UNIVERSITY OF CAPE TOWN

DEPARTMENT OF MATHEMATICAL STATISTICS

****

A DETAILED INVESTIGATION OF THE LINEAR MODEL AND SOME OF ITS UNDERLYING ASSUMPTIONS

BY

D. COUTSOURIDES

****

A thesis prepared under the supervision of Professor C.G. Troskie in fulfilment of the requirements for the degree of Master of Science

****

P R E F A C E

The purpose of this thesis is to provide a study of the linear model. The whole work has been split into 6 chapters.

In Chapter 1 we define and examine the two linear models, i.e. the regression and the correlation model. More specifically we show that the regression model is the conditional version of the correlation model.

In Chapter 2 we deal with the problem of multicollinearity. We investigate the sources of near singularities, we give some methods of detecting the multicollinearity, and we state briefly methods for overcoming this problem.

In Chapter 3 we consider the least squares method with restrictions, and we dispose of some tests for testing the linear restrictions. The theory concerning the sign of least squares estimates is discussed, then we deal with the method for augmenting existing data.

Chapter 4 is mainly devoted to ridge regression. We state methods for selecting the best estimate for k. Some extentions are given dealing with the shrinkage estimators and the linear transforms of the least squares.

In Chapter 5 we deal with the principal components, and we give methods for selecting the best subset of principal components. Much attention was given to a method called fractional rank and latent root regression analysis.

In Chapter 6 comparisons were performed between estimators previously mentioned. Finally the conclusions are stated.

I wish to take this opportunity of expressing my deepest gratitude to Professor C.G. Troskie, Head of the Department of Mathematical Statistics of the University of Cape Town, without whose guidance, encouragement and support this thesis would not have been possible.

I would also like to express my thanks to Associate-Professor A.H. Money, who encouraged me at the start of my studies.

Thanks are also due to Mrs. M.I. Cousins for her efficient and excellent typing.

I would further like to acknowledge the scholarship received from the University of Cape Town.

D. Coutsourides

C O N T E N T S

		page
CHAPTER ONE :	THE REGRESSION AND CORRELATION MODELS	1
1.1	Introduction	1
1.2	The Regression Model (Model A)	2
1.3	The Correlation Model or Model B	5
1.4	Estimation of the regression coefficients and of σ^2	8
1.5	Distribution Theory	12
1.6	Hypothesis testing	26
CHAPTER TWO :	THE PROBLEM OF MULTICOLLINEARITY	38
2.1	Introduction	38
2.2	The sources of multicollinearity	39
2.3	The effects of multicollinearity	40
2.4	Linear combination of regression variables	42
2.5	Geometric picture	45
2.6	Detection of multicollinearity	48
2.7	Proposed solutions	51
CHAPTER THREE :	ORDINARY LEAST SQUARES WITH LINEAR CONSTRAINTS AUGMENTING EXISTING DATA IN LINEAR REGRESSION	54
3.1	Introduction	54
3.2	True restrictions in linear regression	54
3.3	Tests for linear restrictions	56
3.4	The sign of restricted O.L.S. estimates	63
3.5	Augmenting data in linear regression	66
3.6	Development of the augmenting procedure	67
CHAPTER FOUR :	RIDGE REGRESSION AND EXTENSIONS	74
4.1	Introduction	74
4.2	Properties of ridge regression	75
4.3	Ridge trace	77
4.4	Geometric picture of ridge regression	79
4.5	Mean square properties of ridge regression	81
4.6	Generalizations of the Mean Square error	90
4.7	The method of Hoerl and Kennard (1970)	94
4.8	The method of Marquardt (1970) and Marquardt and Snee (1973)	95
4.9	The method of Mallows (1973)	95
4.10	The method of McDonald and Galarleau (1975)	95
4.11	An explicit solution for generalized ridge regression W. Hemmerle (1975)	97
4.12	The method of Guilkey and Murphy (1975)	105

page

4.13 The method of Hoerl and Kennard (1976) 108
4.14 The method of Lawless and Wang (1976) 110
4.15 New Ridge regression - H. Vinod (1976) 111
4.16 A critical view of Ridge regression 116
4.17 Shrinkage estimators 118
4.18 Linear transforms of $\hat{\beta}$ 128

CHAPTER FIVE : PRINCIPAL COMPONENT REGRESSION 134

5.1 Introduction 134
5.2 Derivation of principal components from the correlation matrix 134
5.3 The method of Massy (1965) 139
5.4 The method of Kendall (1968) 140
5.5 The method of Marquardt (1970) 140
5.6 Fractional and modified fractional rank 150
5.7 The method of Greenberg (1975) 156
5.8 Latent root regression analysis 157

CHAPTER SIX : COMPARISONS AND CONCLUSIONS 167

6.1 Introduction 167
6.2 Comparing the ridge estimator and the generalized ridge estimator 167
6.3 Comparing the O.L.S. estimator and the L.R.R.A. 168
6.4 Comparing the fractional rank and the generalized ridge estimator 170
6.5 A geometric portrayal of some biased estimator when the predictors are two 171
6.6 Conclusions 173

BIBLIOGRAPHY 178

C H A P T E R 1

THE REGRESSION AND CORRELATION MODELS

1.1 Introduction

When analysing data by what is conveniently termed "regression analysis", we must very often make an implicit choice. We must decide at least for ourselves whether to view the independent variables as constants or as realizations of random variables. However, we are aware that our decision matters little because the analysis is essentially equivalent. This chapter deals mainly with the relationship between these two models and the important differences are highlighted.

Several authors have already considered this equivalence eg. Bartlett (1933), Rao (1952), Press (1972), Dempster (1969), and more recently excellent discussions are available by Sampson (1974) and Troskie (1976).

We restrict our attention to the cases where the randomness of the model has a multivariate normal distribution. When we deal with constant independent variables we refer to the model as the regression model or Model A. When we view the independent variables as realizations of random variables, we refer to the model as the correlation model or Model B.

1.2 The Regression Model (Model A)

Consider the density function $f(y,x_1,\ldots,x_p, \beta_0,\beta_1,\ldots,\beta_p)$ of a random variable y which depends on p known quantities $x_1,\ldots,x_p$ and $(p+1)$ unknown parameters $\beta_0,\beta_1,\ldots,\beta_p$ which we call the regression coefficients.

We assume that

$$(1.2.1) \quad E(y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p,$$

$$V(y) = \sigma^2$$

and σ^2 does not depend on the x's or β's.

Consider a random selection of y, say $(y_1,\ldots,y_n)$ from pre-selected x's, say (x_{ij}); $i = 1,\ldots,n$, $j = 1,\ldots,p$. Here the x's are fixed or non-random variables. Although the set of x's may vary from sample to sample, we do not expect that this variation will have any effect on the distribution of y. Strictly speaking the matrix (x_{ij}); $i = 1,\ldots,n$, $j = 1,\ldots,p$ should be considered fixed even in repeated samples.

It is customary to write the model (1.2.1) in the following form

$$(1.2.2) \quad Y = X\beta + e$$

where $Y' = (y_1,\ldots,y_n)$, $X = (x_{ij})$; $i = 1,\ldots,n$, $j = 0,1,\ldots,p$ where $x_{io} = 1$, $i = 1,\ldots,n$; $\beta' = (\beta_0,\beta_1,\ldots,\beta_p)$ and $e' = (e_1,\ldots,e_n)$. Furthermore

(1.2.3) $E(e) = 0,\ E(ee') = \sigma^2 I$

For testing purposes we make the additional assumption that e is distributed as multivariate normal, $e \sim N(0,\sigma^2 I)$ or, in fact $Y \sim N(X\beta,\sigma^2 I)$.

The above is generally considered as the general linear model, and detailed discussions can be found in many textbooks, notably that of Graybill (1967).

Two important versions of the above model will play an important role in what follows. The first one is the central model, which we will call Model Ac. Let the ith equation of (1.2.2) be denoted by

$$Y_i = \beta_0 + x_{i1}\beta_1 + \ldots + x_{ip}\beta_p + e_i.$$

Let $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and $\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{i.j}$, $\bar{X}' = (\bar{X}_1,..,\bar{X}_p)$. Then the above equation can be written as

(1.2.4) $$Y_i - \bar{Y} = \beta_1(\bar{X}_{i1} - \bar{X}_1) + \ldots + \beta_p(\bar{X}_{ip} - \bar{X}_p) + e_i$$

$$\cdot\ i = 1,\ldots,n$$

where $$\beta_0 = \bar{Y} - \beta_1\bar{X}_1,\ldots,-\beta_p\bar{X}_p.$$

For convenience sake we will also write this model as

$$Y = X\beta + e$$

but refer to it as Model Ac (i.e. the centered model). Here X is a $(n\times p)$ matrix while β is a $(p\times 1)$ vector. Assuming this model, it is equivalent to assume that the Model A,

given by (1.2.2), passes through the origin that is $\beta_o = 0$.

The second version is the standardized model named Model As. Let $X^*_{ij} = \frac{X_{ij}-\bar{X}_j}{S_{jj}^{\frac{1}{2}}}$, $Y^*_i = \frac{Y_i-\bar{Y}}{S_{YY}^{\frac{1}{2}}}$ where

$S_{jj} = \sum_{i=1}^{n}(X_{ij}-\bar{X}_j)^2$, $S_{yy} = \sum_{i=1}^{n}(Y_i-\bar{Y})^2$. The model (1.2.2) then takes the form

(1.2.5) $\quad Y^* = X^*\beta^* + e,$

where $\quad \beta^* = \begin{pmatrix}\beta_1^* \\ \vdots \\ \beta_p^*\end{pmatrix}$.

Here $X^{*\prime}X^*$ is a "correlation matrix" between the independent variables $(x_1,\ldots,x_p)$ and $X^{*\prime}Y^*$ is a vector whose elements are the "correlation coefficients" between Y and $(x_1,\ldots,x_p)$. Obviously the word "correlation" is ambiguous since $(x_1,\ldots,x_p)$ are not random variables, but it will be used for convenience sake. The β_j^* are often called the <u>Beta coefficients</u>.

If β^* and β are estimated by the Ordinary Least Sqaures Method (O.L.S.) then the relationships between them will be

(1.2.6) $\quad \hat{\beta}_j = \hat{\beta}_j^* \frac{S_{YY}^{\frac{1}{2}}}{S_{jj}^{\frac{1}{2}}}, \quad j = 1,\ldots,p$

and $\quad \hat{\beta}_o = \bar{Y} - \sum_{j=1}^{p}\hat{\beta}_j\bar{X}_j$.

1.3 The Correlation Model or Model B.

In Model B or the correlation (or random) model we assume that the vector $Z' = (\underset{\sim}{Y}, X_1, \ldots, X_p)$ has a multivariate normal distribution with mean μ and covariance matrix Σ, i.e. $Z \sim N(\mu, \Sigma)$. ($\underset{\sim}{Y}$ is scalar random variable to avoid confusion with the vector of observations Y.)

Let

$$(1.3.1) \qquad \mu = \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix}.$$

Now the conditional distribution of $\underset{\sim}{Y}$ given $X_1 = x_1, \ldots, X_p = x_p$ is also normal with mean

$$(1.3.2) \qquad \begin{aligned} E(\underset{\sim}{Y}/x) &= \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \\ &= \beta_0 + \beta_x^{(2)\prime} x \\ &= \mu_y + \beta^{(2)\prime}(x - \mu_x) \end{aligned}$$

where

$$(1.3.3) \qquad \beta^{(2)} = \Sigma_{xx}^{-1}\Sigma_{xy} \quad \text{and} \quad \beta_0 = \mu_y - \beta^{(2)\prime}\mu_x$$

and variance

$$(1.3.4) \qquad V(\underset{\sim}{Y}/x) = \sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy} = \Sigma_{yy \cdot x} = \sigma_B^2.$$

We let β be defined as

$$(1.3.5) \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta^{(2)} \end{pmatrix}$$

The important difference between Model A and Model B is

that in Model B we are working with the conditional distribution of $\underset{\sim}{Y}$ given $(X_1 = x_1,\ldots,X_p = x_p)$. In repeated samples not only is the variability of $\underset{\sim}{Y}$ of interest, but also the variability of $(X_1,\ldots,X_p)$. In a practical situation it is impossible to deal only with the conditional distribution since one cannot sample from the conditional distribution. For example one makes observations from the multivariate normal $N(\mu,\Sigma)$ and not from the conditional distribution of $\underset{\sim}{Y}/x$ which is $N(\beta_0+\beta^{(2)'}x,\sigma_B^2)$.

The following example by Sampson (1974) further illustrates this point.

> "Often, for experiments that result in a set of data vectors, a linear relationship is sought among the variables which can be used for inference and prediction.
>
> For example, we may wish to study the relationship between a student's college grade point average (GPA) and his high school GPA and "college board" scores. Typically the high school GPA and "college board" scores are considered to be independent variables; then the college GPA is called the dependent variable.
>
> The regression analysis treatment of such data requires viewing the independent variables as fixed and the dependent variable as being random (or as being the realization of a random variable once the data is collected). On the other hand, the multivariate analysis of regression treatment views the triplet (college GPA, high school

GPA, "college board" scores) as a trivariate random variable (or, again, as a realization thereof once the data is collected).

In this example, to treat the independent variables as fixed and the dependent variables as random appears debatable. There seems to be no reason to impose a qualitative difference between college GPA and high school GPA - they are both the same kind of variable. While this lack of qualitative difference apparently dictates use of multivariate analysis of regression, data of this sort is often approached using regression analysis (e.g. Draper and Smith). In fact, it is common to see either type of analysis used to analyze such data.

To this end, a restriction is imposed on the independent variables. We do not want the independent variables to be predetermined, i.e., we cannot beforehand choose the values at which to observe the dependent variables. If this could be done, it would make no sense to view the independent variables as the realizations of a (non-trivial) random variable."

For Model B the regression coefficients β_i, $i = 1,\ldots,p$ are the components of $\Sigma_{xx}^{-1}\Sigma_{xy}$ while $\beta_0 = \mu_y - \Sigma_{yx}\Sigma_{xx}^{-1}\mu_x$. The linear function $\Sigma_{yx}\Sigma_{xx}^{-1}x$ is the best linear predictor in the sense of minimum variance and the correlation between $\underset{\sim}{Y}$ and $\Sigma_{yx}\Sigma_{xx}^{-1}x$ is the multiple correlation coefficient (see Anderson 1958, p.32).

When we speak of "regression coefficients" we mean β for Model A and $\Sigma_{xx}^{-1}\Sigma_{xy}$ for Model B. Model A is considered the conditional version of Model B.

If in Model A we assume that the regression plane passes through the origin, i.e. $\beta_0 = 0$, or if we assume the centered Model A_c, then the matrix X is the matrix of deviates of the x_{ij} observations from their means and no column of X can be all ones.

In Model B this assumption is equivalent to assuming that

$$E(Z) = \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix} = \mu = 0$$

1.4 Estimation of the regression coefficients and of σ^2

In this section it is shown that the maximum likelihood (ML) estimates obtained under both models are the same while the ML estimators necessarily differ, being defined on different sample spaces. Here we make the usual distinction between estimate and estimator. Simply, an estimator is a function of random variables; an estimate is that number obtained by evaluating the corresponding estimator at the realizations of the random variable.

