterplots of the columns of Z, or W can be drawn, the quantities w_{ij}^2 are referred to as the leverage components (LC), and the scatterplots of W as the LC plots.

Walker (1989) proposed a diagnostic very similar to Hadi's second diagnostic (h_{ii}, (9.2.49)). By using the SVD of X, Walker (1989) expressed the Hat matrix as

$$\begin{aligned} H &= X(X'X)^{-1}X' \\ &= UDV'(VD^2V')^{-1}VDU' \\ &= UU' \qquad (9.2.50) \end{aligned}$$

which is the same as (9.2.48), as $W = ZD^{-1} = XVD^{-1} = UDV'VD^{-1} = U$. The diagonal element of H as

$$h_{ii} = \sum_{j=1}^{p} u_{ij}^2 \qquad (9.2.51)$$

Velleman and Ypeelar (1980), call the u_{ij}'s the orthogonal leverage components. Diagnostics proposed by Walker (1989) to detect collinearity influential points, involve the set of squared elements of the matrix U

$$u_{ij}^2 \quad i = 1,\ldots,n, \quad j = 1,\ldots,p \tag{9.2.52}$$

Walker (1989) suggests a cut-off value for high leverage is taken as p/n. Once 'high' u_{ij}^2's, say u_{kj}^2, are identified, he fits the model again with the k-th case deleted and compares the condition indices.

9.3 Influential Variables

Extending the notion of collinearity-influential points, Schall and Dunne (1987a) also define collinearity-influential variables.

Consider the following augmented model:

$$Y = [X \ A]\begin{bmatrix}\beta \\ \lambda\end{bmatrix} + \epsilon \tag{9.3.1}$$

where A:nxk is an arbitrary set of variables not necessarily dummy variables. Typically, A would consist of multiplicative interaction terms of the variables already in the model, or extra polynomial regression terms instead of the p already fitted, or the covariates in ANACOVA.

The decision whether to include the variables A is usually based on the F-statistic associated with the variables. Even if this F-test indicates non-significance, A should be included if it has an effect on the estimation of the parameters already in the model. Thus, when deciding whether a particular variable should be included in the model, Schall and Dunne (1987a) advocate the additional use of Cook's distance

$$C_A = \frac{(\hat{\beta} - \tilde{\beta})'X'X(\hat{\beta} - \tilde{\beta})}{\hat{\sigma}^2} \; \frac{n - p}{p} \tag{9.3.2}$$

where $\hat{\beta}$ is the OLSE of model (1.1) and $\tilde{\beta}$ is the LSE of β in the augmented model (9.3.1). If the statistic (9.3.2) indicates high influence of the variables A, it should be included in the model even if the F-test indicates non-significance. The variance of the estimates will increase, but large bias, due to the high influence of the variables, will be removed.

A second influence measure proposed by them is based on the AP-statistic, let

$$X_1^* = [X\ Y]$$

$$X_2^* = [X\ A^*\ Y] \quad \text{with } A^* = A(A'A)^{-\frac{1}{2}}$$

then the AP-statistic is defined as:

$$AP_A = |X_2^{*\prime}X_2^*|/|X_1^{*\prime}X_1^*| \tag{9.3.3}$$

Alternative and computational forms of (9.3.2) and (9.3.3) involve factorization as in (9.2.16) and (9.2.22), yielding

$$AP_A = (RSS - SSA)/RSS)\cdot\ |A(I-X(X'X)^{-1}X')A|/|A'A| \tag{9.3.4}$$

and

$$C_A = \frac{SSA}{RSS}\left[\frac{\tilde{\lambda}'A'A\tilde{\lambda}}{SSA} - 1\right]\frac{n-p}{p} \tag{9.3.5}$$

where SSA is the extra sum of squares due to fitting A after X. SSA is similar to the Q_k defined in (9.2.12), SSA = $\hat{\epsilon}'A\{A'[I - X(X'X)^{-1}X']A\}^{-1}A'\hat{\epsilon}$, and $\tilde{\lambda}$ is the LSE of λ under the model (9.3.1), $\tilde{\lambda} = \{A'[I - X(X'X)^{-1}X']A\}^{-1}A'\hat{\epsilon}$.

The first factor in (9.3.5) is a monotonic function of the F-statistic, as in (9.2.13), and is thus a measure of the statistical significance of the variables. The first factor of (9.3.4) is a decreasing function of the F-statistic. The second factors of (9.3.4) and (9.3.5) are measures of the potential influence of the variables A.

When k = 1 (A:nx1) the second factor of (9.3.5) simplifies to

$$\left[\frac{\tilde{\lambda}'A'A\tilde{\lambda}}{SSA} - 1\right] = \frac{\hat{\epsilon}'A\{A'[I-X(X'X)^{-1}X']A\}^{-1}A'A\{A'[I-X(X'X)^{-1}X']A\}^{-1}A'\hat{\epsilon}}{\hat{\epsilon}'A\{A'[I - X(X'X)^{-1}X']A\}^{-1}A'\hat{\epsilon}} - 1$$

$$= A'A/\{A'[I - X(X'X)^{-1}X']A\} - 1 \qquad (9.3.6)$$

The authors define the quantity (1 - $\{A'[I - X(X'X)^{-1}X']A\}/A'A$) as the leverage of the variable A, extending the notion of the leverage of an observation described in §1.7.

The VIF associated with the variable A in the model (9.3.1) is given by

$$VIF_A(X) = \frac{V[A\ X]}{V[A]}$$

$$= \frac{\begin{bmatrix} A'A & A'X \\ X'A & X'X \end{bmatrix}^{-1}_{11}}{[A'A]^{-1}_{11}}$$

If A:nx1, then

$$VIF_A(X) = \frac{[A'A]}{A'[I-X(X'X)^{-1}X']A} \qquad (9.3.7)$$

Note that the second factor of (9.3.4) and (9.3.5), and the leverage, are monotonic functions of the variance inflation factor. Thus, they are also monotonic functions of the collinearity index associated with the variable A. Cook's distance and the AP-statistic therefore contain collinearity measures.

The measure of the influence of the variable A on the variance inflation factor is given by

$$R_{i,(A)} = \frac{VIF_i(X_{-i}|A)}{VIF_i(X_{-i})}$$

$$= \frac{V[X_i \;\; X_{-i} \;\; A] \; V[X_i]}{V[X_i \;\; A] \;\; V[X]} \; \frac{V[X_i]}{V[X_i]}$$

$$= \frac{VIF_i[X_{-i} \;\; A]}{VIF_i[X_{-i}] \; VIF_i[A]} \qquad (9.3.8)$$

The retention of variables with high influence seems to contradict a common method of handling collinearity in a regression model, namely dropping variables from the model. The statistics (9.3.4) and (9.3.5) take into account the bias of dropping a variable as well as the influence it would have on the estimates. In the extreme case when A is orthogonal to X its 'dropping' will have no influence in the estimates. However when A is exactly or nearly collinear to some of the variables, its effect is confounded with a linear combination of the variables. Because collinearity is a group phenomenon (Stewart (1987)) it is dangerous to drop variables from the model. Using influence statistics, offers at least in some cases a method to 'regularize' a collinear design (Stewart (1987)). It should be empasised that the statistic (9.3.2) is a diagnostic and not a variable selection criterion.

To summarize the following guidelines of Schall and Dunne (1987a) are useful in applying Cook's distance for variables:

Statistic (9.3.2) has two roles: Firstly it should be used as a new type of collinearity diagnostic: it indicates not only the potential for harmful effects of collinearity, but is also quantifies the actual influence of a variable. By using the decomposition of (9.3.5) the computation of (9.3.2) requires no extra computations when F-tests and VIF's are computed.

Secondly the statistic (9.3.2) can also be used for model-checking purposes. A small value of (9.3.2) is reassuring and in the case of a large value, corrective action of some sort must be considered, like mixed regression techniques (Belsley *et al.* (1980, Chapter 4)).

9.4 Further Remarks on Influence

We finally present research fields that were not covered in §9.1 to §9.3.

9.4.1 Weighted Least Squares

An influential observation can be set artificially to zero by using weighted least squares. In weighted least squares the modified normal equation is

$$X'WX\tilde{\beta} = X'WY \qquad (9.4.1)$$

where $W = \mathrm{Diag}[1,,\ldots,1,w_i,1,\ldots 1]$. Then the i-th observation can be 'deleted' or 'downweighted' by making w_i arbitrary small. For reference to these techniques see Belsley *et al.* (1980), Cook and Weisberg (1982) and Cook (1986).

9.4.2 Influence measures in Ridge Regression

The notion of an influential observation in OLSE is extended to RRE by authors including Walker and Birch (1988), Lichtenstein and Velleman (1983) and Chalton (1990). Walker and Birch (1988) show that when RRE is used, the influence of each case changes is a function of the shrinkage parameter k. This change is mainly because of the behaviour of the residual as a function of the k value:

$$\begin{aligned}\hat{\epsilon}_R &= Y - X\hat{\beta}_R \\ &= Y - X(X'X + kI)^{-1}X'X\hat{\beta} \quad \text{(from 4.3.1)} \qquad (9.4.1)\end{aligned}$$

where $\hat{\epsilon}_R$ is the residual vector when the estimation is with the RRE.

Therefore when using RRE one could not rely on the influence measures obtained for OLSE (k = 0). The authors suggested that 'once the value of k is determined (by any method), influence measures should be computed for that k'. They then defined influence measures to use with RRE, by extending h_{ii}, DFFITS and Cook's distance to include RRE.

The concept of subset selection in OLSE is extended to biased estimators by Hoerl *et al.* (1986). Their results indicate that there is potential in using biased estimation to select subsets. They recommend that when applying RR, suspected superfluous variables may be deleted. However they warn that insignificance in an LS model does not necessarily imply that a variable is superfluous, particularly with collinear data.

9.5 Summary

This chapter introduced the concept of influential points and variables. Methods of determining collinearity-influential points were defined and the concept of influence in ridge regression was introduced.

Chapter 10

SIMULATION STUDY

10.1 Introduction

The purpose of this simulation study is to compare the performances of 13 different biased estimators, as well as OLSE on the simulated data discussed in §10.2. In §10.3 we discuss the program, giving a summary of the performance efficiencies of different estimators and tabulated results in §10.4, and some comments on the results in §10.5.

10.2 Data

The simulation study of this thesis follows that of McDonald and Galarneau (1975) and Wichern and Churchill (1978). The data sets were obtained from Chalton (1990) who generated them for a simulation study in his Ph.D thesis.

Chalton (1990), considers a five parameter model, with a sample size of 30 and the predictor variables generated from the following relationship:

For $j = 1,2,3$ and $i = 1,2,\ldots,30$

$$X_{ij} = (1 - \alpha_1^2)^{\frac{1}{2}} Z_{ij} + \alpha_1 Z_{i6} \qquad (10.2.1)$$

For $j = 4,5$ and $i = 1,2,\ldots,30$

$$X_{ij} = (1 - \alpha_2^2)^{\frac{1}{2}} Z_{ij} + \alpha_2 Z_{i6} \qquad (10.2.2)$$

where

(i) Z_{ij} are independent $N(0,1)$ variates generated by the SAS-function RANNOR. The seeds were not recorded by Chalton(1990), as the RANNOR function derives these from the time clock of the computer.

(ii) The parameters a_1 and a_2 determine the degree of collinearity between the predictor variables: a_1^2 is the theoretical correlation between any pair of the variables X_1, X_2 and X_3, a_1a_2 is the theoretical correlation between the variables X_1, X_2, X_3 and X_4 or X_5, and a_2^2 is the theoretical correlation between X_4 and X_5.

Five different combinations of a_1^2 and a_2^2 were considered, and two choices (orientations) of β, suggested by Newhouse and Oman (1971), namely the eigenvectors corresponding to the largest and smallest eigenvalues, denoted by β_L and β_S. For these 10 combinations three different values of σ where considered, namely 0.01, 1.0 and 5.0.

Chalton (1990) generated the Y-vector as sets of 30 data points from the model (10.2.3) for each of the (10×3=30) combinations of orientations, a_1^2 and a_2^2 values, and variance values.

For i = 1,2,...,30

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \epsilon_i \qquad (10.2.3)$$

where the X_{ij} are unstandardized, β_0 is zero, and the ϵ_i are independent $N(0,\sigma^2)$ variates. For the simulation study in this chapter, β_0=10 because we added 10 to the Y's generated by the author. For each combination of X, β and σ, 100 replications of the (30×1) ϵ-vector were generated

The eigenvalues and condition numbers corresponding to the five different combinations of a_1^2 and a_2^2 are shown in Table 10.1. and the coefficients of β_L and β_S are shown in Table 10.2.

10.3 Estimation Programs

The program to obtain the different estimates is given in appendix B. It was written in Fortran 5 and ran on a PC. Double precision was used throughout although we have found in trial runs that it did not make much of a difference. The X matrix was first standardized before any calculations were performed, then the SVD was computed. After any particular standardized estimate was obtained we transformed back to unstandardized parameter estimates before calculating the particular statistic of interest, as discussed later.

The SVD and the OLSE's were computed by using the subroutine SVDCMP and SVBKSB of Press *et al.* (1985). To obtain all the biased estimates, SVBKSB was modified for each particular estimation procedure.

One set of data (for one response vector Y) was run for comparative purposes by Dr. D. Chalton on an IBM4381 using SAS. The results for 12 of the 14 estimators were basically the same as those obtained with Fortran (agreeing to 6 and even 8 digits in estimating the 4 leading decimal digits of the β's. For the two estimators FGRPC1 and FGRPC2 (to be explained below), we found that the estimates differ from the third and sometimes the second digit.

To avoid dividing by zero in these last two estimation procedures (FGRPC1 and FGRPC2) the k_i's were flagged as soon as the delta's (δ's) became smaller than 10^{-10} and in the subroutine that calculates the estimators, the delta's was then set equal to zero.

The criterion of observed TMSE was used as a basis for comparing estimators, and the comparisons were effected by means of ratios.

The program in appendix B was tailored to find the summary statistics given in §10.4 and in the final runs only these statistics are printed. For those users wanting the estimates the obvious print lines should be added and the program should be modified to suit their needs.

10.4 Simulation Results

The definitions of the 14 estimators used are found in previous Chapters, especially the summary tables in Chapter 8. In Table 10.3 various abbreviations are explained. The FPC estimators require further clarification. In Chapter 5 the fractions for the FPCI and FPCV estimators (5.2.15) and (5.2.17) were described as iterative estimators. Because Lee and Birch (1988) observed that a 1-step version of both (5.2.15) and (5.2.17) exhibited already improved estimation properties over other biased estimators we considered only the following non-iterative FPC estimators:

FGRPC1: one step version of FPCI estimator, in (5.2.15) the $[\hat{\delta}_K(t)]_j^2$ is replaced by $[\hat{\delta}_{PC}]_j$, where the vector $\hat{\delta}$ consists of the $\hat{\delta}$'s of the PC1 estimator, and where s^2 is the estimate of σ_i^2 using the PC1 estimator. Basically then FGRPC1 is the GRRE, where k_i is estimated with the $\hat{\delta}$'s and s^2 of the PC1 estimators instead of the OLSE's as used by Hoerl, Kennard and Baldwin.

FGRPC2: the same as FGRPC1, but instead of using known PC1 estimates, we used the known PC2 estimates.

FRPC1 and FRPC2: similar to the ridge estimation procedure, but instead of using Hoerl, Kennard and Baldwin's k, we now estimate k and s^2 using the PC1 and PC2 estimators. These procedures are one step versions of (5.2.17).

The performance of each estimator over the thirty combinations and the 100 repetitions was summarized by computing

$$\sum_{j=1}^{5} \sum_{i=1}^{100} (\tilde{\beta}_{ji} - \beta_j)^2 \qquad (10.4.1)$$

where $\tilde{\beta}_{ji}$ is the j-th element of $\tilde{\beta}$, the estimate of β in the i-th repetition. The results of this simulation study are given in Tables 10.4, 10.5, and 10.6 for σ = 0.01, 1.0 and 5.0 respectively. The evaluations are based on the relative efficiency of each estimator compared to $\hat{\beta}$. Thus the tabulated relative efficiency values are

$$\sum_{j=1}^{5} \sum_{i=1}^{100} (\hat{\beta}_{ji} - \beta_j)^2 / \sum_{j=1}^{5} \sum_{i=1}^{100} (\tilde{\beta}_{ji} - \beta_j)^2 \qquad (10.4.2)$$

where $\tilde{\beta}$ is one of the estimators given in Table 10.3. Entries marked with a **, are very small values ($\leq$ 0.0007).

Table 10.1: Eigenvalues and condition numbers of X′X (X is standardized).

Correlations $a_1^2:a_2^2$	eigenvalues of X′X: λ_i (without β_0)	condition number λ_1/λ_5
.99:.99	(4.920,0.026,0.021,0.013,0.011)	435
.99:.10	(3.157,1.128,0.668,0.031,0.016)	201
.90:.90	(4.215,0.430,0.154,0.126,0.075)	56
.90:.10	(2.755,1.215,0.173,0.168,0.148)	19
.70:.30	(2.283,1.049,0.871,0.496,0.301)	8

Table 10.2: β used in generating Y

Correlations $a_1^2:a_2^2$	β'	Eigenvectors of X′X				
.99:.99	β_L'	[0.4474	0.4473	0.4481	0.4470	0.4463]
	β_S'	[0.2846	0.4760	-0.8302	0.0548	0.0163]
.99:.10	β_L'	[0.5534	0.5542	0.5510	0.1705	0.2323]
	β_S'	[-0.7755	0.1675	0.6084	0.0190	-0.0094]
.90:.90	β_L'	[0.4125	0.4547	0.4649	0.4383	0.4636]
	β_S'	[0.1821	-0.3973	0.5673	0.3037	-0.6284]
.90:.10	β_L'	[0.5634	0.5489	0.5644	0.2354	0.0862]
	β_S'	[-0.0162	0.6908	-0.7051	0.0655	0.1451]
.70:.30	β_L'	[0.5177	0.5611	0.5226	0.0286	0.3785]
	β_S'	[-0.6600	0.7049	0.0099	0.1956	-0.1708]

Table 10.3 Estimators and abbreviations

OLS	Ordinary least squares estimator
RHK	Ridge regression, k is estimated *via* Hoerl, Kennard and Baldwin
RLW	Ridge regression, k is estimated *via* Lawless and Wang
GRHK	Generalized ridge regression, K is estimated *via* Hoerl, Kennard and Baldwin
GRCAS	Generalized ridge regression, K is estimated *via* Troskie
AUGRR	almost unbiased generalized ridge regression
AUORR	almost unbiased operational ridge regression
PC1	principal component regression, delete the smallest singular value
PC2	principal component regression, delete the two smallest singular values
FGRPC1	one step version of FPCI estimator, where the OLSE is replaced by the PC1 estimator.
FGRPC2	one step version of FPCI estimator, where the OLSE is replaced by the PC2 estimator.
FRPC1	one step version of FPCV estimator, where the OLSE is replaced by the PC1 estimator.
FRPC2	one step version of FPCV estimator, where the OLSE is replaced by the PC2 estimator.
SHE	shrinkage estimator

Table 10.4 Relative efficiencies of biased estimators to OLSE (σ = 0.01)

Combinations of correlations and orientations of the β's										
$\alpha_1^2:\alpha_2^2$	99:99		99:10		90:90		90:10		70:30	
β	β_L	β_S	β_L	β_S	β_L	β_S	β_L	β_S	β_L	β_S
OLS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
RHK	1.00	0.99	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00
RLW	1.00	0.07	1.00	0.05	1.00	0.94	1.00	0.99	1.00	1.00
GRHK	1.35	1.33	1.37	1.03	1.05	0.99	0.84	0.93	0.97	0.96
GRCAS	1.32	1.25	1.33	1.05	1.04	1.01	0.89	0.97	0.97	1.00
AUGRR	1.11	1.14	1.10	1.00	1.04	0.98	0.93	0.95	1.00	0.96
AUORR	1.33	0.93	1.38	0.89	1.00	0.99	1.01	1.00	1.00	1.00
PC1	0.78	**	2.49	**	0.03	**	0.24	**	0.04	**
PC2	0.91	**	1.56	**	0.03	**	0.05	**	0.01	**
FGRPC1	0.96	**	2.41	**	0.03	**	0.24	**	0.04	**
FGRPC2	1.01	**	1.56	**	0.03	**	0.05	**	0.01	**
FRPC1	1.00	**	1.00	**	1.00	**	1.00	**	1.00	**
FRPC2	1.00	**	1.00	**	1.00	**	1.00	**	1.00	**
SHE	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 10.5 Relative efficiencies of biased estimators to OLSE (σ = 1.0)

Combinations of correlations and orientations of the β's										
$a_1^2:a_2^2$	99:99		99:10		90:90		90:10		70:30	
β	β_L	β_S	β_L	β_S	β_L	β_S	β_L	β_S	β_L	β_S
OLS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
RHK	3.43	2.88	4.52	3.08	2.49	1.45	2.13	1.14	1.47	0.86
RLW	6.45	7.65	10.82	5.20	1.88	1.30	1.73	0.91	1.28	0.65
GRHK	2.23	1.98	2.16	1.76	2.03	1.39	2.12	1.26	1.94	1.33
GRCAS	1.75	1.65	1.72	1.54	1.66	1.35	1.70	1.25	1.61	1.25
AUGRR	1.39	1.32	1.36	1.23	1.31	1.12	1.34	1.08	1.30	1.16
AUORR	4.26	3.21	6.31	3.61	3.75	1.47	3.22	1.13	1.98	0.76
PC1	1.59	1.41	2.90	1.79	1.79	0.67	1.78	0.47	1.48	0.24
PC2	2.85	2.24	49.55	4.67	2.95	0.83	5.15	0.57	2.32	0.25
FGRPC1	3.58	2.73	6.05	2.67	3.83	0.85	3.61	0.55	2.76	0.26
FGRPC2	6.66	3.86	93.90	4.90	6.08	0.96	9.34	0.61	3.95	0.27
FRPC1	5.05	4.21	10.36	4.26	3.46	1.24	2.55	0.79	1.57	0.37
FRPC2	9.07	5.53	40.05	5.07	4.15	1.18	3.15	0.68	1.64	0.33
SHE	2.52	2.48	2.02	2.01	1.60	1.56	1.41	1.23	1.20	1.04

Table 10.6 Relative efficiencies of biased estimators to OLSE (σ = 5.0)

Combinations of correlations and orientations of the β's										
$a_1^2:a_2^2$	99:99		99:10		90:90		90:10		70:30	
β	β_L	β_S	β_L	β_S	β_L	β_S	β_L	β_S	β_L	β_S
OLS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
RHK	3.40	3.81	4.49	5.59	3.64	3.66	3.35	3.57	2.59	2.43
RLW	360.75	162.19	79.16	61.83	12.68	14.59	6.61	7.16	2.99	3.05
GRHK	1.98	2.09	1.97	2.12	2.14	2.20	1.98	2.21	1.83	1.84
GRCAS	1.64	1.70	1.64	1.71	1.71	1.74	1.64	1.76	1.57	1.58
AUGRR	1.31	1.32	1.29	1.31	1.35	1.36	1.30	1.38	1.25	1.28
AUORR	4.14	4.56	5.95	7.40	4.21	4.09	3.91	3.94	2.95	2.62
PC1	1.65	1.58	2.45	3.13	1.77	1.66	1.64	1.65	1.45	1.18
PC2	3.41	3.19	50.07	35.37	3.32	3.30	5.18	3.75	2.50	1.97
FGRPC1	3.41	3.38	4.47	6.40	3.83	3.56	3.12	3.34	2.59	2.00
FGRPC2	7.46	7.39	85.81	56.71	7.18	6.91	9.03	6.24	4.28	3.10
FRPC1	6.05	6.48	9.93	14.43	6.50	6.13	5.02	5.37	3.57	3.02
FRPC2	14.81	16.22	128.40	75.83	12.37	10.80	12.17	8.86	5.33	4.36
SHE	2.70	2.99	2.39	2.65	2.87	2.99	2.57	2.94	2.40	2.48

10.5 Discussion of the results

Features of Tables 10.4 - 10.6 are:

1. From Table 10.4 we note that the OLSE performs satisfactorily when the collinearity is modest (< 56). As the collinearity increases a slight improvement over OLS was obtained by GRHK, GRCAS, AUGRR and AUORR (except for the second orientation of β). The PC1, PC2, FGRPC1, FGRPC2, FRPC1 and FRPC2 perform disastrously for the 2-nd orientation of the β's (the β's corresponding to the smallest eigenvalue of $X'X$). Only for the 99:10 combination did PC1, PC2, FGRPC1, FGRPC2, FRPC1 and FRPC2 perform clearly better than the OLS.
2. As σ increases the relative efficiency of estimators increased with respect to the OLSE except for combinations 90:90, 90:10 and 70:30, under orientation 2, in Table 10.5, where we again picked up the disastrous pattern of Table 10.4.
3. We observed the same pattern that Lee and Birch (1988) noted: The efficiencies of FPC estimators are proportional to the efficiencies of the PC1 and PC2 estimates. Furthermore when $\sigma \geq 1.0$ the fractional generalized ridge estimators (FGRPC1 and FGRPC2) outperformed the generalized ridge estimators. The fractional ridge estimators (FRPC1 and FRPC2) were also better relative to OLSE than the RHK estimators but for the RLW estimators a 'see-saw' situation is noted.
4. When $\sigma = 5$ some of the biased estimators (e.g. RLW and FRPC2) appears to be exceptionally good. Further investigation with respect to the different orientations and combinations may explain this phenomenon.
5. The AUORR estimator generally performs better than RHK for $\sigma = 1.0$ and 5.0. Contradicting Nomura (1988) we found that AUORR is not better than RLW for $\sigma = 1$. When $\sigma = 5$ RLW outperforms AUORR relative to OLS. RLW was always better (except for 3 orientations) than RHK relative to OLS.
6. We found that the AUGRR was always inferior to the GRRE (both GRHK and GRCAS). This phenomenon was also found by Ohtani (1986).
7. The shrinkage estimators were always similar to or relatively better than OLSE, but rather inferior to other biased estimators.

8. In conclusion then: among the biased estimators, no single estimator appears to dominate the others, in terms of TMSE, across all conditions.

10.6 Recommendations and further research

According to Belsley *et al.* (1980) the above collinearity indices ($\sqrt{435}$, $\sqrt{201}$,...,$\sqrt{8}$), are generally weak with perhaps one being moderate. For further research we recommend that at least one higher collinearity index (say near 100) should be included. Though the results obtained in this simulation suggest the biased estimators are more efficient than OLS, we seek to compare such estimators under a variety of combinations.

The extremely small values in Table 10.4 and 10.5 should be explored as well as the extremely high values in Table 10.6. Furthermore fixing the ϵ_i vector across various σ's, the sampling distributions of β_i for each σ, sampling distributions of $\sum_{j=1}^{5} (\hat{\beta}_j - \beta_j)^2$ for each σ and the estimated covariance of $\hat{\beta}_i$ should be investigated.

In this simulation study we do not report β_0, as β_0 is simply estimated as the mean of the Y vector in the case of the biased estimators, following a practice used elsewhere in the literature.

10.7 Summary

This chapter consisted of a simulation study and its general findings. The data, the program and the results were given and discussed, and areas for further research were described. Generally it appears that no biased class of estimators consistently outperforms other classes on the criterion of relative efficency. Specific conditions seem to be associated with the optimality of specific estimators, but it is not yet possible to define those conditions.

REFERENCES

ALLEN D.M.(1974): The relationship between variable selection and data augmentation and a method for prediction. *Technometrics,* 16, no. 1, February 1974, 125-127.

ANDREWS D.F. AND PREGIBON D.(1978): Finding the outliers that Matter. *Journal of the Royal Statistical Society*, *Series* B, 40, No. 1, 85-93.

BEATON A.E. RUBIN D.B. AND BARONE J.L.(1976): The acceptability of regression solutions: Another look at computational accuracy. *Journal of the American Statistical Association*, 71, 158-168.

BELSLEY D.A.(1982): Assessing the presence of harmful collinearity and other forms of weak data through a test for signal-to-noise. *Journal of Econometrics*, 20, 211-253.

BELSLEY D.A.(1984): Demeaning conditioning diagnostics through centering, and Reply. *The American Statistician*, 38, 73-77 and 90-93.

BELSLEY D.A.(1986): Centering, the constant, first-differencing, and assessing collinearity, in: D.A. Belsley and E. Kuh (Eds.), *Model Reliability* (MIT Press. Cambridge, MA, 1986).

BELSLEY D.A.(1987): Well-conditioned collinearity indices (comment on a paper by G.W. Stewart). *Statistical Science*, 2, 86-91.

BELSLEY D.A., KUH E. AND WELSCH R.E.(1980): *Regression diagnostics: identifying influential data and sources of collinearity.* John Wiley & Sons, New York.

BELSLEY D.A. AND OLDFORD R.W.(1986): The general problem of ill conditioning and its role in statistical analysis. *Computational Statistics and data analysis*, 4, 103-120.

CHALTON D.O.(1990): Contributions to influence, outliers and Bayesian analysis in the multiple linear regression model. Ph.D. Thesis, University of Cape Town.

CHAMBERS J.M.(1977): *Computational Methods for Data Analysis.* John Wiley & Sons, New York.

CHENG D.C. AND IGLAISH H.J.(1976): Principal component estimators in regression analysis. *Review of Economics and Statistics*, 58, 229-234.

COOK R.D.(1977): Detection of influential observations in linear regression. *Technometrics*, 19, No. 1, February 1977, 15-18.

COOK R.D.(1986): Assessment of local influence (with discussion). *Journal of the Royal Statistical Society, Series* B - *Methodological,* 48, 133-169.

COOK R.D. AND WEISBERG S.(1982): *Residuals and influence in regression.* Chapman and Hall, New York.

CROCKER D.C.(1971): Letter to the Editor. *The American Statistician*, 25, no. 3, 55.

DANIEL C. AND WOOD F.S.(1980): *Fitting equations to data,* 2nd ed. John Wiley & Sons, New York.

DORSETT D.(1982): Resistant M-Estimators in the presence of influential points. Ph.D. Dissertation. Dept. of Statistics, Southern Methodist University.

DRAPER N.R. AND JOHN J.A.(1981): Influential observations and outliers in regression. *Technometrics*, 23, No. 1, February 1981, 21-26.

DWIVEDI T.P., SRIVASTAVA V.K. AND HALL R.L.(1980): Finite sample properties of ridge estimators. *Technometrics*, 22, 205-212.

ECKART G. AND YOUNG G.(1936): The approximation of one matrix by another of lower rank. *Psychometrika*, 1, 211-218.

FARRAR D.E. AND GLAUBER R.R.(1967): Multicollinearity in regression analysis: The problem revisited. *Review of Economics and Statistics*, 49, 92-107.

FRISCH R.(1934): *Statistical Confluence Analysis by Means of Complete Regression Systems.* Oslo: Universitetets Okonomiske Institutt, Oslo, Norway.

GALPIN J.S.(1978): An investigation of methods of ridge regression. *Technical Report*, CSIR, Pretoria.

GENTLEMAN J.F. AND WILK M.B.(1975): Detecting outliers II. Supplementing the direct analysis of residuals. *Biometrics,* 31, 387-410.

GOLUB G.H., HOFFMAN A. AND STEWART G.W.(1987): A Generalization of the Eckart-Young-Mirsky matrix approximation theorem. *Linear Algebra and its Applications,* 88/89: 317-327.

GOLUB G.H., KLEMA V. AND STEWART G.W.(1976): Rank degeneracy and least squares problems. *Technical Report* TR-751, Dept. Computer Science, Univ. Maryland.

GOLUB G.H. AND VAN LOAN C.F.(1980): An analysis of the total least squares problem. *SIAM Journal on Numerical Analysis*, 17, 883-893.

GOLUB G.H. AND VAN LOAN C.F.(1983): *Matrix Computations.* Johns Hopkins University Press. Baltimore, MD, 1983.

GOODNIGHT J. AND WALLACE T.D.(1972): Operational techniques and tables for making weak MSE tests for restrictions in regression. *Econometrica*, 40, 699-709.

GRAYBILL F.A.(1969): *Introduction to Matrices with Applications in Statistics*. Wadsworth Publishing Company, Belmont, CA.

GRAYBILL F.A.(1976): *Theory and Applications of the Linear Model.* Duxbury press, Belmont, California.

GUNST R.F.(1983): Regression analysis with multicollinear predictor variables: definition, detection, and effects. *Communications in Statistics, Part A - Theory Methods,* 12, no. 19, 2217-2260.

GUNST R.F. AND MASON R.L.(1980): *Regression analysis and its application. A data-oriented approach.* Statistics: Textbooks and Monographs, 34. Marcel Dekker, New York.

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.

HADI A.S.(1988): Diagnosing collinearity-influential observations. *Computational Statistics & Data Analysis* , 7, 143-159

HAWKINS D.M.(1973): On the investigation of alternative regressions by Principal Component Analysis. *Applied Statistics*, 22, 275-286.

HILL R.C., FOMBY T.B. AND JOHNSON S.R.(1977): Component selection norms for principal components regression. *Communications in Statistics, Part A - Theory and Methods,* 6, 309-334.

HINKLEY D.V.(1977): Jackknifing in unbalanced situations. *Technometrics*, 19, No. 3, 285-292.

HOAGLIN D.C. AND WELSCH R.E.(1978): The Hat matrix in regression and ANOVA. *The American Statistician*, 32, 17-22.

HOERL A.E. AND KENNARD R.W. (1970a): Ridge regression: Applications to nonorthogonal problems. *Technometrics* 12, 69-82.

HOERL A.E. AND KENNARD R.W. (1970b): Ridge regression: Baised estimation for nonorthogonal problems. *Technometrics* 12, 55-69.

HOERL A.E., KENNARD R.W. AND BALDWIN K.F.(1975): Ridge regression: some simulations. *Communications in Statistics*, 4, 105-123.

HOERL R.W., SCHUENEMEYER J.H. AND HOERL A.E.(1986): A simulation of biased estimation and subset selection regression techniques. *Technometrics*, 28, 369-380.

HORN R.A. AND JOHNSON C.R.(1987): *Matrix Analysis.* Cambridge University Press, Cambridge.

HSIANG T.C.(1976): A Bayesian view on ridge regression. *The Statistician*, 24, 267-268.

JAMES W. AND STEIN C.(1961): Estimation with quadratic loss, in Neyman J.(ed.). *Proceedings of the Fourth Berkeley Symposium*, Los Angeles: University of California Press, 1961, 361-379.

JOHNSON T. AND WALLACE T.D.(1969): Principal Components and Multicollinearity. Department of Economics, Econometrics Workshop Discussion Paper, North Carolina State University, Raleigh, North Carolina.

JOHNSTON J.(1963): *Econometric Methods.* McGraw-Hill, New York.

JOLLIFFE I.T.(1972): Discarding variables in a principal component analysis. I. Artificial data. *Applied Statistics*, 21, 160-173.

JOLLIFFE I.T.(1973): Discarding variables in a principal component analysis. II. Real data. *Applied Statistics*, 22, 21-31.

JOLLIFFE I.T.(1982): A Note on the use of principal components in Regression. *Applied Statistics*, 31, 300-303.

KADIYALY K.(1984): A class of almost unbiased and efficient estimators of regression coefficients. *Economics Letters,* 16, 293-296.

KALMAN R.E.(1984): We can do something about multicollinearity! *Communications in Statistics, Part A - Theory and Methods*, 13, no. 2, 115-125.

KASHYAP A.K., SWAMY P.A.V.B., MEHTA J.S., AND PORTER R.D.(1984): Estimating distributed lag relationships using near-minimax procedures. *Special Studies Paper*, Federal Reserve Board, Washington, D.C.

KEMPTHORNE P.J.(1985): Assessing the influence of single cases on the condition number of a design matrix. Memorandum NS-509, Department of Statistics, Harvard University.

KENDALL M.G.(1957): *A Course in Multivariate Analysis.* Griffin, London.

KETELLAPPER R.H.(1983): On estimating parameters in a simple Linear Errors-in-Variables Model. *Technometrics*, 25, no. 1, 43-47.

KUMAR T.K.(1975): Multicollinearity in regression analysis. *Review of Economics and Statistics*, 57, 365-366.

LAWLESS J.F. AND WANG P.(1976): A simulation study of ridge and other regression estimators. *Communications in Statistics*, 5, 307-323.

LAWSON C.L. AND HANSON R.J.(1974): *Solving Least-Squares problems.* Prentice-Hall, Inc., Englewood Cliffs, N.J.

LEE W.W.(1986): Fractional principal components regression: a general approach to biased estimators. Unpublished Ph.d. dissertation. Dept. of Statistics, Virginia Polytechnic Institute and State University.

LEE W.W. AND BIRCH J.B.(1988): Fractional principal components regression: A general approach to biased estimators. *Communications in Statistics*, *Part B - Simulation and Computation,* 17, 713-727.

LESAGE J.P. AND SIMON S.D.(1988): Centering and scaling of regression algorithms in the face of ill-conditioning. *Journal of Statistical Computation and Simulation,* 30, 273-283

LICHTENSTEIN C.H. AND VELLEMAN P.F.(1983): The effects of ridge regression on high leverage points in the data. Unpublished manuscript.

LINDLEY D.V. AND SMITH A.F.M.(1972): Bayes estimates for the linear model. (With discussion). *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.

LOTT W.F.(1973): Optimal set of principal component restrictions on a least squares regression. *Communications in Statistics*, 2, 449-464.

MALLOWS C.L.(1964): Choosing variables in a linear regression: A Graphical aid, presented at the Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, Kansas, May 7-9, 1964.

MARQUARDT D.W.(1970): Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation. *Technometrics*, 12, 591-612.

MARQUARDT D.W.(1980): You should standardize the predictor variables in your regression models (discussion of a paper by G. Smith and F. Campbell). *Journal of the American Statistical Association,* 75, 87-91.

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

MASON R.L. AND GUNST R.F.(1985): Outlier-induced collinearities. *Technometrics* 27, 401-407.

MASON R.L. GUNST R.F. AND WEBSTER J.T.(1975): Regression analysis and problems of multicollinearity. *Communications in Statistics*, 4, 277-292.

MASSY W.F.(1965): Principal component 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. AND GALARNEAU D.I.(1975): A Monte-Carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association*, 70, 407-416.

MIRSKY L.(1960): Symmetric gauge functions and unitarily invariant norms. *Quarterly Journal of Mathematics*, 11, 50-59.

MULLET G.M.(1976): Why regression coefficients have the wrong sign. *Journal of Quality Technology*, 8, 121-126

NETER J. AND WASSERMAN W.(1974): *Applied Linear Statistical Models.* Richard D. Irwin, Inc. Ontario.

NEWHOUSE J.P. AND OMAN S.D.(1971): An evaluation of Ridge estimators. *Technical report* No. R-716-PR, The Rand Corporation, Santa Monica, Calif.

NOMURA M.(1988): On the almost unbiased ridge regression estimator. *Communications in Statistics, Series B - Simulation and Computation*, 17, 729-743.

OHTANI K.(1986): On small sample properties of the almost unbiased generalized ridge estimator. *Communication in Statistics, Part A - Theory and Methods*, 15, 1571-1578.

O'HAGEN J. AND McCABE B.(1975): Test for the severity of multicollinearity in regression analysis: A comment. *Review of Economics and Statistics*, 57,368-370.

PRESS W.H., FLANNERY B.P., TEUKOLSKY S.A. AND VETTERLING W.T. (1985): *Numerical Recipes: The Art of Scientific Computing.* Cambridge University Press.

QUENOULLE M.H.(1956): Notes on bias in estimation. *Biometrika*, 43, 353-360.

RANDALL J.H. AND RAYNER A.A.(1987): The accuracy of least squares calculations with the Cholesky algorithm. *Technical report*, University of Natal.

RAWLINGS J.(1988): *Applied regression analysis: a research tool.* Wadsworth & Brooks/Cole: Pacific Grove, California.

SASTRY M.V.R.(1970): Some limits in the theory of multicollinearity. *The American Statistician,* 24, 39-40.

SCHALL R. AND DUNNE T.T.(1987a): Influential variables in linear regression. (to appear in *Technometrics*, 1990).

SCHALL R. AND DUNNE T.T.(1987b): Variance inflation and collinearity in regression. *Technical Report* 5/87, Institute for Biostatistics of the South African Medical Research Council, Tygerberg, Republic of South Africa.

SCHNEEWEISS H.(1976): Consistent estimation of a regression with errors in variables. *Metrica,* Band 23, 101-115.

SCLOVE S.L.(1968): Improved estimators for coefficients in linear regression. *Journal of the American Statistical Association*, 63, 596-606.

SEARLE S.R.(1971): *Linear Models.* John Wiley & Sons, New York.

SILVEY S.D.(1969): Multicollinearity and imprecise estimation. *Journal of the Royal Statistical Society*, *Series* B, 31, 539-552.

SIMON S.D. AND LESAGE J.P.(1988): The impact of collinearity involving the intercept term on the numerical accuracy of regression. *Computer Science in Economics and Management* 1, 137-152.

SINGH B., CHAUBEY Y.P. AND DWIVEDI T.D.(1986): An almost unbaised ridge estimator. *Sankhya, Series B,* 48, 342-346.

SMITH G. AND CAMPBELL F.(1980): A critique of some ridge regression methods. *Journal of the American Statistical Association*, 75, no. 369, 74-103.

STEWART G.W.(1973): *Introduction to matrix computations.* Academic Press, New York.

STEWART G.W.(1977): On the perturbation of pseudo-inverses, projections, and linear least squares problems. *SIAM Review*,. 19, 634-666.

STEWART G.W.(1987): Collinearity and least squares regression. With discussion by D.A. Belsley, A.S. Hadi, D.W. Marquardt, P.F. Velleman, R.A. Thisted, and with a reply by the author. *Statistical Science.* 2, no. 1, 68-100.

SWAMY P.A.V.B., MEHTA J.S., THURMAN S.S. AND IYENGAR N. S.(1985): A generalized multicollinearity index for regression analysis. *Sankhya, Series* B, 47, no. 3, 401-431.

THISTED R.A.(1980): Comment on a paper by Smith G. and Campbell F. *Journal of the American Statistical Association*, 75, 81-86.

THURMAN S.S., SWAMY P.A.V.B. AND MEHTA J.S.(1984): An examination of distributed lag model coefficients estimated with smoothness priors. *Special Studies Paper*, Federal Reserve Board, Washington. DC.

TROSKIE C.G.(1990): Personal communication.

TUKEY J.W.(1958): Bias and confidence in not quite large samples. *The Annals of Mathematical Statistics,* 29, 614.

VAN HUFFEL S.(1990): Personal communication.

VAN HUFFEL S. AND VANDEWALLE J.(1985): The Use and Applicability of the Total Least Squares Technique in Linear Regression Analysis, Internal Report, Esat Lab., Dept. of Electrical Engineering, K.U. Leuven Belgium, 1985.

VAN HUFFEL S. AND VANDEWALLE J.(1987): Algebraic relationships between classical regression and total least-squares estimation. *Linear Algebra and its Applications,* 93, 149-160.

VAN HUFFEL S., VANDEWALLE J. AND STAAR J.(1984): The total linear least squares problem: properties, applications and generalization. Submitted to *SIAM Journal on Numererical Analysis.*

VELLEMAN P.F. AND YPELAAR M.A.(1980): Constructing regressions with controlled features: a method of probing regression performance. *Journal of the American Statistical Association*, 75, no. 372, 839-844.

VINOD H.D. AND ULLAH A.(1981): *Recent advances in regression Methods.* Marcell Dekker Inc., New York.

WALKER E.(1989): Detection of collinearity-influential observations. *Communications in Statistics, Part A - Theory and Methods,* 18, 1675-1690.

WALKER E. AND BIRCH J.B.(1988): Influence measures in ridge regression. *Technometrics,* 30, 221-227.

WALLACE T.D.(1972): Weaker criteria and tests for linear restrictions in regression. *Ecomometrica,* 40, 689-698.

WAMPLER R.H.(1970): A Report on the accuracy of some widely used least squares computer programs. *Journal of the American Statistcal Association,* 65, 549-565.

WAMPLER R.H.(1980): Test procedures and problems for least-squares algorithms. *Journal of Econometrics*, 12, 3-22.

WEBSTER J.T., GUNST R.F. AND MASON R.L.(1974): Latent root regression analysis. *Technometrics*, 16, 513-522.

WEDIN P.A.(1969): On pseudo-inverses of perturbed matrices, Lund Univ. Comput. Sci. Tech. Rep., Lund, Sweden.

WEDIN P.A.(1973): Perturbation theory for pseudo-inverses. *BIT*. 13, 217-232.

WETHERILL G.B., DUNCOMBE P., KENWARD M., KOLLERSTROM, J., PAUL S.R. AND VOWDEN B.J.(1986): *Regression analysis with applications*. Chapmann & Hall, London-New York.

WILKINSON J.H.(1965): *The Algebraic Eigenvalue Problem.* Oxford University Press, London.

WU C.F.J.(1986): Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, Vol. 14, No. 4, 1261-1295.

YANCEY T.A., JUDGE G.G. AND BOCK M.E.(1973): Wallace's weak mean square error criterion for testing linear restrictions in regression: A Tighter Bound. *Econometrica,* 41, 1203-1206.

Appendix A

USEFUL FORMULAE AND DERIVATIONS

A.1 If $x \sim N(\mu,V)$ then for A symmetric and conformable

$$E(x'Ax) = tr(AV) + \mu'A\mu$$

$$V(x'Ax) = 2tr(AV)^2 + 4\mu'AVA\mu$$

(Searle (1971, pp55-57))

A.2 Let A be an nxn matrix with eigenvalues $\lambda_1,\lambda_2,\ldots\lambda_n$, then

$$tr(A) = \Sigma\ \lambda_i .$$

If A is symmetric and $\lambda_i > 0 \quad \forall i$

then $tr\ (A^-) = \Sigma\ \lambda_i^{-1}$

(Graybill (1969, pp223-225))

A.3 Let A be an nxn symmetric matrix with eigenvalues

$$\alpha_1 \geq \alpha_2 \cdots \geq \alpha_n$$

Let k be an integer, $1 \leq k \leq n$. Let B be the (n-1) x (n-1) symmetric matrix obtained by deleting the k-th row and column from A. Then the ordered eigenvalues β_i of B interlace with those of A as follows:

$$\alpha_1 \geq \beta_1 \geq \alpha_2 \geq \beta_2 \geq \cdots \geq \beta_{n-1} \geq \alpha_n$$

(Lawson and Hanson (1974, p24) and the proof can be found in Wilkinson (1965, pp99-109)).

A.4 Let $A = [a_1 \ldots a_p]$ be a column partitioning of A:mxp, and denote the i-th singular value of A by $\sigma_i(A)$. If $A_r = [a_1 \ldots a_r]$ then for $r = 1,\ldots,p-1$

$$\sigma_1(A_{r+1}) \geq \sigma_1(A_r) \geq \sigma_2(A_{r+1}) \geq \ldots \geq \sigma_r(A_{r+1}) \geq \sigma_r(A_r) \geq \sigma_r(A_{r+1})$$

Equivalently, adding a column to a matrix increases the largest singular value and diminishes the smallest (Golub and Van Loan (1983, p286).

Appendix B

PROGRAM

The program code is given in this appendix. Some of the comments and documentation were added using T^3 to relate the programming to the development of the theory in Chapters 1 to 10. The program-code is printed in bold and the comments unbold. Various abbreviations for the estimators were summarized in Table 10.3 of Chapter 10. It is suggested that §10.3 should be read with this program.

```
***************************************************************************
*                             MAIN PROGRAM                                *
*                                                                         *
*  AIM OF THIS PROGRAM                                                    *
*                                                                         *
*  This program computes the 14 estimators for the simulation study of    *
*  Chapter 10.  The estimation of the TMSE for 5 regression coefficients  *
*  is computed for each estimator and summed over 100 replicate samples.  *
*                                                                         *
*  DESCRIPTION OF THE VARIABLES                                           *
*                                                                         *
*  Most variable names are self-explanatory.  Comments are added after    *
*  the declaration statements, where necessary or useful.                 *
*                                                                         *
***************************************************************************

C  Declaration of variables

      parameter (maxr=30,maxc=7)   physical row/column dimension
      character*14  FNAME   input data file
      character*14  OUTFNA   output data file
      double precision XSIM(30,105)  simulated data matrix
      double precision X(maxr,maxc)   unstd X, without a column of ones
```

double precision **TEMPX(maxr,maxc)** unstd X, with a column of ones

double precision **STDX(maxr,maxc)** std X matrix (correlation form)

double precision **Y(maxr)** Y-vector

double precision **U(maxr,maxc), W(maxc), V(maxc,maxc)** SVD std X

double precision **UU(maxr,maxc), UW(maxc), UV(maxc,maxc)** SVD unstd X

double precision **TW(maxc)** λ_i's, smallest set equal to 0 for PC

double precision **DELTA(maxc)** $\hat{\delta}$

double precision **B(maxc)** std $\hat{\beta}$

double precision **BU(maxc)** unstd $\hat{\beta}$

double precision **AVE(maxc)** vector - means of columns of X

double precision **SS(maxc)** vector - sums of squares of columns of X

double precision **C(maxr)** vector send to SVDSORT

double precision **BB(maxc)** Vector send to SVDSORT

double precision **RES(maxr)**

double precision **KPC1(maxc),KPC2(maxc)** $\hat{K}$, *via* PC1/PC2

double precision **KHK(maxc)** $\hat{K}$ *via* HKB

double precision **KCAS(maxc)** $\hat{K}$ *via* Troskie

double precision **sse, ssepc1, ssepc2** $\hat{\sigma}$ *via* OLS/PC1/PC2

double precision **TEMPB(maxc)**

double precision **faclw** $\hat{k}$ *via* Lawless and Wang

double precision **fachk** $\hat{k}$ *via* Hoerl, Kennard and Baldwin

double precision **hkpc2, hkpc1** $\hat{k}$ *via* PC2/PC1

double precision **nomura** $\hat{k}$ *via* Nomura

double precision **sinf** flag, marking infinity

double precision **meany** mean of Y

double precision **F** F used in Troskie's method

double precision **shrink** shrinkage factor

double precision **sum**

double precision **stemp** $\hat{\beta}'\hat{\beta}$

double precision **tmse** trace of variance of OLSE

double precision **tbeta(maxc)** β

```
      double precision sols(maxc), sauorr(maxc)
      double precision srhk(maxc), srlw(maxc)
      double precision sgrhk(maxc), sgrcas(maxc)
      double precision saugrr(maxc), ssh(maxc)
      double precision spc1(maxc), spc2(maxc)
      double precision sfgrp1(maxc), sfgrp2(maxc)
      double precision sfrpc1(maxc), sfrpc2(maxc)
```

$$\sum_{j=1}^{100}(\tilde{\beta}_{ij}-\beta_i)^2$$

for i = 1,...,5

```
      double precision tols, trhk
      double precision rlw, tgrhk
      double precision tgrcas, taugrr
      double precision tpc1, tpc2
      double precision tfgrp1, tfgrp2
      double precision tfrpc1, tfrpc2
      double precision tauorr, tssh
```