Under Model A let $Y' = (Y_1,\ldots,Y_n)$ be a random sample of $\underset{\sim}{Y}$ and let $X = \{x_{ij}\}$, $i = 1,\ldots,n$, $j = 0,\ldots,p$ be the matrix of independent observations where $x_{io} = 1$, i.e. the first

column of X consist of one's). Then the ML estimators of β and σ^2 are

$$(1.4.1) \qquad \hat{\beta}_A = (X'X)^{-1}X'Y, \qquad \text{and}$$

$$(1.4.2) \qquad \hat{\sigma}_A^2 = \frac{1}{n}(Y'Y - \hat{\beta}_A'X'X\hat{\beta}_A)$$

Under Model B assume that $(Z_1,\ldots,Z_n)$ is a random sample of $Z \sim N(\mu,\Sigma)$. For comparison with Model A we write the sample as

$$(1.4.3) \qquad \begin{pmatrix} Z_1' \\ Z_n' \end{pmatrix} = \begin{pmatrix} Y_1, & X_{11},\ldots,X_{1p} \\ Y_n, & X_{n_1},\ldots,X_{np} \end{pmatrix} = (Y,X)$$

Then if $A = \sum_{\alpha=1}^{n}(Z_\alpha - \bar{Z})(Z_\alpha - \bar{Z})'$ then from Anderson (1958) if follows immediately that the ML estimators of β and σ^2 under Model B are

$$(1.4.4) \qquad \hat{\beta}_B = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_B^{(2)} \end{pmatrix} = \begin{pmatrix} \bar{Y} - \hat{\beta}_B^{(2)\prime}\bar{X} \\ A_{xx}^{-1} \quad A_{xy} \end{pmatrix}$$

$$(1.4.5) \qquad \hat{\sigma}_B^2 = \Sigma_{yy\cdot x} = \frac{1}{n}(a_{yy} - A_{yx}A_{xx}^{-1}A_{xy})$$

where

$$(1.4.6) \qquad A = \begin{pmatrix} a_{yy} & A_{yx} \\ A_{xy} & A_{xx} \end{pmatrix}$$

$$\bar{Y} = \frac{1}{n}\Sigma Y_i, \quad \bar{X}' = (\bar{X}_1,\ldots,\bar{X}_p),$$

with $\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ji}$.

Notice that the estimators $\hat{\beta}_A^2$ and $\hat{\sigma}_A^2$ are functions of only one random variable, namely Y, while $\hat{\beta}_B^2$ and $\hat{\sigma}_B^2$ are functions of random variables (Y,X), i.e. $\bar{Y}$, $\bar{X}$, and A.

Theorem 1.4.1

Based on the realizations of samples under Model A and Model B the ML estimates of β^2 and σ^2 are the same.

Proof

Partition X as $X = (i, X_1)$ where $i' = (1,\ldots,1)$ and $X_1 = \{x_{\ell j}\}$, $\ell = 1,\ldots,n$ $j = 1,\ldots,p$, i.e. X is an $n \times p$ matrix.

Let $\beta = \begin{pmatrix} \beta_0 \\ \beta_{(2)} \end{pmatrix}$, $\hat{\beta}_A = \begin{pmatrix} \hat{\beta}_{OA} \\ \hat{\beta}_A^{(2)} \end{pmatrix}$

Then

$$X'X = \begin{pmatrix} i' \\ X_1' \end{pmatrix} (i, X_1) = \begin{pmatrix} i'i & i'X_1 \\ X_1'i & X_1'X_1 \end{pmatrix}$$

$$= \begin{pmatrix} n & i'X_1 \\ X_1'i & X_1'X_1 \end{pmatrix}$$

Let $M = (I - ii'/n)$.

Applying the partitioned inverse rule

$$\begin{pmatrix} E & F \\ G & H \end{pmatrix}^{-1} = \begin{pmatrix} E^{-1}(I+FD^{-1}GE^{-1}), & -E^{-1}FD^{-1} \\ -D^{-1}GE^{-1} & , \quad D^{-1} \end{pmatrix}, \; D = H - GE^{-1}F$$

with $E = n$, $F = i'X_1$, $G = X_1'i$, $H = X_1'X_1$

and $D = X_1'X_1 - X_1'in^{-1}i'X_1$

$= X_1'(I - ii'/n)X_1$

$= X_1'M\,X_1$

to $(X'X)^{-1}$ we obtain

$$(X'X)^{-1} = \begin{pmatrix} n^{-1}(I+i'X_1'(X_1'M\,X_1)^{-1}X_1'in^{-1}), & -n^{-1}i'X_1(X_1'MX_1)^{-1} \\ -(X_1'MX)^{-1}X_1'in^{-1} & , \quad (X_1'MX_1)^{-1} \end{pmatrix}$$

Now

$$\hat{\beta}_A = (X'X)^{-1}X'Y = (X'X)^{-1}\begin{pmatrix} i'Y \\ X_1'Y \end{pmatrix}$$

$$= \begin{pmatrix} n^{-1}(I+iX_1(X_1'M\,X_1)^{-1}X_1'in^{-1})i'Y - n^{-1}i'X_1(X_1'MX_1)^{-1}X_1'Y \\ -(X_1'MX_1)^{-1}X_1'in^{-1}i'Y + (X_1'MX_1)^{-1}X_1'Y \end{pmatrix}$$

$$= \begin{pmatrix} n^{-1}(i'Y - i'X_1(X_1'MX_1)^{-1}(X_1'MY) \\ (X_1'MX_1)^{-1}X_1'MY \end{pmatrix}$$

$$= \begin{pmatrix} \hat{\beta}_{OA} \\ \hat{\beta}_A^{(2)} \end{pmatrix}$$

Now

$$\hat{\beta}_{OA} = n^{-1}(i'Y - i\bar{X}_1\hat{\beta}_A^{(2)})$$

$$= \frac{1}{n}\left(\textstyle\sum_{j=1}^{n} Y_j - \left(\sum_{j=1}^{n} X_{j1}, \ldots, \sum_{j=1}^{n} X_{jp}\right)\hat{\beta}_A^{(2)}\right)$$

$$= \bar{Y} - \hat{\beta}_A^{(2)\prime}\bar{X}$$

Now

$$MX_1 = (I - ii'/n)X_1$$

$$= X_1 - i(\bar{X}_1, \ldots, \bar{X}_p)$$

$$= \begin{pmatrix} X_{11} - \bar{X}_1, \ldots\ldots, & X_{1p} - \bar{X}_p \\ X_{n1} - \bar{X}_1, \ldots\ldots, & X_{np} - \bar{X}_p \end{pmatrix}$$

$$MY = \begin{pmatrix} Y_1 - \bar{Y} \\ \cdot \\ \cdot \\ \cdot \\ Y_n - \bar{Y} \end{pmatrix}$$

also $M'M = M$. Thus

$$X_1'MX_1 = X_1'M'MX_1$$

$$= \begin{bmatrix} \Sigma(X_{1\alpha}-\bar{X}_1)^2, \ldots, \Sigma(X_{1\alpha}-\bar{X}_1)(X_{p\alpha}-\bar{X}_p) \\ \vdots \\ \Sigma(X_{p\alpha}-\bar{X}_p), \ldots, \Sigma(X_{p\alpha}-\bar{X}_p)^2 \end{bmatrix}$$

$$= A_{xx}$$

Similarly

$$X_1'M_1Y = X_1'M_1'M_1Y = A_{xy}$$

Thus

$$\hat{\beta}_A^{(2)} = A_{xx}^{-1}A_{xy} = \hat{\beta}_B^{(2)}$$

and $\hat{\beta}_{OA} = \hat{\beta}_{OB}$.

To prove that theorem for σ^2, note that

$$\begin{aligned} n\hat{\sigma}_A^2 &= (Y-X\hat{\beta}_A)'(Y-X\hat{\beta}_A) \\ &= Y'Y - (\hat{\beta}_{OA}, \hat{\beta}_A^{(2)})' \begin{pmatrix} i' \\ X_1' \end{pmatrix} Y \\ &= Y'Y - \hat{\beta}_{OA}i'Y - \hat{\beta}_A^{(2)'}X_1'Y \\ &= Y'Y - \hat{\beta}_{OA}\Sigma Y_i - A_{yx}A_{xx}^{-1}X_1'Y \\ &= Y'Y - \Sigma Y_i(\bar{Y}-\hat{\beta}_1\bar{X}_1-\ldots-\hat{\beta}_p\bar{X}_p) - A_{yx}A_{xx}^{-1}X_1'Y \\ &= (Y'Y-n\bar{Y}^2) - A_{yx}A_{xx}^{-1}(X_1'Y-\Sigma Y_i\bar{X}) \\ &= \Sigma(Y_i-\bar{Y})^2 - A_{yx}A_{xx}^{-1}A_{xy} \\ &= a_{yy} - A_{yx}A_{xx}^{-1}A_{xy} \\ &= n\hat{\sigma}_B^2 \end{aligned}$$

1.5 Distribution Theory

In the following theorem the standard distribution theory results are summarized for Model A.

Theorem 1.5.1

If $Y = X\beta+e$ with $e \sim N(O,\sigma^2 I)$ then

(i) $\hat{\beta}_A \sim N(\beta,\sigma^2(X'X)^{-1})$

(ii) $\dfrac{(n-p-1)S_A^2}{\sigma^2} = (Y-X\hat{\beta}_A)'(Y-X\beta_A)/\sigma^2$
is distributed as χ^2_{n-p-1}

(iii) $\hat{\beta}_A^2$ and S_A^2 are independent.

The proof of the above theorem can be found in Graybill (1961) p.113.

In the multivariate analysis case, i.e. Model B, the distribution theory results from the next slightly more general theorem.

Theorem 1.5.2

Suppose Σ and A are partitioned into q and $p-q$ rows and columns

(1.5.1) $$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \quad A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

Let $\Sigma_{11.2} = \Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. If $A \sim W(\Sigma,p,n)$ then the following results hold.

(i) $A_{11.2} = A_{11}-A_{12}A_{22}^{-1}A_{21} \sim W(\Sigma_{11.2},n-(p-q)$

(ii) $A_{22} \sim W(\Sigma_{22},p-q,n)$

(iii) $A_{22}^{-1}A_{21}$ for given $A_{22} = a_{22}$ is distributed

$N(\Sigma_{22}^{-1}\Sigma_{21}, \Sigma_{11.2} * a_{22}^{-1})$ where $A*B$ is the Kronecker product of matrices A and B.

(iv) $A_{11.2}$ is independent of $A_{22}^{-1}A_{21}, A_{22}$

Proof ((i) and (iv)

From the partitioned inverse rule

$$\Sigma^{-1} = \begin{pmatrix} \Sigma_{11.2}^{-1} & -\Sigma_{11.2}^{-1}\Sigma_{12}\Sigma_{22}^{-1} \\ -\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11.2}^{-1} & \Sigma_{22}^{-1}+\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11.2}^{-1}\Sigma_{12}\Sigma_{22}^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} \Sigma_{11.2}^{-1} & -\Sigma_{11.2}^{-1}\beta' \\ -\beta\Sigma_{11.2}^{-1} & \Sigma_{22}^{-1}+\beta\Sigma_{11.2}^{-1}\beta' \end{pmatrix}$$

where

$$\Sigma_{11.2} = \Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

$$\Sigma_{22.1} = \Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

$$\beta = \Sigma_{22}^{-1}\Sigma_{21} .$$

Using the formula $(I+LM)^{-1} = I-L(I+ML)^{-1}M$ one can show that

$$\Sigma_{22.1}^{-1} = (I-\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}\Sigma_{22}^{-1}$$

$$= \Sigma_{22}^{-1}+\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11.2}^{-1}\Sigma_{12}\Sigma_{22}^{-1}$$

$$= \Sigma_{22}^{-1}+\beta\Sigma_{11.2}^{-1}\beta'$$

Now $$tr(\Sigma^{-1}A) = tr(\Sigma_{11.2}^{-1}A_{11}-\Sigma_{11.2}^{-1}\beta'A_{21}-\beta\Sigma_{11.2}^{-1}A_{12}+$$

$$+\Sigma_{22}^{-1}A_{22}+\beta\Sigma_{11.2}^{-1}\beta'A_{22})$$

$$= tr(\Sigma_{11.2}^{-1}A_{11})+tr(\Sigma_{22}^{-1}A_{22})+tr(\beta\Sigma_{11.2}^{-1}\beta'A_{22})$$

$$-tr(\Sigma_{11.2}^{-1}\beta'A_{21})-tr(\beta\Sigma_{11.2}^{-1}A_{12})$$

Using the formula $A_{11.2}+A_{12}A_{22}^{-1}A_{21} = A_{11}$

we have

$$\begin{aligned} tr(\Sigma^{-1}A) = {} & tr(\Sigma_{11.2}^{-1}A_{11.2})+tr(\Sigma_{11.2}^{-1}A_{12}A_{22}^{-1}A_{21}) \\ & +tr(\beta\Sigma_{11.2}^{-1}\beta'A_{22}) \\ & -tr(\Sigma_{11.2}^{-1}\beta'A_{21}) \\ & -tr(\beta\Sigma_{11.2}^{-1}A_{12}) \\ & +tr(\Sigma_{22}^{-1}A_{22}) \end{aligned}$$

if we set $B = A_{22}^{-1}A_{21}$ then

$$\begin{aligned} (1.5.2) \quad tr(\Sigma^{-1}A) = {} & tr(\Sigma_{11.2}^{-1}A_{11.2})+tr(B-\beta)'\Sigma_{11.2}^{-1}(B-\beta)A_{22} \\ & +tr(\Sigma_{22}^{-1}A_{22}) . \end{aligned}$$

The density of A is

$$\frac{(2\pi)^{-\frac{1}{2}np}\Gamma_p^{-1}(\frac{1}{2}n)}{|\Sigma|^{(n/2)}} |A|^{(n-p-1)/2}\exp(tr(-\tfrac{1}{2}\Sigma^{-1}A)) =$$

$$= \frac{C}{|\Sigma|^{(n/2)}} |A|^{(n-p-1)/2}\exp(tr(-\tfrac{1}{2}\Sigma^{-1}A)) .$$

The distribution of A is the same as the distribution of A_{11}, A_{21}, and A_{22} and by (1.5.2) it can be written as

$$\begin{aligned} (1.5.3) \quad & \frac{C}{|\Sigma_{22}|^{(n/2)}|\Sigma_{11.2}|^{(n/2)}} |A_{22}|^{(n-p-1)/2}|A_{11.2}|^{(n-p-1)/2} \\ & \cdot\exp\{(-\tfrac{1}{2})tr(\Sigma_{11.2}^{-1}A_{11.2})\} \\ & \cdot\exp\{(-\tfrac{1}{2})tr(\Sigma_{22}^{-1}A_{22}) \\ & \cdot\exp\{(-\tfrac{1}{2})tr(B-\beta)'\Sigma_{11.2}^{-1}(B-\beta)A_{22}\} \\ & dA_{11}dA_{21}dA_{22} \end{aligned}$$

because of $|\Sigma| = |\Sigma_{22}||\Sigma_{11\cdot 2}|$ and $|A| = |A_{22}||A_{11\cdot 2}|$.

If we transform from A_{11} to $A_{11\cdot 2}$ and A_{21} to B, so the Jacobians are $J(A_{11}\to A_{11\cdot 2}) = 1$ and $J(A_{21}\to B) = |A_{22}|^{(p-q)}$ respectively.

The joint distribution of A_{22}, $A_{11\cdot 2}$, and B therefore becomes

$$(1.5.4)\quad W(\Sigma_{22},p-q,n)\; W(\Sigma_{11\cdot 2},q,n-p-q)$$

$$\cdot\; \frac{\exp\{(-\tfrac{1}{2})\operatorname{tr}(B-\beta)'\Sigma_{11\cdot 2}^{-1}(B-\beta)A_{22}\}dA_{22}dA_{11\cdot 2}dB}{|A_{22}\Gamma^{(p-q)/2}|\Sigma_{11\cdot 2}|^{(q/2)}(2\pi)^{[q(p-q)/2]}}$$

Observe that the joint density of $A_{11\cdot 2}$, A_{22} and B splits into parts, one containing only $A_{11\cdot 2}$ and the other containing A_{22} and B. This, therefore, shows that $A_{11\cdot 2} \sim W(\Sigma_{11\cdot 2},q,n-p-q)$ independently of A_{22} and B.

(ii) A is distributed as $\sum_{\alpha=1}^{n-1} Z_a Z_a'$ where the Z_a are independent each with the distribution $N(0,\Sigma)$. Partition Z_a into subvectors of q and $p-q$ components

$$Z_a = \begin{bmatrix} Z_a^{(1)} \\ Z_a^{(2)} \end{bmatrix}$$

Then $Z_a^{(2)}$ are independent each with the distribution $N(0,\Sigma_{22})$, and $A_{22} \sim \sum_{\alpha=1}^{n-1} Z_a^{(2)} Z_a^{(2)'}$ which has the distribution $W(\Sigma_{22},p-q,n)$.

(iii) The joint distribution of A_{22} and B is as already found in (1.5.4).

$$(1.5.5)\quad W(\Sigma_{22},p-q,n)\,\frac{\exp\{(-\tfrac{1}{2})\operatorname{tr}(B-\beta)'\Sigma_{11.2}^{-1}(B-\beta)A_{22}\}dA_{22}\,dB}{|A_{22}|^{-(p-q)/2}|\Sigma_{11.2}|^{(q/2)}(2\pi)^{[q(p-q)/2]}}$$

But from (ii) $A_{22} \sim W(\Sigma_{22},p-q,n)$, and so the conditional distribution of B, when A_{22} is fixed, is

$$(1.5.6)\quad |A_{22}|^{(p-q)/2}|\Sigma_{11.2}|^{-(q/2)}(2\pi)^{-[q(p-q)/2]}$$
$$\cdot\exp\{-(\tfrac{1}{2})\operatorname{tr}\Sigma_{11.2}^{-1}(B-\beta)A_{22}(B-\beta)'\}dB\ .$$

We will show that the distribution in (1.5.6) is normal.

Let $A_{22} = MM'$ where M is a lower triangular matrix. Transform now from B to the U by the relation

$$(1.5.7)\quad U = BM \quad \text{or} \quad B = UM^{-1}.$$

The Jacobian of this transformation is

$$J(B\to U) = |M^{-1}|^{p-q} = |A_{22}|^{-(p-q)/2}$$

The distribution of U is therefore

$$(1.5.8)\quad \frac{1}{(2\pi)^{[q(p-q)/2]}|\Sigma_{11.2}|^{(q/2)}}\exp\{-(\tfrac{1}{2})\operatorname{tr}\Sigma_{11.2}^{-1}$$
$$(U-\beta M)(U-\beta M)'\}dU\ .$$

Every column of U has a $(p-q)$ variate normal distribution, with mean given by the corresponding column of βM, and variance covariance matrix $\Sigma_{11.2}$; all the q columns of U are independently distributed. This is the conditional

distribution of U when A_{22} is fixed.

When $A_{22} = a_{22}$ i.e. fixed, we have

$E(U|M) = \beta M$; $\text{var}(U|M) = (\Sigma_{11.2} * I)$. From (1.5.7), therefore

$$E(B|A_{22} = a_{22}) = E(UM^{-1}|M) = \beta MM^{-1} = \beta = \Sigma_{22}^{-1}\Sigma_{21}$$

and

$$\begin{aligned} \text{var}(B|A_{22} = a_{22}) &= \text{var}(UM^{-1}|M) \\ &= \Sigma_{22.1} * M'^{-1} I M^{-1} \\ &= \Sigma_{22.1} * (MM')^{-1} \\ &= \Sigma_{22} * a_{22}^{-1} \end{aligned}$$

The elements of $B = A_{22}^{-1}A_{21}$ are linear function of the elements of U, which have a multivariate normal distribution, so B is distributed normally.

This very appealing and useful theorem was first obtained by Bartlett (1933), who credits Wishart as the stimulus for this work. Much later proofs were given by Dempster, Stein and Sylvan (1969). This proof was also used implicitly in Anderson's (1958) derivation of the form of the Wishart distribution.

We now show that the joint distribution of the M.L. estimators of the parameters under Model A is just the conditional distribution of the M.L. estimators under Model B given that $A_{xx} = a_{xx}$.

Theorem 1.5.3

The joint distribution of the M.L. estimators of the parameters under Model A, i.e. $F_A(\hat{\beta}_A^{(2)}, S_A^2)$ is just the conditional distribution of the M.L. estimators of Model B given $A_{xx} = a_{xx}$, i.e. $F_A(\hat{\beta}_A^{(2)}, S_A^2) = F_B(\hat{\beta}_B^{(2)}, S_B^2 | A_{xx} - a_{xx})$.

Proof

It follows from the distributions of Model A and Model B that $F_A(\hat{\beta}_A^{(2)}, S_A^2) = F_B(\hat{\beta}_B^{(2)}, S_B^2 | X)$. But from Theorem 1.5.1 we have that $F_A(\hat{\beta}_A^{(2)}, S_A^2) = F_A(\hat{\beta}_A^{(2)}) F_A(S_A^2)$ depends on the fixed X only through X'X, i.e. A_{xx}. Thus

$$F_B(\hat{\beta}_B^{(2)}, S_B^2 | X) = F_B(\hat{\beta}_B, S_B^2 | X'X).$$

Corollary 1.5.4

Given Theorem 1.5.1 the following results can be shown to hold.

(i) $a_{yy} - A_{yx}A_{xx}^{-1}A_{xy} \sim W(\Sigma_{yy.x}, n-((p+1)-1))$

(ii) $(A_{xx}^{-1}A_{yx} | A_{xx} = a_{xx}) \sim N(\Sigma_{xx}^{-1}\Sigma_{yx}, \Sigma_{yy.x}a_{xx}^{-1})$

(iii) $a_{yy} - A_{yx}A_{xx}^{-1}A_{xy}$ is independent of $A_{xx}^{-1}A_{xy}, A_{xx}$

(iv) $A_{xx} \sim W(\Sigma_{xx}, n)$

Hint: The proof follows straight forward from Theorem 1.5.2 for $q = 1$ and $(p+1)$ independent variables.

We need the following lemma:

Lemma 1.5.5

If Y and T are random variables then

(i) $E(Y) = \underset{T}{E}(E(Y|T))$

(ii) $var(Y) = \underset{T}{E}(var(Y|T)) + \underset{T}{var}(E(Y|T))$.

Proof

(i) We set $E(Y|T) = g(t)$ then

$$\underset{T}{E}(E(Y|T)) = \underset{T}{E}(g(t))$$

$$= \int_{-\infty}^{\infty} g(t) f_T(t)\,dt$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Y\, f_{Y|t}(y|t) f_T(t)\,dt\,dy$$

$$= \int_{-\infty}^{\infty} Y \int_{-\infty}^{\infty} f_{Y,T}(y,t)\,dt\,dy$$

$$= \int_{-\infty}^{\infty} Y\, f_Y(y)\,dy$$

$$= E(Y)$$

(ii) For the unconditional variable Y with $E(Y) = \mu$ we have

$$E(Y-\mu)^2 = var(Y) = E(Y^2) - (E(Y))^2$$

so $$var(Y|T = t) = E(Y^2|T = t) - (E(Y|T = t))^2.$$

Taking expectation with respect to the T variable we have

$$\underset{T}{E}(var(Y|T = t)) = \underset{T}{E}(E(Y^2|T = t)) - \underset{T}{E}(E(Y|T = t))^2$$

$$= var(Y) - \underset{T}{var}(E(Y|T = t))$$

or

$$var(Y) = \underset{T}{E}(var(Y|T = t)) + \underset{T}{var}(E(Y|T = t)) .$$

Theorem 1.5.6

Consider the Model B then

(i) $(\hat{\beta}_B^{(2)} | A_{xx} = a_{xx}) \sim N(\beta_B^{(2)}, \Sigma_{yy.x} a_{xx}^{-1})$

(ii) $E(\hat{\beta}_B^{(2)}) = \beta_B^{(2)}$

(iii) $\text{var}(\hat{\beta}_B^{(2)}) = (\Sigma_{yy.x}/n-p-1)\Sigma_{xx}^{-1}$

(iv) The joint unconditional density of $\hat{\beta}_B = (\hat{\beta}_{OB}, \hat{\beta}_B^{(2)})'$ is given by (1.5.14).

Proof

(i) We know that the M.L. estimator for $\beta_B^{(2)}$ is given by $\hat{\beta}_B^{(2)} = A_{xx}^{-1} A_{yx}$ then we use the result (ii) of Corollary 1.5.4.

(ii) Using the first result of Lemma 1.5.5 we have

$$\begin{aligned} E(\hat{\beta}_B^{(2)}) &= E(E(\hat{\beta}_B^{(2)} | A_{xx} = a_{xx})) \\ &\quad a_{xx} \\ &= E(\beta_B^{(2)}) \\ &= \beta_B^{(2)} \end{aligned}$$

(iii) Using the second result of Lemma 1.5.5 and the fact that if $A_{xx} \sim W(\Sigma_{xx}, n)$ then

$$E(A_{xx}^{-1}) = (n-p-1)^{-1}\Sigma_{xx}^{-1}$$

and the result follows.

(iv) Before we give the joint density we need the following results.

(A) If $(v_1, v_2, \ldots, v_p) = v \quad -\infty < v < \infty$ then

$$\int \exp\{-\tfrac{1}{2}[(x-v)'M^{-1}(x-v) + v'H^{-1}v]\}dv =$$

$$= (2\pi)^{p/2}|MH|^{\frac{1}{2}}|H+M|^{-\frac{1}{2}}\exp\{-\tfrac{1}{2}x'(H+M)^{-1}x\}.$$

(B) $$\int \exp\{-\tfrac{1}{2}\mathrm{tr}(\Sigma^{-1}A)\}dA =$$

$$= (2\pi)^{(pn)/2}|\Sigma|^{(n/2)}2^{p}[\Pi_{i=1}^{p}C(n-p+i)]^{-1}$$

where $C(n) = \dfrac{2\pi^{n/2}}{\Gamma(n/2)}$ is the surface area of a sphere of unit radius in n-dimensional space.

(1.5.9) $$Z = (\underset{\sim}{Y}, X_1, \ldots, X_p)' \sim N(\mu, \Sigma)$$

$$\bar{Z} = (\bar{Y}, \bar{X}_1, \ldots, \bar{X}_p)' = (\bar{Y}, \bar{X})'$$

and $$\Sigma^{-1} = \begin{pmatrix} \sigma^{YY} & \Sigma^{Yx} \\ \Sigma^{xY} & \Sigma^{xx} \end{pmatrix}$$

where $$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad \bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{i,j} \qquad j = 1,2,\ldots,p.$$

The joint density of the sample dispersion matrix A and the sample mean vector $\bar{Z}$ is given by

(1.5.10) $$g(A,\bar{Z}) = c_1\exp\{-\tfrac{1}{2}\mathrm{tr}\Sigma^{-1}[A+n(\bar{Z}-\mu)(\bar{Z}-\mu)']\}|A|^{\frac{1}{2}(n-p-1-2)}$$

because A and $\bar{Z}$ are independently distributed. The constant is given by

(1.5.11) $$c_1 = n^{\frac{1}{2}(p+1)}(2\pi)^{-(p+1)n/2}|\Sigma|^{-\frac{1}{2}n}\Pi_{i=1}^{p+1}C(n-p-2+i)\,2^{-(p+1)}$$

we set

$$(1.5.12)\quad A_{yy.x} = a_{yy} - A_{yx}A_{xx}^{-1}A_{xy}$$

$$\hat{\beta}_{OB} = \bar{Y} - \hat{\beta}_B^{(2)\prime}\bar{X} \quad \text{and} \quad \hat{\beta}_B^{(2)} = A_{xx}^{-1}A_{xy}$$

Then (1.5.10) becomes

$$(1.5.13)\quad g(A_{xx}, A_{yy.x}, \hat{\beta}_B^{(2)}, \bar{X}, \hat{\beta}_{OB}) =$$

$$= c_1 \exp\{-\tfrac{1}{2}\mathrm{tr}\Sigma_{xx}^{-1}A_{xx} - \tfrac{1}{2}\mathrm{tr}\sigma^{yy}A_{yy.x}$$

$$-\tfrac{1}{2}\mathrm{tr}(\hat{\beta}_B^{(2)}-\beta_B^{(2)})'\sigma^{yy}(\hat{\beta}_B^{(2)}-\beta_B^{(2)})A_{xx}$$

$$-\tfrac{1}{2}(\bar{X}-\mu_x)'\Sigma_{xx}^{-1}(\bar{X}-\mu_x)$$

$$-\tfrac{1}{2}n(\hat{\beta}_{OB}-\beta_{OB}+(\hat{\beta}_B^{(2)}-\beta_B^{(2)})'\bar{X})'\sigma^{yy}(\hat{\beta}_{OB}-\beta_{OB}+(\hat{\beta}_B^{(2)}-\beta_B^{(2)})'\bar{X})\}$$

$$|A_{xx}|^{-\frac{1}{2}(n-p-1-2)}\,|A_{yy.x}|^{-\frac{1}{2}(n-p-1-2)}$$

If we integrate with respect to $\bar{X}$, $A_{yy.x}$ and A_{xx} by using the (A) and (B) we obtain

$$(1.5.14)\quad g(\hat{\beta}_{OB}, \hat{\beta}_B^{(2)}) = C_2\,|\Sigma_{xx}^{-1} + (\hat{\beta}_B^{(2)}-\beta_B^{(2)})\sigma^{yy}(\hat{\beta}_B^{(2)}-\beta_B^{(2)})'|^{-(n+1)/2}$$

$$\cdot\exp\{-\tfrac{1}{2}n(\hat{\beta}_{OB}-\beta_{OB}) + (\hat{\beta}_B^{(2)}-\beta_B^{(2)})\mu_x)'((\hat{\beta}_B^{(2)}-\beta_B^{(2)},$$

$$\Sigma_{xx}(\hat{\beta}_B^{(2)}-\beta_B^{(2)})$$

$$+ (\sigma^{yy})^{-1})^{-1}((\hat{\beta}_{OB}-\beta_{OB}) + (\hat{\beta}_B^{(2)}-\hat{\beta}_B)'\mu_x)$$

where

$$C_2 = (2\pi)^{-\frac{1}{2}n^{\frac{1}{2}}}(\sigma^{yy})^{\frac{1}{2}(p-2)}\,|\Sigma_{xx}|^{-\frac{1}{2}(n-2)}$$

$$\cdot\Pi_{i=1}^{p}C(n-p-1-1+i)\,(\Pi_{i=1}^{p}C(n+1-p-1+i))^{-1}$$

Theorem 1.5.6

The marginal density of $\hat{\beta}_B^{(2)}$ is given by

$$g(\hat{\beta}_B^{(2)}) = k\left|\Sigma_{xx}^{-1} + (\hat{\beta}_B^{(2)}-\beta_B^{(2)})'\sigma^{yy}(\hat{\beta}_B^{(2)}-\beta_B^{(2)})\right|^{-\frac{1}{2}n}$$

where $$k = \Sigma_{xx}^{-\frac{1}{2}(n-1)}\left|\sigma^{yy}\right|^{-\frac{1}{2}p}\pi^{p}$$

$$\cdot\Pi_{i=1}^{p}\Gamma(\tfrac{1}{2}(n+1-i))\Gamma(\tfrac{1}{2}(n-p-i))$$

$$\cdot[\Pi_{i=1}^{p+1}\Gamma(\tfrac{1}{2}(n-i))]^{-1}$$

Proof

On integrating (1.5.14) with respect to $\hat{\beta}_{OB}$.

An alternative proof follows immediately from the <u>represent ation 1.5.8</u> by observing that

$$\left(\frac{n-p+1}{\Sigma_{yy.x}}\right)^{\frac{1}{2}}(\hat{\beta}_B^{(2)}-\beta_B^{(2)})\,|\,A_{xx} = a_{xx} \sim N(0,n-p+1\ a_{xx}^{-1})$$

where $A_{xx} \sim W(\Sigma_{xx},n)$, and using the theorem 1.5.9. In other words, unconditionally we have $\left(\frac{n-p+1}{\Sigma_{yy.x}}\right)^{\frac{1}{2}}(\hat{\beta}_B^{(2)}-\beta_B^{(2)}) \sim T_{n-p+1}$ $(0,\Sigma_{xx,p}^{-1})$. Note we suppose that $E(Z) = 0$ (see p.8).