$$\sum_{i=1}^{5}\sum_{j=1}^{100}(\tilde{\beta}_{ij}-\beta_i)^2$$

```
      real tsigma     σ = 0.1 or 1.0 or 5.0
      integer N,P,K,1,M,R1,C1,R2,C2,R,rep

C  Read in file names

      write (*,*) 'Enter filename as name.dat, 1-14 characters'
      read (*,85) FNAME
85    format (A14)
      write (*,*) 'Enter output  filename as name.out, 1-14 characters'
      read (*,85) OUTFNA

C  Set logical dimension of matrix, and define value near infinity
C  Note:  change here for new data sets, or change program to read in
C  these dimensions.

      N = 30
      P = 6
      sinf = 1.0d-10
```

```
C  Open files, output file exist, change here if either is a new file

        open (15, file=FNAME, status='OLD')
        open (16, file=OUTFNA, status='OLD')

C  Read true sigma and betas, and write to output file

       write(16,85) OUTFNA
       read (15,*) tsigma
       write(16,*) 'This are the results of  ', FNAME
       write(16,*) 'sigma =', tsigma
       read(15,*) (tbeta(i),i = 2,6)
       write(16,*) 'True betas'
       write(16,*) (tbeta(i),i = 2,6)
       tbeta(1) = dble(10.0)

C  Set trace sums for the 14 estimators equal to zero

       do 10 i = 1,p
           sols(i) = dble(0.0)
           srhk(i) = dble(0.0)
           srlw(i) = dble(0.0)
           sgrhk(i) = dble(0.0)
           sgrcas(i) = dble(0.0)
           saugrr(i) = dble(0.0)
           sauorr(i) = dble(0.0)
           spc1(i) = dble(0.0)
           spc2(i) = dble(0.0)
           sfgrp1(i) = dble(0.0)
           sfgrp2(i) = dble(0.0)
           sfrpc1(i) = dble(0.0)
           sfrpc2(i) = dble(0.0)
           ssh(i) = dble(0.0)
10     continue
```

```
C  Read in simulated data matrix

      do 90 i = 1,N
          read(15,*) (XSIM(i,l), l = 1,105)
90    continue

C  Add a column of ones to the TEMPX matrix and copy the other columns to X

      do 92 i = 1,N
          TEMPX(i,1) = dble(1.0)
          do 91 j = 2,P
               TEMPX(i,j) = XSIM(i,(100 + j -1))
               X(i,j-1) = TEMPX(i,j)
91        continue
92    continue

c  Make a copy of TEMPX into UU

      CALL COPY (TEMPX, UU, maxr, maxc, N, P)

c   Compute the SVD, unsorted

      CALL SVDCMP(UU, N, P, maxr, maxc, UW, UV)
      write(16,*) 'SVDCMP COMPUTED UNSTANDARDIZED MATRIX'

C  Sort the SVD

      CALL SVDSORT(UU, UW, UV, N, P, maxr, maxc, C, BB)

C  Standardize the matrix, and  get back STDX

      IP = P
      P = P - 1
      CALL STAND(X, maxr, maxc, N, P, STDX, AVE, SS)
```

```
C  Copy STDX into U

         CALL COPY(STDX, U, maxr, maxc, N, P)

C  Decompose matrix U

         CALL SVDCMP(U, N, P, maxr, maxc, W, V)

C  Sort the SVD

         CALL SVDSORT(U, W, V, N, P, maxr, maxc, C, BB)
         write(16,*) ' singular values of std X'
         write (16,*) (W(j), j=1,P)

C  Split the Y and X part in XSIM

C  SVD is computed, pick up all the Y's (100 Y-vectors), and calculate the
C  corresponding estimators, and efficiency totals

        do 500 rep = 1,100
             meany = dble(0.0)          * calculate the mean of Y
              do 94 i = 1,N
                   Y(i) = XSIM(i,rep) + dble(10.0)
                   meany = meany + Y(i)
94           continue
             meany = meany/dble(N)

C  Compute the OLS estimator, using the SVD of the unstandardized X matrix

             CALL SVBKSB(UU, UW, UV, N, IP, maxr, maxc, Y, B)
             do 100 i = 1,IP
                  sols(i) = sols(i) + (B(i) - tbeta(i))**2
100          continue
```

```
C  Compute the estimate of sigma (OLS), unstandardized X

         CALL SIGMA(TEMPX, maxr, maxc, N, IP, Y, B, RES, sse)

C  Compute shrinkage estimator from unstandardized data using the d in
C  equation (5.1.10)

         tmse = DBLE(0.0)
         stemp = DBLE(0.0)

         do 110 i = 2,IP
              stemp = stemp + B(i)**2
              tmse = tmse + DBLE(1)/(UV(i)**2)
110      continue
         shrink = stemp/((sse * tmse) + stemp)
         do 125 i = 1,IP
              ssh(i) = ssh(i) + ((B(i) * shrink) - tbeta(i))**2
125      continue

C  Compute the OLS estimator - from standardized X matrix

         CALL SVBKSB(U, W, V, N, P, maxr, maxc, Y, B)

C  Compute the DELTA'S - OLS

         CALL VTBETA(V, maxc, B, P, DELTA)

C  Compute the K and k values for RRE, GRRE, and the others

         nomura = DBLE(0.0)
         do 200  i=1,p
              KHK(i) = sse/(delta(i)**2)
              F = ((w(i)**2) * (DElTA(i)**2))/sse
              KCAS(i) = (w(i)**2)/(DBLE(1) + F)
              sum = DSQRT(DBLE(1) + ((W(I)**2)*(DELTA(I)**2))/sse)
              nomura = nomura + (DELTA(I)**2)/(DBLE(1) + sum)
```

```
200    continue
          nomura = DBLE(P) * sse/nomura

C  Compute the Hoerl, Kennard and Baldwin  and the Lawless and Wang constant

          fachk = DBLE(0.0)
          faclw = DBLE(0.0)
          do 220  i = 1,P
              fachk = fachk + B(i) ** 2
              faclw = faclw + (DELTA(i)**2) * (V(i)**2)
220       continue
          fachk = (DBLE(P) * sse)/fachk
          faclw = (DBLE(P) * sse)/faclw

C  Compute the PC estimators, by setting the smallest singular value equal
C  to 0 for PC1

          do 250 i = 1,4
              TW(i) = W(i)
250       continue

          TW(5) = DBLE(0.0)  *  set smallest singular value to zero

C  Compute PC1 estimator

          CALL SVBKSB(U, TW, V, N, P, maxr, maxc, Y, B)

C  Compute the estimate of sigma via PC1

          CALL SIGMA(STDX, maxr, maxc, N, P, Y, B, RES, ssepc1)

          ssepc1 = DBLE(n-p) * ssepc1 - DBLE(N) * (meany**2)  correct for Ȳ
          ssepc1 = ssepc1/(DBLE(N-P-1))
```

```
C  Compute the ridge constant, k, using PC1 estimator

          hkpc1 = DBLE(0.0)
          do 260 i = 1,P
               hkpc1 = hkpc1 + B(i)**2
260       continue
          hkpc1 = DBLE(p) * ssepc1/hkpc1

C  Unstandardize the PCE's and compute efficiency

          CALL BETA(SS, AVE, B, BU, maxc, P, meany)
          do 262 i = 1,(P+1)    here $\hat{\beta}_0-\beta$ was also computed, but not used
               spc1(i) = spc1(i) + (BU(i) - tbeta(i))**2
262       continue

C  Compute the deltas (V'b) to obtain the constants for fractional
C  rank estimators  $\hat{\delta} = V'\hat{\beta}_{pc}$

          CALL VTBETA(V, maxc, B, P, DELTA)

C  Compute vector K, using PC1 estimator

          do  280 i = 1,P
               if ((DELTA(i)**2).ge.sinf) then
                    KPC1(i) = ssepc1/(DELTA(i)**2)
                    else
                    KPC1(i) = DBLE(9999.9)
               endif
280       continue

C  Set the second smallest eigenvalue equal to zero

          TV(4) = DBLE(0.0)
```

```
C  Compute the PC2 estimator

      CALL SVBKSB(U, TV, V, N, P, maxr, maxc, Y, B)

C  Compute the estimate of sigma using PC2 estimator

      CALL SIGMA(STDX, maxr, maxc, N, P, Y, B, RES, ssepc2)
      ssepc2 = DBLE(N-P)*ssepc2 - DBLE(N) * (meany**2)
      ssepc2 = ssepc2/(DBLE(N-P-1))

C  Compute k, using PC2

      hkpc2 = DBLE(0.0)
      do 290 i = 1,P
         hkpc2 = hkpc2 + B(i)**2
290   continue
      hkpc2 = DBLE(P) * ssepc2/hkpc2

C  Compute the V'b for PC2

      CALL VTBETA(V, maxc, B, P, DELTA)

C  Compute the vector K using PC2

      do 330 i = 1,P
         if ((DELTA(i)**2).ge.sinf) then
            KPC2(i) = ssepc2/(DELTA(i)**2)
            else
            KPC2(i) = DBLE(9999.9)
         endif
330   continue
```

```
C  Unstandardize the PCE's and compute efficiency

          CALL BETA(SS, AVE, B, BU, maxc, P, meany)
          do 335 i = 1,(P+1)
               spc2(i) = spc2(i) + (BU(i) - tbeta(i))**2
335       continue

C  Compute the ridge estimators - using k of Hoerl, Kennard and Baldwin

          CALL RIDGE(U, W, V, N, P, maxr, maxc, Y, B, fachk)

C  Unstandardize the RHKB estimators and compute efficiency

          CALL BETA(SS,AVE,B,BU,MAXC,P,meany)
          do 345 i = 1,(P+1)
               srhk(i) = srhk(i) + (BU(i) - tbeta(i))**2
345    continue

C  Compute the ridge estimators - using k of Lawless and Wang

          CALL RIDGE(U, W, V, N, P, maxr, maxc, Y, B, faclw)

C  Unstandardize the ridge estimators and compute efficiency

          CALL BETA(SS, AVE, B, BU, maxc, P, meany)
          do 355 i = 1,(P+1)
               srlw(i) = srlw(i) + (BU(i) - tbeta(i))**2
355       continue

C  Compute the GRRE via Hoerl, Kennard and Baldwin

          CALL GRRE(U, W, V, N, P, maxr, maxc, Y, B, Khk)
```

```
C   Unstandardize GRRE and compute efficiency

          CALL BETA(SS, AVE, B, BU,maxc, P, meany)
          do 365 i = 1,(P+1)
               sgrhk(i) = sgrhk(i) + (BU(i) - tbeta(i))**2
365       continue

C  Compute the GRRE via Troskie

          CALL GRRE(U, W, V, N, P, maxr, maxc, Y, B, Kcas)

C  Unstandardize GRRE and compute efficiency

      CALL BETA(SS, AVE, B, BU, maxc, P, meany)
          do 375 i = 1,(P+1)
               sgrcas(i) = sgrcas(i) + (BU(i) - tbeta(i))**2
375       continue

C  Compute the AUORR estimator, via nomura

          CALL AUORR(U, W, V, N, P, maxr, maxc, Y, B, nomura)

C  Unstandardize AUORR and compute efficiency

          CALL BETA(SS, AVE, B, BU, maxc, P, meany)
          do 385 i = 1,(P+1)
               sauorr(i) = sauorr(i) + (BU(i) - tbeta(i))**2
385       continue

C  Compute the AUGRR estimator where estimate of K is via Hoerl, Kennard and
C  Baldwin

          CALL AUGRR(U, W, V, N, P, maxr, maxc, Y, B, Khk)
```

```
C  Unstandardize AUGRR and compute efficiency

          CALL BETA(SS, AVE, B, BU, maxc, P, meany)
          do 395 i = 1,(P+1)
               saugrr(i) = saugrr(i) + (BU(i) - tbeta(i))**2
395       continue

C  Compute FPC estimator, using GRR, where K is estimated by PC1

          CALL GRRE(U, W, V, N, P, maxr, maxc, Y, B, Kpc1)

C  Unstandardize the FPCI estimator, and compute efficiency

          CALL BETA(SS, AVE, B, BU, maxc, P, meany)
          do 400 i = 1,(P+1)
               sfgrp1(i) = sfgrp1(i) + (BU(i) - tbeta(i))**2
400       continue

C  Compute the FPC estimator, using GRR, where K is estimated by PC2

          CALL GRRE(U, W, V, N, P, maxr, maxc, Y, B, Kpc2)

C  Unstandardize the FPCI estimator, and compute efficiency

          CALL BETA(SS, AVE, B, BU, maxc, P, meany)
          do 410 i = 1,(P+1)
               sfgrp2(i) = sfgrp2(i) + (BU(i) - tbeta(i))**2
410       continue

C  Compute the FPC estimator, using RR, where k is estimated using PC1

          CALL RIDGE(U, W, V, N, P, maxr, maxc, Y, B, hkpc1)
```

```
C  Unstandardize the FPCV estimator, and compute efficiency

          CALL BETA(SS, AVE, B, BU, maxc, P, meany)
          do 420 i = 1,(P+1)
               sfrpc1(i) = sfrpc1(i) + (BU(i) - tbeta(i))**2
420       continue

C  Compute the FPC estimator, using RR, where k is estimated using PC2

          CALL RIDGE(U, W, V, N, P, maxr, maxc, Y, B, hkpc2)

C  Unstandardize the FPCV estimator, and compute efficiency

          CALL BETA(SS, AVE, B, BU, maxc, P, meany)
          do 430 i = 1,(P+1)
               sfrpc2(i) = sfrpc2(i) + (BU(i) - tbeta(i))**2
430       continue
500   continue

C  Set sum of traces (over 5 regression coefficients) for the relevant
C  estimators equal to zero

      tols = DBLE(0.0)
      trhk = DBLE(0.0)
      trlw = DBLE(0.0)
      tgrhk = DBLE(0.0)
      tgrcas = DBLE(0.0)
      taugrr = DBLE(0.0)
      tauorr = DBLE(0.0)
      tpc1 = DBLE(0.0)
      tpc2 = DBLE(0.0)
      tfgrp1 = DBLE(0.0)
      tfgrp2 = DBLE(0.0)
      tfrpc1 = DBLE(0.0)
      tfrpc2 = DBLE(0.0)
      tssh = DBLE(0.0)
```

```
C  Compute the efficiency totals for the different biased estimators

      do 550 j = 2,6
          tols = tols + sols(j)
          trhk = trhk + srhk(j)
          trlw = trlw + srlw(j)
          tgrhk = tgrhk + sgrhk(j)
          tgrcas = tgrcas + sgrcas(j)
          taugrr = taugrr + saugrr(j)
          tauorr = tauorr + sauorr(j)
          tpc1 = tpc1 + spc1(j)
          tpc2 = tpc2 + spc2(j)
          tfgrp1 = tfgrp1 + sfgrp1(j)
          tfgrp2 = tfgrp2 + sfgrp2(j)
          tfrpc1 = tfrpc1 + sfrpc1(j)
          tfrpc2 = tfrpc2 + sfrpc2(j)
          tssh = tssh + ssh(j)
550   continue

C Write the efficiency and the relative efficiency ratio to output file

      write(16,555) tols, tols/tols
555   format (' OLS', T10, 2F15.8)
      write(16,565) trhk, tols/trhk
565   format (' RHK', T10, 2F15.8)
      write(16,575) trlw, tols/trlw
575   format (' RLW', T10, 2F15.8)
      write(16,585) tgrhk, tols/tgrhk
585   format ('GRHK', T10, 2F15.8)
      write(16,595) tgrcas, tols/tgrcas
595   format ('GRCAS', T10, 2F15.8)
      write(16,605) taugrr, tols/taugrr
605   format ('AUGRR', T10, 2F15.8)
      write(16,615) tauorr, tols/tauorr
615   format ('AUORR', T10, 2F15.8)
      write(16,625) tpc1, tols/tpc1
```

```
625    format ('PC1',T10, 2F15.8)
       write(16,635) tpc2, tols/tpc2
635    format ('PC2', T10, 2F15.8)
       write(16,645) tfgrp1, tols/tfgrp1
645    format ('FGRPC1', T10, 2F15.8)
       write(16,655) tfgrp2, tols/tfgrp2
655    format ('FGRPC2', T10, 2F15.8)
       write(16,665) tfrpc1, tols/tfrpc1
665    format ('FRPC1', T10, 2F15.8)
       write(16,675) tfrpc2, tols/tfrpc2
675    format ('FRPC2', T10, 2F15.8)
       write(16,685) tssh, tols/tssh
685    format ('SHE', T10, 2F15.8)

       close (16)
       CLOSE (15)
       END     *   End of main program   *
```

```
      SUBROUTINE AUGRR (U, W, V, rows, col, maxr, maxc, Y, BETA, K)

C  This subroutine computes the AUGRR estimator and is the same as SVBKSB
C  except for a change in the root part.  The comments under SVBKSB are also
C  relevant here.  The vector K is input value.

      parameter (nmax = 100)
      double precision U(maxr,maxc), W(maxc), V(maxc,maxc)
      double precision Y(maxr), BETA(maxc)
      double precision TMP(nmax), S
      double precision K(maxc)
      integer rows, col

      do 12 j = 1,col
         S = DBLE(0.0)
         if (W(j).NE.DBLE(0.)) then
            do 11 i = 1,rows
                S = S + U(i,j) * Y(i)
11          continue      *  note change to SVBKSB is done here*
            S = S * (W(j)*(DBLE(1.0)+K(j)/(W(j)**2+K(j)))/(W(j)**2+K(j)))
         endif
         TMP(j) = S
12    continue
      do 14 J = 1,col
         S = DBLE(0.0)
         do 13 jj = 1,col
             S = S + V(j,jj) * TMP(jj)
13       continue
      BETA(j) = S
14    continue
      return
      end
      SUBROUTINE AUORR (U, W, V, rows, col, maxr, maxc, Y, BETA, rc)

C  This subroutine computes the AUORR estimator and is the same as SVBKSB
C  except for a change in the root parts (in SVBKSB we change the S part).
C  The comments under SVBKSB are also relevant here.  The ridge constant
C  is denoted by rc.

      parameter (nmax = 100)
      double precision U(maxr, maxc), W(maxc), V(maxc,maxc)
      double precision Y(maxr), BETA(maxc)
      double precision TMP(nmax), S
      double precision rc
      integer rows, col

      do 12 j = 1,col
          S = DBLE(0.0)
          if (W(j).NE.dble(0.0)) then
               do 11 i = 1,rows
                    S = S + U(i,j) * Y(i)
11             continue
                         *changes to SVBKSB are done here*
               S = S *(DBLE(1.0) - rc**2/((W(j)**2 + rc)**2))/W(j)
          endif
      TMP(j) = S
12    continue
      do 14 j = 1,col
          S = DBLE(0.0)
          do 13 jj = 1,col
               S = S + V(j,jj) * TMP(jj)
13        continue
          BETA(j) = S
14    continue
      return
      end
      SUBROUTINE BETA (S, AVE, B, BU, maxc, row, meany)

C  Aim:  Unstandardize the betas
C  Input to this subroutine:
C         S - vector containing the sum of squares of columns of X
C         AVE - vector containing the means of columns of X,
C         meany - mean of Y
C         B - vector containing the standardized betas
C         maxc -  physical dimensions of S, AVE, B, BU
C         row - logical dimension of S, AVE, B, BU
C  Return to main program   BU - vector, of unstandardized betas

      double precision S(maxc), AVE(maxc), B(maxc), BU(maxc)
      double precision meany
      integer row

      BU(1) = DBLE(0.0)
      do 100 i = 1,row
          BU(i+1) = B(i) * DBLE(1)/(DSQRT(S(I)))
          BU(1) = BU(1) + BU(i+1) * AVE(I)
100   continue
      BU(1) = meany - BU(1)
      return
      end
      SUBROUTINE COPY (A, B, maxr, maxc, rA, cA)

C  Copies A(rA,cA) into B and loses previous B

      double precision A(maxr,maxc)
      double precision B(maxr,maxc)
      integer rA,cA

      do 100 i = 1,rA
          do 20 j = 1,cA
               B(i,j) = A(i,j)
20        continue
100   continue
      return
      end
      SUBROUTINE GRRE(U, W, V, rows, col, maxr, maxc, Y, BETA, K)

C  This subroutine computes the generalized ridge regression estimator and
C  is the same as SVBKSB  except for adding $k_i$ to the diagonal elements of
C  W when we divide.  The comments under SVBKSB are also relevant here.  K
C  is the ridge vector.  If (k(i).ge.1x10**10), then DELTA (V'B) is set to 0

      parameter (nmax = 100)
      double precision U(maxr,maxc), W(maxc), V(maxc,maxc)
      double precision Y(maxr), BETA(maxc)
      double precision TMP(nmax), S, K(maxc)
      integer rows, col
      do 12 j = 1,col
          S = DBLE(0.0)
          If (W(j).NE.DBLE(0.0)) then
               do 11 i=1,rows
                    S = S + U(i,j) * Y(i)
11             continue
               S = S * (W(j)/(W(j)**2 + K(j)))  *changes made here*
          endif
          TMP(j) = S
12   continue
      do 13 j = 1,col
          if (K(j).eq.DBLE(9999.9)) then
               TMP(j) = DBLE(0.0)
          endif
13    continue
      do 15 J = 1,col
          S = DBLE(0.0)
          do 14 jj = 1,col
          S = S + V(j,jj) * TMP(jj)
14    continue
      BETA(j) = S
15    continue
      return
      end
      SUBROUTINE RIDGE(U, W, V, ROWS, COL, MAXR, MAXC, Y, BETA, rc)

C  This subroutine computes the ridge estimator and is the same as SVBKSB
C  except for adding k to the diagonal elements of W when we divide.
C  The comments under SVBKSB are also relevant here.  The ridge constant
C  is denoted by rc.

      parameter (nmax = 100)
      double precision U(maxr,maxc), W(maxc), V(maxc,maxc)
      double precision Y(maxr), BETA(maxc)
      double precision TMP(nmax), S, rc
      integer rows, col

      do 12 j = 1,col
          S = DBLE(0.0)
          if (W(j).NE.DBLE(0.)) then
               do 11 i = 1,rows
                    S = S + U(i,j) * Y(i)
11             continue
               S = S * (W(j)/(W(j)**2 + rc))    *changes made here*
          endif
          TMP(j) = S
12    continue
      do 14 j = 1,col
          S = DBLE(0.0)
          do 13 jj = 1,col
               S = S + V(j,jj) * TMP(jj)
13         continue
          BETA(j) = S
14    continue
      return
      end
      SUBROUTINE SIGMA (MAT, maxr, maxc, N, P, Y, B, RES, sse)

C  This subroutine computes an estimate of sigma square, sse.
C  MAT is the X matrix, Y is the Y vector, and B is the vector of
C  regression coefficients.  The physical dimensions are maxr and maxc and
C  the logical dimensions are N and P.

      double precision MAT(maxr,maxc), Y(maxr)
      double precision RES(maxr), sse
      double precision B(maxc), temp
      integer N,P,i,j

      sse = DBLE(0.0)
      do 100 i = 1,N
          temp = DBLE(0.0)
          do 15 j = 1,P
               temp = temp + MAT(i,j) * B(j)
15        continue
          RES(I) = Y(I) - temp
          sse = sse + (RES(I)**2)
100   continue
      sse = sse/DBLE(N-P)
      return
      end
      SUBROUTINE STAND (MAT, maxr, maxc, N, P, STDX, AVE, SS)

C  This subroutine standardizes MAT (physical dimensions maxr, maxc:
C  logical dimensions N and P) by subtracting the mean from each column and
C  then divides the column by its standard deviation.
C  Note the matrix MAT should not contain a column of ones

      double precision MAT(maxr,maxc), AVE(maxc)
      double precision SS(maxc), STDX(maxr,maxc)
      double precision temp
      integer N, P, j, i

      do 6000 j = 1,P
          AVE(j) = DBLE(0.0)
          do 605 i = 1,N
               AVE(j) = AVE(j) + MAT(i,j)
605       continue
          AVE(j) = AVE(j)/DBLE(N)
6000  continue
      do 6010 j = 1,P
          temp = DBLE(0.0)
               do 610 i = 1,N
                    temp = temp + (MAT(i,j) - AVE(j))**2
610            continue
          SS(j) = temp
6010  continue
      do 7000 i = 1,N
          do 6900 j = 1,P
               STDX(i,j) = (MAT(i,j) - AVE(j))/DSQRT(SS(j))
6900      continue
7000  continue
      return
      end
      SUBROUTINE SVBKSB (U, W, V, M, N, MP, NP, B, X)

C  The computation of the estimators is based on this subroutine.  The
C  program-code is from Press et al. (1985).  The only change made to
C  SVBKSB was a switch to double precision.
C
C  AIM of SVBKSB:  Solves AX = B  for a vector X, where A is specified by
C  the arrays U, W, V as returned by SVDCMP.  M and N are the logical
C  dimensions of A.  MP and NP are the physical dimensions of A.  B is the
C  input right-hand side.  X is the output solution vector.  No input
C  quantities are destroyed, so the routine may be called sequentially with
C  different B's

      parameter (nmac = 100)  Maximum anticipated value of N
      double precision U(MP,NP), W(NP), V(NP,NP)
      double precision B(MP), X(NP), TMP(nmax), S
      do 12 j = 1,N    Calculate U'B
         S = DBLE(0.)
         if (W(j).NE.DBLE(0.)) then   Nonzero result only if $W_j \neq 0$
            do 11 i = 1,M
               S = S + U(i,j) * B(i)
11          continue
            S = S/W(j)  division by $W_j$, changes made here for other
                        methods
         endif
         TMP(j) = S
12    continue
      do 14 j = 1,N   Matrix multiply by V
         S = DBLE(0.)
         do 13 jj = 1,N
            S = S + V(j,jj) * TMP(jj)
13    continue
      X(j) = S
14    continue
      return
      end
      SUBROUTINE SVDCMP (A, M, N, MP, NP, W, V)

C  This routine computes the SVD of the matrix A.  The program-code is form
C  Press et al. (1985)  and is not given here.  The only change made to
C  SVDCMP was a switch to double precision.  Remember to make a copy
C  of A  before calling SVDCMP as U replaces A on output.  Furthermore
C  the matrix V is output as V and not the transpose of V.
```

```
      SUBROUTINE SVDSORT(U, W, V, N, M, NP, MP, C, BB)

C  This subroutine sorts the SVD from high to low and is borrowed from
C  Troskie (1990).

      double precision U(NP,MP), W(MP), V(MP,MP)
      double precision C(NP), BB(MP)
      do 90 k = 1,M-1
           do 100 j = 1,M-K
               if (W(j).LT.W(j+1)) then
                    HOLD = W(j)
                    do 110 l = 1,N
                         C(l) = U(l,j)
110                 continue
                    do 120 l = 1,M
                         BB(l) = V(l,j)
120                 continue
                    W(j) = W(j+1)
                    do  130 l = 1,N
                         U(l,j) = U(l,j+1)
130                 continue
                    do 140 l = 1,M
                         V(l,j) = V(l,j+1)
140                 continue
                    W(j+1) = HOLD
                    do 150 l = 1,N
                         U(l,j+1) = C(l)
150                 continue
                    do 160 l = 1,M
                         V(l,j+1) = BB(l)
160                 continue
               endif
100        continue
90    continue
      return
      end
      SUBROUTINE VTBETA(V, maxc, B, P, DELTA)

C  This subroutine computes the delta's, V'B, Here V, B are input matrices
C  and the DELTAS are returned.  Maxc is the physical dimensions, and P the
C  logical dimensions of the matrices.

      double precision V(maxc,maxc), B(maxc), DELTA(maxc)
      double precision sum
      integer maxc, P

      do 180 i = 1,P
          sum = DBLE(0.0)
          do 175 j = 1,P
               sum = sum + V(j,i) * B(j)
175       continue
          DELTA(i) = sum
180   continue
      return
      end
```

Appendix C

NOTATION

This dissertation is written in such a way that as the material develops, the relevant notation is described. Description in context is intended to obviate any confusion. However this summary of notation is also supplied for those readers interested only in particular chapters. The first part of the notation under A, represents some matrix operations on the matrix A.

$A:nxp$ - matrix of order nxp

a - scalar or column-vector

A' - transpose of A

a_i - i-th element of a when a is a vector

A_i - i-th column of A when A is a matrix,

a_{ij} - i-th row and j-th column element of the matrix A

$[A]_{ij}$ - i-th row and j-th column element of the matrix A

A^- - (Moore-Penrose) generalized inverse of A

A^{-1} - inverse of A

$|a|$ - absolute value of a

$|A|$ - determinant of A

$\lambda_i(A'A)$ - i-th eigenvalue of A'A

$||a||$ - vector norm of a

$||a||_2$ - Euclidean norm (or vector 2-norm) of a vector a

$||A||$ - matrix norm of A

$||A||_F$ - Frobenius norm of a matrix A

$||A||_p$ - matrix p-norm of A

N(A) - null space of A

tr[A] - trace of A

r(A) - rank of A

R(A) - range of A

$a' = [a_1,\dots,a_{p+1}]$, where a_i is defined in Chapter 6

$a(0) = \hat{\delta}_{PC}$

a(t) - estimate of δ at the t-th iteration

acc_j - numerical accuracy of the j-th regression coefficient

AP - Andrews-Pregibon

AP^k_{ij} - Andrews-Pregibon statistic for k suspected outliers, cases i,j,...