Definition 1.5.7

A p-variate random vector $X = (X_1,\ldots,X_p)'$ is said to have a (nonsingular) multivariate t distribution with mean vector $\mu = (\mu_1,\ldots,\mu_p)'$ and covariance matrix $n(n-2)^{-1}\Sigma$, $n > 2$, denoted by $T_n(\mu,\Sigma,p)$, if it has the probability

density function given by

$$f(x) = \frac{\Gamma(\frac{1}{2}(n+p))}{(\pi n)^{\frac{1}{2}n}\Gamma(\frac{1}{2}n)|\Sigma|^{\frac{1}{2}}}\left\{1 + \frac{(x-\mu)'\Sigma^{-1}(x-\mu)}{n}\right\}^{-\frac{1}{2}(n+p)} \quad n > 2$$

It is noted that $T_n(0,1,1) = t_n$ is the Student's t-distribution with n degrees of freedom.

It is convenient to represent a multivariate t-vector in the following form:

Representation 1.5.8

Let $X \sim T_n(\mu,\Sigma,p)$. Then X may be written as

$$X = (V^{\frac{1}{2}})^{-1}Y+\mu$$

where $V^{\frac{1}{2}}$ is the symmetric square root of V, i.e.

$$V^{\frac{1}{2}}V^{\frac{1}{2}} = V \sim W(\Sigma^{-1},n+p-1)$$

and $Y \sim N(0,nI)$ independent of V. This implies $X|V \sim N(\mu,nV^{-1})$ where $V \sim W(\Sigma^{-1},n+p-1)$. The Student's t random variable is then represented as

$X|V \sim N(0,na'V^{-1}a/a'\Sigma a)$ for any $a \neq 0$, where

$V \sim W(\Sigma^{-1},n+p-1)$.

Theorem 1.5.9

$X \sim T_n(\mu,\Sigma,p)$ if and only if for any $a \neq 0$

$$(a'\Sigma a)^{-\frac{1}{2}}a'(X-\mu) \sim t_n.$$

Proof

$$X|V \sim N(\mu, nV^{-1})$$

if and only if

$$(a'\Sigma a)^{-\frac{1}{2}}a'(X-\mu)|V \sim N\left(0, \frac{n(a'V^{-1}a)}{a'\Sigma a}\right) \quad \text{for any} \quad a \neq 0$$

if and only if

$$(a'\Sigma a)^{-\frac{1}{2}}a'(X-\mu) \sim t_n$$

since $a'\Sigma a/a'V^{-1}a \sim X^2_n$.

1.6 Hypothesis testing

The question remains as to how the test of hypotheses compare under both models. It is shown that the rejection regions, and the test statistics evaluated at the data are the same for the two models. But the two tests are shown to differ in the power functions. In what follows we test at level α.

Model A

Regression coefficients and hypothesis testing

(i) $H_0 : \beta_A^{(2)} = 0$ against $H_1 : \beta_A^{(2)} \neq 0$

we reject if

$$(1.6.1) \quad F = \frac{\hat{\beta}_A^{(2)'} C_{22}^{-1} \hat{\beta}_A^{(2)}}{(Y'Y - \hat{\beta}_A' X'Y)} \cdot \frac{n-p-1}{p} > F^{(\alpha)}_{p,n-p-1}$$

where

$$C = (X'X)^{-1} = \begin{bmatrix} C_{11} & C_{(1)} \\ C'_{(1)} & C_{22} \end{bmatrix}$$

(ii) To test if the regression plane passes through the origin

$$H_0 : \beta_{OA} = 0 \quad \text{against} \quad H_1 : \beta_{OA} \neq 0 .$$

We reject if

$$(1.6.2) \qquad F = \frac{\hat{\beta}_{OA}^2}{S_A^2 C_{11}} > F_{1,n-p-1}^{(\alpha)}$$

(iii) To test if the entire vector of regression coefficients is zero

$$H_0 : \beta_A = 0 \quad \text{against} \quad H_1 : \beta_A \neq 0 .$$

We reject if

$$(1.6.3) \qquad F = \frac{\hat{\beta}_A X'Y}{(Y'Y - \hat{\beta}_A' X'Y)} \cdot \frac{n-p-1}{p+1} > F_{p+1,n-p-1}^{(\alpha)}$$

(iv) To test if a subvector of

$$\beta_A = (\beta_0, \beta_1, \ldots, \beta_p)' = (\beta^{(q+1)}, \beta^{(p-q)})' \quad \text{is zero}$$

where

$$\beta_A^{(q+1)} = (\beta_0, \ldots, \beta_q)' \quad \text{and} \quad \beta_A^{(p-q)} = (\beta_{q+1}, \ldots, \beta_p)'$$

$$H_0 : \beta_A^{(p-q)} = 0 \quad \text{against} \quad H_1 : \beta_A^{(p-q)} \neq 0.$$

We reject if

$$(1.6.4) \qquad F = \frac{\hat{\beta}_A^{(p-q)\prime} C_{22}^{-1} \hat{\beta}_A^{(p-q)}}{(Y'Y - \hat{\beta}_A' X'Y)} \cdot \frac{n-p-1}{p-q} > F_{p-q,n-p-1}^{(\alpha)}$$

where

$$(X'X)^{-1} = C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}$$

C_{11} is $(q+1) \times (q+1)$ matrix

and

C_{22} is $(p-q) \times (p-q)$ matrix.

In all cases, when we reject, the central F-distribution becomes the non-central F-distribution. For example in case (i) if $\beta_A^{(2)} \neq 0$ then the distribution F is non-central with non-centrality parameter $\frac{1}{2\sigma^2} \beta_A^{(2)\prime} C_{22}^{-1} \beta_A^{(2)}$.

Coefficient of determination and hypothesis testing

The test statistics can also be expressed in terms of multiple coefficient of determination which is defined by

$$(1.6.5) \qquad R^2_{y.x_1,\ldots,x_p} = \frac{\hat{\beta}_A' X'Y - (\Sigma Y_i)^2/n}{\Sigma Y_i^2 - (\Sigma Y_i)^2/n}$$

The multiple coefficient of determination plays the role of the square of the multiple correlation coefficient in regression analysis. (In the present Model A, where the X's are considered non-random, the term "correlation coefficient" cannot be used in the usual way. The coefficient in (1.6.5) measures the linear influence of the explanatory variables $(X_0, X_1, \ldots, X_p)$. It is easy to show that (1.6.1) is equal to

$$(1.6.6) \qquad F = \frac{R^2_{y.x_0,\ldots,x_p}}{1-R^2_{y.x_0,\ldots,x_p}} \cdot \frac{n-p-1}{p} \sim F_{p,n-p-1}$$

Similarly (1.6.4) can be expressed in terms of the multiple coefficient of determination. If $R^2_{y.x_0,\ldots,x_p}$ is the multiple coefficient of determination between Y and $(X_0,\ldots,X_p)$ and $R^2_{y.x_0,\ldots,x_q}$ is the multiple coefficient of determination between Y and $(X_0,\ldots,X_q)$ where $q < p$. Then we have that (1.6.4) is equal to

$$(1.6.7) \qquad F = \frac{R^2_{y.x_0,\ldots,x_p} - R^2_{y.x_0,\ldots,x_q}}{1-R^2_{y.x_0,\ldots,x_p}} \cdot \frac{n-p-1}{p-q} \sim F_{p-q,n-p-1}$$

Confidence bands for the regression surface

The problem of confidence banding the regression surface is to construct two functions $\bar{f}$ and $\underline{f}$ based on the sample data, which lies entirely above and below, respectively, the unknown true regression surface f with probability $1-a$.

If the surface is of the form $f(x_1,x_2,\ldots,x_p) = \beta_1x_1+ \ldots +\beta_px_p$, then the $100(1-a)$ percent confidence band is

$$(1.6.8) \qquad \sum_{i=1}^{p}\beta_ix_i \in \sum_{i=1}^{p}\beta_ix_i \pm (pF^{(a)}_{p,n-p})^{\frac{1}{2}}(x'(X'X)^{-1}x)^{\frac{1}{2}}$$

for all $x = (x_1,\ldots,x_p)$.

For the $f(x_1,\ldots,x_p) = \beta_0+\beta_1x + \ldots +\beta_px_p$ surface a $100(1-a)$ percent confidence band is

$$(1.6.9) \qquad \beta_0+\sum_{i=1}^{p}\beta_ix_i \in \hat{\beta}_0+\sum_{i=1}^{p}\hat{\beta}_ix_i \pm ((p+1)F^{(a)}_{p+1,n-p-1})^{\frac{1}{2}} \cdot S(x'(X'X)^{-1}x)^{\frac{1}{2}}$$

for all $x = (1,x_1,\ldots,x_p)$

where $S = +\sqrt{S^2_A}$.

Model B

(i) Testing the hypothesis

$$H_0 : \beta_B^{(2)} = \Sigma_{xx}^{-1}\Sigma_{xy} = 0 \quad \text{against} \quad H_1 : \beta_B^{(2)} \neq 0.$$

we reject if

$$(1.6.10) \quad F = \frac{\hat{\beta}_B^{(2)\prime} A_{xx}^{-1} \hat{\beta}_B^{(2)}}{a_{yy} - A_{yx}A_{xx}^{-1}A_{xy}} \cdot \frac{n-p-1}{p} > F_{p,n-p-1}^{(\alpha)} .$$

The F in (1.6.10) conditionally and unconditionally is distributed like central F with p and $n-p-1$ degrees of freedom.

(ii) Rao (1949) has proposed the following unconditional test for testing the hypothesis

$$H_0 : \beta_{OB} = 0 \quad \text{against} \quad H_1 : \beta_{OB} \neq 0 .$$

He proposes

$$(1.6.11) \quad \frac{n\hat{\beta}_{OB}^2 (a_{yy} - A_{yx}A_{xx}^{-1}A_{xy})^{-1}}{1+N\bar{X}^{(2)\prime} A_{xx}^{-1} \bar{X}^{(2)}} \; U$$

where U has the beta density given by

$$(1.6.12) \quad g(u) = (B(\tfrac{1}{2}, (n-p-1)/2))^{-1} u^{-\frac{1}{2}} / (1+u)^{n-p} .$$

(iii) It appears difficult to obtain a test statistic for the simultaneous testing of hypotheses about β_0 and $\beta^{(2)}$.

Multiple correlation and hypothesis testing

We define the multiple correlation coefficient as the maximum correlation between Y and a linear combination of

$(X_1,\ldots,X_p) = X^{(2)}$ say $a'(X^{(2)} - \mu_x)$. This maximum correlation is obtained by the regression function $\beta_{OB}+\beta_B^{(2)'}(X^{(2)}-\mu_x)$ and the coefficient itself is given by

$$(1.6.13)\quad \rho_{y.x_1,\ldots,x_p} = \left(\frac{\beta_B^{(2)'}\Sigma_{xx}\beta_B^{(2)'}}{\sigma_{yy}}\right)^{\frac{1}{2}}$$

$$= \left(\frac{\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}}{\sigma_{yy}}\right)^{\frac{1}{2}}$$

where Σ_{xx} is positive definite.

Testing for significant regression, i.e. $\beta_B^{(2)} = 0$ is equivalent to testing whether Y is independent of $(X_1,X_2,\ldots,X_p)$, i.e. $\rho_{y.x_1,\ldots,x_p} = 0$.

The appropriate test statistic is given by

$$(1.6.14)\quad F = \frac{A_{yx}A_{xx}^{-1}A_{xy}}{a_{yy}-A_{yx}A_{xy}^{-1}A_{xy}} \cdot \frac{n-p-1}{p}$$

$$= \frac{R^2_{y.x_1,x_2,\ldots,x_p}}{1-R^2_{y.x_1,x_2,\ldots,x_p}} \cdot \frac{n-p-1}{p}$$

which has the $F_{p,n-p-1}$ distribution. $R^2_{y.x_1,\ldots,x_p}$ is the sample estimate of the square of the multiple correlation coefficient.

Partial multiple correlation and hypothesis testing

Let $(\underset{\sim}{Y},X_1,\ldots,X_p)' = Z \sim N(\mu,\Sigma)$. Suppose that Z is split into three sets of components

$Z = \begin{pmatrix} Y \\ X^{(r)} \\ X^{(s)} \end{pmatrix}$ where $X^{(r)}$ has r-components $(r \geq 1)$ and $X^{(s)}$ has s-components $(s > 0)$.

Let μ and Σ be partitioned according to the partitioning of Z so that

$$(1.6.15) \quad \mu = \begin{pmatrix} \mu_y \\ \mu_r \\ \mu_s \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_{yy} & \Sigma_{yx_r} & \Sigma_{yx_s} \\ \Sigma_{x_r y} & \Sigma_{x_r x_r} & \Sigma_{x_r x_s} \\ \Sigma_{x_s y} & \Sigma_{x_s x_r} & \Sigma_{x_s x_s} \end{pmatrix}$$

The population partial multiple correlation coefficient is defined as

$$(1.6.16) \quad \rho^2_{y.(r|s)} = \frac{\Sigma_{y.x_r} \Sigma^{-1}_{x_r x_r . x_s} \Sigma'_{y.x_r}}{\Sigma_{yy.x_s}}$$

where $\Sigma_{yy.x_s} = \text{var}(Y | X^{(s)} = x^{(s)})$

and $\Sigma_{x_r x_r . x_s} = \text{var}(X^{(r)} | X^{(s)} = x^{(s)})$.

It is supposed that we want to test the hypothesis

$$H_0 : \beta^{(r)} = 0 \quad \text{against} \quad \beta^{(r)} \neq 0.$$

This hypothesis is equivalent to testing whether the partial multiple correlation coefficient between Y and $X^{(r)}$ given $X^{(s)} = x^{(s)}$ is zero.

If the sample estimate of the $\rho^2_{y.(r|s)}$ is the

$R^2_{y.(r|s)} = \dfrac{A_{y.x_r} A^{-1}_{x_r x_r . x_s} A'_{y.x_r}}{A_{yy.x_s}}$ then the appropriate statistic is

$$(1.6.17) \quad F = \frac{R^2_{y.(r|s)}}{1-R^2_{y.(r|s)}} \cdot \frac{n-p-1}{r} \sim F_{r,n-p-1}$$

We note that

$$(1.6.18) \quad R^2_{y.(r|s)} = \frac{R^2_{y.x_1,\ldots,x_p} - R^2_{y.x_{r+1},\ldots,x_p}}{1-R^2_{y.x_{r+1},\ldots,x_p}}$$

The above F statistic becomes

$$(1.6.19) \quad F = \frac{R^2_{y.x_1,\ldots,x_p} - R^2_{y.x_{r+1},\ldots,x_s}}{1-R^2_{y.x_1,\ldots,x_p}} \cdot \frac{n-p-1}{r}$$

A comparison of the hypotheses test between Model A and Model B

(1) We observe that (1.6.1) is algebraicly equivalent to (1.6.10).

Under the null hypothesis, conditionally (i.e. given $A_{xx} = a_{xx}$) the F in (1.6.10) follows the same distribution as the distribution of the F in (1.6.1).

Under the alternative hypothesis:

(i) the F in (1.6.1) is distributed as noncentral $F'_{p,n-p-1}(\delta_1)$ with noncentrality parameter

$$\delta_1 = \frac{1}{2\sigma^2_A} \beta_A^{(2)'} C^{-1}_{22} \beta_A^{(2)}$$

(ii) The density of

(1.6.20) $U_A = \frac{p}{n-p-1} F$, where F as in (1.6.1) is given by

(1.6.21) $(\Gamma(\frac{1}{2}(n-p-1+p))/\Gamma(\frac{p}{2})\Gamma(\frac{1}{2}(n-p-1)))$

$$u^{\frac{1}{2}p-1}(1+u)^{-\frac{1}{2}(n-p-1+p)} e^{-\delta_1} {}_1F_1(\tfrac{1}{2}(n-p-1+p);\ \tfrac{1}{2}p;\ \delta_1;\ u(1+u))$$

where ${}_1F_1$ is a hypergeometric function.

We note that ${}_1F_1(a,b;c;x) = \sum_{j=o}^{\infty} \frac{\Gamma(a+j)}{\Gamma(a)} \frac{\Gamma(b+j)}{\Gamma(b)} \frac{\Gamma(c)}{\Gamma(c+j)} \frac{x^j}{j!}$

(iii) The unconditional density of

(1.6.22) $U_B = \frac{p}{n-p-1} F$, where F is in (1.6.10) is given by

(1.6.23) $(\Gamma(\frac{1}{2}n)/\Gamma(\frac{1}{2}(n-p-1))\Gamma(\frac{1}{2}p))u^{\frac{1}{2}p-1}(1+u)^{-\frac{1}{2}n}(1-\rho^2)^{\frac{1}{2}n}$

$$\cdot {}_2F_1(\tfrac{1}{2}n,\tfrac{1}{2}n;\ \tfrac{1}{2}r;\rho^2 u/(1+u))$$

where (a) ${}_2F_1$ is a hypergeometric function

(b) $\rho^2 = \frac{\beta_B^{(2)\prime} \Sigma_{xx} \beta_B^{(2)}}{\sigma_{yy}}$, and ρ is the population multiple correlation coefficient. We observe that the densities of U_A and U_B are entirely different.

(iv) The F in (1.6.10) conditionally, is also distributed as noncentral $F'_{p,n-p-1}(\delta_2)$ with noncentrality parameter

$$\delta_2 = \tfrac{1}{2} \frac{\beta_B^{(2)\prime} a_{xx} \beta_B^{(2)}}{\Sigma_{yy.x}}$$

$$= \frac{1}{2\sigma_B^2} (\beta_B^{(2)\prime} a_{xx} \beta_B^{(2)})$$

We observe that $\delta_1 = \delta_2$.

The power function $P_A(\beta,\sigma^2)$ of the test (i) of Model A is given by

(1.6.24) $P_A(\beta,\sigma^2) = \text{prob}(f_{p,n-p-1}(\delta) > F^{(\alpha)}_{p,n-p-1})$

where

$$f_{p,n-p-1}(\delta) \sim F'_{p,n-p-1}(\delta_1)$$

and

$$P_A(\beta,\sigma^2) = P_B(\beta,\sigma^2) | A_{xx} = a_{xx})$$

where

$P_B((\beta,\sigma^2)|A_{xx} = a_{xx})$ is the conditional power function of the test (i) of Model B.

Tables of the percentage points of (1.6.20), (i.e. the noncentral F) are available, (see Graybill (1958)).

Unfortunately tables of percentage points of (1.6.22) are not available.

However, (1.6.22) is closely related to the distribution of the square of the multiple correlation coefficient

$U_B = \frac{R^2}{1-R^2}$ and tables have been produced by Yoong-Sin Lee, Biometrika (1972), 59, 1, p.175.

(2) The formula (1.6.19) is algebraically equivalent to (1.6.1) and (1.6.7).

Linear constraints and hypothesis testing

Model A

We treat this problem in Chapter 3.

Model B

Theorem 1.6.1

Let H be a $k \times p$ matrix, $r(H)=p$ such that the column space generated by H is a subspace of the column space generated by $X'X$ where X is the realization of the random matrix.

Let $$L_0^2 = \min_{\beta_B} \|Y - X\beta_B\|^2$$

$$L_1^2 = \min_{H\beta_B = \xi} \|Y - X\beta_B\|^2, \quad \text{where } \xi \text{ is given.}$$

Then we have the following results:

(a) L_0^2 and $L_1^2 - L_0^2$ are independent

(b) $L_0^2 \sim \sigma^2 \chi^2_{n-p}$

(c) $L_1^2 - L_0^2 \sim \sigma^2 \chi^2_k(\delta)$ noncentral $\chi^2_{(k)}$ with

$$\delta = (H\beta_B - \xi)'[H(X'X)^{-1}H']^{-1}(H\beta_B - \xi)/\sigma^2$$

(d) If $H \cdot \beta_B = \xi$ is true then

$$L_1^2 - L_0^2 / L_1^2 \sim \frac{k}{n-p} F_{k,n-p}$$

Here $\beta_B = \beta_B^{(2)}$

Theorem 1.6.2

To test the hypothesis

$H_0 : H\,\Sigma_{xx}^{-1}\Sigma_{xy} = H\,\beta_B^{(2)} = \xi$ against $H_1 : H\,\Sigma_{xx}^{-1}\Sigma_{xy} \neq \xi$

we reject if $\lambda_1^{*2}-\lambda_0^{*2}/\lambda_0^{*2} \geq \frac{k}{n-p} F_{k,n-p}^{(\alpha)}$

where (i) $\lambda_1^{*2} = \min_{H\,\beta=\xi} \|Y-X\beta\|^2$ $\lambda_0^{*2} = a_{yy}-A_{yx}A_{xx}^{-1}A_{xy}$

(ii) Y, X, $a_{yy}, A_{yx}, A_{xy}, A_{xx}^{-1}$ are the realizations of the corresponding matrices.

(iii) H is of full rank.

Hint

Using the principle of conditionality and the result (d) of Theorem 1.6.1, if we derive the distribution of $\lambda_1^2-\lambda_0^2/\lambda_1^2$ under the H_1 it is easy to find the power function of the test as described in Theorem 1.6.2. We note that $\lambda_0^2 = a_{yy}-A_{yx}A_{xx}^{-1}A_{xy}$ and $\lambda_1^2 = \min_{H\,\beta_B=\xi} \|Y-X\beta_B\|^2$. The power function of the test in Theorem 1.6.2 is given by $P_B(P,\Sigma_{xx},\Sigma_{yy.x}) = \text{prob}\{X > (k/n-p)F_{k,n-p}^{(\alpha)}\}$ where (a) $X \sim \chi_{k+2n}^2/\chi_{n-p}^2$ (b) the numerator and the denominator are independent and (c)

$$\text{prob}(n=i) = \frac{\Gamma((n-p+k/2)+i)}{\Gamma(i+1)\Gamma(n-p+k/2)}[1+\tau^2(\Sigma_{xx})]^{-(n-p+k)\frac{1}{2}}\left(\frac{\tau^2(\Sigma_{xx})}{1+\tau^2(\Sigma_{xx})}\right)^i$$

$i = 0,1,2,\ldots$ and $\tau^2(A) = (H\,\beta_B^{(2)}-\xi)'(H\,A^{-1}H')^{-1}(H\,\beta_B^{(2)}-\xi)/\Sigma_{yy.x}$

for all A positive definite.

C H A P T E R 2

THE PROBLEM OF MULTICOLLINEARITY

2.1 Introduction

In order to apply the method of O.L.S. (Ordinary Least Squares) the independent variables should not be perfectly linearly related, i.e. the correlation coefficients between X_i and X_j, $r_{ij} \neq 1$ for $i,j = 1,2,\ldots,p$ and $i \neq j$.

There are two extreme cases.

(i) The variables are such that $r_{ij} = 1$, $i,j = 1,\ldots,p, i \neq j$. In this case it is impossible to obtain numerical values for the $\hat{\beta}_i$, $i = 1,\ldots,p$ and the method of O.L.S. breaks down.

(ii) The variables are such that $r_{ij} = 0$, $i,j = 1,2,\ldots,p$, $i \neq j$, that is the variables are orthogonal and there is no problem in the estimation of the β_i, $i = 1,2,\ldots,p$.

In practice neither of the above extreme cases is often met. Before we examine the problem in detail some definitions are required.

Consider Model A where

(2.1.1) $Y = X\beta+e$, $E(e) = 0$, $E(ee') = \sigma^2 I$.

We say there exist <u>extreme multicollinearity</u> when there exist

at least one linear relationship among the columns of the X matrix; that is rank (X) = r(X) < p.

When there is nearly extreme multicollinearity we have for any $p_1 \leq p$ that

$$(2.1.2) \quad \sum_{j=1}^{p_1} a_j X_j \simeq 0$$

where X_j are the columns of X, and the a_j are not all zero.

2.2 The Sources of Multicollinearity

(i) An over-defined Model

This case occurs when there are more explanatory variables than observations. This problem arises frequently in medical research where a lot of observations are taken on each individual.

(ii) Sampling Techniques

This happens when the experimenter only samples from a subspace of the space of regressor variables. This subspace is approximately an hyperplane defined by one or more of the relationships of the form (2.1.2). For example in an industrial situation when one wishes to predict profits from knowledge of variables such as income and labour costs. If we analyze the data we find a strong positive relationship between income and labour costs; higher labour costs result in higher prices, which in turn results in increased income. This type of multicollinearity due to the sampling technique is not inherent in the model since the data could be collected during a period when prices are constant or decreasing but the labour cost is increasing.

(iii) Physical constraints on the model or in the population

This source exists regardless of the sampling technique and is more or less similar to the previous one. Examples of this occur in chemical analyses where the sum of certain constituents in a solution must be constant; although the values of individuals may vary.

2.3 The effects of multicollinearity

In what follows we make the assumption that $X'X$ is the correlation matrix between the X's, (i.e. we assume model A_s, i.e. the standardized model). We will now examine the effects of ill-conditioned data on the $\text{var}(\hat{\beta}_i)$ and the $\text{cov}(\hat{\beta}_i, \hat{\beta}_j)$.

The variance of the ith element of $\hat{\beta}$ is given by $\text{var}(\hat{\beta}_i) = \sigma^2 (x'x)^{ii}$, where $(x'x)^{ii}$ is the (i,i) element of $(X'X)^{-1}$. From the matrix theory we know that:

$$(2.3.1) \quad (x'x)^{ii} = \frac{(X'X)_{ii}}{|X'X|} = (-1)^{2i} \frac{|(X'X)_{ii}|}{|X'X|}$$

where $(X'X)_{ii}$ is the cofactor of the (i,i) element of $(X'X)$. We distinguish between the following two cases:

(1) The X_i variable is orthogonal to the remaining members of X, then $|(X'X)_{ii}| = |X'X|$ so $(x'x)^{ii} = 1$ and $\text{var}(\hat{\beta}_i) = \sigma^2$.

(2) The X_i variable is perfectly dependent on the remaining X_j, $j = 1,2,\ldots,p$, $i \neq j$ variables; then $|X'X| = 0$. While the numerator of (2.3.1) remains unaffected, the $\text{var}(\hat{\beta}_i) = \infty$.

The above results can also be obtained by using the following proposition.

Proposition 2.3.1

The (i,j) element of $(X'X)^{-1}$ is given by

$$(x'x)^{ij} = \begin{cases} 1/(1-R^2_{i.1,2,\ldots,i-1,i+1,\ldots,p}) & \text{if } i = j \\ -r^2_{ij.1,2,\ldots,i-1,i+1,j-1,j+1,\ldots,p} \Big/ \\ (1-R^2_{i.1,2,\ldots,i-1,i+1,j-1,j+1,\ldots,p}) \\ (1-R^2_{j.1,2,\ldots,i-1,i+1,j-1,j+1,\ldots,p}) & \text{if } i \neq j \end{cases}$$

where $r_{ij.1,2,\ldots,i-1,i+1,j-1,j+1,\ldots,p}$ is the partial correlation coefficient between i and j keeping the others fixed and $R_{i.1,2,\ldots,i-1,i+1,j-1,j+1,\ldots p}$ is the coefficient of determination of the ith variable on the remaining (p-2) variables.

Proof

We give the proof for the case $i = j$. From the partitioned inverse rule (see Chapter 1)

let $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$ then

$$A^{-1} = \begin{bmatrix} A_{11}^{-1}+A_{11}^{-1}A_{12}(A_{22}-A_{21}A_{11}^{-1}A_{12})^{-1}A_{21}A_{11}^{-1} & \cdots & -A_{11}^{-1}A_{12}(A_{22}-A_{21}A_{11}^{-1}A_{12})^{-1} \\ \vdots & & \\ -(A_{22}-A_{21}A_{11}^{-1}A_{12})^{-1}A_{21}A_{11}^{-1} & \cdots & (A_{22}-A_{21}A_{11}^{-1}A_{12})^{-1} \end{bmatrix}.$$

Let the matrix $X = (X^* : X_i)$ be, where X_i is the ith variable and X^* is $(p-1)\times(P-1)$ matrix.

Then $$X'X = \begin{bmatrix} X^{*\prime}X^* & X^{*\prime}X_i \\ X_i'X^* & X_i'X_i \end{bmatrix}.$$

Then $$(x'x)^{ii} = \frac{1}{X_i'X_i - \hat{\beta}'X^{*\prime}X^*\hat{\beta}} = \frac{1}{(1-R^2_{i.1,2,\ldots,i-1,i+1,\ldots,p})X_i'X_i}$$

$$= \frac{1}{(1-R^2_{i.1,2,\ldots,i-1,i+1,\ldots,p})}$$

where $\hat{\beta}$ is the estimates for the regression coefficients when we regress the X_i on the remaining variables and $X_i'X_i = 1$ because X is the standardized matrix.

2.4 Linear combinations of regression variables

In this subsection we will discuss the estimability of a linear combination of the parameters.

We will also see that the estimated regression coefficients can be very poor estimates of the individual parameters. In the following let us assume that p_1 variables, $p_1 \leq p$, are involved in the multicollinearity.

Proposition 2.4.1

The linear combination $c'\beta$ is estimable if and only if

c can be expressed as a linear combination of the columns (or rows) of $X'X$.

Theorem 2.4.2

The linear function $c'\beta$ is estimable if and only if c is a linear combination of latent vectors of $X'X$ corresponding to non-zero latent roots of this matrix.

Proof

Let P be the orthogonal matrix whose rows are the latent vectors of $X'X$ and Λ is the diagonal matrix whose elements are the latent roots.

We have $X = (XP')P\beta = Za$, then

$$(2.4.1)\quad Z'Z = (XP')'XP' = PX'XP' = \Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_p \end{pmatrix},$$

$$\lambda_1 \geq \lambda_2 \ldots \geq \lambda_p .$$

Suppose that the matrix $X'X$ has a latent root 0 of multiplicity j. Then the j-columns of XP' are zero say the last j.

It follows that the last j-components of a are annihilated, we cannot estimate $a_{p-j+1},\ldots,a_p$ from our observations. On the other hand we can estimate $a_1,a_2,\ldots,a_{p-j}$ or linear combinations of these.

Now $c'\beta = c'P'P\beta = (Pc)'P\beta = (Pc)'a$.

Hence we can estimate $c'\beta$ iff the last j-components of Pc are zero. If we set $Pc = \delta$ then $c = P'\delta$.

We must note that the poor precision in the estimation procedure does not imply that the estimated model is a poor predictor. From (2.4.1) it follows that

$$(2.4.2) \quad (X'X)^{-1} = \sum_{i=1}^{p} \frac{1}{\lambda_i} v_i v_i'$$

where v_i' is $(p\times 1)$ latent vector corresponding to the latent root λ_i. The smallest root λ_p identifies the latent vector describing the multicollinearity; so if $\lambda_p \simeq 0$ then $X'Xv_p = \lambda_p v_p \simeq 0$ but $X'Xv_p \simeq 0$ iff $Xv_p \simeq 0$. Consequently the multicollinearity is attributable to the fact that every row of X is orthogonal to v_p.