AP_A - Andrews-Pregibon statistic for variables A

ANACOVA - analysis of covariance

ANOVA - analysis of variance

AUGRRE - almost unbiased generalized ridge regression estimator

AUORRE - almost unbiased operational ridge regression estimator

β:px1 - vector of regression coefficients that must be estimated

$\beta = [\beta_a' \ \beta_b']'$

$\beta_a : p_a x1$

$\beta_b : p_b x1$

$\beta_b^* : p_b x1$ - any p_b vector

β_i - i-th element of β

β_I - sub-vector of β, where I $\subset$ $\{1,...,p\}$, indicates the index set specifying the subset

β_0 - intercept term

$\tilde{\beta}$ - any estimator of β, usually biased

$\tilde{\beta}$ - LSE of β in the augmented model

$\hat{\beta}$ - OLSE of β

$\delta\hat{\beta}$ - perturbation vector of $\hat{\beta}$

$\hat{\beta}_a$ - OLSE of β_a

$\hat{\beta}_b$ - OLSE of β_b

$\hat{\beta}_c$ - constrained LSE of β

β_{basic} - unique solution of smallest norm

$\hat{\beta}_{\text{CLS}}$ - corrected least square estimator

$\hat{\beta}_{-i}$ - OLSE of β with the i-th case (i-th row) deleted

$\hat{\beta}_{-I}$ - OLSE of β with the rows indexed by I deleted, $I \subset \{1,2,\ldots,n\}$

$\beta_{(i)}$ - subset of β that corresponds to the $X_{(i)}$ part of X

$\hat{\beta}_{\text{FPC}}$ - fractional principal component estimator of β

$\hat{\beta}_{G}$ - posterior mean of β

$\hat{\beta}_{J}$ - jackknife estimator of β

$\hat{\beta}_{JW}$ - weighted jackknife estimator of β

$\hat{\beta}_{K}$ - generalized ridge regression estimator of β

$\hat{\beta}_{LR}$ - latent root estimator of β

$\hat{\beta}_{PC}$ - principal component estimator of β

$\hat{\beta}_{R}$ - ridge regression estimator of β

$\hat{\beta}_{SH}$ - shrunken estimator of β

$\hat{\beta}_{SSH}$ - Sclove's modified shrunken estimator of β

$\hat{\beta}_{TLS}$ - total least square estimator of β

$\hat{\beta}_{\hat{r}}$ - estimate of β, obtained when $X_{\hat{r}}$ is used instead of X

$\hat{\beta}_s$ - biased estimator of β

$(\hat{\beta}_s)_{-h}$ - biased estimator of the truncated model

$(\beta|Y)$ - β given Y, Bayesian posterior random variable

b - bias vector

b_1, b_2 - two competing estimators

b - estimator of β in Chapter 8

b_{ji} - j-th element of b, in the i-th Monte Carlo repetition

$b_{j,LS}$ - j-th element of the OLSE

$\tilde{b}$ - OLSE for β, where the X matrix is perturbed ($\tilde{X} = X + \delta X$)

$\tilde{b}$ - value of a coefficient as computed free from round-off errors

b_i - OLSE when X_i is regressed on the remaining columns of X

$B_1 = X_1(X_1'X_1)^{-1/2}$

BLUE - best linear unbaised estimator

$c_i = u_i'Y\sqrt{\lambda_i}$ element in describing the estimators of β

$c' = [c_1, c_2, \ldots c_p]$, vector of constants

c - number of collinearities

$c_i = \sum_{j=1}^{p} (z_{ij}/\lambda_j)^2$

C - Cook's squared distance

C_i - Cook's squared distance after i-th case deleted

C_{ij}.. - Cook's squared distance when cases i,j,... are deleted

C_A - Cook's squared distance when variables A are deleted

C_p - Mallows' C_p

CLS - corrected least squares

Cov(•,•) - covariance between two random variables

d - deterministically (fixed) or stochasically defined constant

Δ:nxp - diagonal matrix of the singular values of X

$\Delta = [\Delta_1\ \Delta_2]$, Δ_1:(p-r)x(p-r)

Δ_e:nx(p+1) - error matrix, in the errors-in-variables model

$[\Delta_e]_i$ - i-th row of the error matrix

$\Delta_{\hat{r}}$:nx$\hat{r}$ - diagonal matrix of the singular values of $X_{\hat{r}}$

df - degrees of freedom

Diag(•) - diagonal matrix of elements of a vector argument

diag(•) - vector of diagonal elements of a matrix argument

D^2 - Cook's squared distance

$D_a = \mathrm{Diag}(\sqrt{\lambda_1},\sqrt{\lambda_2}\ldots\sqrt{\lambda_k})$ - the non-zero singular values of X

$\hat{D}_a$ - diagonal matrix consisting of the singular values of $\hat{X}$

$\tilde{\Delta}$ - diagonal matrix of the singular values of the augmented matrix [X Y]

$\delta = V'\beta$

δ_i - i-th element of δ

δ_{max} - largest element of δ

$\delta = [\delta_1' \ \delta_2']'$, δ_1:rx1 and δ_2:(p-r)x1

$\hat{\delta}$ - OLSE of δ

$\hat{\delta}_J$ - jackknife ridge estimator of δ

$\hat{\delta}_{JW}$ - weighted jackknife ridge estimator of δ

$[\hat{\delta}_{JW}]_i$ - i-th element of $\hat{\delta}_{JW}$

$\hat{\delta}_K$ - GRRE δ, K is non-stochastic

$\hat{\delta}_{\hat{K}}$ - GRRE of δ, K is stochastic

$[\hat{\delta}_{\hat{K}}]_i$ or $(\hat{\delta}_{\hat{K}})_i$ i-th element of $\hat{\delta}_{\hat{K}}$

$[\hat{\delta}_{\hat{0}}]_i$ - i-th element of operational AUGRRE

$[\hat{\delta}_K(t)]_i$ - GRRE of δ_i at t-th iteration

$[\hat{\delta}_K(0)]_i = \hat{\delta}_i$

$\hat{\delta}_R$ - ridge estimator of δ

$\hat{\delta}_R(t)$ - ridge estimate of δ at t-th iteration

$\hat{\delta}_R(0) = \hat{\delta}$

$\hat{\delta}_{FPC}$ - fractional principal component estimator of δ

$\hat{\delta}_{FPCI} = F_{PCI}\hat{\delta}$

$\hat{\delta}_{FPCV} = F_{PCV}\hat{\delta}$

∂_i - shrinkage factor

$\delta\bullet$ - perturbation in $\bullet$

$DFFIT_i$ - change of fit on forecasting after i-th observation is deleted

$DFFITS_i$ - standardized change in fitted value of i-th case after deletion

E - Mayer and Willke's class of estimators

ϵ:nx1 - nx1 vector of uncorrelated random error variables

$\epsilon = [\epsilon_a \; \epsilon_b]$

ϵ_a:(n-k)x1

ϵ_b:kx1

ϵ - scalar controlling collinearity in modified Wampler data set

$\hat{\epsilon}$:nx1 - nx1 residual vector

$\hat{\epsilon} = [\hat{\epsilon}_a \ \hat{\epsilon}_b]$

$\hat{\epsilon}_a$:(n-k)x1

$\hat{\epsilon}_b$:kx1

$\hat{\epsilon}_i$ - residual term for i-th observation

$\hat{\epsilon}_{-i}$:(n-1)x1 - vector of residuals estimator after i-th case deleted

$\hat{\epsilon}_{\hat{r}}$ - residuals due to fitting $X_{\hat{r}}$

$\hat{\epsilon}_R$ - residual vector under RRE

$\hat{\epsilon}_s$ - residuals due to using $\hat{\beta}_s$

$(\hat{\epsilon}_s)_{-h}$ - residuals due to using $(\hat{\beta}_s)_{-h}$

E(.) - expectation of the scalar, vector or matrix argument

e_j - error in the j-th column of X

e_i - residual vector of X_i when regressed on X_{-i}

e_i^{-j} - residual vector of X_i when the regressors are $[X_{-i} \ u_j]$

e_{ij} - j-th element of e_i

$\eta_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij}$ - expected value from fitted equation

$\eta_i = \sqrt{\lambda_1}/\sqrt{\lambda_i}$ - i-th condition index

$\eta_k(-i)$ - k-th condition index computed without the i-th case

η_i - efficiency of the LSE of δ_i relative to the ridge estimator

exp - exponent

f(x) - density function of random variable X

f(X) - density function of random vector X

f(•) - function of a random variable, vector or a matrix

$F(n_1,n_2:\gamma)$ - non-central F-distribution with n_1 and n_2 df and non-centrality parameter γ

$F(\alpha:1,n-q-2)$ - tabulated 100α percentile F-value for 1 and n-q-2 df

F_L - lowest partial F-test value

F_0 - preselected critical value

F_{in} - constant value chosen as criterion for inclusion of a variable

$F_i = \lambda_i \hat{\delta}_i^2/\sigma^2$

$F = \mathrm{Diag}[f_1,f_2,\ldots,f_p]$ - fraction matrix

$F_{PCI} = \mathrm{Diag}(f^*_{1,PC},\ldots,f^*_{p,PC})$

$F_{PCV} = \mathrm{Diag}(f^*_{1,PCV},\ldots,f^*_{p,PCV})$

f_j - fractions (j-th diagonal element of F)

f^0_j - j-th optimal fraction

$f_{j,GR}(t+1)$ - fraction at t-th iteration in FPCI, using GRRE

$f_{j,PC}(t+1)$ - fraction at t-th iteration of FPCE, using PCE

$f_{j,R}(t+1)$ - fraction at t-th iteration of FPCV, using RR

$f_{j,PCV}(t+1)$ - fraction at t-th iteration of FPCV, using PCE

$f^*_{j,PC} = \lim_t \,[f_{j,PC}(t+1)]$

$f^*_{j,PCV}(t+1) = \lim_t f_{j,PCV}(t+1)$

FPC - fractional principal component

FPCI - iterative fractional principal component estimator, *via* the GRR method

FPCV - iterative fractional principal component estimator, *via* the RR method

$G(\bullet)$ - Mayer and Willkes's term for TMSE

γ - non-centrality parameter

γ - scalar controlling collinearity in (2.5.3)

$\gamma_1(k) = \mathrm{tr}(V(\hat{\beta}_R))$ - total variance of the ridge estimator

$\gamma_2(k) = \beta'(Z - I)'(Z - I)\beta$ - total squared bias of ridge estimator

γ_{ij} - j-th eigenvalue of $X'_{-i}X_{-i}$

$\Gamma(\bullet)$ - gamma function

$\Gamma_p = [\sum_{i=1}^{n} (\nu_i - \eta_i)^2 + \sum_{i=1}^{n} Var(\hat{Y}_i)]/\sigma^2$ - Mallows' Γ_p

GRRE - generalized ridge regression estimator

GVIF - generalized variance inflation factor

$GVIF_i$ - i-th generalized variance inflation factor

H:nxn - Hat matrix

H_o - null hypothesis

$h_i = h_{ii} = x_i'(X'X)^{-1}x_i$ - leverage values

I - identity matrix

I_p - (pxp) identity matrix

IMP_j - importance of the j-th variable

L_j - level of importance

$inf(X) \overset{def}{=} \min \|X\nu\|$, for $\|\nu\| = 1$

inf(X) - spectral norm of the smallest matrix E such that X + E is exactly collinear

JRE - jackknifed ridge estimator

$k = p-1$ - number of independent variables, excluding the column of one's

k - ridge constant

k - number of influential or outlier observations

$K{:}pxp$ - diagonal matrix with ridge values k_i on the diagonal

$k_i(opt)$ - optimum value for k_i

k_h - harmonic mean of k_i

$\hat{k}_h$ - estimator of harmonic mean of k_i

$\hat{K}$ - estimator of K

$\hat{k}_i$ - i-th element of $\hat{K}$

$\hat{k}_{LW}$ - Lawless and Wang estimate of k

$K = \{F, \Omega, \Lambda^*\}$ - condition triple

$K(X) = \sqrt{\lambda_1}/\sqrt{\lambda_p}$ - condition number of X (also indicated by η_p)

$K1_i$, $K2_i$ - boundary values of k_i

κ_i - i-th collinearity index

L - unique positive lower triangular matrix of order p

L_1 - Euclidean distance from $\hat{\beta}$ to β

λ_i - i-th eigenvalue of $X'X$

λ_{max} - max eigenvalue of X′X

λ_{min} - min eigenvalue of X′X

$\tilde{\lambda}$ - LSE of λ

Λ - range set

Λ^* - range set

LC - leverage components

LRR - latent root regression

LRRE - latent root regression estimator

LSE - least square estimator

$M_a = I - X_a(X_a'X_a)^{-1}X_a$

$\tilde{m}$ - Swamy's measure of collinearity

m_L - lower bound of $\tilde{m}$

m_U - upper bound of $\tilde{m}$

$mci = \sum_{i=1}^{p} \lambda_p^2/\lambda_i^2$ - ratio of the squares of eigenvalues

$M_{X_0} = \text{plim}(n^{-1}X_0{'}X_0)$

$M_X = \text{plim}(n^{-1}X'X)$

$M_{\Delta_e} = \text{plim}(n^{-1}\Delta_e'\Delta_e)$

MDFFIT - statistic DFFITS extended to deletion of more than one data point

MLE - maximum likelihood estimator

MS - mean square

MSE - mean square error random variable in ANOVA

MSE - matrix mean square error (expectation)

MSR - mean square of regression

μ - expectation of random variable X

$\boldsymbol{\mu}$ - expectation of random vector X

$\hat{\mu}$ - estimate of μ

MVIF - marginal variance inflation factor

n - scalar, usually the number of rows of Y or X

$N(\mu,\sigma^2)$ - univariate normal distribution

$N(\boldsymbol{\mu},\sigma^2 I)$ - multivariate normal distribution with independence and homoscedasticity

$\nu_i = \nu(X_{1i},X_{2i},\ldots)$ - expected value from true equation for the conditional expectation of $(Y_i|X_{1i}\ldots X_{pi})$

$\nu = (n - p)$ df for residual sum of squares

0 - vector or matrix whose elements are zero

OLSE - ordinary least square estimator

ω_i - i-th singular value of [X Y]

Ω - domain

1 - column vector of ones

P - projection matrix

$[P_{-i}]$ - projection matrix onto column space of variables in X_{-i}

p - number of independent variables

p_a - number of independent variables in the subset X_a

$\phi_{ij} = v_{ij}^2/\lambda_i$ - i,j-th variance-decomposition proportion of variance of j-th regression coefficient associated with i-th component

$\Pi_{ij} = \phi_{ij}/\phi_{TOTAL}$ - variance decomposition proportions

π - pi

Π:pxp - permutation matrix

P_i - pseudo-value

PC - principal component

PC(i) - principal component estimator, with i components deleted

PCE - principal component estimator

PCR - principal component regression

plim - probability limit

PMSE - predicted mean square error

Pr(•) - probability

psd - positive semi-definite

PVIF - partial variance inflation factor

Q:nxn - orthogonal-matrix in QR-decomposition of X

Q_k - extra sum of squares

Q_i - weighted pseudo-value

$Q(X,Y,\beta)$ - criterion function to be minimized

ρ^2 - population value estimated by R^2

R:nxp - upper triangular matrix in QR-decomposition of X

R_1 pxp - non-zero part of upper triangular matrix of QR-decomposition

$R:rxr = [0 \ I_r]$ - restriction matrix

R^p - real p-dimensional space

R_{ij} - statistic for detection of collinearity-influential points

R_{IJ} - generalization of R_{ij}

$R_{i,(A)}$ - measure of influence of variable A on variance inflation factor

R - coefficient of multiple correlation

R^2 - coefficient of multiple determination

$R^2_{y\varphi}$ - coefficient of multiple determination where φ is set of independent variables included in model

$R^2_{yi.\phi}$ - coefficient of partial determination where Φ denotes set of regressor X variables already in model prior to fitting X_i.

$R_{ij\bullet}$ - coefficient of partial correlation between i-th and j-th columns of X, while all other columns are held constant

R^2_a - adjusted coefficient of multiple determination

R^2_i - coefficient of determination from regression of X_i on other independent variables

R^2_s - modified coefficient of multiple determination

$(R^2_s)_{-h}$ - modified coefficient of multiple determination after deletion of the h-th column

$R(X)$ - column space (range-space) of X

$\hat{r}$ - estimate of $r(X) = r$

r - number of restrictions

r_i - standardized residual

r_{12} - observed coefficient of correlation between the variables represented by the first two columns of X

r_{ij} - coefficient of correlation between i-th and j-th variables

$r(X)$ - rank of X

RB - relative bias

RE - relative efficiency

RLS - restricted least squares

RMSE - relative mean square error

RR - ridge regression

RRE - ridge regression estimator

RSS - residual sum of squares

s - number of disconnected subsets

s - number of non-predictive near-singularities

S - positive semi-definite matrix

S_i - scalar quantity

σ^2 - population variance scalar

$\sigma^2_{j,LS}$ - theoretical variance of $b_{j,LS}$

$\sigma^2_{Y_e}$:1x1 $= E[(Y_e)_i^2]$

$\sigma_{Y_eX_e}$:px1 $= E[(X_e)_i[(Y_e)_i+\epsilon_i]]$

σ_ϵ^2:1x1 = $E[\epsilon_i^2]$

$\sigma_\mu^2 = \text{plim}(n^{-1}\mu'\mu)$

$\hat{\sigma}^2_{CLS}$ - consistent estimator of σ_μ^2

$(\hat{\sigma}_\epsilon^2 + \hat{\sigma}^2_{Y_e})_{CLS}$ - consistent least square estimator of $(\sigma_e^2 + \sigma^2_{Y_e})$

σ_υ^2 - scalar used in errors-in-variables model

$\hat{\sigma}^2$ - estimate of σ^2

$\hat{\sigma}(i)$ - estimated error variance when the i-th row of X and Y have been deleted

Σ_{X_e}:pxp = $E[(X_e)_i(X_e)_i']$

Σ_{Δ_e} - covariance matrix of error variables of the errors-in-variables model

s^2 - OLS estimate of SSE

s^2 - Swamy's estimate of σ^2

s_y^2 - sample variance of Y

SEMSE - standardized empirical mean square error

SH - shrunken

SHE - shrunken estimator

SS - sum of squares

SSA - extra sum of squares due to fitting A after X

SSB_p - sum of squared bias

SSE - error sum of squares or residual sums of squares

$SSE(X_1,X_2,\ldots,X_p)$ - SSE of all independent regressor variables ((...)) included in model

SSR - regression sum of squares

SSTO - total sum of squares (usually corrected for mean)

SX-OLS - subset selection on X where method of estimation is OLS

SX-TLS - subset selection on X where method of esitmation is TLS

SXY-TLS - subset selection on [X Y] where method of estimation is TLS

SXY-VTLS - subset selection on [X Y] where method of estimation is TLS, with a variant

SVD - singular value decomposition of a matrix .

$\tau_j = [(n-p)^{1/2}\kappa_j e_j]/\|X_j\|$

τ^2 (or τ) - signal-to-noise ratio

τ_i^2 - non-centrality parameter of F-distribution associated with δ_i

τ_*^2 - value of τ^2 under null hypothesis

τ^{-1} - coefficient of variation

θ:nx1 - residual vector for alternative model

$\theta = [\theta_a \ \theta_b]$

$\hat{\theta}$, $\hat{\theta}_a$, $\hat{\theta}_b$: estimates of θ, θ_a, θ_b

θ - sum of product terms for two columns of X

TLS - total least squares

TLSE - total least squares estimator

TMSE - total mean square error

TMSE(Y_f) - total mean square error of prediction

tr[•] - trace of a matrix argument

t_n - central t-distribution with n degrees of freedom

t^2 - t-ratio

u - statistic

u_o - a Lagrangian multiplier

U:nxp - left singular vectors of X

$U_{\hat{r}}$:nx$\hat{r}$ - left singular vectors of $X_{\hat{r}}$

u_i:nx1 - i-th left singular vector of X

u_{ij} - j-th element of i-th left singular vector of X, or so-called leverage components

$u_i = y_i - z_i'\hat{\delta}_K$ (§4.7.1)

u_j - j-th unit vector in the space $\mathbb{R}^n$

$\tilde{U}$:nx(p+1) - left singular vectors of [X Y]

$\tilde{u}_i$:nx1 - i-th left singular vector of [X Y]

V_J - distribution-free estimate of variance for jackknife estimator

V_{JW} - distribution-free estimate of variance for weighted jackknife estimator

$\mathbf{V}$ - variance-covariance matrix, usually $\mathbf{V} = \sigma^2 I$

V:pxp - right singular vectors of X

$V_{\hat{r}}$ - right singular vectors of $X_{\hat{r}}$

v_i:px1 - i-th right singular vector of X

v_{ij} - j-th element of the i-th eigenvector

$\tilde{V}$:(p+1)x(p+1) - right singular vectors of [X Y]

$\tilde{v}_i$:(p+1)x1 - i-th right singular vector of [X Y]

$\tilde{v}_{i,j}$ - j-th component of i-th right singular vector $\tilde{v}_i$

$\tilde{v}_i^0$:px1 - vector containing the first p components of i-th right singular vector $\tilde{v}_i$ of [X Y]

Var(.), V(.) - variance matrix

$V(\hat{\beta}_b|X)$ - variance-covariance matrix of $\hat{\beta}_b$ conditional on X

φ - random variable, vector or a matrix

VIF - variance inflation factor

$VIF_A(X)$ - VIF associated with variable A in model (9.3.1)

VIF_i or $VIF_i(X_{(i)})$ - i-th variance inflation factor

$VIF_i(X_1|X_2)$ - partial variance inflation factor of X_i in $X_{(i)} = [X_1\ X_2]$

$VIF_i(X_{(i)}|1)$ - variance inflation factor of β_i obtained from mean centered data

$VIF_i(X_1)$ - marginal variance inflation factor

VIF_I - generalized variance inflation factor

VIF_i^{-j} - i-th variance inflation factor with j-th observation deleted

VIF_I^{-J} - generalized variance inflation factor with cases indexed by J deleted

$W = Diag[1,\ldots,1,w_i,1,\ldots,1]$ - matrix used in weighted LSE

$w_i = z_i'A^{-1}z_i$

WA - Wampler accuracy of b

$(W'\Delta_1^{-1}W)_{-h',-h}$ - submatrix of $(W'\Delta_1^{-1}W)$ after deleting h-th row and column

χ^2, $\chi^2(n,0)$ or χ_n^2, central χ^2- distribution with n df

$\chi^2(n,\gamma)$ - non-central χ^2- distribution with n df and non-centrality parameter γ

X:nxp - matrix of fixed regressors or explanatory variables.

X:nxp - observed matrix of p explanatory variables

X_0:nxp - true but unobservable variables

X_e:nxp - measurement or observation errors (unobserved)

$(X_e)_i'$:1xp - i-th row of X_e

$\tilde{X}$:nxp - perturbed X matrix, $\tilde{X} = X + E$ or $X + \delta X$

$X^\dagger = (X'X)^{-1}X'$ - pseudo-inverse or generalized inverse of X

$X_{\hat{r}}$ - rank $\hat{r}$ matrix approximation of X

X_i:nx1 - i-th column of X

$\bar{x}_j$ - mean of j-th column

X_{ij} - i-th element of j-th column

$X_{(i)}$ - X-matrix with i-th column deleted

X_I - sub-matrix of X containing columns indexed by I

$(X'X)^{-1}_{ii}$ - i-th diagonal elements of $(X'X)^{-1}$

$(X'X)_{II}$ - sub-matrix of X formed by columns and rows indexed by I

x_i':1xp - i-th row of X

X_{-i} - X matrix with i-th row deleted

$\{X_\phi\}$ - set of regressor variables already in model

$\hat{X}$ - approximation to X that satisfies $\text{rank}(\hat{X}) \leq k = \text{rank}(X)$

[X Y]:nx(p+1) - augmented matrix

$[X\ \hat{Y}]$ - LS approximation of [X Y] with orthogonal projection $\hat{Y}$ of Y onto the column space R(X) of X

$[\tilde{X}\ \tilde{Y}]$ - TLS approximation of [X Y]

$[\Delta\tilde{X}\ \Delta\tilde{Y}] = [X - \tilde{X}, Y - \tilde{Y}]$ - TLS approximation error

Y:nx1 - vector of observed response

Y_i - i-th element of Y

$[\hat{Y}(-i)]$ - estimated Y obtained by using $\hat{\beta}_{-i}$

$[\hat{Y}(-i)]_i$ - i-th element of $[\hat{Y}(-i)]$

$\bar{Y}$ or $\bar{y}$ - mean of Y

Y^*:nx1 - standardized Y values, corrected for mean, and scaled

$\hat{Y}$:nx1 - fitted vector of Y

$\hat{Y}_i$ - i-th value of fitted responses

Y_{-i} - sub-vector of Y after deletion of i-th element

Y_0:nx1 - true but unobservable variables

Y_e:nx1 - measurement or observation errors (unobserved)

Z:nxp - matrix of principal components (Z = XV)

$\|\bullet\|$ - norm of a matrix argument

ABDELMALEK N.N.(1974): On the solution of the linear least squares problem and pseudo-inverses. *Computing*, 13, 215-228.

ABOLONTSEV Y.I. AND KILDISHEV G.S.(1984): Statistical adequacy of regression models and the problem of collinearity. *Ekonomika i Matematiceskie Metody,* 20, no. 6, 1078-1083.

AFIFI A.A. AND ELASHOFF R.M.(1966): Missing observations in multivariate statistics, I. Review of the Literature. *Journal of the American Statistical Association*, 61, 595-604.

AFRIAT S.N.(1957): Orthogonal and oblique projectors and the characteristics of pairs of vector spaces. *Proc. Cambridge Philos. Soc.*, 53, 800-816.

AHAMAD B.(1967): An analysis of crimes by the method of principal component analysis. *Applied Statistics*, 16, 17-35.

AIGNER D.J.(1974): MSE dominance of least-squares with error-of-observation. *Journal of Econometrics*, 2, 365-372.

AITKIN M.A.(1969): Miscellanea - Some tests for correlation matrices. *Biometrika* 56, 443.

AKAIKE H.(1973): Information theory and the extension of the maximum likelihood principle. In 2*nd International Symposium on Information Theory.* (B.N. Petrov and F. Csaki, Eds.), pp. 267-281. Akailseoniai-Kindo, Budapest.

AKAIKE H.(1974): A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19, 723.

AKDENIZ F. AND OZTURK F.(1981): The effects of multicollinearity - a geometric view. *Communications de la Faculte des Sciences de l'Universite d'Ankara Serie A,* 30, no. 3, 17-26 (1982).

ALLDREDGE J.R. AND GILB N.S.(1976): Ridge regression: An annotated bibliography. *International Statistical Review*, Vol. 44, 355-360.

ALLEN D.M.(1971): Mean square error of prediction as a criterion for selecting variables. *Technometrics,* Vol. 13, 469-475.

ALLEN D.M.(1971): The prediction sum of squares as a criterion for selecting predictor variables. Univ. of Ky., Dept. of Statistics, *Technical Report* 23.

ALLEN D.M.(1972): Baised prediction using multiple linear regression. Univ. of Ky., Dept of Statistics, *Technical Report* 36.

ALLEN D.M.(1974): The relationship between variable selection and data augmentation and a method for prediction. *Technometrics,* 16, no. 1, February 1974, 125-127.

ALLEN D.M. AND CADY F.B.(1982): *Analyzing Experimental Data by Regression.* Lifetime Learning Publications, Belmont, California, 1982.

AMEMIYA T.(1980): Selection of regressors. *International Economic Review*, 21, 331-354.

ANDERSON R.L. AND BANCROFT T.A.(1952): *Statistical Theory in Research.* McGraw-Hill, New York.

ANDERSON T.W.(1958): *An introduction to multivariate statistical analysis.* John Wiley & Sons, New York.

ANDERSON T.W.(1984): Estimating linear statistical relationships. *Annals of Statistics*, 12, 1-45.

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

ANDREWS D.F. AND PREGIBON D.(1978): Finding the outliers that matter. *Journal of the Royal Statistical Society, Series* B, 40, No. 1, 85-93.

ANSCOMBE F.J.(1973): Graphs in statistical analysis. *The American Statistician*, 27, 17-21.

ANSCOMBE F.J. AND TUKEY J.W.(1963): The examination and analysis of residuals. *Technometrics,* 14, 141-160.

ARCHETTI F. AND CUGIANI M.(1980): Numerical techniques for stochastic systems. Papers based on lectures presented at the Conference held at Gargnano, September 1979. Edited by Francesco Archetti and Marco Cugiani.

ARKIN V.I., SHIRAEV A. AND WETS R.(1984): Stochastic optimization. Proceedings of the international conference held in Kiev, September 1984.

ASKIN R.G.(1982): Multicollinearity in regression: Review and examples. *Journal of Forecasting,* 1, 281-292.

ASKIN R.G. AND MONTGOMERY D.C.(1980): Augmented robust estimators. *Technometric*s, 22, 333-341.

ATKINSON A.C.(1981): Two graphical displays for outlying and influential observations in regression. *Biometrika,* 68, no. 1, 13-20.

AVULA X.J.R., KALMAN R.E., LIAPIS A.I. AND RODIN E.Y.(1984): Mathematical modelling in science and technology. Proceedings of the fourth international conference on mathematical modelling held in Zurich, August 15-17, 1983. Edited by Xavier J. R. Avula, Rudolf E. Kalman, Anthanasios I. Liapis and Ervin Y. Rodin.

AYDELOTTE W.O.(1966): Quantification in history. *Amer. Hist. Rev.*, 71, 814-833.

BALAKRISHNAN A.V.(1963): An operator theoretic formulation of a class of control problems and a steepest descent method of solution. *SIAM Journal on Control*, 1, 109-127.

BANERJEE K.S. AND CARR R.N.(1971): A Comment on ridge regression: Biased estimation for non-orthogonal problems. *Technometrics*, 13, 895-898.

BARANCHIK A.J.(1970): A Family of minimax estimators of the mean of a multivariate normal distribution. *Annals of Mathematical Statistics*, 41, 2, 642-645.

BARNETT V.D.(1970): Fitting straight lines - the linear functional relationship with replicated observations. *Applied Statistics*, 19, 135-144.

BARNETT V.D. AND LEWIS T.(1978): *Outliers in Statistical Data.* John Wiley & Sons, New York.

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

BASKERVILLE J.C. AND TOOGOOD J.H.(1982): Guided regression modeling for prediction and exploration of structure with many explanatory variables. *Technometrics*, 24, 9-17.

BEALE E.M.L., KENDALL M.G. AND MANN D.W.(1967): The discarding of variables in multivariate analysis. *Biometrika*, 54, 357-366.

BEATON A.E. RUBIN D.B. AND BARONE J.L.(1976): The acceptability of regression solutions: Another look at computational accuracy. Research Bullitin 72-44, Princeton, N.J.: Educational Testing Service, 1972.

BEATON A.E. RUBIN D.B. AND BARONE J.L.(1976): The acceptability of regression solutions: Another look at computational accuracy. *Journal of the American Statistical Association*, 71, 158-168.

BECKMAN R.J., NACHTSHEIM, C.J. AND COOK R.D.(1987): Diagnostics for mixed-model analysis of variance. *Technometrics*, 29, no. 4, 413-426.

BECKMAN R.J. AND TRUSSELL H.J.(1974): The distribution of an arbitrary studentized residual and the effects of updating in multiple regression. *Journal of the American Statistical Association*, 69, 199-201.

BEHNKEN D.W. AND DRAPER N.R.(1972): Residuals and their variance patterns. *Technometrics*, 14, 102-111.

BELSLEY D.A.(1969): *Industry Production Behavior: The Order-Stock Distinction.* Amsterdam: North-Holland Publishing Co, 1969.

BELSLEY D.A.(1980): The statistical problem associated with collinearity and a test for its presence. *Technical report no.* 6, Center for Computational Research in Economics and Management Science, MIT, Cambridge, MA.

BELSLEY D.A.(1981): Forecasting and collinearity. *Technical report no.* 27, Center for Computational Research in Economics and Management Science, MIT, Cambridge, MA.

BELSLEY D.A.(1982): Assessing the presence of harmful collinearity and other forms of weak data through a test for signal-to-noise. *Journal of Econometrics*, 20, 211-253.

BELSLEY D.A.(1983): Centering, the constant, first-differencing and diagnosing collinearity. *Technical Report* No. 33, Center for Computational Research in Economics and Management Science, Massachusetts Institute of Technology, Cambridge, 1983.

BELSLEY D.A.(1983): Conditioning in models with logs. *Technical Report* No. 36. Center for Computational Research in Economics and Management Science, MIT, Cambridge, Mass.

BELSLEY D.A.(1983): Elements in assessing the conditioning of estimators of nonlinear models. *Preprints.* IFAC/IFORS Conference on Modeling, Washington, D.C., June.

BELSLEY D.A.(1984): Collinearity and forecasting. *Journal of Forecasting*, Vol. 3, 183-196.

BELSLEY D.A.(1984): Demeaning conditioning diagnostics through centering, and Reply. *The American Statistician*, 38, 73-77 and 90-93.

BELSLEY D.A.(1984): Eigenvector weaknesses and other topics in conditioning. *Technometrics*, 26, 297-299.

BELSLEY D.A.(1986): Centering, the constant, first-differencing, and assessing collinearity, in: D.A. Belsley and E. Kuh (Eds.), *Model Reliability* (MIT Press. Cambridge, MA, 1986).

BELSLEY D.A.(1987): Well-conditioned collinearity indices (comment on a paper by G.W. Stewart). *Statistical Science*, 2, 86-91.

BELSLEY D.A., KUH E. AND WELSCH R.E.(1980): *Regression diagnostics: identifying influential data and sources of collinearity.* John Wiley and Sons, New York.

BELSLEY D.A., KUH E. AND WELSCH R.E.(1989): Review of "Regression diagnostics: Identifying influential data and sources of collinearity". *Journal of Applied Econometrics*, 4, 97-99

BELSLEY D.A. AND OLDFORD R.W.(1986): The general problem of ill conditioning and its role in statistical analysis. *Computational Statistics and data analysis*, 4, 103-120.

BELSLEY D.A. AND WELSCH R.E.(1988): Comment on "Combining Robust and Traditional Least Squares Methods" by M.A. Janson. *Journal of Business and Economic Statistics,* 6, 442-447.

BENDEL R.B.(1986): The effect of centering on the condition number of polynomial regression models. *Proceedings of the SAS Users Group International Conference.* SAS Institute, Cary, NC.

BEN-ISRAEL A.(1966): On error bounds for generalized inverses. *SIAM Journal on Numererical Analysis,* 3, 585-592.

BEN-ISRAEL A. AND GREVILLE T.N.E.(1974): *Generalized Inverses: Theory and Applications.* John Wiley, New York.

BEREANU B., GRIGORESCU, S., IOSIFESCU, M. AND POSTELNICU, T.(1981): Proceedings of the Sixth Conference on Probability Theory. Held in Brasov, September 10 - 15, 1979. Editura Academiei Republicii Socialiste Romania, Bucharest.

BERGER J.(1982): Selecting a minimax estimator of a multivariate normal mean. *Annals Statist,* 10, 81-92.

BERGER J.O.(1985): *Statistical Decision Theory and Bayesian Analysis,* 2nd ed. Springer-Verslag, New York.

BERK K.N.(1977): Tolerance and condition in regression computations. *Journal of the American Statistical Association*, 72, 863-866.

BERK K.N.(1978): Gauss-Jordan v. Choleski, in *Comput. Science and Statist:* 11-*th Annual Symposiom on the Interface.* Inst. of Statist., N. Carolina State Univ., 321-324.

BERKSON J.(1950): Are there two regressions? *Journal of the American Statistical Association,* 45, 164-180.

BINKLEY J.K.(1982): The effect of variable correlation on the efficiency of seemingly unrelated regression in a two-equation model. *Journal of the American Statistical Association*, 77, no. 380, 890-895.

BINKLEY J.K. AND NELSON C.H.(1988): A note on the efficiency of seemingly unrelated regression. *The American Statistician,* 42, 137-139.

BJÖRCK Å.(1967): Solving linear least squares problems by Gram-Schmidt orthogonalization. *BIT* 7, 1-21.

BJÖRCK Å.(1967): Iterative refinement of linear least squares solutions I. *BIT* 7, 257-278.

BJÖRCK Å.(1978): Comment on the iterative refinement of least-squares solutions. *Journal of the American Statistical Association,* 73, no. 361, 161-166.

BJÖRK Å. AND GOLUB G.H.(1973): Numerical methods for computing angles between linear subspaces. *Mathematics of Computations*, 27, 579-594.

BLAFIELD E.(1983): Multicollinearity in equation systems. *Acta Univ. Tamper. Ser. A,* 153, 1-11.

BLAYLOCK N.W.(1987): P*itman nearness comparison of regression estimators.* Masters thesis at the University of Texas at San Antonio, TX.

BOARDMAN T.J.(1981): The future of statistical computing on desktop computers. *American Statistician*, 36, 49-58.

BOOTH G.W., BOX G.E.P., MULLER M.E. AND PETERSON T.I.(1959): Forecasting by Generalized Regression Methods, Non-linear Estimation. (Princeton-IBM), February 1959, International Business Machines Corp., Mimeo.

BOULLION T.L. AND ODELL P.L.(1971): *Generalized Inverse Matrices.* John Wiley & Sons, New York.

BOX G.E.P.(1979): Robustness in the strategy of scientific model building. In *Robustness in Statistics*, eds. R.L. Launer and G.N. Wilkinson. New York: Academic Press, 201-236.

BOX G.E.P. AND DRAPER N.R.(1975): Robust design. *Biometrika*, 62, 347-352.

BOX G.E.P., HUNTER W.G. AND HUNTER J.S.(1978): *Statistics for experimenters*. New York: Wiley-Interscience.

BOX G.E.P. AND MULLER M.E.(1958): A Note on the generation of random normal deviates. *Annals of Mathematical Statistics*, 29, 610-611.

BOX G.E.P. AND TIAO G.C.(1973): *Bayesian Inference in Statistical Analysis.* John Wiley & Sons, New York.

BOX M.J. AND DRAPER N.R.(1971): Factorial designs, the $|X'X|$ criterion, and some related matters. *Technometrics*, 13, 731-742.

BRADLEY R.A. AND SRIVASTAVA S.S.(1979): Correlation in polynomial regression. *The American Statistician*, 33, 11-14.

BRANHAM R.L.Jr(1987): Are orthogonal transformations worthwhile for least squares problems? ACM SIGNUM Newsleteer 22, January, 14-18.

BRITT H.I AND LUECKE R.H.(1973): The estimation of parameters in Nonlinear, Implicit Models. *Technometrics* 15, 233-247.

BROECKX F.C.M.(1983): Bayesian estimation of parameters in a linear regression model with normally distrubuted prior information. *Time series analysis: theory and practice,* 4. North-Holland, Amsterdam-New York.

BROOK R.J. AND FLETCHER R.H.(1981): Optimal significance levels of prior tests in the presence of multicollinearity. *Communications in Statistics, Part A - Theory and Methods,* 10, no. 14, 1401-1413.

BROOK R.J. AND MOORE T.(1980): On the expected length of the Least Squares Coefficient Vector. *Journal of Econometrics*, 12, 245-246.

BROWN P.J.(1977): Centering and scaling in ridge regression. *Technometrics*, 19, 35-36.

BROWN R.L., DURBIN J. AND EVANS J.M.(1975): Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society,* B37, 149-163.

BROWN W.G. AND BEATTIE B.R.(1975): Improving estimates of economic parameters by use of ridge regressiion with production function applications. *American Journal of Agricultural Economics*, 57, 21-32.

BROWNE M.W.(1969): Factor analysis models and their application to prediction problems. Ph.D. Thesis, University of South Africa.

BROWNE M.W. AND ROCK D.A.(1978): The choice of additive constants in ridge regression. *South African Statistical Journal*, Vol. 12, No 1, 65-74.

BROWNLEE K.A.(1965): *Statistical Methodology in Science and Engineering,* Second Edition, John Wiley & Sons, Inc., New York.

BRUCKER P. AND PAULY R.(1985): IX symposium on operations research. Part II. Sections 5-8. Proceedings of a symposium held at the University of Osnabruck, Osnabruck, August 27-29, 1984. Verlagsgruppe Athenaum/Hain/Hanstein, Konigstein.

BUCK S.F.(1966): A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society*, *Series B*, 22, 302-360.

BUTCHER J.C.(1960): Random sampling from the normal distribution. *Computer J.*, 3, 251-253.

BUTTIMER D.J.(1972-1973): Supply response in the Irish dairy and beef herds, 1953-1970: an econometric exercise. *Irish Journal of Agricultural Economics and Sociology*, 4.

CADY F.B. AND ALLEN D.M.(1972): Combining experiments to predict future yield data. *Agronomy Journal*, 64, 211-214.

CARNES B.A. AND SLADE N.A.(1988): The use of regression for detecting competition with multicollinear data. *Ecology,* 69, 1266-1274.

CASELLA G.(1980): Minimax ridge regression estimation. *Annals of Statistics*, 8, 1036-1056.

CASELLA G.(1985): An introduction to empirical Bayes data analysis. *The American Statistician*, 39, 83-87.

CERDAN S.V.(1989): A note on the behaviour of augmented principal-component plots in regression. *Communications in Statistics, Part A - Theory and Methods*, 18, 331-342

CHALTON D.O.(1990): Contributions to influence, outliers and Bayesian analysis in the multiple linear regression model. Ph.D. Thesis, University of Cape Town.

CHAMBERS J.M.(1973): Linear regresseon computations: Some numerical and statistical aspects. Proceedings of the 39th Session of the International Statistical Institute. *Bulletin of the International Statistical Institute* 45, Part 4, 245-254.

CHAMBERS J.M.(1977): *Computational Methods for Data Analysis.* John Wiley & Sons, New York.

CHANDLER D. AND KAHAN W.M.(1970): The rotation of eigenvectors by a perturbation III. *SIAM Journal on Numererical Analysis,* 7, 1-46.

CHATTERJEE S. AND HADI A.S.(1988): *Sensitivity analysis in linear regression*. John Wiley & Sons, New York.

CHATTERJEE S. AND PRICE B.(1977): *Regression Analysis by Example.* John Wiley & Sons, New York.

CHENG D.C. AND IGLAISH H.J.(1976): Principal component estimators in regression analysis. *Review of Economics and Statistics*, 58, 229-234.

CHIPMAN J.S.(1964): On least squares with insufficient observations. *Journal of the American Statistical Association,* 59, 1078-1111.

CHIPMAN J.S.(1976): Estimation and aggregation in econometrics: An application of the theory of generalized inverses. *Generalized Inverses and Applications*, ed. M. Zuhair Nashed, Academic Press, New York, 549-769.

CHRISTENSEN R.(1987): *Plane answers to complex questions. The theory of linear models.* Springer-Verlag, New York-Berlin.

CHRISTOPEIT N., HELMES K. AND KOHLMANN, M.(1986): Stochastic differential systems. Proceedings of the third Bad Honnef conference held in Bad Honnef, June 3-7, 1985. Springer-Verlag, Berlin-New York.

CHUN D.(1968): A Note on a regression transformation for smaller roundoff error. *Technometrics*, 10, 393-396.

CLARKE G.P.Y.(1980): Moments of the least-squares estimators in a non-linear regression model. *Journal of the Royal Statistical Society, Series* B, 42, 227-237.

CLAYTON D.G.(1971): Algorithm AS46: Gram-Schmidt orthogonalization. *Applied Statistics*, 20, 335-337.

CLUTTON-BROCK M.(1965): Using the observations to estimate prior distribution. *Journal of the Royal Statistical Society*, *Series* B, 27, 17-27.

COCHRAN W.G.(1953): *Sampling Techniques.* John Wiley & Sons, New York.

COCHRAN W.G.(1968): Errors of measurements in statistics. *Technometrics*, 10, 637-666.

COHEN J. AND COHEN P.(1975): *Applied multiple Regression/Correlation analysis for the Behavioral Sciences*, John Wiley & Sons, New York.

COLE R.(1969): Data errors and forecasting accuracy. In *Economic Forecasts and Expectations* (J. Mincer, ed.). National Bureau of Economic Research, New York.

CONIFFE 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 observations in linear regression. *Technometrics*, 19, No. 1, February 1977, 15-18.

COOK R.D.(1979): Influential observations in linear regression. *Journal of the American Statistical Association,* 74, 169-174.

COOK R.D.(1986): Assessment of local influence (with discussion). *Journal of the Royal Statistical Society*, *Series* B, *Methodological,* 48, 133-169.

COOK R.D. AND WEISBERG S.(1980): Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics*, 22, 1980, 495-508.

COOK R.D. AND WEISBERG S.(1982): *Residuals and influence in regression.* Chapman and Hall, New York.

COPAS J.B.(1969): Compound decisions and empirical Bayes. (With discussion). *Journal of the Royal Statistical Society, Series* B, *Methodological*, 31, 397-425.

COPAS J.B.(1983): Regression, prediction and shrinkage. *Journal of the Royal Statistical Society, Series* B, *Methodological*, 45, 311-354.

CORNELL J.A.(1981): *Experiments with Mixtures: Designs, Models and the Analysis of Mixture Data.* John Wiley & Sons, New York.

COXE K.(1975): Do principal components solve multicollinearity? The Longley data revisited. Presented at joint annual meetings of Biometric Soc., Amer. Statist. Assoc., & Inst. of Math. Statist., Atlanta, Georgia, August 1975.

CRAMER N.(1946): *Mathematical methods of statistics.* Princeton University Press, Princeton, NJ.

CRITCHLEY F.(1985): Influence in principal components analysis. *Biometrika,* 72, no. 3, 627-636.

CROCKER D.C.(1971): Letter to the Editor. *The American Statistician*, 25, no. 3, 55.

CROXTON F.E., COWDEN D.J. AND BOLEH B.W.(1969): *Practical Business Statistics,* 4th Ed., Englewood Cliffs, Prentice-Hall, Inc., New Jersey

DAGENAIS G.(1973): The use of incomplete observations in multiple regression analysis, A Generalized least squares Approach. *Journal of Econometrics*, 1, 317-328.

DAGENAIS M.G.(1983): Extension of the ridge regression technique to nonlinear models with additive errors. *Econom. Lett.,* 12, no. 2, 169-174.

DANIEL C. AND WOOD F.S.(1980): *Fitting equations to data,* 2nd ed. John Wiley & Sons, New York.

DALING J.R. AND TAMURA H.(1970): Use of orthogonal factors for selection of variables regression equation - an illustration. *Applied Statistics*, 19, 260-268.

DAVID N.A. AND STEWART G.W.(1982): Significance testing in a functional model. *Technical Report* 1204, Dept. Computer Science, Univ. Maryland.

DAVID N.A. AND STEWART G.W.(1986): Hypothesis testing with errors in the variables. *Technical Report* TR-1735, Dept. Computer Science, Univ. Maryland.

DAVIES R.B. AND HUTTON B.(1975): The effect of errors in the independent variables in linear regression. *Biometrika*, 62, 383-391.

DAVIS C. AND KAHAN W.M.(1970): The rotation of eigenvectors by a perturbation III, *SIAM Journal on Numerical Analysis*, 7, 1-46.

DEEGAN J. Jr(1976): A Test of the numerical accuracy of some matrix inversion algorithms commonly used in least squares programs. *Journal of Statistical Computations and Simulation*, 4, 269-278.

DE GRUTTOLA V., WARE J.H. AND LOUIS T.A.(1987): Influence analysis of generalized least squares estimators. *Journal of the American Statistical Association,* 82, no. 399, 911-917.

DEKEN J.G.(1983): Approximating conditional moments of the multivariate normal distribution. *SIAM Journal on Scientific and Statistical Computing,* 4, no. 4, 720-732.

DELANEY N.J. AND CHATTERJEE S.(1986): Use of the bootstrap and cross-validation in ridge regression. *Journal of Business and Economic Statistics,* 4, 255-262.

DEL PINO G.E.(1984): Linear restrictions and two step least squares with applications. *Statistics & Probability Letters*, 2, no. 4, 245-248.

DEL RIO M.(1988): On the potential in the estimation of linear functions in Regression. *Communications in Statistics, Part A - Theory and Methods,* 17, 729-738.

DEMING W.E.(1946): *Statistical adjustment of data.* John Wiley & Sons, New York.

DE MOOR B.(1984): First order perturbation analysis of the singular value decomposition. *Internal Report,* ESAT laboratory, K.U.Leuven.

DEMPSTER A.P.(1971): Model searching and estimations in the logic of inference. In *Foundations of Statistical Inference* (V. P. Godambe and D.A. Sprott, eds)

DEMPSTER A.P.(1973): Alternatives to least squares in multiple regression, in D. Kabe and R.P Gupta, eds., *Multivariate Statistical Inference*, North-Holland Publishing Co., Amsterdam, 1973, 25-40.

DEMPSTER A.P. AND RUBIN D.B.(1983): Rounding error in regression: the appropriateness of Shepard's corrections. *Journal of the Royal Statistical Society*, *Series* B, *Methodological*, 45, no. 1, 51-59.

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.

DENT W.T. AND CAVANDER D.C.(1977): More on computational accuracy in regression. *Journal of the American Statistical Association,* 72, 598-600.

DENTON F.T. AND KUIPER J.(1965): The effect of measurement errors on parameter estimates and forecasts. *Rev. Econ. Statist.*, 47, 198-206.

DESOER C.A. AND WHALEN B.H.(1963): A note on pseudoinverses. *J. SIAM* 11, 442-447.

DEY D.K. AND BERGER J.O.(1983): On truncation of shrinkage estimators in simultaneous estimation of normal means. *Journal of the American Statistical Association,* 78, 865-869.

DICKEY J.M.(1968): Three multidimensional-integral identities with Bayesian applications. *Annals of Mathematical Statistics*, 39, 1615-1627.

DICKEY J.(1971): The Bayesian alternatives to the F-Test. SUNY at Buffalo *Research Report* no 50.

DICKEY J.M.(1974): Bayesian Alternatives to the F-test and least squares estimates in the normal linear model. In *Studies in Bayesian Econometrics and Statistics,* Eds. A. Zellner And S. Fienberg, Amsterdam: North-Holland Publishing Co.

DONALDSON J.R. AND SCHNABEL R.B.(1987): Computational experience with confidence regions and confidence intervals for nonlinear least squares. *Technometrics*, 29, 67-82.

DORSETT D.(1982): Resistant M-Estimators in the Presence of Influential Points. Ph.D.Dissertation. Dept. of Statistics, Southern Methodist University.

DORSETT D. AND GUNST R.F.(1982): Bounded-Leverage weights for robust regression estimators. *Technical Report* 171. Southern Methodist University, Dept. of Statistics.

DORSETT D., GUNST R.F. AND GARTLAND E.C. Jr(1983): Multicollinear effects of weighted least squares regression. *Statistics and Probability Letters.* 1, 207-211.

DRAPER J.(1964): Some statistical problems in research and development. *The Statistician*, 14, 311-318.

DRAPER N.R.(1961): Missing values in response surface designs. *Technometrics*, 3, 389-398.

DRAPER N.R. AND JOHN J.A.(1981): Influential observations and outliers in regression. *Technometrics*, 23, No. 1, February 1981, 21-26.

DRAPER N.R. AND SMITH H.(1981): *Applied Regression Analysis.* John Wiley & Sons, New York. (2nd ed).

DRAPER N.R. AND STONEMAN D.M.(1966): Testing for the inclusion of variables in linear regression by a randomisation technique. *Technometrics*, 8, 695-699.

DRAPER N.R. AND VAN NOSTRAND R.C.(1978): Ridge regression - Is it worthwhile? University of Wisconsin Statistics Dept. *Technical Report* No 501.

DRAPER N.R. AND VAN NOSTRAND R.C.(1979): Ridge regression and James-Stein estimation: Review and comments. *Technometrics,* 21, 451-466.

DRISCOLL M.F. AND BOARDMAN T.J.(1986): Collinearity and points of expansion in polynomial regression. Computer Science and Statistics: Proceedings of the 18th Symposium on the Interface.

DWIVEDI T.P., SRIVASTAVA V.K. AND HALL R.L.(1980): Finite sample properties of ridge estimators. *Technometrics*, 22, 205-212.

DWYER P.S.(1951): *Linear Computations.* John Wiley & Sons, New York.

DYKSTRA O. Jr.(1971): The Augmentation of experimental data to maximize $|X'X|$. *Technometrics*, 13, 682-688.

DYSHIN O. A.(1988): Noise immunity of the selection criteria for regression models with correlated perturbations. *Soviet Journal of Automation and Information Sciences,* 21, no. 3, 16-24. *Avtomatika,* 1988, no. 3, 17-25, 93.

ECKART G. AND YOUNG G.(1936): The approximation of one matrix by another of lower rank. *Psychometrika*, 1, 211-218.

EDWARDS A.W.F.(1969): Statistical methods in scientific inference. *Nature, Lond.* 222, 1233-1237.

EDWARDS J.B.(1969): The relation between the F-test and R^2. *The American Statistician,* 23, 28.

EFRON B. AND MORRIS C.(1971): Limiting the risk of Bayes and Empirical Bayes estimators - Part I: The Bayes case. *Journal of the American Statistical Association*, 66, 807-815.

EFRON B. AND MORRIS C.(1972): Empirical Bayes on vector observations - an extension of Stein's method. *Biometrika*, 59, 335-347.

EFRON B. AND MORRIS C.(1972): Limiting the risk of Bayes and Empirical Bayes estimators - Part II: The Empirical Bayes Case. *Journal of the American Statistical Association*, 67, 130-139.

EFRON B. AND MORRIS C.(1973): Stein's Estimation rule and its Competitors - an Empirical Bayes approach. *Journal of the American Statistical Association*, 68, 117-130.

EFRON B. AND MORRIS C.(1973): Combining possibly related estimation problems. *Journal of the Royal Statistical Society*, *Series* B, 35, 3 (1973), 379-421.

EFRON B. AND MORRIS C.(1975): Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70, 311-319.

EFROYMSON M.A.(1960): Multiple regression analysis. Chapter 17 in *Mathematical Methods for Digital Computers.* Edited by A. Ralston and H.S. Wilf, John Wiley & Sons, New York.

EFROYMSON M.A.(1965): Multiple regression analysis. In *Mathematical Methods for Digital Computers*, 191-203.

EPLETT W.J.R.(1978): A Note about the multipliers in latent root regression. *Journal of the Royal Statistical Society*, *Series* B, 40, 184-185.

ERICSON W.A.(1969): Subjective Bayesian models in sampling finite populations (with Discussion). *Journal of the Royal Statistical Society*, *Series* B, 31, 195-233.

EZEKIEL M.(1924): A Method of handling curvilinear correlation for any number of Variables. *Journal of the American Statistical Association,* 19, 431-453.

EZEKIEL M. AND FOX K.A.(1959): *Methods of Correlation and Regression Analysis.* John Wiley & Sons, New York.

FABRYCY M.Z.(1975): Multicollinearity caused by specification errors. *Applied Statistics*, 24, 250-254.

FAREBROTHER R.W.(1972): Principal component estimators and minimum mean square error criteria in regression analysis. *Review of Economics and Statistics*, 54, 332-336.

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

FAREBROTHER R.W.(1979): Estimation with aggregated data. *Journal of Econometrics,* 10, no. 1, 43-55.

FAREBROTHER R.W. AND BERRY G.(1974): Remark AS R12: A remark on algorithm AS6: triangular decomposition of a symmetric matrix. *Applied Statistics,* 23, 447.

FARRAR D.E. AND GLAUBER R.R.(1967): Multicollinearity in regression analysis: The problem revisited. *Review of Economics and Statistics*, 49, 92-107.

FEARN T.(1983): A misuse of ridge regression in the calibration of a near infrared reflectance instrument. *Applied Statistics*, 32, 73-79.

FELDSTEIN M.S.(1973): Multicollinearity and the mean squared error of the alternative estimators. *Econometrica,* 41, No. 4, 337-346.

FIENBERG S.F.(1967): Cell estimates for one-way and two-way analysis of variance tables. *Memorandum* NS-69, Department of Statistics, Harvard University.

FIENBERG S.F.(1971): Discussion of a paper by H.O. Hartley and R.R. Hocking on incomplete data analysis. *Biometrics,* 27, 813-817.

FISHER R.A.(1925): *Statistical Methods for Research Workers.* Oliver and Boyd, Edinburgh and London, 1925.

FLACK V.F.(1989): Predictability measures for ridge regression models. *Communications in Statistics*, *Part A - Theory and Methods*, 18, 755-766.

FLORENS J.P., MOUCHART M., RAOULT J.P., SIMAR L. AND SMITH A.F.M.(1983): Specifying statistical models. From parametric to nonparametric, using Bayesian or non-Bayesian approaches. Proceedings of the Second Franco-Belgian Meeting of Statisticians held at Louvain-la-Neuve, October 15-16, 1981. Lecture Notes in Statistics, 16.

FLORENS J.P., MOUCHART M., AND RICHARD J.F.(1974): Bayesian Inference in error-in-variables models. *Journal of Multivariate Analysis,* 4, 419-452.

FLURY B.W.(1989): Understanding partial statistics and redundancy of variables in regression and discriminant analysis. *The American Statistician*, 43, 27-31

FOMBY T.B. AND HILL R.C.(1979): Multicollinearity and the minimax conditions of the Bock Stein-like estimator. *Econometrica,* 47, no. 1, 211-212.

FOMBY T.B., HILL R.C. AND JOHNSON S.R.(1978): An optimal property of principal components in the context of restricted least squares. *Journal of the American Statistical Association*, 73, 191-193.

FORSYTHE G.E.(1970): Pitfalls in computation, or why a math book isn't enough. *The American Mathematical Monthly*, 77, 931-956.

FORSYTHE G. AND MOLER C.(1967): *Computer solution of linear algebraic systems.* Prentice-Hall, Englewood Cliffs, NJ.

FOURGEAUD C., GOURIEROUX C. AND PRADEL J.(1984): Some theoretical results for generalized ridge regression estimators. *Journal of Econometrics*, 25, no. 1-2, 191-203.

FOX J.(1984): *Linear statistical models and related methods with applications to social research.* John Wiley & Sons, New York.

FOX K.A. AND COONEY J.F.(1954): Effects of Intercorrelations upon Multiple Correlation and regression measures. U.S. Dept. of Agriculture, Agricultural Marketing Service, Washingtion, D.C.

FOX L.(1950): Practical methods for the solution of normal equations and the inversion of matrices. *Journal of the Royal Statistical Society, Series B*, 12, 120-136.

FRANE J.W.(1978): Detecting and describing statistical and numerical ill-conditioning. Proceedings of the Statictical Computing Section, American Statistical Association, 68-77.

FRANCIS I.S.(1983): Invited discussion of paper by E.B. James, Microcomputers: the coming revolution in Statistics. *Bulletin of the International Statistical Institute,* 50, 3, 140.

FREUND R.J.(1963): A warning of roundoff errors in regression. *The American Statistician*, 17, 5, 13-15.

FREUND R.J. AND MINTON P.D.(1979): *Regression methods. A tool for data analysis.* Statistics: Textbooks and Monographs, 30. Marcel Dekker, New York.