Since $\lambda_p \simeq 0$ implies that $1/\lambda_p$ is large and this is the reason for large diagonal and off-diagonal elements in the p rows or columns of $(X'X)^{-1}$. The estimate of $W_i\beta = \sum_{j=1}^{p} x_{ij}\beta_j$ (where W_i is the ith row of X) is given by

$$(2.4.3) \quad W_i\hat{\beta} = W_i(X'X)^{-1}X'Y \ .$$

Then the $\operatorname{var}(W_i\hat{\beta}) = W_i \operatorname{var}(\hat{\beta})W_i'$

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

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

$$= \sigma^2 W_i\left(\sum_{i=1}^{p} \frac{1}{\lambda_i} v_i v_i'\right)W_i' \ .$$

Since $W_i v_p \simeq 0$ the ill effects of λ_p^{-1} being large are cancelled. So the linear combination $\sum_{j=1}^{p} x_{ij}\beta_j$ can be

estimated quite well although it may not be true for the individual parameters.

2.5 Geometric picture

(i) Two independent variables (general picture)

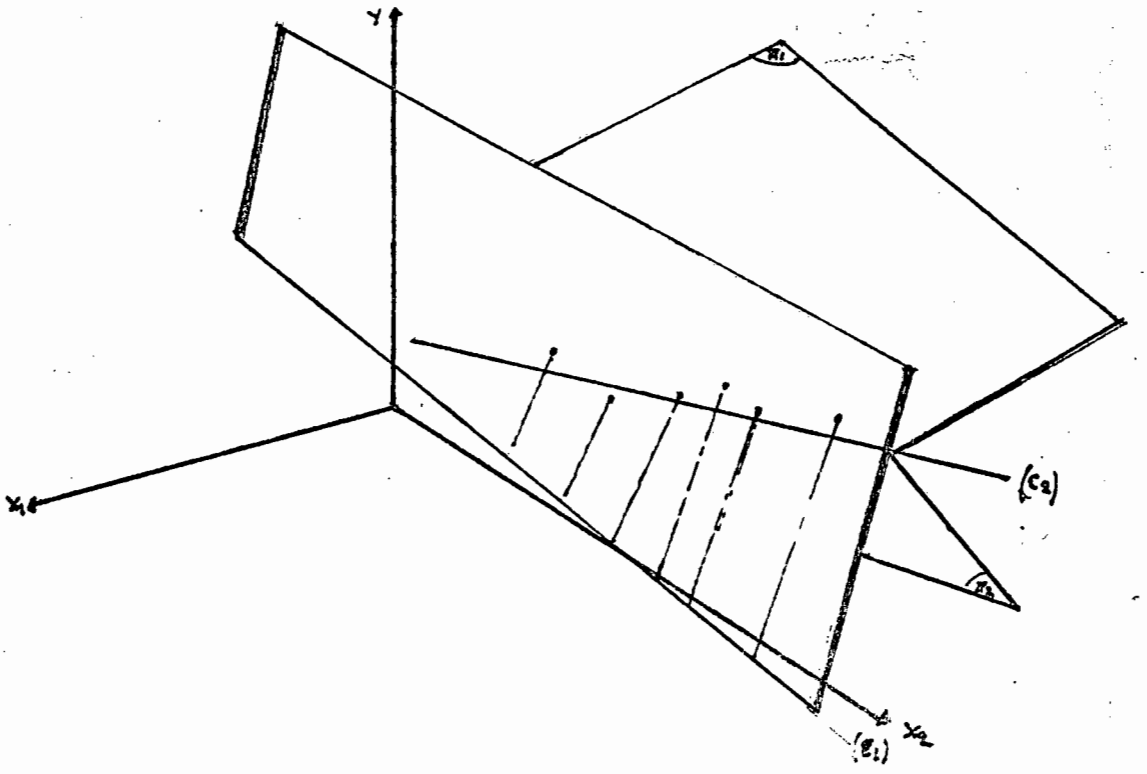


Figure 2.5.1

Consider the Figure 2.5.1 where the values of the independent variables X_1 and X_2 are bunched up on a line (ε_1).

The least squares fit of Y is not any more a plane but rather a line (ε_2). If we try to explain Y with a plane - rather than a line (ε_2) - there exist an infinite number of planes passing through (ε_2). Each of these planes yields the same R.S.S. (R.S.S. - Residual sum of squares).

(ii) Two independent variables. (Picture shows the inflation in variance.)

Consider the model

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \\ X_{31} & X_{31} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} : e_i \sim N(0,\sigma^2) \quad i = 1,2.$$

The dimension of the sample space is 3. The vector of observations Y define a vector $\vec{OY}$ from the origin to the point Y with coordinates (Y_1, Y_2, Y_3). The jth column of X defines a vector $\vec{OX}_j$ in the sample space. The vectors $\vec{OX}_1$ and $\vec{OX}_2$ define a subspace of 2 dimension called the estimation space. This subspace is represented by a plane in which the vector $\hat{Y} = X\hat{\beta}$ must lie. The sphere of y observations is centered at the mean $E(Y)$, which is in the plane π_1 generated by $\vec{OX}_1$ and $\vec{OX}_2$. Let $\hat{Y}$ be perpendicular to the plane π_1 then $\hat{Y}$ is the shortest distance from Y to any point of the estimation space, and $\hat{Y}$ becomes the estimate of $E(Y)$. For simplicity we suppose that $\vec{OX}_1$ and $\vec{OX}_2$ is of unit length. Now the estimate $\hat{\beta}_1$ of the true population coefficient can be found by projecting $\hat{Y}$ along $\vec{OX}_2$ onto $\vec{OX}_1$. Similarly for $\hat{\beta}_2$ we project $\hat{Y}$ along $\vec{OX}_1$ onto $\vec{OX}_2$. The $E(Y)$ is fixed while the disc around it represents possible fitted values of $\hat{Y}$ corresponding to the possible observed Y's falling in the sphere. The projection $\gamma_1\delta_1$ of the whole disc along $\vec{OX}_2$ onto $\vec{OX}_1$ represents the interval of possible $\hat{\beta}_1$.

In the following Figures 2.5.2 and 2.5.3 the regressors are

orthogonal and non-orthogonal. In the second case the interval $\gamma\delta$ becomes very large.

Another geometric picture utilizing the latent vectors and latent roots will be given in Section 5.8 on latent root regression analysis.

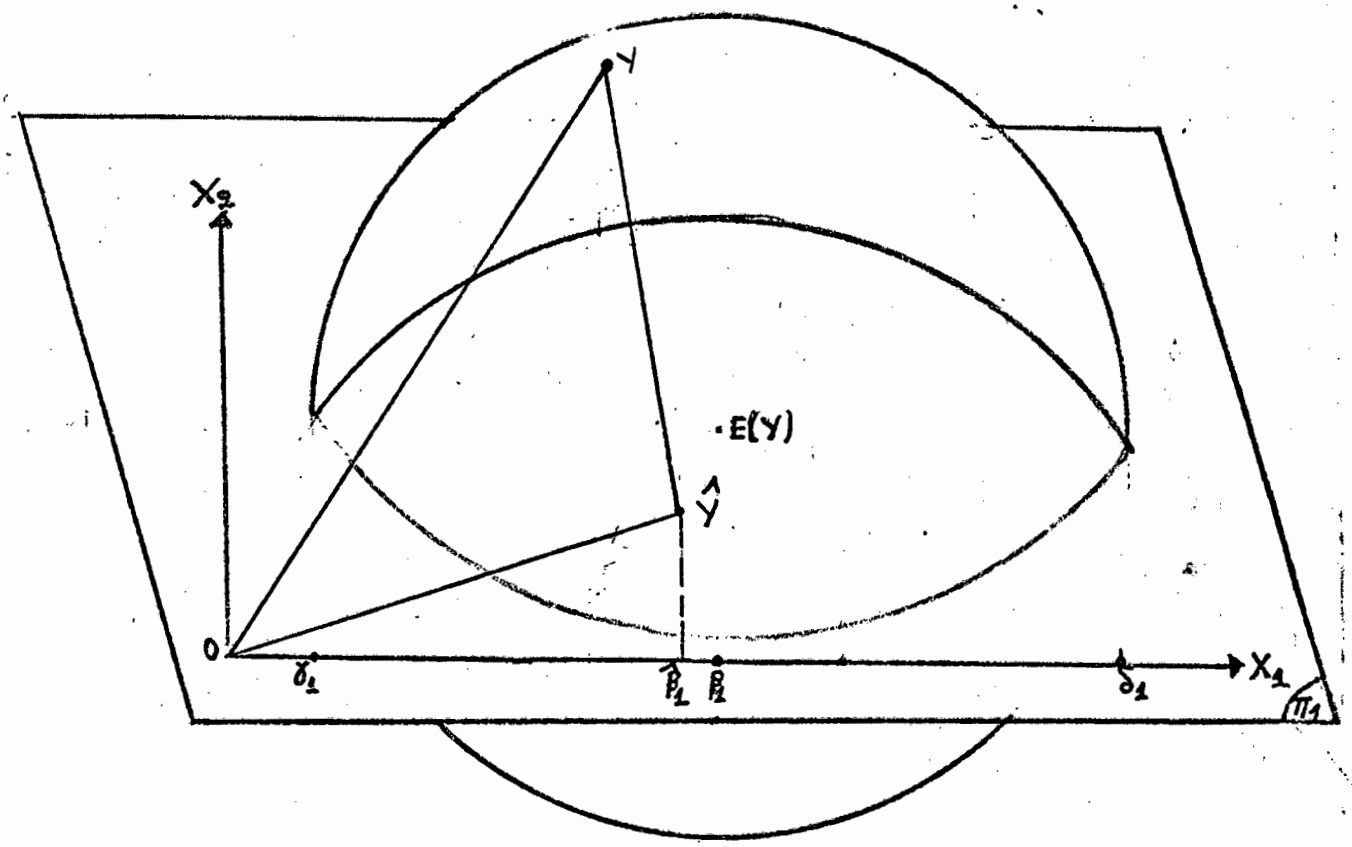


Figure 2.5.2

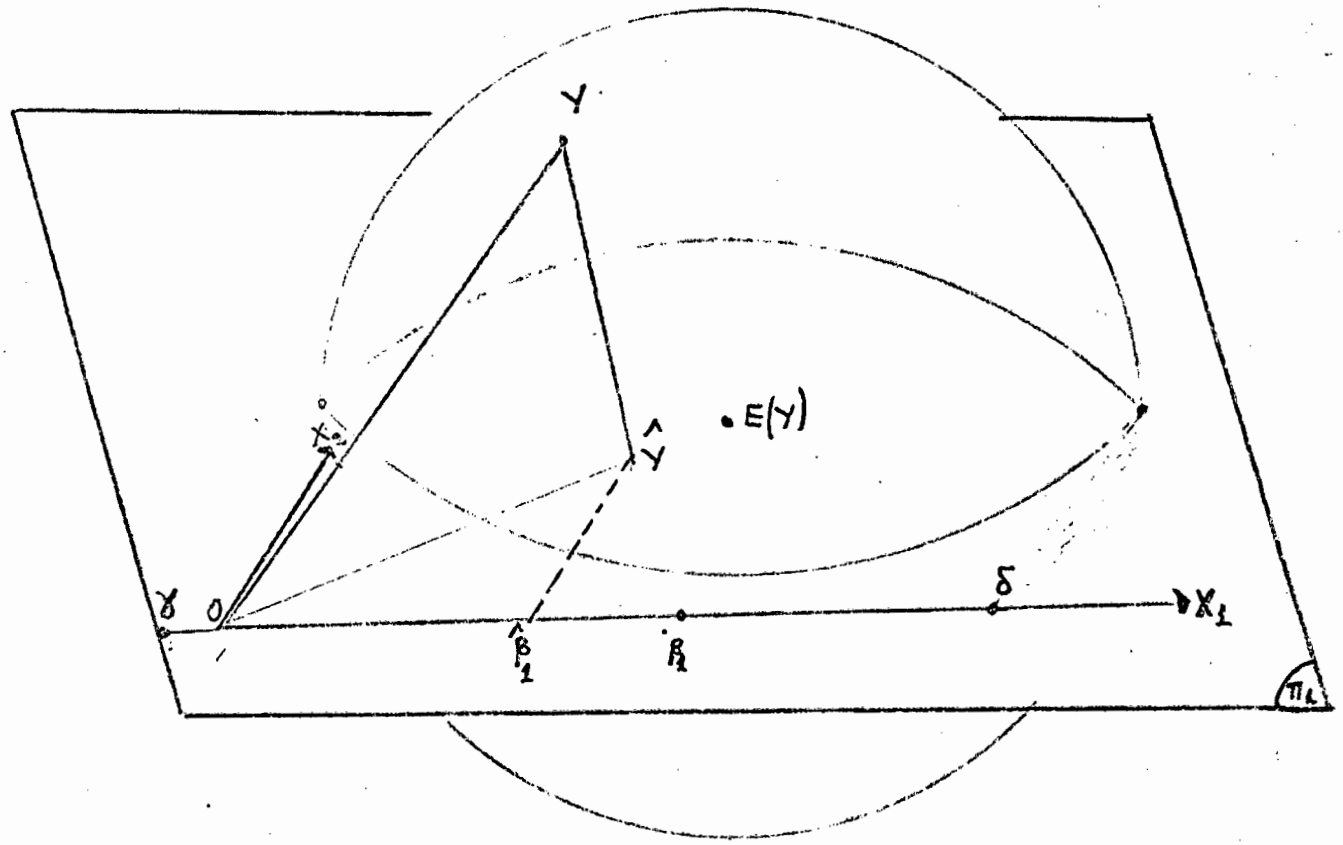


Figure 2.5.3

2.6 Detection of multicollinearity

(a) If X_1 and X_2 are two independent variables and if r_{ij} the correlation between them, and $R_{y.1,2}$ is the coefficient of determination then the multicollinearity is said to be "harmful" if $r_{ij} > R_{y.1,2}$. This method on extention to multiple dimensions breaks down.

(b) If $R^2_{y.1,2...,p}$ is the multiple coefficient of correlation between Y and all the regressor variables, and $R^2_{y.1,2,...,j-1,j+1,...,p}$ is the largest coefficient of determination regressing the Y on the independent variables except

X_j; then if a high degree of multicollinearity is present in the data

$$(2.6.1) \quad R^2_{y.1,2,\ldots,p} - R^2_{y.1,\ldots,j-1,j+1,\ldots,p}$$

will be small. But if the difference in (2.6.1) is small it does not imply that multicollinearity exists. It probably reveals the worthlessness of X_j as a predicator; moreover it does not give an indication which variables are linearly related.

(c) Suppose that $X'X$ is the correlation matric between the X's then $0 \leq |X'X| \leq 1$. If $|X'X| = 0$ an exact linear relationship exists between the X_j columns. If $|X'X| = 1$ then the X_j's are orthogonal.

In all other cases there exists some degree of multicollinearity which becomes severe as $|X'X| \simeq 0$. This method does not cast light on the nature of the linear relaships.

(d) For formal purposes we suppose $(Y,X_1,X_2,\ldots,X_p)' \sim N(\mu,\Sigma)$.

Stage 1

Test for the presence and severity of multicollinearity

$$H_0 : |X'X| = 1 \qquad H_1 : 0 \leq |X'X| < 1.$$

We use the following statistic proposed by Bartlett:

$$X^2_{|X'X|} = -(n-1 - \tfrac{1}{6}(2p+5))\log_e|X'X|$$

where $X^2_{|X'X|} \sim X^2_{\frac{1}{2}(p(p-1))}$.

If the observed value of $X^2_{|X'X|} > X^2_{(p(p-1))\frac{1}{2}}$ then reject H_0 otherwise accept H_0.

Stage 2

Test for localization. For $j = 1,\ldots,p$ test the hypothesis

$$H_0 : \rho^2_{j.1,2,\ldots,j-1,j+1,\ldots,p} = 0$$

$$H_1 : \rho^2_{j.1,2,\ldots,j-1,j+1,\ldots,p} \neq 0$$

Use the $$F' = \frac{(R^2_{j.1,2,\ldots,j-1,j+1,\ldots,p})/(p-1)}{(1-R^2_{j.1,2,\ldots,j-1,j+1,\ldots,p})/n-p} \sim F_{p-1,n-p}$$

if F' observed value is greater than the critical value reject H_0.