FRIEDMAN D.J., MONTGOMERY D.C.(1985): Evaluation of the predictive performance of biased regression estimators. *Journal of Forecasting,* 4, 153-163.

FRISCH R.(1934): *Statistical Confluence Analysis by Means of Complete Regression Systems.* Oslo: Universitetets Okonomiske Institutt, Oslo, Norway.

FULLER W.A.(1980): Properties of some estimators for the errors in variable model. *The Annals of Statistics,* 8, 407-422.

GALPIN J.S.(1978): An investigation of methods of ridge regression. *Technical Report*, CSIR, Pretoria.

GARDNER J.R. AND HYMANS S.H.(1978): An econometric model of the U.S. monetary sector. *RSQE Research Report.* The University of Michigan, Ann Arbor, MI.

GARNHAM N.F.J.(1979): Some aspects of the use of principal components in multiple regression. Dissertation for M.Sc. in Statistics an University of Kent at Canterbury.

GARSIDE M.J.(1965): The best subset in multiple regression analysis. *Applied Statistics*, 14.

GARSIDE M.J.(1971): Some computational procedures for the best subset problem. *Applied Statistics*, 20, 8-15.

GAUSS C.F.(1821): Theroria combinationis observationum erroribus minimus obnoxiae, in Werke IV, Koniglichen Gessellschaft der Wissenschaften zu Gottingen, 1821, pp 1-26.

GAYLOR D.W. AND MERRILL J.A.(1968): Augmenting existing data in multiple regression. *Technometrics,* 10, 73-81.

GEIGENMULLER U., TITULAER U.M. AND FELDERHAF B.U.(1983): The approximate nature of the Onsager-Casimir reciprocal relations. *Phys. A,* 119, no. 1-2, 53-66.

GENTLE J.E.(1978) Computations for least absolute values estimation. *Communications in Statistics, Part B - Simulation and Computation,* 6, no. 4, 720-732.

GENTLEMAN J.F.(1980): Finding the K most likely outliers in two-way tables. *Technometrics*, 22, 591-600.

GENTLEMAN J.F. AND WILK M.B.(1975): Detecting outliers II. Supplementing the direct analysis of residuals. *Biometrics*, 31, 387-410.

GENTLEMAN W.M.(1973): Least squares computations by Givens transformations without square roots. *J. Inst. Maths. Applics*, 12, 329-236.

GENTLEMAN W.M.(1975): Error analysis of QR decomposition by Givens transformations. *Linear Algebra and its Applications,* 10, 189-197.

GERHOLD G.A.(1969): Least squares adjustment of weighted data to a general linear equation. *Amer. J. Phys.,* 37, 156-161.

GIBBONS D.G.(1981): A simulation study of some ridge estimators. *Journal of the American Statistical Association*, 76, 131-139.

GILL P.E. AND MURRAY W.(1979): Computation of Lagrange multiplier estimates for constrained minimization. *Mathematical Programming*, 17, 32-60.

GLESER L.J.(1981): Estimation in a multivariable errors-in-variables regression model : Large sample results. *The Annals of Statistics*, 9, 24-44.

GLESER L.J. AND WATSON G.S.(1973): Estimation of a linear transformation. *Biometrika* 60, 525-534.

GODAMBE V.P.(1966): A new approach to sampling from finite populations. I. Sufficiency and linear estimation. *Journal of the Royal Statistical Society, Series* B, 28, 310-319.

GOHBERG I.C. AND KREIN M.G.(1969): Introduction to the theory of Nonself-adjoint operators. *American Mathematical Society, Providence, R.I.*

GOLDBERGER A.S.(1964): *Econometric theory.* John Wiley & Sons, New York.

GOLDBERGER A.S.(1968): *Topics in Regression Analysis.* The Macmillan Company, London.

GOLDER E.R.(1976): The spectral test for the evaluation of congruential Pseudo-random generators. *Applied Statistics*, 25, 173-180.

GOLDMAN A.J. AND ZELEN M.(1964): Weak generalized inverses and minimum variance linear unbiased estimation. *J. Res. Nat'l. Bur. St'ds.* 68b, 151-172.

GOLDSTEIN M. AND SMITH A.F.M.(1974): Ridge-type estimations for regression analysis. *Journal of the Royal Statistical Society, Series*, B, 36, 2, 284-291.

GOLUB G.H.(1965): Numerical methods for solving linear least squares problems. *Numer. Math.* 7, 206-216.

GOLUB G.H.(1969): Matrix decompositions and statistical calculations. *Statistical Computation.* R.C. Milton and J.A. Nelder, eds. Academic Press, New York, pp. 365-397.

GOLUB G.H.(1973): Some modified eigenvalue problems. *SIAM Review,* 15, 318-344.

GOLUB G.H., HOFFMAN A. AND STEWART G.W.(1987): A Generalization of the Eckart-Young-Mirsky matrix approximation theorem. *Linear Algebra and its Applications,* 88/89: 317-327.

GOLUB G.H., KLEMA V. AND STEWART G.W.(1976): Rank degeneracy and least squares problems. *Technical Report* TR-751, Dept. Computer Science, Univ. Maryland.

GOLUB G.H. AND KAHAN W.(1965): Calculating the singular values and Pseudo-inverse of a matrix. *SIAM* Journal on *Numer*ical *Ana*lysis*.,* Ser. B, 2, 205-224.

GOLUB G.H. AND PEREYRA V.(1973): The differentiation of pseudoinverses and nonlinear least squares problems whose variables separate. *SIAM Journal on Numerical Analysis*, 10, 413-432.

GOLUB G.H. AND REINSCH C.(1970): Singular-Value decomposition and least-squares solutions. *Numerische Mathematik*, 14, 403-420.

GOLUB G.H. AND VAN LOAN C.(1979): Total least squares. In *Smoothing Techniques for curve estimation.* T. Gasser and M. Rosenblatt, eds. Springer-Verslag, New York, 69-76.

GOLUB G.H. AND VAN LOAN C.(1980): An analysis of the total least squares problem. *SIAM Journal on Numerical Analysis*, 17, 883-893.

GOLUB G.H. AND VAN LOAN C.F.(1983): *Matrix Computations.* Johns Hopkins University Press. Baltimore, MD, 1983.

GOLUB G.H. AND WILKINSON J.H.(1966): Note on the iterative refinement of least squares solution. *Numer. Math.*, 9, 139-148.

GONIN R. AND DU TOIT S.H.C.(1987): Numerical algorithms for solving nonlinear L_p-norm estimation problems. II. A mixture method for large residual and ill-conditioned problems. *Communications in Statistics*, *Part A - Theory and Methods,* 16, no. 4, 969-986.

GOODNIGHT J. AND WALLACE T.D.(1972): Operational techniques and tables for making weak MSE tests for restrictions in regression. *Econometrica*, 40, 699-709.

GORMAN J.W.(1970): Fitting equations to mixture data with restraints on compositions. *Journal of Quality Technology,* 2, 186-194.

GORMAN J.W. AND TOMAN R.J.(1966): Selection of variables for fitting equations to data. *Technometrics,* 8, 27-51.

GOSLING B.J. AND PUTERMAN M L.(1985): Ridge estimation in regression problems with autocorrelated errors: A Monte Carlo study. *Communications in Statistics*, *Part B - Simulation and Computation,* 14, 577-613.

GRAYBILL F.A.(1969): *Introduction to Matrices with Applications in Statistics*. Wadsworth Publishing Company, Belmont, CA.

GRAYBILL F.A.(1976): *Theory and Applications of the Linear Model.* Duxbury press, Belmont, California.

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

GREVILLE T.N.E.(1959): The pseudo-inverse of a rectangular or singular matrix and its application to the solution of systems of linear equations. *SIAM Review* 1, 38-43.

GREVILLE T.N.E.(1960): Some applictions of the pseudo-inverse of a matrix. *SIAM Review* 2, 15-22.

GRILICHES Z. AND INTRILIGATOR M.D.(1983): *Handbook of econometrics.* Vol. I. North-Holland Publishing Co., Amsterdam-New York.

GROSSMANN W., MOGYORODI J., VINCZE I. AND WERTS W.(1988): Probability theory and mathematical statistics with applications. Proceedings of the Fifth Pannonian Symposium on Mathematical Statistics, held in Visegrad, May 20-24, 1985.

GRUBER, J.(1984): *Multicollinearity and biased estimation.* Proceedings of the conference held at the University of Hagen, Hagen, September 8-10, 1980. Edited by Josef Gruber.

GUERARD J.B. Jr AND BEIDLEMAN C.R.(1986): Composite forecasting of annual earnings: An application of biased regression techniques. *Journal of Statistical Computation and Simulation,* 24, 1-16.

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.(1983): Regression analysis with multicollinear predictor variables: definition, detection, and effects. *Communications in Statistics, Part A - Theory and Methods,* 12, no. 19, 2217-2260.

GUNST R.F.(1984): Comment: Toward a balanced assessment of collinearity diagnostics. *The American Statistician,* 38, 79-82.

GUNST R.F. AND MASON R.L.(1973): Some additional indices for selecting variables in regression. Presented at joint annual meetings of Biometric Soc., Amer. Statist. Assoc., & Inst. of Math. Statist., New York, Dec. 27-30. 1973.

GUNST R.F. AND MASON R.L.(1976): Generalized mean square error properties and regression estimators. *Communications in Statistics*, *Part A - Theory and Methods,* 5, 1501-1508

GUNST R.F. AND MASON R.L.(1977): Advantages of examining multicollinearities in regression analysis. *Biometrics,* 33, 249-260.

GUNST R.F. AND MASON R.L.(1977): Baised estimation in regression: an evaluation using mean squared error. *Journal of the American Statistical Association*, 72, 616-628.

GUNST R.F. AND MASON R.L.(1979):. Some considerations in the evaluation of alternate prediction equations. *Technometrics,* 21, no. 1, 55-63.

GUNST R.F. AND MASON R.L.(1980): *Regression analysis and its application. A data-oriented approach.* Statistics: Textbooks and Monographs, 34. Marcel Dekker, New York.

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.

GUNTHER-JURGENS G., ENDEBROCK P. AND KLATTE R.(1987): RESI: practical experience in computer arithmetic.

GUPTA R.P.(1973): A note on multicollinearity and imprecise estimation. *Statist. Hefte (N.F.),* 14, 84-87.

HADI A.S.(1986): *The Prediction Matrix: Its Properties and Role in Data Analysis.* Proceedings of the Business and Economic Statistics Section, American Statistical Association, Washington, DC.

HADI A.S.(1988): Diagnosing collinearity-influential observations. *Computational Statistics & Data Analysis* , 7, 143-159

HADI A.S AND VELLEMAN P.F.(1987): Diagnosing near collinearities in least squares regression (comment on a paper by G.W. Stewart). *Statistical Science,* 2, 93-98.

HAITOVSKY Y.(1968): Missing data in regression analysis. *Journal of the Royal Statistical Society, Series* B, 30, 67-82.

HAITOVSKY Y.(1969): Multicollinearity in regression analysis: Comment. *The Review of Economics and Statistics,* Vol 51, 486-489.

HAITOVSKY Y.(1972): On errors of measurement in regression analysis in economics. *International Statistical Review,* 40, 23-45.

HALD A.(1952): *Statistical Theory with Engineering Applications.* John Wiley & Sons, New York.

HALPERIN M. AND GURIAN J.(1971): A note on estimation in straight line regression when both variables are subject to error. *Journal of the American Statistical Association*, 66, 587-589.

HAMILTON D.(1987): Sometimes $R^2 > r^2(yx_1) + r^2(yx_2)$: Correlated variables are not always redundant. *The American Statistician,* 41, 129-132.

HAMMER G. AND PALLASCHKE D.(1984): Contributions to operations research and mathematical economics. Vol. II.

HAMPEL F.R., RONCHETTI F.M., ROUSSEEVW P.J. AND STAHEL V.A.(1986): *Robust Statistics: The Approach Based on Influence Functions.* John Wiley & Sons, New York.

HANSON R.J. AND LAWSON C.L.(1969): Extensions and applications of the Householder Algorithm for solving linear least squares problems. *Mathematics of Computation* 23, 787-812.

HARTLEY H.O.(1961): The modified Gauss-Newton method for the fitting of non-linear regression functions by least-squares. *Technometrics* 3, 269-280.

HARVEY A.C.(1977): Some comments on multicollinearity in regression. *Applied Statistics*, 26, 188-191.

HAWKINS D.M.(1973): On the investigation of alternative regressions by Principal Component Analysis. *Applied Statistics*, 22, 275-286.

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

HAWKINS D.M.(1980): *Identification of Outliers.* Chapman and Hall, London.

HAWKINS D.M. AND EPLETT W.J.R.(1982): The Cholesky factorization of the inverse correlation or covariance matrix in multiple regression. *Technometrics*, 24, 191-198.

HEALY M.J.R.(1963): Fitting a quadratic. *Biometrics*, 19, 362-363.

HEALY M.J.R.(1963): Programming multiple regression. *The Computer Journal*, 6, 57-61.

HEALY M.J.R.(1968): Algorithm AS6: triangular decomposition of a symmetric matric: algorithm AS7: inversion of a positive semi-definite matrix. *Applied Statistics*, 18, 195-199.

HEARON J.Z. AND EVANS J.W.(1968): Differentiable generalized inverses. *J. Res. Nat. Bur. Stand., Sect.* B, 72B 109-113.

HEIKKILA E.(1988):. Multicollinearity in regression models with multiple distance measures. *Journal of Regional Science*, 28, 345-361

HEMMERLE W.J.(1975): An explicit solution for generalized ridge regression. *Technometrics*, 17, 309-314.

HEMMERLE W.J. AND BRANTLE T.F.(1978): Explicit and constrained generalized ridge estimation. *Technometrics*, 20, 109-120.

HEMMERLE W.J. AND CAREY M.B.(1983): Some properties of generalized ridge estimators. *Communications in Statistics*, *Part B - Simulation and Computation*, B12, 239-253.

HENDERSON H.V. AND VELLEMAN P.F.(1981): Building multiple regression models interactively. *Biometrics,* 37, 391-411.

HENDRY D.F.(1980): Econometrics - alchemy or science: *Economica* 47, 387-406.

HETTMANSPERGER T.P. AND McKEAN J.W.(1977): A Robust alternative based on ranks to least squares in analyzing linear models. *Technometrics*, 19, 275-284.

HIGHAM N.J. AND STEWART G.W.(1987): Numerical linear algebra in statistical computing. The state of the art in numerical analysis (Birmingham, 1986).

HILL B.M.(1969): Foundations for the theory of least squares. *Journal of the Royal Statistical Society, Series B*, 31, 89-97.

HILL R.W.(1977): Robust regression when there are outliers in the carriers. Unpublished Ph.D.dissertation, Harvard University, Dept. of Statistics.

HILL R.C., FOMBY T.B. AND JOHNSON S.R.(1977): Component selection norms for principal components regression. *Communications in Statistics, Part A - Theory and Methods,* 6, 309-334.

HILL R.C. AND JUDGE G.G.(1987): Improved prediction in the presence of multicollinearity. *Journal of Econometrics,* 35, no.1, 83-100. 1987

HILL R.C. AND JUDGE G.G.(1990): Improved estimation under collinearity and squared error loss. *Journal of Multivariate Analysis*, 32, 296-312.

HILL R.C. AND ZIEMER R.F.(1983): Missing regressor values under conditions of multicollinearity. *Communications in Statistics, Part A - Theory and Methods*, 12, no. 22, 2557-2573.

HIMMELBLAU D.M.(1970): *Process Analysis by Statistical methods.* John Wiley & Sons, New York.

HINKLEY D.V.(1976): Robust jackknife correlation. Stanford Univ., Biostat. *Technical Report*, 19.

HINKLEY D.V.(1977): Jackknife confidence limts using Student-t approximations. *Biometrika,* 64, 21-28.

HINKLEY D.V.(1977): Jackknifing in unbalanced situations. *Technometrics*, 19, No. 3, 285-292.

HOAGLIN D.C. AND WELSCH R.E.(1978): The Hat matrix in regression and ANOVA. *The American Statistician*, 32, 17-22.

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

HOCKING R.R.(1983): Developments in linear regression methodology: 1959-1982 (with discussion). *Technometrics*, 25, 219-249.

HOCKING R.R.(1984): Discussion of K-clustering as a detection tool for influential subsets in regression. By J.B. Gray and R.F. Ling, *Technometrics,* 26, 321-323.

HOCKING R.R AND DUNN M.R.(1982): Collinearity, influential Data and ridge Regression, Paper presented at University of Delaware Symposium on Ridge Regression.

HOCKING R.R AND LESLIE R.N.(1967): Selection of the best subset in regression analysis. *Technometrics*, 9, 531-540.

HOCKING R.R. AND PENDLETON O.J.(1983): The regression dilemma. *Communications in Statistics, Part A - Theory and Methods,* 12, 497-527.

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.(1959): Optimum solution of many variable equations. *Chem. Engr. Prog.*, 55, 69-78.

HOERL A.E.(1962): Application of ridge analysis to regression problems. *Chem. Engr. Prog.*, 58, 54-59.

HOERL A.E.(1964): Ridge analysis. *Chem. Engr. Prog. Symposium Series* 60, 67-77.

HOERL A.E. AND KENNARD R.W.(1968): On regression analysis and biased estimation. *Techometrics,* 10, 422-423. Abstract.

HOERL A.E. AND KENNARD R.W.(1970): Ridge regression: Applications to nonorthogonal problems. *Technometrics* 12, 69-82.

HOERL A.E. AND KENNARD R.W.(1970): Ridge regression: Baised estimation for nonorthogonal problems. *Technometrics* 12, 55-69.

HOERL A.E. AND KENNARD R.W.(1976): Ridge regression: iterative estimation of the biasing parameter. *Communication in Statistics*, A5, 77-88.

HOERL A.E. AND KENNARD R.W.(1982): Ridge Regression, Bibliography Update. *Accession Report* 16487, E.I. du Pont, Wilmington, DE.

HOERL A.E., KENNARD R.W. AND BALDWIN K.F.(1975): Ridge regression: some simulations. *Communications in Statistics*, 4, 105-123.

HOERL A.E., KENNARD R.W. AND HOERL R.W.(1985): Practical use of ridge regression: A challange met. *Applied Statistics,* 34, 114-120.

HOERL R.W., SCHUENEMEYER J.H. AND HOERL A.E.(1986): A simulation of biased estimation and subset selection regression techniques. *Technometrics*, 28, 369-380.

HOLDEN K.(1969): The effect of revisions to data on two econometric studies. *Manchester School*, 37, 23-37.

HOOPER J.W. AND THEIL H.(1958): The extension of Wald's method of fitting straight lines to multiple regression. *Rev. Int. Statist. Inst.,* 26, Part 1, 37-47.

HORN R.A AND JOHNSON C.R.(1987): *Matrix Analysis.* Cambridge University Press, Cambridge.

HORTON R.L. AND GUERARD J.B. Jr(1985): The management of executive compensation in large, dynamic firms: A further look. *Communications in Statistics, Part B - Simulation and Computation,* 14, 441-448.

HöSCHEL H.P. AND PENEV S.(1980): Least squares curve-fitting for nonlinear models with errors-in-variables and globally convergent Gauss-Newton-procedures. Discussion Paper 8025, Center for Operation Research and Econometrics, Universite Catholique de Louvain.

HOSMANE B. AND HUA T.A.(1985): Multicollinear effect in logistic regression. American Statistical Association Proceedings of the Statistical Computing Section.

HOTELLING H.(1957): The relations of the newer multivariate statistical methods to factor analysis. *Brit. J. Statist. Psychol.,* 10, 69-79.

HOUSEHOLDER A.S (1964): *The theory of Matrices in Numerical Analysis.* Blaisdell, New York.

HOWELL J.A.(1971): Algorithm 406. Exact solutions of linear equations using residual arithmetic [F4]. *Communications of the ACM* 14, 180-184.

HSIANG T.C.(1976): A Bayesian view on ridge regression. *The Statistician,* 24, 267-268.

HSUAN, F.C.(1981): Ridge regression from principal component point of view. *Communications in Statistics, Part A - Theory and Methods,* 10, no. 1981-1995.

HUANG D.S.(1970): *Regression and Econometrics.* The Macmillan Company, New York.

HUBER P.J.(1972): Robust statistics: a review. *Annals of Mathematical Statistics*, 43, 1041-1067.

HUBER P.J.(1973): Robust regression: Asymptotics, Conjectures and Monte Carlo. *Annals of Statistics.* 1, 789-821.

HUBER P.J.(1975): Robustness and designs. In *A Survey of Statistical Design and Linear Models*, ed. J.N. Srivastava. North-Holland, Amsterdam.

HUBER P.J.(1981): *Robust Statistics.* John Wiley & Sons, New York.

IMAN R.L. AND CONOVER W.J.(1979): The use of the rank transform in regression. *Technometrics,* 21, 499-510.

IGLARSH H.J. AND CHENG D.C.(1979): Weighted estimators in regression with multicollinearity. *Journal of Statistical Computation and Simulation,* 10, no. 2, 103-112.

JACKSON J.E. AND HEARNE F.T.(1973): Relationships among coefficients of vectors used in Principal Components. *Technometrics,* 15, 601-610.

JACKSON P.H., NOVICK M.R. AND THAYER D.T.(1971): Estimating regressions in m-groups. *Brit. J. Math. Statist. Psychol.*, 24, 129-153.

JAHN W. AND RIEDEL M.(1984): Reduction of the dimension in the linear model with stochastic regressors. *Commentationes Mathematicae Universitatis Carolinae*, 25, no.4, 747-761.

JAMES W. AND STEIN C.(1961): Estimation with quadratic loss, in Neyman J.(ed.). *Proceedings of the Fourth Berkeley Symposium*, Los Angeles: University of California Press, 1961, 361-379.

JEFFERS J.N.R.(1965): Correspondence. *The Statistician*, 15, 207-208.

JEFFERS J.N.R.(1967): Two case studies in the application of principal component analysis. *Applied Statistics*, 16, 225-236.

JEFFERS J.N.R.(1981): Investigation of alternative regressions: some practical examples. *The Statistician,* 30, 79-88.

JEFFREYS H.(1961): *Theory of Probability.* Third Edition, Oxford University Press, London, Chapter III.

JOHN J.A. AND DRAPER N.R.(1978): On testing for two outliers or one outlier in two-way tables. *Technometrics*, 20, 69-78.

JOHNSON S.R., REIMER S.C. AND ROTHROCK T.P.(1973): Principal components and the problem of multicollinearity. *Metroeconomica,* 25, 306-317.

JOHNSON T. AND WALLACE T.D.(1969): Principal Components and Multicollinearity. Department of Economics, Econometrics Workshop Discussion Paper, North Carolina State University, Raleigh, North Carolina.

JOHNSTON J.(1963): *Econometric Methods.* McGraw-Hill, New York.

JOHNSTONE I.(1987): On the admissibility of some unbiased estimates of loss. In *Statistical Decision Theory and Related Topics* IV (U.S. Gupta and J. Berger, Eds.), Vol. 1, pp 281-297. Springer-Verlag, New York.

JOLLIFFE I.T.(1972): Discarding variables in a principal component analysis. I. Artificial data. *Applied Statistics*, 21, 160-173.

JOLLIFFE I.T.(1973): Discarding variables in a principal component analysis. II. Real data. *Applied Statistics*, 22, 21-31.

JOLLIFFE I.T.(1982): A Note on the use of principal components in regression. *Applied Statistics*, 31, 300-303.

JONES D.A.(1978): Nonlinear autoregressive processes. *Proceedings of the Royal Society of London, Series A,* 360, no. 1700, 71-95.

JONES S.(1988): *GAUSS, Version 2.0, System and Graphics Manual.* Aptech Systems, Kent, WA.

JORDAN T.L.(1968): Experiments on error growth associated with some linear least-squares procedures. *Mathematics of Computation,* 22, 579-588.

JUDGE G.G. AND BOCK M.E.(1976): A comparison of traditional and Stein rule estimators under weighted squared error loss. *International Economic Review*, 17, 234-240.

JUDGE G.G. AND BOCK M.E.(1978): *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics.* North-Holland Publishing Co., New York.

JUDGE G.G., GRIFFITHS W.E., HILL R.C. LÜTKEPOHL H. AND LEE T.C.(1980): *The Theory and Practice of Econometrics.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York. (2-nd edition 1985).

JUDGE G.G., HILL R.C. AND BOCK M.E.(1990): An adaptive empirical Bayes estimator of the multivariate normal mean under quadratic loss. *J. Econom.* in press.

JUDGE G.G., YI G., YANCEY T. AND TERÄSVIRTA T.(1987): The extended Stein procedure for simultaneous model selection and parameter estimation. *J. Econom.*, 35, 375-392.

KADIYALY K.(1984): A class of almost unbiased and efficient estimators of regression coefficients. *Economics Letters,* 16, 293-296.

KAGIWADA H., KALABA R. AND MEASE K.(1977): The flabbiness and instability of regression analysis and computational methods for improvement. Applications of statistics (Proc. Sympos., Wright State Univ., Dayton, Ohio, 1976)

KAKWANI N.C.(1965): Note on the use of prior information in forecasting with a Linear Regression Model. *Sankhya, Series A,* 27, 101-104.

KALMAN R.E.(1984): We can do something about multicollinearity! *Communications in Statistics, Part A - Theory and Methods*, 13, no. 2, 115-125.

KARLIN S. AND STUDDEN W.J.(1966): Optimal experimental design. *Annals of Mathematical Statistics*, 37, 783-816.

KASHYAP A.K., SWAMY P.A.V.B., MEHTA J.S., AND PORTER R.D.(1984): Estimating distributed lag relationships using near-minimax procedures. *Special Studies Paper*, Federal Reserve Board, Washington, D.C.

KATO T.(1966): *Perturbation Theory for Linear Operators.* Springer, Berlin.

KEATING J.P. AND MASON R.L.(1985): Practical relevance of an alternative criterion in estimation. *The American Statistician,* 39, 1868-1875.

KEATING J.P. AND MASON R.L.(1988): James-Stein estimation from an alternative perspective. *The American Statistician*, 42, 160-164.

KEIFER J.(1959): Optimum experimental designs. *Journal of the Royal Statistical Society, Series B*, 21, 272-319.

KEMPTHORNE O.(1957): *An Introduction to Genetic Statistics.* John Wiley & Sons, New York. Reprinted in 1968 by the Iowa State University Press.

KEMPTHORNE P.J.(1985): Assessing the influence of single cases on the condition number of a design matrix. Memorandum NS-509, Department of Statistics, Harvard University.

KENDALL M.G.(1957): *A Course in Multivariate Analysis.* Griffin, London.

KENDALL M.G. AND STUART A.(1968): *The Advanced Theory of* Statistics (2nd ed.), Vol. 3. Griffen, London.

KENNARD R.V AND STONE L.(1969): Computer aided design of experiments. *Technometrics,* 11, 137-148.

KENNEDY V.J. Jr AND GENTLE J.E.(1980): *Statistical computing.* Statistics: Textbooks and Monographs, 33. Marcel Dekker, Inc., New York.

KETELLAPPER R.H.(1982): The Relevance of large-sample properties of estimators for the errors-in-variables model: A Monte Carlo study. *Communications in Statistics, Series* B, 11, 625-634.

KETELLAPPER R.H.(1983): On estimating parameters in a simple Linear Errors-in-Variables Model. *Technometrics,* 25, no. 1, 43-47.

KHATRI, C.G.(1961): Simultaneous confidence bounds on the departures from a particular kind of multicollinearity. *Annals of the Institute of Statistical Mathematics*, 13, 239-242.

KLEIN L.R.(1962): *An Introduction to Econometrics.* The MacMillan Co., New York.

KLEIN L.R. AND NAKAMURA M.(1962): Singularity in the equation systems of Econometrics, Some Aspects of Multicollinearity. *International Economic Review,* 3, 274-299.

KLEINBAUM D.G., KUPPER L.L. AND MULLER K.E.(1988): *Applied regression analysis and other multivariable methods.* PWS-KENT Publishing company, Boston. (Second edition.)