Stage 3

Examination of the pattern of multicollinearity.

Let $\rho_{i,j|p-2}$ be the population partial correlation between X_i and X_j keeping the other variables fixed. We want to test the hypothesis $H_0 : \rho_{i,j|p-2} = 0$ against

$H_1 : \rho_{i,j|p-2} \neq 0$.

Use the $$t' = \frac{(r_{i,j|p-2})(n-p)^{\frac{1}{2}}}{(1-r^2_{i,j|p-2})^{\frac{1}{2}}} \sim t_{n-p}$$

where $r_{i,j|p-2}$ is the sample partial correlation.

If the observed value t is larger than the critical value one rejects H_0, i.e. accept that X_i and X_j are responsible for the multicollinearity.

(e) Let $A = (Y^*, X^*)$ be the matrix of standardized dependent and independent variables. Find the eigenvalues and eigenvectors of $A'A$. If the eigenvalue is less than 0,03 and the last component of the corresponding eigenvector less than 0,10 then there exists a near singularity and that does not have predictive value, (refer to Section 5.8).

2.7 Proposed solutions

(i) Augmentation. This method is most effectively utilized when the multicollinearity can be identified as a result of sampling a subspace of the independent variables or for an over-defined model.

Many times, however, additional data cannot be collected for the following reasons:

(a) economic restraints

(b) changes in the population under study

(c) unavailability.

There are instances, moreover, when augmentation could change the population under study. For example, requiring that data be collected outside the region of the multicollinearity could result in a poor prediction equation since the additional data points may be rare in the population.

Including these rare data points in a proportion not representative of their presence in the population could in fact cause the "outliers" to heavily influence the estimated model, (refer to Chapter 3).

(ii) Least squares estimation with restrictions. As we have seen the least squares predictor may be adequate provided one only predicts in the region of the multicollinearity, i.e. only when $X_1,\ldots,X_{p_1}$ satisfy (2.1.2). If the prediction is restricted to this region, one may obtain satisfactory results. Possibly additional information may enable one to look at the values of the estimated parameters, (refer to Chapter 3).

(iii) The choice of variables to be included in the model has a direct influence on the problem of multicollinearity. If variables involved in a multicollinearity are not included in the model the problem will not arise. This also could have disasterous effects if the regressors excluded from the model are the best predictors.

In the case of the over-defined model it is desirable to use this method. But when multicollinearity is due to the sampling technique, variable selection procedures must be performed with great care.

If the source of multicollinearity is the physical constraints on the model then it matters little as far as prediction is concerned which of the variables involved in the

multicollinearity is removed from the estimated model.

(iv) Other methods

(1) The least square regression on the latent vectors of X'X, (refer to Chapter 5).

(2) Ridge regression : This method is consisting of adding small positive quantities to diagonal elements of X'X, (refer to Chapter 4).

(3) Shrunken-estimation which is some way is an extension of Ridge regression, (refer to Chapter 4).

(4) Latent Root regression analysis which is a modification of the principal component method, (refer to Chapter 5).

C H A P T E R 3

ORDINARY LEAST SQUARES WITH LINEAR CONSTRAINTS - AUGMENTING EXISTING DATA IN LINEAR REGRESSION

3.1 Introduction

In the first part of this chapter we begin the consideration of least squares problems in which the variables are required to satisfy specified linear constraints. Such problems arise in a variety of applications, e.g. fitting curves to data. A result on the sign of restricted least squares estimates is stated. We note that if X is an ill-conditioned matrix where $r(X) = m$ then we can improve our estimates by imposing $(1+p)-m$ restrictions.

From §3.5 to §3.6 we examine the problem, "given a set of non-orthogonal data, how should additional observations be added to remove the correlation among the independent variables, and under that condition, could minimum variance estimates be provided of the regression coefficients."

3.2 True restrictions in linear regression

Consider the Model A $Y = X\beta+e$ in (1.2.4). It is well known that $\hat{\beta} = (X'X)^{-1}X'Y$ is distributed as $N(\beta,\sigma^2C^{-1})$ where $C = X'X$. Suppose there is an exact structural form of the β, i.e. suppose one is certain that the underlying data are generated in such a way that one can constrain the para-

meter space according to (3.2.1) $H\beta = h$ where H is an $m \times (p+1)$ matrix of known constants with $r(H) = m \leq (p+1)$ and h is $(m\times 1)$ vector of known constants.

The problem of least squares with restrictions can be stated as follows:

"$\text{Min}_{\beta}(Y-X\beta)'(Y-X\beta)$ subject to $H\beta-h = 0$."

Therefore we minimize

$$F = (Y-X\beta)'(Y-X\beta)-2\lambda'(H\beta-h)$$

where λ is a $(m\times 1)$ vector of Lagrange multipliers, with respect to β and λ.

Setting the derivative of F with respect to β equal to zero gives for the minimizing value $\hat{\beta}_c$

$$\tfrac{1}{2}\frac{\partial F}{\partial \beta} = -X'Y + X'X\hat{\beta}_c - H'\lambda^* = 0$$

so

$$\hat{\beta}_c = \hat{\beta} + (X'X)^{-1}H'\lambda^* \tag{3.2.2}$$

Premultiplying by H we have

$$H\hat{\beta}_c = H\hat{\beta} + H(X'X)^{-1}H'\lambda^*$$

imposing the restriction $H\hat{\beta}_c = h$

we have

$$h = H\hat{\beta}_c = H\hat{\beta} + H(X'X)^{-1}H'\lambda^*$$

whence

$$\lambda^* = (H(X'X)^{-1}H')^{-1}(h-H\hat{\beta}). \tag{3.2.3}$$

So finally we have from (3.2.2) and (3.2.3)

(3.2.4) $\hat{\beta}_c = \hat{\beta} - C^{-1}H'(HC^{-1}H')^{-1}(H\hat{\beta}-h)$

where $\hat{\beta}$ is the O.L.S. estimator of the Model A.

The following are the properties of the constrained estimator $\hat{\beta}_c$

(1) (3.2.5) $\hat{\beta}_c \sim N(\beta - C^{-1}H'(HC^{-1}H')^{-1}(H\beta-h), \sigma^2[I-C^{-1}H'(HC^{-1}H')^{-1}H]C^{-1})$

(2) $\hat{\beta}_c$ is unbiased estimator of β if the restriction in equation (3.2.1) is true.

(3) $\hat{\beta}_c$ has smaller variance than $\hat{\beta}$, because

(3.2.6) $\text{var}(\hat{\beta}) - \text{var}(\hat{\beta}_c) = \sigma^2 C^{-1}H'(HC^{-1}H')^{-1}HC^{-1}$ is a positive semi-definite matrix of order $(p+1)$ and rank $m \leq 1+p$.

(4) The $R.S.S.(\hat{\beta}_c) > R.S.S.(\hat{\beta})$.

We note that if the restrictions are not true, then $\hat{\beta}_c$ is a biased estimator with smaller variance than the O.L.S. estimator $\hat{\beta}$.

3.3 Tests for linear restrictions

Theorem 3.3.1

The statistic $U_1 = (R.S.S.(\hat{\beta}_c) - R.S.S.(\hat{\beta}))/m \Big/ \frac{R.S.S.(\hat{\beta})}{n-p-1}$ $\sim F_{m,n-p-1}(\lambda)$ - noncentral F with noncentrality parameter $\lambda = (H\beta-h)'(HC^{-1}H')^{-1}(H\beta-h)/2\sigma^2$.

Proof

Considering the numerator of the U_1 statistic we have

$$(3.3.1)\quad \frac{Q_0}{\sigma^2} = (R.S.S.(\hat{\beta}_c) - R.S.S.(\hat{\beta}))/\sigma^2$$
$$= \frac{1}{\sigma^2}[Y-XC^{-1}H'(HC^{-1}H')^{-1}h]'M_0[Y-XC^{-1}H'(HC^{-1}H')h]$$

where $M_0 = XC^{-1}H'(HC^{-1}H')^{-1}HC^{-1}X'$ is an idempotent matrix of rank m.

We see that Q_0/σ^2 is quadratic form in the random vector $d_1 = \frac{1}{\sigma}[Y-XC^{-1}H'(HC^{-1}H')^{-1}h]$ where $d_1 \sim N(\frac{1}{\sigma}[X\beta-XC^{-1}H'(HC^{-1}H')^{-1}h],I)$ so $d_1'M_0d_1 = Q_0/\sigma^2 \sim \chi^2_m(\lambda)$ noncentral χ^2 with noncentrality parameter

$$(3.3.2)\quad \lambda = \frac{1}{2\sigma^2}(H\beta-h)'(HC^{-1}H')^{-1}(H\beta-h).$$ Similarly

$$(3.3.3)\quad Q_1/\sigma^2 = \frac{R.S.S.(\hat{\beta})}{\sigma^2}$$
$$= \frac{1}{\sigma^2}[Y-XC^{-1}H'(HC^{-1}H')^{-1}h]'M_1[Y-XC^{-1}H'(HC^{-1}H')^{-1}h]$$

is distributed as $\chi^2_{n-p-1}(\lambda)$ where $M = I-XC^{-1}X'$ is idempotent matrix of rank n-p-1. Independence of two quadratic forms involving a multivariate normal vector with mean μ and variance-covariance I requires that the product of the matrix of one quadratic form, and the matrix of the other quadratic form yields the null matrix. But

$$(3.3.4)\quad M_0M_1 = XC^{-1}H'(HC^{-1}H')^{-1}HC^{-1}X'[I-XC^{-1}X'] = O$$

so Q_0/σ^2 and Q_1/σ^2 are independent. Finally $U_1 = Q_0/m/Q_1/n-p-1 \sim F_{m,n-p-1}(\lambda)$ where λ as in (3.3.2).

The test $\lambda = 0$, tests the validity of the restrictions $H\beta = h$, and the distribution of U_1 becomes central F under

the null hypothesis. The F test has been used in empirical work to choose between two sets of estimators. For example, suppose one is fitting a regression model to cross section and time series data. It has been common in practice to test the hypothesis that the regression parameters are the same, say, at all time periods.

Acceptance of the hypothesis at some predetermined level is then used to justify constraining the parameter space.

But if one is willing to accept some bias in trade for a reduction in variance, then even if the restrictions are not true, one might still prefer the restricted estimator $\hat{\beta}_c$. For this reason we give some alternative tests.

(3.3(a)) Strong mean square error criterion (S.M.S.E.C.)

Definition 3.3.2

Consider two estimators $\tilde{\beta}_1$ and $\tilde{\beta}_2$ of β. We say that $\tilde{\beta}_1$ is better than $\tilde{\beta}_2$ according to S.M.S.E.C., if and only if

(3.3.5) $M.S.E.(\ell'\tilde{\beta}_1) \leq M.S.E.(\ell'\tilde{\beta}_2)$ for all $\ell \neq 0$

or equivalently

(3.3.6) $E(\tilde{\beta}_2-\beta)(\tilde{\beta}_2-\beta)' - E(\tilde{\beta}_1-\beta)(\tilde{\beta}_1-\beta)'$ - positive semi-definite.

Theorem 3.2.3. A necessary and sufficient condition for $\hat{\beta}_c$ to be better than $\hat{\beta}$ according to S.M.S.E.C. is

$$[(H\beta-h)'(HC^{-1}H')^{-1}(H\beta-h)/2\sigma^2] \leq \tfrac{1}{2} .$$

Proof

(3.3.7) $$\text{The M.S.E.}(\hat{\beta}_c) = \sigma^2[I-C^{-1}H'(HC^{-1}H')^{-1}H]C^{-1} + C^{-1}H'(HC^{-1}H')^{-1}(H\beta-h)(H\beta-h)'(HC^{-1}H')^{-1}HC^{-1}$$

(3.3.8) $$\text{M.S.E.}(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

(3.3.9) $$\text{M.S.E.}(\hat{\beta}_c)-\text{M.S.E.}(\hat{\beta}) = \sigma^2 C^{-1}H'(HC^{-1}H')^{-1} \cdot [HC^{-1}H' - \frac{1}{\sigma^2}(H\beta-h)(H\beta-h)'](HC^{-1}H')^{-1}HC^{-1}.$$

The matrix in (3.3.9) is of the form σ^2ABA' where $A = C^{-1}H'(HC^{-1}H')^{-1}$ is $(p+1)\times m$ - matrix and $r(A) = m \leq p+1$. Therefore ABA' is positive semi-definite iff B is positive semi-definite.

Consequently $\hat{\beta}_c$ is "better" than $\hat{\beta}$ iff

(3.3.10) $$\ell'[HC^{-1}H' - \frac{1}{\sigma^2}(H\beta-h)(H\beta-h)']\ \ell \geq 0;\quad \ell \neq 0.$$

Since $HC^{-1}H'$ is positive definite matrix

(3.3.10) is equivalent to

(3.3.11) $$Q = (\ell'(H\beta-h)(H\beta-h)'\ell/\sigma^2\ell'HC^{-1}H'\ell) \leq 1$$

But (3.3.11) satisfies the conditions of a version of the Cauchy-Schwarts inequality (see Rao "Linear Statistical Inference and its Applications" 2nd edition p.60).

So we have

(3.3.12) $\sup_{\ell} Q = Q_0$

where $Q_0 = (H\beta-h)'(HC^{-1}H')^{-1}(H\beta-h)/\sigma^2$

and the supremum is attained at the value

(3.3.13) $\ell_0 = (HC^{-1}H')^{-1}(H\beta-h)$.

Now we show that B is positive semi-definite if and only if

(3.3.14) $Q_0 \leq 1$.

We have that if (3.3.10) is satisfied for all $\ell \neq 0$ then (3.3.11) is satisfied for all $\ell \neq 0$ therefore for the particular ℓ_0 at (3.3.13). Now if $Q_0 \leq 1$ is satisfied, because Q_0 is the supremum of Q for all ℓ (3.3.11) must be satisfied for all ℓ. We can rewrite (3.3.14) in the form

(3.3.15) $\lambda = ((H\beta-h)'(HC^{-1}H')^{-1}(H\beta-h)/2\sigma^2) \leq \frac{1}{2}$.

Lehman in his book "Testing Statistical Hypotheses" shows that the U_1 test statistic can be used to provide a uniformly most powerful test for the S.M.S.E.C.

The density $f_\lambda(U)$ of the noncentral F in Theorem 3.3.1 is given by

$$(3.3.16)\quad f_\lambda(u) = \sum_{i=0}^{\infty} \Gamma\left(\frac{2i+m+q}{2}\right)\left(\frac{m}{q}\right)^{\frac{1}{2}(2i+m)} \lambda^i e^{-\lambda} u^{\frac{1}{2}(2i+m-2)} \Big/ \Gamma(q/2)\,\Gamma\left(\frac{2i+m}{2}\right) i! \left(1+\frac{mu}{q}\right)^{\frac{1}{2}(2i+m+q)}$$

where $q = n-p-1$.

It can be shown that $f_{\lambda_1}(u)/f_{\lambda_0}(u)$ is a nondecreasing function of the real valued function $w = \frac{mu}{q+mu}$ for all $\lambda_0 < \lambda_1$ where w is a monotone increasing function of u and vice versa.

The density function of w, $h_\lambda(w)$ is given by

$$(3.3.17)\quad h_\lambda(w) = \sum_{i=o}^{\infty} \frac{\lambda^i e^{-\lambda}}{i!} \frac{\Gamma((2i+m+q)/2)}{\Gamma(q/2)\Gamma((2i+m)/2)} w^{\frac{1}{2}(2i+m-2)} (1-w)^{(q-2)/2}$$

where $0 \leq w \leq 1$.