KMENTA J.(1971): *Elements of Econometrics.* The MacMillan Company, New York.

KNUTH D.E.(1969): *The art of computer programming*, Vol. 2, Reading, Mass: Addison Wesley Publishing Co.

KOPITZKE R., BOARDMAN T.J. AND GRAYBILL F.A.(1975): Least squares programs - a look at the square root procedure. *The American Statistician,* 29, 64-66.

KRASKER W.S. AND WELSCH R.E.(1982): Efficient bounded-influence regression estimation. *Journal of the American Statistical Association,* 77, 595-604.

KSHIRSAGAR A.M.(1972): *Multivariate analysis.* Marcel Dekker, New York.

KUKS J. AND OLMAN V.(1972): Minimax Linear estimation of regression Coefficients, II, *Iswestija Akademija Nauk Estonskoj*, SSR 21, 66-72.

KUMAR T.K.(1975): Multicollinearity in regression analysis. *Review of Economics and Statistics*, 57, 365-366.

KUNG E.C. AND SHARIF T.A.(1980): Multi-regression forecasting of the Indian summer monsoon with antecedent pattern of the large scale circulation. In *WMO Symposium on Probabilistic and statistical Methods in Weather Forecasting.*, 295-302.

KUNUGI T., TAMURA T. AND NAITO T.(1961): New acetylene process uses hydrogen dilution. *Chemical Engineering Progress*, 57, 43-49.

KUPPER L.L AND MEYDRECH E.F.(1973): A new approach to mean squared error estimation of response surfaces. *Biometrika,* 60, 573-579.

LAI T.L.W. AND CHING Z.(1986): On the concept of excitation in least squares identification and adaptive control. *Stochastics,* 16, no. 3-4, 227-254.

LAIRD R.J. AND CADY F.B.(1969): Combined analysis of yield data from fertilizer experiments. *Agronomy Journal* 61, 829-834.

LARSEN W.A. AND McCLEARY S.J.(1972): The use of partial residual plots in regression analysis. *Technometrics,* 14, 781-790.

LAWLESS J.F.(1981): Mean square error properties of generalized ridge estimators. *Journal of the American Statistical Association*, 76, 462-466.

LAWLESS J.F. AND WANG P.(1976): A simulation study of ridge and other regression estimators. *Communications in Statistics*, 5, 307-323.

LAWSON C.L. AND HANSON R.J.(1974): *Solving Least-Squares problems.* Prentice-Hall, Inc., Englewood Cliffs, N.J.

LEAMER E.E.(1973): Multicollinearity: A Bayesian interpretation. *The Review of Economics and Statistics*, 55, 371-380.

LEAMER E.E.(1978): *Specification Searches.* John Wiley & Sons, New York.

LEBOWITZ J.L.(1981): Fourth international conference on collective phenomena. Held in Moscow, April 12-14, 1981. New York Academy of Sciences, New York.

LEE K. AND CAMPBELL D.B.(1985): Selecting the optimum k in ridge regression. *Communications in Statistics Part A - Theory and Methods,* 15, 1589-1604.

LEE A.H. AND SILVAPULLE M.J.(1988): Ridge estimation in logistic regression. *Communications in Statistics*, *Part B - Simulation and Computation,* 17, 1231-1257.

LEE T.S.(1987): Algorithm AS 223: optimum ridge parameter selection. *Applied Statistics*, 36, 112-118.

LEE T.S. AND CAMPBELL D.B.(1985): Selecting the optimum k in ridge regression. *Communications in Statistics*, *Part A - Theory and Methods,* 14, 1589-1594.

LEE W.W.(1986): Fractional principal components regression: a general approach to biased estimators. Unpublished Ph.D. dissertation. Dept. of Statistics, Virginia Polytechnic Institute and State University.

LEE W.W.(1987): A new deletion criterion of principal components regression with orientations of the parameters. *Journal of the Korean Statistical Society,* 16, 55-70

LEE W.W. AND BIRCH J.B.(1988): Fractional principal components regression: A general approach to biased estimators. *Communications in Statistics, Part B - Simulation and Computation,* 17, 713-727.

LEIFMAN L.J.(1983): (Selected translations). A translation of the mathematics section of Vestnik Leningrad. Univ. Mat. Mekh. Astronom. 1978, vyp. 1, 7, 13, 19. Edited by Lev J. Leifman. Vestnik Leningrad Univ. Math. 11 (1983).

LEHMER E.(1944): Inverse tables of probabilities of errors of the second kind. *Annals of Mathematical Statistics* 15, 388-398.

LESAGE J.P. AND SIMON S.D.(1985): Numerical accuracy of statistical algorithms for microcomputers. *Computational Statistics and Data Analysis* 3, 47-57.

LESAGE, J.P. AND SIMON S.D.(1988): Centering and scaling of regression algorithms in the face of ill-conditioning. *Journal of Statistical Computation and Simulation,* 30, 273-283

LEWIS T.O. AND ODELL P.L.(1966): A Generalization of the Gauss-Markov theorem. *Journal of the American Statistical Association,* 61, 1063-1066.

LICHTENSTEIN C.H.(1981): Ridge regression and its effect on high leverage points in the data. M. S. Thesis, Cornell University, Ithaca, N.Y.

LICHTENSTEIN C.H. AND VELLEMAN P.F.(1983): The effects of Ridge regression on high leverage points in the data. Unpublished manuscript.

LIN K. AND KMENTA J.(1982): Ridge regression under alternative loss criteria. *Review of Economics and Statistics*, 64, 488-494.

LINDLEY D.V.(1947): Regression lines and the linear functional relationship. *Journal of the Royal Statistical Society, Series B*, 9, 218-244.

LINDLEY D.V. AND SMITH A.F.M.(1972): Bayes estimates for the linear model. (With discussion). *Journal of the Royal Statistical Society*, *Series B,* 34, 1-41.

LINNIK I.(1961): *Method of Least squares and Principles of the Theory of Observations.* Pergamon Press, New York.

LIVIATAN J.(1961): Errors-in-variables and Engel curve analysis. *Econometrica,* 29, 336-362.

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

LONGLEY J.W.(1981): Least squares computations and the condition of the matrix. *Communications in Statistics, Part B - Simulation and Computation,* 10, no. 6, 593-615.

LONGLEY J.W.(1981): Modified Gram-Schmidt process vs. classical Gram-Schmidt. *Communications in Statistics, Part B - Simulation and Computation,* 10, 517-527.

LORD F.M. AND NOVICK M.R.(1968): Statistical Thoeories of Mental Test Scores. Addison-Wesley Publishing Co., Reading, Mass.

LOTT W.F.(1973): Optimal set of Principal Component Restrictions on a Least Squares regression. *Communications in Statistics*, 2, 449-464.

LOWERRE J.M.(1974): On the mean square error of parameter estimates for some biased estimators. *Technometrics,* 16, 461-464.

LUSH J.L.(1937): *Animal Breeding Plans.* Iowa State University Press, Ames, Iowa.

LUND R.E.(1975): Tables for an approximate test for outliers in linear models. *Technometrics*, 17, 473-476.

MAASOUMI E.(1980): A ridge-like method for simultaneous estimation of simultaneous equations. *Journal of Econometrics*, 12, no. 2, 161-176.

MACGREGOR J.F., HARRIS T.J. AND WRIGHT, J.D.(1984): Duality between the control of processes subject to randomly occurring deterministic disturbances and ARIMA stochastic disturbances. *Technometrics*, 26, no. 4, 389-397.

MADANSKY A.(1959): The fitting of straight lines when both variables are subject to error. *Journal of the American Statistical Association*, 54, 173-205.

MADDALA G.S.(1977): *Econometrics.* McGraw-Hill, New York.

MAGEL R.C. AND HERTSGAARD D.(1987): A collinearity diagnostic for nonlinear regression. *Communications in Statistics, Part B - Simulation and Computation,* 16, 85-97.

MAGNUS J.R. AND NEUDECKER H.(1987): Matrix differential calculus with applications in statistics and econometrics.

MAINDONALD J.H.(1976): Least squares programs - a second look. *The American Statistician,* 30, 202-203.

MAINDONALD J.H.(1977): Least squares computations based on the Cholesky decomposition of the correlation matrix. *Journal of Statistical Computation and Simulation,* 5, 247-258.

MALINVAUD E.(1970): *Statistical Methods of Econometrics, 2nd ed.* North-Holland, Amsterdam.

MALLOWS C.L.(1964): Choosing variables in a linear regression: A Graphical aid, presented at the Central Regional meeting of the Institute of Mathematical Statistics, Manhattan, Kansas, May 7-9, 1964.

MALLOWS C.L.(1973): Some comments on C_p. *Technometrics*, 15, 661-675.

MANDEL J.(1982): Use of the singular value decomposition in regression analysis. *The American Statistician*, 36, 15-24.

MANDEL J.(1985): The regression analysis of collinear data. *Journal of Research of the National Bureau of Standards,* 90, 465-478

MANSFIELD E.R.(1975): Principal component approach to handling multicollinearity in regression analysis. Ph.D. dissertation, Department of Statistics, Southern Methodist University, Dallas, Texas.

MANSFIELD E.R. AND CONERLY M.D.(1987):. Diagnostic value of residual and partial residual plots. *The American Statistician,* 41, 107-116.

MANSFIELD E.R. AND HELMS B.P.(1982): Detecting multicollinearity. *The American Statistician,* 36, 158-160.

MANSFIELD E.R., WEBSTER J.T. AND GUNST R.F.(1977): An analytic variable selection technique for principal component regression. *Journal of the Royal Statistical Society, Series C*, 26, no. 1, 34-40.

MANTEL N.(1970): Why Stepdown procedures in variable selection. *Technometrics* 12, 621-625.

MANTEL N.(1987): Coping with collinearities using prior estimates of regression coefficients. *Rivista Di Statistica Applicata,* 20, 357-363.

MARDIA K.V., KENT J.T. AND BIBBY J.M.(1979): *Multivariate Analysis.* Academic Press, London.

MARONNA R., BUSTOS O. AND YOHAI V.(1979): Bias- and efficiency- robustness of general M-estimators for regression with random carriers. Smoothing techniques for curve estimation (Proc. Workshop, Heidelberg, 1979), 91-116.

MARQUARDT D.W.(1963): An algorithm for least-squares estimation of nonlinear parameters. *J. Soc. Indust. Appl. Math.,* 11, 431-441.

MARQUARDT D.W.(1970): Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation. *Technometrics*, 12, 591-612.

MARQUARDT D.W.(1980): You should standardize the predictor variables in your regression models (discussion of a paper by G. Smith and F. Campbell). *Journal of the American Statistical Association,* 75, 87-91.

MARQUARDT D.W., BENNETT R.G. AND BURRELL E.J.(1961): Least-squares analysis of Electron Paramagnetic Resonance Spectra. *Jour. Molec. Spectroscopy* 7, 269-279.

MARQUARDT D.W. AND SNEE R.D.(1974): Test statistics for mixture models. *Technometrics*, 16, 533-537.

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

MARQUARDT D.W. AND STANLEY R.M.(1979): Biased estimators for mixture models and smooth regression: Examples of driving toward the null hypothesis. Unpublished manuscript.

MASON R.L.(1986): Latent root regression: a biased regression methodology for use with collinear predictor variables. *Communications in Statistics, Part A - Theory and Methods,* 15, no. 9, 2651-2678.

MASON R.L. AND GUNST R.F.(1985): Outlier-induced collinearities. *Technometrics* 27, 401-407.

MASON R.L. AND GUNST R.F.(1985): Selecting principal components in regression. *Statistics & Probability Letters* , 3, 299-301.

MASON R.L. GUNST R.F. AND WEBSTER J.T.(1975): Regression analysis and problems of multicollinearity. *Communications in Statistics*, 4, 277-292.

MASON R.L., KEATING J.P., SEN P.K. AND BLAYLOCK N.W.(1989): Comparison of regression estimators using Pitman's measure of closeness. *Journal of Statistical Planning and Inference.* to appear.

MASSY W.F.(1965): Principal component 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.

McCABE G.P.(1978): Evaluation of regression coefficients using α-acceptability. *Technometrics*, 20, 131-139.

McCABE G.P.(1984): Principal Variables. *Technometrics*, 26, 137-144.

McCALLUM B.T.(1970): Artificial orthogonalization in regression analysis. *Review of Economics and Statistics*, 52, 110-113.

McCANN R.C.(1984): *Introduction to Linear Algebra.* Harcourt Brace Jonanovich, New York.

McCULLAGH P. AND NELDER J.A.(1983): *Generalized Linear Models.* Chapman and Hall, London.

McDONALD G.C.(1980): Some algebraic properties of ridge coefficients. *Journal of the Royal Statistical Society, Series B,* 42, no. 1, 31-34.

McDONALD G.C. AND GALARNEAU D.I.(1975): A Monte-Carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association*, 70, 407-416.

McDONALD G.C. AND SCHWING R.C.(1973): Instabilities of regression estimates relating air pollution to mortality. *Technometrics*, 15, 463-481.

McGIFFEN M.E. Jr, CARMER S.G. RUESINK W.G.(1988): Diagnosis and treatment of collinearity problems and variable selection in least-squares models. *Journal of Economic Entomology,* 81, 1265-70

MEETER D.A.(1966): On a theorem used in nonlinear least squares. *SIAM Journal on Applied Mathematics*, 14, 1176-1179.

MICKEY M.R., DUNN O.J. AND CLARK V.(1967): Note on the use of stepwise regression in detecting outliers. *Computers and Biomedical Research*, 1, 105-111.

MILLER A.J.(1990): *Subset Selection in Regression.* Chapman and Hall.

MILLER R.G.(1974): An unbalance jackknife. *Annals of Statistics*, 2, 880-891.

MILLER R.G.(1974): The jackknife: a review. *Biometrika*, 61, 1-15.

MILOVANOVIC G. V.(1984): Numerical methods and approximation theory. Papers from the conference held at the University of Nis, Faculty of Electronic Engineering, Nis, September 26-28, 1984.

MIRSKY L.(1960): Symmetric gauge functions and unitarily invariant norms. *Quarterly Journal of Mathematics*, 11, 50-59.

MITRA A. AND ALAM K.(1980): Measurement error in regression analysis. *Communinications in Statistics, Part A - Theory and Methods,* 9, no. 7, 717-723.

MOGYORODI J., VINCZE I. AND WERTS W.(1984): Statistics and probability. Proceedings of the third Pannonian symposium on mathematical statistics held in Visegrad, September 13-18, 1982.

MONTGOMERY D.C.(1982): *Introduction to linear regression analysis.* John Wiley & Sons, New York.

MONTGOMERY D.C. AND ASKIN R.G.(1981): Problems of nonnormality and multicollinearity for forecasting methods based on least squares. *AIIE Transactions,* 13, no. 2, 102-115.

MONTGOMERY D.C. AND PECK E.A.(1982): *Introduction to linear regression analysis.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York.

MONTGOMERY D.C., **MARTIN** E. AND PECK E.A.(1980): Interior analysis of the observations in multiple linear regression. *Journal of Quality Technology,* 12, 165-173.

MORAN P.A.P.(1971): Estimating structural and functional relationships. *Journal of Multivariate Analysis*, 1, 232-255.

MOSER C.A. AND SCOTT W.(1961): *British Towns.* Oliver & Boyd., Edinburgh.

MOSTELLER F. AND TUKEY J.(1977): *Data Analysis and Regression.* Addison-Wesley, Reading, Mass..

MUIRHEAD C.R.(1986): Distinguishing outlier types in time series. *Journal of the Royal Statistical Society*, *Series B*, 48, no. 1, 39-47.

MULLETT G.M.(1976): Why regression Coefficients have the wrong sign. *Journal of Quality Technology*, 8 121-126.

MULLETT G.M. AND **MURRAY** T.W.(1971): A New method for examining rounding error in least squares regression Computer programs. *Journal of the American Statistical Association,* 66, 496-498.

MUNDLAK Y.(1981): On the concept of nonsignificant functions and its implications for regression analysis. *J. Econom.*, 16, 139-150.

MURPHY J.L.(1973): Corrective procedures for selected cconometric problems. *Introductory Econometrics.* (Illinois: R.D. Irwin Inc., 1973)

MYERS R.H.(1971): *Response Surface Methodology.* Allyn and Bacon, Boston.

MYERS R.H.(1986): *Classical and Modern Regression With Applications.* Duxbury Press, Boston.

NAES T. AND MARTENS H.(1985): Comparison of prediction methods for multicollinear data. *Communications in Statistics, Series B - Simulation and Computation,* 14, 545-576.

NELDER J.A.(1972): Discussion of a paper by D.V. Lindley and A.F.M. Smith. *Journal of the Royal Statistical Society, Series B,* 34, 18-20.

NETER J. AND WASSERMAN W.(1974): *Applied Linear Statistical Models.* Irwin, Inc. Illinois.

NEWHOUSE J.P. AND OMAN S.D.(1971): An evaluation of ridge estimators. *Technical report* No. R-716-PR, The Rand Corporation, Santa Monica, Calif.

NICHOLLS D.F. AND QUINN B.G.(1980): The estimation of random coefficient autoregressive models. I. *Journal of Time Series Analalysis*, 1, no.1, 37-46.

NOMURA M.(1988): On the almost unbiased ridge regression estimator. *Communications in Statistics, Series B - Simulation and Computation*, 17, 729-743.

NOMURA M. AND OHKUBO T.(1985): A note on combining ridge and principal component regression. *Communications in Statistics, Series A - Theory and Methods,* 14, 2489-2493.

NOVICK M.R., JACKSON P.H., THAYER D.T. AND COLE N.S.(1972): Estimating multiple regressions in m-group; a cross-validation study. *British Journal of Mathematical and Statistical Psychology,* 25.

NYQUIST H.(1988): Applications of the jackknife procedure in ridge regression. *Computational Statistics & Data Analysis* , 6, 177-183.

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

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

OBENCHAIN R.L.(1977): Letter to the editor. *Technometrics*, 19, 348-349.

OBENCHAIN R.L.(1978): Good and optimal ridge estimators. *Annals of Statistics*, 6, 1111-1121.

OBENCHAIN R.L.(1980): Formulas for generalized ridge regression computation. Unpublished manuscript.

OBENCHAIN R.L. AND VINOD H.D.(1974): Estimates of partial derivatives from ridge regression on ill-conditioned data. NBER-NSF Seminar on Bayesian Inference in Econometrics, Ann Arbor, Mich.

O'HAGAN J. AND McCABE B.(1975): Tests for the severity of multiclollinearity in regression analysis: A comment. *The Review of Economics and Statistics*, 57, 368-370.

OHTANI K.(1986): On small sample properties of the almost unbiased generalized ridge estimator. *Communication in Statistics, Part A - Theory and Methods*, 15, 1571-1578.

OLDFORD R.W,(1987): On the N-dimensional geometry of regression diagnostics. *Communications in Statistics*, *Part A - Theory and Methods*, 16, 2517-2540.

O'LEARY D.P. AND RUST B.W.(1986): Confidence intervals for inequality-constrained least squares problems, with applications to ill-posed problems. *Society for Industrial and Applied Mathematics. Journal on Scientific and Statistical Computing,* 7, no. 2,473-489.

OMAN S.D.(1978): A Bayesian comparison of some estimators used in linear regression with multicollinear data. *Communications in Statistics*, *Part A - Theory and Methods*, 7, no. 6, 517-534.

OMAN S.D.(1981): A confidence bound approach to choosing the biasing parameter in ridge regression. *Journal of the American Statistical Association,* 76, no. 374, 452-461.

OMAN S.D.(1984): A different empirical Bayes interpretation of ridge and Stein estimators. *Journal of the Royal Statistical Society*, *Series B*, 46, 544-557.

O'NEILL M., SINCLAIR J.G. AND SMITH F.J.(1969): Polynomial curve fitting when abscissas and ordinates are both subject to error. *Comput. J.*, 12, 52-56.

ORRIS J.B.(1982): The role of microcomputers in statistical computing. In J.S. Rustagi and D.A. Wolfe (Eds.), *Teaching of Statistics and Statistical Computing,* Academic Press, New York.

OSBORNE M.R.(1985): *Finite algorithms in optimization and data analysis.* John Wiley & Sons, Chichester.

OSBORNE M.R. AND WATSON G.A.(1985): An analysis of the total approximation problem in separable norms, and an algorithm for the total l_1 problem. Society for Industrial and Applied Mathematics. *Journal on Scientific and Statistical Computing,* 6, no. 2, 410-424.

OUELLETTE D.V.(1981): Schur complement and statistics. *Linear Algebra Appl.*, 36, 187-295.

OZAKI T.(1981): Nonlinear phenomena and time series models. Proceedings of the 43rd session of the International Statistica Insitute, Vol. 3 (Buenos Aires, 1981). With a discussion. *Bulletin of the International Statistical Institute,* 49, no. 3, 1193-1210, 1225-1230.

OZAKI T.(1982): The statistical analysis of perturbed limit cycle processes using nonlinear time series models. *Journal of Time Series Analysis,* 3, no. 1, 29-41

OZTURK F.(1984): A discrete shrinking method as alternative to least squares. Universite d'Ankara. Faculte des Sciences. Communications. Serie A_1. *Mathematiques,* 33, no. 22, 179-185 (1986).

PARK S.H.(1981): Collinearity and optimal restrictions on regression parameters for estimating responses. *Technometrics* 23, no. 3, 289-295.

PARLETT B.N.(1980): *The Symmetric Eigenvalue Problem.* Prentice Hall, Englewood Cliffs, N.J.

PAVEL-PARVU M. AND KORGANOFF A.(1969): Iteration functions for solving polynomial equations. *Constructive Aspects of the Fundamental Theorem of Algebra.* B. Dejon and P. Henrici, eds., John Wiley & Sons, New York.

PEARCE D.K. AND REITER S.A.(1985): Regression strategies when multicollinearity is a problem: A methodological note. *Journal of Accounting Research,* 23, 405-407.

PEARSON K.(1901): On lines and planes of closest fit to points in space. *Phil. Mag.*, 2, 559-572.

PEDDADA S.D.(1985): A Short note on Pitman's measure of nearness. *The American Statistician*, 39, 298-299.

PEDDADA S.D., NIGAM A.K. AND SAXENA A.K.(1989): On the inadmissibility of ridge estimator in a linear model. *Communications in Statistics, Part A - Theory Methods*, 18, 3571-3585.

PEELE L. AND RYAN T.(1982): Minimax regression estimators with Application to Ridge Regression. *Technometrics*, 24, 157-159.

PEMBERTON J. AND TONG H.(1981): A note on the distributions of nonlinear autoregressive stochastic models. *Journal of Time Series Analysis*, 2, no 1, 49-52.

PENROSE R.(1955): A generalized inverse for matrices. *Proc. Cambridge Philos. Soc.*, 51, 506-513.

PENROSE R.(1956): On best approximate solution of linear matrix equations, *Proc. Cambridge Philos. Soc.,* 52, 17-19.

PEREYRA V.(1969): Stability of general systems of linear equations. *Aequationes mathematicae,* Vol. 2, 194-206.

PETERS G. AND WILKINSON J.H.(1970): The least squares problem and pseudo-inverses. *The Computer Journal*, Vol 13, 309-316.

PHLIPS L.(1969): Business pricing policies and inflation: Some evidence from EEC countries. *Journal of Industrial Economics,* 18, no. 1, 1-14.

PHILLIPS P.C.B.(1984): The exact distribution of the Stein-rule estimator. *Journal of Econometrics*, 25, 591-612.

PITMAN E.J.G.(1937): The closest estimates of the statistical parameters. *Proc. of Cambridge Phil. Soc.*, 33, 212-223.

PLACKETT R.L.(1950): Some theorems in least squares. *Biometrika,* 37, 149-157.

POIRIER D.J.(1976): *The Econometrics of Structural Change.* North-Holland Publishing Company, Amsterdam.

POLASEK W.(1984): Regression diagnostics for general linear regression models. *Journal of the American Statistical Association,* 79, no. 386, 336-340.

POLASEK W.(1987): Bounds on rounding errors in linear regression models. *The Statistician*, 36, 221-227.

POPE P.T. AND WEBSTER J.T.(1972): The use of an F-statistic in stepwise regression procedures. *Technometrics,* 14, 327-340.

POWELL D.R. AND MacDONALD J.R.(1972): A Rapidly convergent iterative method for the solution of the generalized nonlinear least squares problem. *The Computer Journal* 15, 148-155.

PREGIBON D.(1981): Logistic regression diagnostics. *Annals of Statistics,* 9, no. 4, 705-724.

PRESS S.J.(1987): The MISER criterion for imbalance in the analysis of covariance. *Journal of Statistical Planning and Inference,* 17, no. 3, 375-388.

PRESS W.H., FLANNERY B.P., TEUKOLSKY S.A. AND VETTERLING W.T. (1985): *Numerical Recipes: The Art of Scientific Computing.* Cambridge University Press.

PRICE C.M.(1964): The matrix pseudoinverse and minimal variance estimates. *SIAM Review* 6, 115-120.

QUENOUILLE M.H.(1956): Notes on bias in estimation. *Biometrika,* 43, 353-360.

RADUCHEL W.J.(1971): Multicollinearity once again. Harvard Institute of Economic, *Research Paper* No. 205.

RAIFFA H. AND SCHLAIFER R.(1961): *Applied Statistical Decision Theory,* Harvard University, Boston, Chapters 11 and 13.

RANDALL J.H. AND RAYNER A.A.(1987): The accuracy of least squares calculations with the Cholesky algorithm. *Technical report*, University of Natal.

RAO C.R.(1962): A Note on a Generalized Inverse of a Matrix with Applications to Problems in Mathematical Statistics. *Journal of the Royal Statistical Society, Series B,* 24, 152-158.

RAO C.R.(1973): *Linear Statistical Inference and Its Applications, 2nd Edition.* John Wiley & Sons. New York.

RAO C.R., KEATING J.P. AND MASON R.L.(1986): Pitman nearness criterion and its determination. *Communications in Statistics, Part A - Theory and Methods*, 15, 3173-3191.

RAO C.R. AND MITRA S.K.(1971): *Generalized Inverse of Matrices and Its Applications.* John Wiley & Sons, New York.

RAWLINGS J.(1988): *Applied regression analysis: a research tool.* Wadsworth & Brooks/Cole: Pacific Grove, California.

REEDS J.A.(1978): Jackknifing maximum likelihood estimates. *The Annals of Statistics*, 6, 727-739.

REILLY P.M. AND PATINO-LEAL H.(1981): A Bayesian study of the error-in-variables models. *Technometrics,* 23, no. 3, 221-231.

REVANKAR M.S.(1974): Some finite sample results in the context of two seemingly unrelated regression equations. *Journal of the American Statistical Association*, 69, 187-190.

RICHARDSON D.H. AND DE-MIN W.(1970): Least squares and grouping method estimators in the errors-in-variables models. *Journal of the American Statistical Association*, 65, 724-748.

RIEDER H.(1987): Robust regression estimators and their least favorable contamination curves. *Statistics & Decisions*, 5, no. 3-4, 307-336.

RIGGS D., GUARNIERI J. AND ADELMAN S.(1978): Fitting straight lines when both variables are subject to error. *Life Science*, 22, 1305-1360.

RILEY J.D.(1955): Solving systems of linear equations with a positive definite, symmetric, but possibly ill-conditioned matrix. *Mathematics of Computation*, 9, 96-101.

RINNE H.(1984): A method of choosing additional sets of observations in multiple linear regression models to overcome multicollinearity. In *Multicollinearity and Biased Estimation,* edited by Josef Gruber.

ROBERTSON A.(1955): Prediction equations in quantitative genetics. *Biometrics,* 11, 95-98.

ROLLER V.F.(1988):. Adjusted variables: An important tool for teaching regression in the applications curricula. *Journal of Applied Statistics,* 15, 85-95

ROLPH J.E.(1976): Choosing Shrinkage estimators for regression problems. *Communications in Statistics, Part A - Theory and Methods*, 5, 789-801.

RONNER A.E.(1983): Perturbation and duality in linear models. VII. Symposium on operations research, Sektionen 4-9 (St. Gallen, 1982).

RUPERT D. AND CARROLL R.J.(1980): Trimmed least squares estimation in the linear model. *Journal of the American Statistical Association,* 75, 828-838.

RUSHTON S.(1951): On least squares fitting by orthogonal polynomials using the Cholesky method. *Journal of the Royal Statistical Society, Series B,* 13, 92-99.

RUTISHAUSER H.(1968): Once Again: The Least Square Problem. *Linear Algebra and Its Applications* 1, 479-488.

RYAN T.A.Jr, JOINER B.L. AND RYAN B.F.(1976): *Mini-tab Student Handbook.* Duxbury press, North Scituate, Mass.

SACHS W.H.(1976): Implicit multifunctional nonlinear regression analysis. *Technometrics,* 18, 161-173.

SAGER T.W. AND THISTED R.A.(1982): Maximum likelihood estimation of isotonic modal regression. *Annals of Statistics*, 10, no. 3, 690-707.

SARHAN A.E., GREENBERG B.G. AND ROBERTS E.(1962): Modified square root method of matrix inversion. *Technometrics*, 4, 282-287.

SASTRY M.V.R.(1970): Some limits in the theory of multicollinearity. *The American Statistician,* 24, 39-40.

SAXENA A.K.(1980): Principal components and its use in regression analysis: the problem revisited. *Statistica (Bologna),* 40, no. 3, 363-368.

SCHAEFER R.L.(1986): Alternative estimators in logistic regression when the data are collinear. *Journal of Statistical Computation and Simulation,* 25, 75-91.

SCHALL R. AND DUNNE T.T.(1987): Influential variables in linear regression. (to appear in *Technometrics*, 1990).

SCHALL R. AND DUNNE T.T.(1987): Variance inflation and collinearity in regression. *Technical Report* 5/87, Institute for Biostatistics of the South African Medical Research Council, Tygerberg, Republic of South Africa.

SCHEFFE H.(1959): *The Analysis of Variance.* John Wiley & Sons, New York.

SCHNEEWEISS H.(1976): Consistent estimation of a regression with errors-in-variables. *Metrica,* Band 23, 101-115.

SCHOENSTADT A.L., FAULKNER F.D., FRANKER. AND RUSSAK I.B.(1980): Information linkage between applied mathematics and industry. II. Proceedings of the Second Annual Workshop held in Monterey, Calif., February 22-24, 1979.

SCHWETLICK H. AND TILLER V.(1985): Numerical Methods for estimating Parameters in nonlinear models with errors in the variables. *Technometrics*, 27, 17-24.

SCLOVE S.L.(1968): Improved estimators for coefficients in linear regression. *Journal of the American Statistical Association*, 63, 596-606.

SCOTT D.T. BRYCE G.R. AND ALLEN D.M.(1985): Orthogonalization-triangulation methods in statistical calculations. *The American Statistician*, 39, 128-135.

SCOTT J.T. Jr(1966): Factor analysis and regression. *Econometrica*, 34, 552-562.

SEARLE S.R.(1971): *Linear Models.* John Wiley & Sons, New York.

SEARLE S.R.(1981): *Matrix Algebra Useful for Statistics.* John Wiley & Sons, New York .

SEBER G.A.F.(1977): *Linear Regression Analysis.* John Wiley & Sons, New York.

SEIFERT H.G.(1977): Multicollinearity and the prediction error. *Statistiche Hefte,* 18, no. 4, 233-253.

SHAO J.(1987): *On resampling methods for Variance estimation and related topics*, unpublished Ph.D. thesis, University of Wisconsin-Madison, Dept. of Statistics.

SHAO J.(1987): Sampling andd resampling: An efficient approximation to Jackknife variance estimators. *Technical Report* 799, University of Wisconsin-Madison, Dept. of Statistics.

SHAO J. AND WU C.F.J.(1986): Some general theory for the Jackknife. *Technical Report* 797, University of Wisconsin-Madison, Dept. of Statistics.

SHILLER R.J.(1973): A distributed lag estimator derived from smoothness priors. *Econometrica,* 41, 775-788.

SIDIK S.M.(1975): Comparison of some biased estimation methods (including Ordinary Subset Regression) in the Linear Model. Technical Report No. NASA TN D-7932, National Aeronautics and Space Administration, Lewis Research Center, Cleveland.

SILVEY S.D.(1969): Multicollinearity and imprecise estimation. *Journal of the Royal Statistical Society*, *Series* B, 31, 539-552.

SIMON S.D. AND LESAGE J.P.(1988): The impact of collinearity involving the intercept term on the numerical accuracy of regression. *Computer Science in Economics and Management* 1, 137-152.

SIMON S.D. AND LESAGE J.P.(1988): Benchmarking numerical accuracy of statistical algorithms, forthcoming in *Computational Statistics and Data Analysis.*

SIMONOFF J.S. AND TSAI C.(1989): The use of guided reformulations when collinearities are present in nonlinear regression. *Journal of the Royal Statistical Society*, *Series C*, 38, no. 1, 115-126.

SINGH B. AND CHAUBEY Y.P.(1987): On some improved ridge estimators. *Statistiche Hefte,* 28, 53-67.

SINGH B., CHAUBEY Y.P. AND DWIVEDI T.D.(1986): An almost unbaised ridge estimator. *Sankhya, Series B,* 48, 342-346.

SIOTANI M., HAYAKAWA T. AND FUJIKOSHI Y.(1985): *Modern multivariate statistical analysis: a graduate course and handbook.* American Sciences Press, Columbus, Ohio.

SLOTTJE D.J. AND BASMANN R.L.(1986): *Innovations in quantitative economics:* essays in honor of Robert L. Basmann. Edited and with an introduction by Daniel J. Slottje.

SMITH A.F.M. AND SPIEGELHALTER D.J.(1980): Bayes Factors and choice criteria for linear models. *Journal of the Royal Statistical Society, Series B*, 42, 213-220.

SMITH G.(1974): Multicollinearity and forecasting. Cowles Foundation Discussion Paper No. 383.

SMITH G.(1980): An example of ridge regression difficulties. *Canadian Journal of Statistics*, 8, 217-225.

SMITH G. AND BRAINARD W.(1976): The Value of a priori information in estimating a financial model. *Journal of Finance*, 31, 1299-1322.

SMITH G. AND CAMPBELL F.(1980): A critique of some ridge regression methods. With comments by Ronald A. Thisted, Donald W. Marquardt, R. Craig Van Nostrand,D. V. Lindley, Robert L. Obenchain, Lawrence C. Peele, Thomas P. Ryan, H. D. Vinod and Richard F. Gunst, and with a reply by the authors. *Journal of the American Statistical Association,* 75, no. 369, 74-103.

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 I. Developing prediction equations. *Journal of Quality Technology*, 5, 67-79.

SNEE R.D.(1977): Validation of regression models: methods and examples. *Technometrics*, 19, 415-428.

SNEE R.D.(1983): Review of *Regression diagnostics: Identifying Influential data and Sources of Collinearity*, by D.A. Belsley, E.Kuh and R.E. Welsch. *Journal of Quality Technology*, 15, 149-153.

SNEE R.D.(1983): Discussion of "Developments in Linear Regression Methodology: 1959-1982" by R.R. Hocking. *Technometrics*, 25, 230-237.

SNEE R.D. AND MARQUARDT D.W.(1984): Collinearity diagnostics depend on the domain of prediction, the model, and the data (Comment on "Demeaning condition diagnostics through centering" by D.A. Belsley). *The American Statistician,* 38, 83-87.

SNEE R.D. AND RAYNER A.A.(1982): Assessing the accuracy of Mixture Model Regression calculations. *Journal of Quality Technology,* 14, 67-79.

SPARKS R.S.(1987): Evaluating prediction procedures in multivariate regression: A re-sampling approach. *South African Statistical Journal*, Vol. 21, 63-98.

SPJOTVOLL E.(1972): Multiple comparison of regression functions. *Annals of Mathematical Statistics*, 43, 1076-1088.

SPRENT P.(1966): A generalized least-squares approach to linear functional relationships. *Journal of the Royal Statistical Society, Series B,* 28, 278-297.

STARKS T.H. AND FANG J.H.(1982): On the estimation of the generalized covariance function. *J. Internat. Assoc. Math. Geol.,* 14, no. 1, 57-64.

STEIN C.(1960): "Multiple Regression", in *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling,* ed. Ingrim Olkin, Stanford University Press., Stanford, Calif.

STEIN C.M.(1962): Confidence sets for the mean of a multivariate normal distribution. *Journal of the Royal Statistical Society, Series* B, 24, 265-296.

STEWART G.W.(1969): On the continuity of the generalized inverse. *SIAM Journal of Applied Mathematics*, 17, 33-45.

STEWART G.W.(1973): Error and perturbation bounds for subspaces associated with certain eigenvalue problems. *SIAM Review,* 15, 727-764.

STEWART G.W.(1973): *Introduction to matrix computations.* Academic Press, New York.

STEWART G.W.(1977): On the perturbation of pseudo-inverses, projections, and linear least squares problems. *SIAM Review*,. 19, 634-666.

STEWART G.W.(1977): Sensitivity coefficients for the effects of errors in the independent variables in a linear regression. Technical Report TR-571, Department of Computer Science, University of Maryland, College Park MD.

STEWART G.W.(1979):. The effects of rounding error on an algorithm for downdating a Cholesky factorization. *J. Inst. Math. Appl.,* 23, no. 2, 203-213.

STEWART G.W.(1983): A Nonlinear version of Gauss's minimum variance theorem with applications to an errors-in-the-variables model. Computer Science *Technical Report* TR-1263, Univ. of Maryland, 1983.

STEWART G.W.(1984): On the asymptotic behavior of scaled singular value and QR decompositions. *Mathematics of Computation,* 43, no 168, 483-489.

STEWART G.W.(1984): On the invariance of perturbed null vectors under column scaling. *Numer. Math.*, 44, no. 1, 61-65.

STEWART G.W.(1984): Rank degeneracy. *SIAM Journal on Scientific and Statistical Computing*, 5, 403-413.

STEWART G.W.(1987): Collinearity and least squares regression. With discussion by D.A. Belsley, A.S. Hadi, D.W. Marquardt, P.F. Velleman, R.A. Thisted, and with a reply by the author. *Statistical Science.* 2, no. 1, 68-100.

STEWART G.W. AND ALLEN D.M.(1986): Collinearity, scaling, and rounding error. Computer Science and Statistics: Proceedings of the 17th Symposium on the Interface.

ST. JOHN R.C.(1984): Experiments with mixtures, ill-conditioning and ridge regression. *Journal of Quality Technology,* 16, 81-96.

STONE R.(1945): The analysis of market demand. *Journal of the Royal Statistical Society, Series A,* 108, 286-382.

STRAWDERMAN W.E.(1978): Minimax adaptive generalized ridge regression estimators. *Journal of the American Statistical Association*, 73, 623-627.

STROUD W.F.(1972): Comparing conditional means and variances in a regression model with measurement errors of known variances. *Journal of the American Statistical Association*, 407-412.

SWAMY P.A.V.B., MEHTA J.S., THURMAN S.S. AND IYENGAR N. S.(1985): A generalized multicollinearity index for regression analysis. *Sankhya, Series* B, 47, no. 3, 401-431.

SWAMY P.A.V.B. AND MEHTA J.S.(1983): Ridge regression estimation of the Rotterdam model. *Journal of Econometrics,* 22, 365-390.

SWAMY P.A.V.B. AND MEHTA J.S.(1985): On a neglected measure of multicollinearity. *Special Studies Paper*, Federal Reserve Board, Washingtion, DC.

SWAMY P.A.V.B., MEHTA J.S. AND RAPPOPORT R.N.(1978): Two methods of evaluating Hoerl and Kennard's ridge regression. *Communications in Statistics*, A7, 1133-1165.

SWINDEL B.F.(1976): Good ridge estimators based on prior information. *Communications in Statistics,* 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.

SWINDEL B.F. AND CHAPMAN D.D.(1973): Good ridge estimators. Proc. Joint Statistical meetings, Dec 1973, pp 126.

TAO G.C. AND ZELLNER A.(1964): Bayes theorem and the use of prior information in regression analysis. *Biometrika*, 51, 219-230.

TAUBMAN S.B.(1978): A comparison of the accuracy of certain least squares procedures, *Proceedings of the Statistical Computing Section, American Statistical Association*, 165-166.

TAWIL J.J.(1972): The linear structural relationship. Unpublished paper 1972.

TAYLOR J.M.G.(1989): A note on the cost of estimating the ratio of regression parameters after fitting a power transformation. *Journal of Statistical Planning Inference,* 21, 223-230

THEIL H.(1963): On the use of incomplete prior information in regression analysis. *Journal of the American Statistical Association*, 58, 401-414.

THEIL H.(1971): *Principles of Econometrics.* John Wiley & Sons,, New York.

THEIL H.(1975): *Theory and Measurement of Consumer Demand,* 1. North-Holland Publishing Company.

THEIL H. AND GOLDBERGER A.S.(1961): Pure and mixed statistical estimation in economics. *International Economic Review*, 2, 65-78

THEOBALD C.M.(1974): Generalization of mean square error applied to ridge regression. *Journal of the Royal Statistical Society, Series B*, 36, 103-106.

THISTED R.A.(1976): Ridge regression, minimax estimation and empirical Bayes methods. Ph.D. Thesis, Stanford University, Dept. of Statistics.

THISTED R.A.(1978): Multicollinearity, Information, and Ridge Regression. *Technical Report* No. 66, University of Chicago, Dept. of Statistics.

THISTED R.A.(1980): Comment. *Journal of the American Statistical Association,* 75, 81-86.

THISTED R.A. AND MORRIS C.N.(1979): Theoretical results for adaptive ordinary ridge regression estimators. *Technical Report* No. 94, University of Chicago, Dept.of Statistics.

THOMPSON B. AND BORRELLO G.M.(1985): The importance of structure coefficients in regression research. *Educational and Psychological Measurement,* 45, 203-209.

THOMPSON M.L.(1978): Selection of variables in multiple regression: Part I. A review and evaluation. *International Statistical Review*, 46. 1-20.

THOMPSON M.L.(1978): Selection of variables in multiple regression: Part II. Chosen procedures, computations and examples. *International Statistical Review*, 46. 129-146.

THURMAN S.S., SWAMY P.A.V.B. AND MEHTA J.S.(1984): An examination of distributed lag model coefficients estimated with smoothness priors. *Special Studies Paper*, Federal Reserve Board, Washington. DC.

TIAO G.C. AND ZELLNER A.(1964): Bayes's theorem and the use of prior knowledge in regression analysis. *Biometrika,* 51, 219-230.

TIHONOV A.N., KUHNERT F., KUZNECOV N.N., MOSZYNSKI K. AND WAKULICZ A.(1978): Mathematical models and numerical methods. Papers from the Fifth Semester held at the Stefan Banach International Mathematical Center, Warsaw, February - June 1975. Banach Center Publications, 3.

TINTNER G., RAO J.N.K. AND STRECKER H.(1978): *New results in the variate difference method.* Vandenhoeck & Ruprecht, Gottingen.

TORO-VIZCARRONDO C.E. AND WALLACE T.D.(1968): A test of the mean square error criterion for restrictions in linear regression. *Journal of the American Statistical Association*, 63, 558-572.

TOUTENBURG H.(1982): *Prior information in linear models.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York.

TRENKLER G.(1980): Generalized mean square error comparisons of biased regression estimators. *Communications in Statistics, Part A - Theory and Methods,* 9, 1247-1259.

TRENKLER G.(1984): Some further remarks on multicollinearity and the minimax conditions of the Bock-Stein-like estimator. *Econometrica*, 52, no. 4, 1067-1069.

TRENKLER D. AND TRENKLER G.(1984): Minimum mean square error ridge estimation. *Sankhya*, *Series* A, 46, no. 1, 94-101.

TRENKLER D. AND TRENKLER G.(1984): On the Euclidean distance between biased estimators. *Communications in Statistics, Part A - Theory and Methods,* 13, no. 3, 273-284.

TRIPP R.E.(1983): Nonstochastic ridge regression and effective rank of the regressors matrix. Unpublished Ph.D. dissertation, Virginia Polytechnic Institute and State University, Dept. of Statistics.

TROSKIE C.G. AND CONRADIE W.J.(1986): The distribution of the ratios of characteristic roots (condition numbers) and their applications in principal component or ridge regression. *Linear Algebra and its Applications,* 82 *255-279*

TSAI C.L.(1986): Score test for the first-order autoregressive model with heteroscedasticity. *Biometrika,* 73, no. 2, 455-460.

TUKEY J.W.(1958): Bias and confidence in not quite large samples. *The Annals of Mathematical Statistics,* 29, 614.

TUKEY J.W.(1972): Data analysis, computation and mathematics. *Quarterly of Applied Mathematics*, 30, 51-65.

TUKEY J.W.(1972): Some graphic and semigraphic displays, in *Statistical papers in Honor of George W. Snedecor,* ed. T.A. Bancroft, Iowa State University Press., Ames, Iowa.

TUKEY J.W.(1975): Instead of Gauss-Markov least squares, what?, in *Applied Statistics*, ed R.P. Gupta, Amsterdam: North-Holland Publishing Co., 352-372.

TUKEY J.W.(1977): *Exploratory data analysis.* Addison Welsley Publishing Co., Reading, Mass.

TURING A.M.(1948): Rounding-off errors in matrix processes. *Quart. J. Mech. Appl. Math.* 1, 287-308.

ULLAH A., VINOD H.D. AND KADIYALE R.K.(1981): A family of improved shrinkage factors for the ordinary ridge estimator. E.G. Charatsis ed., *Proceedings of the Econometric Society, European Meeting* 1979. Amsterdam: North-Holland, 259-277.

VAN DEN BOS, A.(1981): Degeneracy in nonlinear least squares. *Proc. IEE-D,* 128, no. 3, 109-116.

VAN DER MEER R., LINSSEN H.N. AND GERMAN A.L.(1978): Improved methods of estimating Monomer reactivity ratios in Copolymerization by considering experimental errors in both variables. *J. Polymer SCI. PCE*, 16, 2915-2930.

VAN DER SLUIS A.(1969): Condition numbers and equilibration of matrices. *Numer. Math.* 14, 14-23.

VAN HUFFEL S.(1985): A reliable, efficient deconvolution technique based on total linear least squares for calculating the renal retention function. Master's Thesis in Biomedical Engineering, Katholieke Universiteit Leuven.

VAN HUFFEL S. AND VANDEWALLE J.(1985): The use and applicability of the total least squares technique in linear regression analysis. Internal Report, Esat Lab., Dept. of Electrical Engineering, K.U. Leuven Belgium, 1985.

VAN HUFFEL S. AND VANDEWALLE J.(1985): The use of total least squares techniques for identification and parameter estimation. Preprints Proceedings of the 7th IFAC Symposium on Identification and System parameter Estimation, York, U.K., 3-7 July, 1167-1172.

VAN HUFFEL S. AND VANDEWALLE J.(1987): Algebraic relationships between classical regression and total least-squares estimation. *Linear Algebra and its Applications*, 93, 149-160.

VAN HUFFEL S. AND VANDEWALLE J.(1987): Analysis and solution of the nongeneric total least-squares problem. *SIAM Journal on Matrix Analysis and Application*, 9, No. 3, 360-372.

VAN HUFFEL S., VANDEWALLE J. AND STAAR J.(1984): The total linear least squares problem: properties, applications and generalization. Submitted to *SIAM Journal on Numerical Analysis.*

VAN LOAN C.(1979): On Stewart's singular value decomposition for partitioned orthogonal matrices. Department of Computer Science Report STAN-CS-79-767. Stanford University, Stanford CA.

VAN NOSTRAND R.C.(1977): *Some Distributional Properties and Comparisons of Shrinkage Estimators*. PhD thesis, University of Wisconsin-Madison, Dept. of Statistics.

VELLEMAN P.F. AND VELLEMAN A.Y.(1969): *The Data Desk Handbook*. Data Description, Ithaca, N.Y.

VELLEMAN P.F. AND WELSCH R.E.(1981): Efficient computing of regression diagnostics. *The American Statistician*, 35, 234-242.

VELLEMAN P.F. AND YPELAAR M.A.(1980): Constructing regressions with controlled features: a method of probing regression performance. *Journal of the American Statistical Association*, 75, no. 372, 839-844.

VILLEGAS C.(1966): On the asymptotic efficiency of least squares estimators. *Annals of Mathematical Statistics*, 37, 1676-1683.

VINOD H.D.(1976): 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.(1976): Canonical ridge and econometrics of joint production. *Journal of Econometrics*, 4, 147-166.

VINOD H.D.(1976): Simulation and extension of the minimum mean square error estimator in comparison with Stein's. *Technometrics*, 18, 491-496.

VINOD H.D.(1978): A Survey of ridge regression and related techniques for improvements over ordinary least squares. *The Review of Economics and Statistics,* 60, 121-131.

VINOD H.D.(1982): Enduring regression estimator. In *Time series analysis: theory and practice, 1* (Valencia, 1981).

VINOD H.D. AND ULLAH A.(1981): *Recent Advances in Regression Methods*. Marcell Dekker Inc., New York.

VUCHKOV I.N. AND BOYADZHIVA L.N.(1978): Regression analysis of Recycle Systems Data, Paper presented at the 6th Int. Congress of Chemical Engineering, Praha, Czechoslovakia.

WAHBA G.(1977): A Survey of some smoothing problems and the method of generalized Cross-Validation for solving them. *Applications of Statistics*, ed. Paruchuri R. Krishnaiah, New York: North-Holland Publishing Co., 507-523.

WALKER E.(1989): Detection of collinearity-influential observations. *Communications in Statistics*, *Part A - Theory and Methods,* 18, 1675-1690.

WALKER E. AND BIRCH J.B.(1988): Influence measures in ridge regression. *Technometrics,* 30, 221-227.

WALKER M.A.(1967): Some critical comments on "An analysis of crimes by the method of principal component analysis" by B. AHAMAD. *Applied Statistics,* 16, 36-39.

WALLACE T.D.(1972): Weaker criteria and tests for linear restrictions in regression. *Econometrica,* 40, 689-698.

WALLACE T.D. AND ASHAR V.G.(1972): Sequential methods in model construction. *Review of Economics and Statistics,* 54, 172-178.

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

WALLS R.C. AND WEEKS D.L.(1969): A Note on the variance of a predicted response in regression. *The American Statistician*, 23, 24-26.

WALSH J.E.(1959): Computer-feasible general method for fitting and using regression functions when data are incomplete. Report SP-71, System Development Corporation, Santa Monica, California.

VAMPLER R.H.(1970): A Report on the accuracy of some widely used least squares computer programs. *Journal of the American Statistical Association,* 65, 549-565.

VAMPLER R.H.(1980): Test procedures and problems for least-squares algorithms. *Journal of Econometrics*, 12, 3-22.

VARE J.H. (1972): The fitting of straight lines when both variables are subject to error and the ranks of the means are known. *Journal of the American Statistical Association*, 67, 891-897.

VARGA A.(1989): Experimental design in tests of linear factor models. *Journal of Business and Economic Statistics*, 7, 191-198.

VEBSTER J.T., GUNST R.F. AND MASON R.L.(1973): Recent developments in stepwise regression procedures. *Proc. Univ. Kentucky Conf. on Regression with a Large Number of Predictor Variables.* Lexington, Ky., Oct. 11-12, 1973, 34-53.

VEBSTER J.T., GUNST R.F. AND MASON R.L.(1974): Latent root regression analysis. *Technometrics*, 16, 513-522.

VEDIN P.A.(1969): On pseudo-inverses of perturbed matrices, Lund Univ. Comput. Sci. Tech. Rep., Lund, Sweden.

VEDIN P.A.(1973): Perturbation theory for pseudo-inverses. *BIT.* 13, 217-232.

VEISBERG S.(1980): *Applied Linear Regression.* John Viley & Sons, New York.

VEISBERG S.(1983): Some principles for regression diagnostics and influence analysis. *Technometrics*, 25, 240-244.

VELSCH R.E. AND KUH E.(1977): Linear regression diagnostics. Vorking paper 173, Cambridge, Mass.: National Bureau of Economic Research.

VERMUTH N.(1972): *An Empirical Comparison of Regression Methods.* Unpublished Ph.D. thesis, Department of Statistics, Harvard University, Cambridge, Mass., 1972.

VERMUTH N.(1972): *APL-Functions for Data Simulation, Regression Methods and Data Analysis Techniques.* Research Report CP-15, Department of Statistics, Harvard University, Cambridge, Mass., 1972.

VETHERILL G.B., DUNCOMBE P., KENWARD M., KOLLERSTROM, J., PAUL S.R. AND VOWDEN B.J.(1986): *Regression analysis with applications.* Chapman & Hall, London-New York.

VICHERN D.W. AND CHURCHILL G.A.(1978): A comparison of ridge estimators. *Technometrics*, 20, 301-311.

WILKINSON J.H.(1963): Rounding errors in algebraic processes. Prentice-Hall, Englewood Cliffs, N.J.

WILKINSON J.H.(1965): *The Algebraic Eigenvalue Problem.* Oxford University Press, London.

WILKINSON J.H.(1967): The solution of ill-conditioned linear equations in A. Ralston and H.S. Wilf, eds., *Mathematical Methods for Digital Computers,* Vol. 2, John Wiley & Sons, Inc., New York, 65-93.

WILKS S.S.(1932): Moments and distributions of estimates of population parameters from Fragmentary samples. *Annals of Mathematical Statistics*, 3, 163-195.

WILKS S.S.(1962): *Mathematical statistics.* John Wiley & Sons,, New York.

WILLAN A.R. AND WATTS D.G.(1978): Meaningful multicollinearity measures. *Technometrics*, 20, 407-412.

WOLD S., RUHE A., WOLD H. AND DUNN W.J.(1984): The collinearity problem in linear regression, the partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing,* 5, no. 3, 735-743.

WONNACOTT R.J. AND WONNACOTT T.H.(1979): *Econometrics,* 2nd ed. John Wiley & Sons, New York.

WOOD F.S.(1973): The use of individual effects and residuals in fitting equations to data. *Technometrics*, 15, 677-686.

WOOD F.S.(1984): Effect of centering on collinearity and interpretation of the constant (Comment on "Demeaning condition diagnostics through centering" by D.A. Belsley). *The American Statistician,* 38, 88-90.

WU C.F.J.(1986): Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, Vol. 14, No. 4, 1261-1295.

YANCEY T.A., JUDGE G.G. AND BOCK M.E.(1973): Wallace's weak mean square error criterion for testing linear restrictions in regression: A Tighter bound. *Econometrica,* 41, 1203-1206.

YANCEY T.A., JUDGE G.G. AND BOCK M.E.(1974): A Weak square error test when stochastic restrictions are used in regression. *Communications in Statistics*, 3, 755-769.

YORK D.(1966): Least squares fitting of a straight line. *Canad. J. Physics,* 44, 1079-1086.

YOSHIOKA S.(1986): Multicollinearity and avoidance in regression analysis. *Behaviormetrika,* 19, 103-120.

YOUNG A.S.(1982): The Bivar criterion for selecting regressors. *Technometrics,* 24, no. 3, 181-189.

ZARKOVICH S.S.(1966): Quality of statistical data. (Food and Agricultural Organization of the United nations), Rome.

ZELLNER A.(1962): An efficient method of estimating seemingly unrelated regressions and tests for Aggregation Bias. *Journal of the American Statistical Association*, 57, 348-368.

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

ZELLNER A.(1986): On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques* (P. Goel and A. Zellner, Eds,), pp 233-243. North-Holland, Amsterdam.

ZELLNER A. AND HUANG D.(1962): Further properties of efficient estimators for seemingly unrelated regression equations. *International Economic Review*, 3 300-313.

ZELLNER A. AND VANDAELE W.(1972): Bayes-Stein estimators for k-means regression and simulataneous equations. *H.G.B. Alexander Research Foundation*, Graduate School of Business, University of Chicago, 1972.