For $\lambda = 0$, (3.3.17) is the beta distribution. For $\lambda > 0$, $h_\lambda(w)$ is the noncentral beta. The family of noncentral F densities has the monotone likelihood ratio property in w, so the uniformly most powerful test of the S.M.S.E.C. is given by

$$H_0 : \lambda \leq \tfrac{1}{2} \quad \text{against} \quad H_1 : \lambda > \tfrac{1}{2}$$

Accept H_0 : if $W^* < W_a$

Reject H_0 : if $W^* \geq W_a$

where W_a is determined by

$$\int_o^{W_a} h_{\frac{1}{2}}(w)\, dw = 1-a$$

Tables have been produced by Wallace (1969) in the "Journal of the American Statistical Association, Vol. 64, p.1649-1663."

(3.3(b)) Some other criteria can be found by the same author in "Econometrica, Vol. 40, No. 4, 1972, p. 689-709." The following Table A summarizes the various criteria and tests of linear restrictions.

TABLE A

Criterion	Critical value of λ	Test: Compute u_1 in (Theorem 3.2.1) and compare it to the critical value of:
The test of restrictions $H\beta = h$ is true	$\lambda = 0$	The usual F distribution
$E(\hat{\beta}-\beta)(\hat{\beta}-\beta)'-E(\hat{\beta}_c-\beta)(\hat{\beta}_c-\beta)'$ is non-negative definite definite matrix	$\lambda \leq \frac{1}{2}$	Noncentral $F(\frac{1}{2})$ Tables : JASA (Vol. 64 (1969) pp. 1649-1663).
$E(\hat{\beta}-\beta)'(\hat{\beta}-\beta)-E(\hat{\beta}_c-\beta)'(\hat{\beta}_c-\beta)$ is positive scalar	$\lambda \leq \theta$	Noncentral $F(\theta)$ $\theta=\frac{1}{2}\lambda_p tr(X'X)^{-1}H'(H(X'X)^{-1}H')^{-1}H(X'X)^{-1}$ where λ_p is the smallest eigenvalue of X'X. Compute probability of larger F from approximation given in "Econometrica"
$E(\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta)-E(\hat{\beta}_c-\beta)'X'X(\hat{\beta}_c-\beta)$ - positive scalar	$\lambda \leq m/2$	Noncentral $F(m/2)$ Tabulated in "Econometrica"

"Econometrica" vol. 40, (1972) pp. 699-709.

3.4 The sign of restricted O.L.S. estimates

Least squares estimates of the coefficients of a linear regression model often have signs that are regarded by the researcher to be "wrong."

In an effort to obtain right signs, statistically insignificant variables are sometimes dropped from the equation. Surprisingly enough, there can be no change in sign of any coefficient that is more significant than the coefficient of the omitted variable. We know that

(3.4.1) $\hat{\beta}_i \sim N(\beta_i, \sigma^2 C^{ii})$ and the t-statistic for testing $\beta_i = 0$ is given by

$$(3.4.2) \quad t_i = \frac{\hat{\beta}_i}{\hat{\sigma}(C^{ii})^{\frac{1}{2}}}$$

If we constrain β_i to the h_1-scalar, then the H-matrix is a row vector with one in the ith column and zeroes elsewhere. We can rewrite (3.2.4) as follows:

$$(3.4.3) \quad \hat{\beta}_c = \hat{\beta} - C^{-1}_{(i)}(\hat{\beta}_i - h_1)/C^{ii}$$

where $C^{-1}_{(i)}$ is the ith column (row) of C^{-1}.

Lemma 3.4.1

The least-squares estimate of β_j and the constrained least-squares estimate of β_j, with $\beta_i = 0$, have the same sign

if the t-statistic of β_i is less in absolute value than the t-statistic of β_j.

Proof

We will show that if $t_i^2 \leq t_j^2$ then $\hat{\beta}_j$ and $\hat{\beta}_{c,j}$ have the same sign. Using (3.4.3) we have that

$$\hat{\beta}_{c,j} = \hat{\beta}_j - c^{ji}\hat{\beta}_i/c^{ii} \tag{3.4.4}$$

$$= \hat{\sigma}(c^{jj})^{\frac{1}{2}}[t_j - c^{ji}t_i/(c^{ii}c^{jj})^{\frac{1}{2}}]$$

C^{-1} is positive definite implies that

$$-1 \leq c^{ji}/(c^{ii}c^{jj})^{\frac{1}{2}} \leq 1. \tag{3.4.5}$$

More specifically $\hat{\beta}_j$ and $\hat{\beta}_{c,j}$ have the same sign iff

$$0 \leq \hat{\beta}_{c,j}\hat{\beta}_j = \hat{\sigma}^2 c^{jj} \quad [t_j^2 - c^{ji}t_i t_j/(c^{ii}c^{jj})^{\frac{1}{2}}]. \tag{3.4.6}$$

But by hypothesis $|t_j| > |t_i|$, and

$$t_j^2 \geq |t_i t_j| \geq |c^{ji}t_i t_j/(c^{ii}c^{jj})^{\frac{1}{2}}| \geq c^{ji}t_i t_j/(c^{ii}c^{jj})^{\frac{1}{2}} .$$

Corollary 3.4.2

The least-squares estimate of $\beta_j - c$ and the constrained least squares estimate of $\beta_j - c$, with β_i set to h_1, have the same sign if the t-statistic for testing $\beta_i = h_1$ is less in absolute value than the t-statistic for testing $\beta_j = c$.

Hint: We must prove that if $t_i^2 \leq t_j^2$ then $0 \leq (\hat{\beta}_{c,j}-c)(\hat{\beta}_j-c)$ where $\hat{\beta}_{c,j} = \hat{\beta}_j - c^{ji}(\hat{\beta}_i-h_1)/c^{ii}$

Theorem 3.4.3

The constrained least squares estimate of β_j must lie in the interval $(\hat{\beta}_j-(c^{jj})^{\frac{1}{2}}|t|\hat{\sigma}, \hat{\beta}_j+(c^{jj})^{\frac{1}{2}}|t|\hat{\sigma})$ where t is the t-statistic for testing the univariate restrictions.

Proof

We will show that

$$(3.4.7) \quad \hat{\beta}_j-(c^{jj})^{\frac{1}{2}}|t|\hat{\sigma} \leq \hat{\beta}_{cj} \leq \hat{\beta}_j+(c^{jj})^{\frac{1}{2}}|t|\hat{\sigma}$$

or equivalently

$$(3.4.8) \quad -(c^{jj})^{\frac{1}{2}}|t|\sigma^2 \leq -\frac{c^{ji}(\hat{\beta}_i-h_1)}{c^{ii}} \leq +(c^{jj})^{\frac{1}{2}}|t|\hat{\sigma}$$

or equivalently

$$(3.4.9) \quad -(c^{jj})^{\frac{1}{2}}|t|\hat{\sigma} \leq (-c^{ji}t(c^{ii})^{\frac{1}{2}}\sigma/c^{ii}) \leq (c^{jj})^{\frac{1}{2}}|t|\hat{\sigma}$$

or equivalently

$$(3.4.10) \quad -|t| \leq (-c^{ji}t/(c^{ii})^{\frac{1}{2}}(c^{jj})^{\frac{1}{2}}) \leq |t|$$

or equivalently

$$(3.4.11) \quad |c^{ji}t/(c^{ii}c^{jj})^{\frac{1}{2}}| \leq |t|$$

But (3.4.11) is true because of (3.4.12).

$$(3.4.12)\quad \left| \frac{C^{ji}}{(C^{ii}C^{jj})^{\frac{1}{2}}} \right| \leq 1$$

Theorem 3.4.4

The least squares estimate and the constrained least squares estimate of β_j, $\hat{\beta}_j$ and $\beta_{c,j}$ respectively, with the constraint $H = h$ satisfy the inequality

$$|\hat{\beta}_j - \hat{\beta}_{c,j}| \leq \hat{\sigma}(C^{jj})^{\frac{1}{2}} m^{\frac{1}{2}} U^{\frac{1}{2}}$$

where $r(H) = m$, U_1 is the statistic for testing the constraint and $\hat{\sigma}(C^{jj})^{\frac{1}{2}}$ is the standard error of $\hat{\beta}_j$.

The likelihood ellipsoid $(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})$ evaluated at the constrained least squares estimate $\hat{\beta}_c$ (3.2.4) takes on the value

$$s^2 = (H\hat{\beta}-h)'(HCH')^{-1}(H\hat{\beta}-h)$$

The estimate, $\hat{\beta}_c$, lies on the ellipsoid $(\hat{\beta}_c-\hat{\beta})'X'X(\hat{\beta}_c-\hat{\beta}) = s^2$. Projecting this ellipsoid onto the jth axis yields the desired inequality.

3.5 Augmenting data in linear regression.

Our main objective is to reduce the pairwise correlations among the independent variables. In addition, we choose to obtain the minimum variance estimates of the regression coefficients under the condition of orthogonality. This present-

ation is limited to the centered Model A_c, see (1.2.4), but here $i = 1,2,\ldots,n+1,\ldots,n+m = N$, where m points are added to the original n-data points.

In a study of this type, it is necessary to choose a criterion for judging whether one set of added experimental observations is superior to another set of observations. We choose to minimize the determinant of the inverse of the matrix of the corrected sum of squares and cross-products of the independent variables, $|(X'X)^{-1}|$, subject to the condition that the off-diagonal elements of $(X'X)$ are zero after adding data points. Minimizing $|(X'X)^{-1}|$ is equivalent to maximizing $|X'X|$. It will be shown that this minimizes the volume of the confidence region for the regression coefficients, and minimizes the maximum variance of a predicted value, $\text{var}(\hat{Y})$, in the experimental region.

3.6 Development of the augmenting procedure

Suppose the independent variables are transformed so that the experimental limits for each variable range from -1 to $+1$. That is, the experimental region is a hyper-cube centered about the origin and bounded by the hyper-planes at -1 and $+1$ in each dimension. The solutions obtained here, based on maximizing $|(X'X)|$ are invariant under change of scale. Thus without loss of generality we make the linear transformation

$$(3.6.1) \qquad Z_{ij} = [X_{ij}-(X_{Lj}+X_{uj})/2]/(X_{uj}-X_{Lj})/2$$

where X_{Lj} and X_{uj} are the lower and upper limits for X_j.

For orthogonal design we have

$$(3.6.2) \quad \text{var}(\hat{Y}) = \sigma^2[\frac{1}{N} + \sum_{j=1}^{p} \frac{(Z_j-\bar{Z}_j)^2}{\sum_{i=1}^{n}(Z_{ij}-\bar{Z}_j)^2}]$$

Proposition 3.6.1

The maximum $\text{var}(\hat{Y})$ (see (3.6.2)) in the experimental region is minimized by adding points to the 2^p corners of the region such that $\bar{Z}_j = 0$ for all $j = 1,2,\ldots,p$.

Proof

The maximum var(Y) is at one or more of the corners of the experimental region, i.e. $Z_j = \pm 1$ for all j. Thus adding points such that $\bar{Z}_j = 0$ for all j, minimize the maximum values of the $(Z_j-\bar{Z}_j)^2$ in the numerator of (3.6.2). We have for the denominator

$$(3.6.3) \quad \sum_{i=1}^{N}(Z_{ij}-\bar{Z}_j)^2 = \sum_{i=1}^{N}Z_{ij}^2 - N\bar{Z}_j^2$$

$$= \sum_{i=1}^{n}Z_{ij}^2 + \sum_{i=n+1}^{n+m} Z_{ij}^2 - N\bar{Z}_j^2$$

The value $\sum_{i=1}^{n}Z_{ij}^2$ is fixed. The maximum values that the augmented data points can assume are ± 1. Thus the sum of squares is maximised by placing all m augmented points at ± 1 such that $\bar{Z}_j = 0$. Then

$$(3.6.4) \quad \max\sum_{i=1}^{N}(Z_{ij}-\bar{Z}_j)^2 = \sum_{i=1}^{n}Z_{ij}^2 + m$$

is achieved.

Repeating this procedure in all p-dimensions places the

augmented points in the 2^p corners of the region such that $\bar{Z}_j = 0$ for all j. This solution can be accomplished if $\sum_{i=1}^{n} Z_{ij}$ are integers for all j, since only integral values of ± 1 are added.

Remarks

(1) As the number of additional points, m, increase the value of $\bar{Z}_j$ will tend to zero and the minimum variance will tend to be more closely approximated.

(2) The data points are added such that $\bar{Z}_j \simeq 0$ with

(3.6.5) $\sum_{i=1}^{n} Z_{ij}Z_{ij'} \simeq 0$ for all $j \neq j'$.

Proposition 3.6.2

$|(X'X)|$ is maximized when the determinant of the transformed variables (see 3.6.1) is maximized.

Hint: Use the remark 2 and the fact that

(3.6.6) $|(X'X)| = (\prod_{J=1}^{p}(X_{uj}-X_{Lj})/2)\prod_{J=1}^{p}[\sum_{i=1}^{n}(Z_{ij}-\bar{Z}_j)^2]$.

(a) Two independent variables: It is desired to add m additional points to the existing n points such that

(3.6.7) $\bar{Z}_1=\bar{Z}_2= \sum_{i=1}^{N} Z_{i1}Z_{i2} = 0$

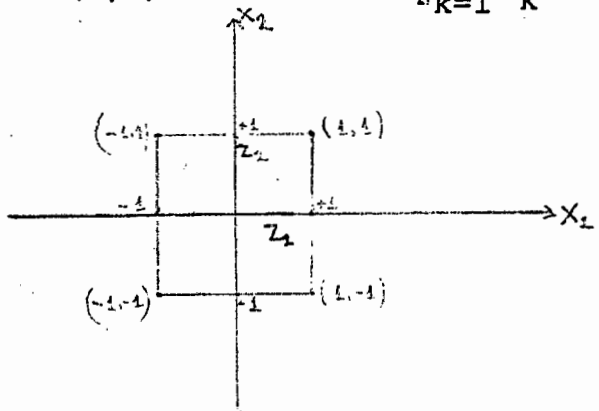
or

(3.6.8) $\sum_{i=1}^{m} Z_{i1}+\sum_{i=n+1}^{n+m} Z_{i1} = \sum_{i=1}^{n} Z_{i2}+\sum_{i=n+1}^{n+m} Z_{i2} = \sum_{i=n}^{n} Z_{i1}Z_{i2}$

$+ \sum_{i=n+1}^{n+m} Z_{i1}Z_{i2} = 0$ where $n+m = N$.

Identify the number of new points, n_k, to be added to the kth corner as follows:

Rule A: n_1 at $(-1,-1)$, n_2 at $(+1,-1)$, at n_3 at $(-1,1)$ at n_4 at $(1,1)$ such that $\sum_{k=1}^{4} n_k = m$



The values of Z_{ij} for $i = n+1,\ldots,n+m$ in (3.6.8) may be replaced by ± 1 according to the above rule A giving

$$(3.6.9)\quad \sum_{i=1}^{n} Z_{i1} - n_1 + n_2 - n_3 + n_4 = 0$$

$$\sum_{i=1}^{n} Z_{i2} - n_1 - n_2 + n_3 + n_4 = 0$$

$$\sum_{i=1}^{n} Z_{i1} Z_{i2} + n_1 - n_2 - n_3 + n_4 = 0$$

$$n + \sum_{i=1}^{4} n_k = n+m = N$$

The above system can be solved with respect to n_k's and thus we have the number of points to be added to each corner

$$(3.6.10)\quad n_1 = \tfrac{1}{4}\left(m+\sum_{i=1}^{n} Z_{i1}+\sum_{i=1}^{n} Z_{i2}-\sum_{i=1}^{n} Z_{i1}Z_{i2}\right)$$

$$n_2 = \tfrac{1}{4}\left(m-\sum_{i=1}^{n} Z_{i1}+\sum_{i=1}^{n} Z_{i2}+\sum_{i=1}^{n} Z_{i1}Z_{i2}\right)$$

$$n_3 = \tfrac{1}{4}\left(m+\sum_{i=1}^{n} Z_{i1}-\sum_{i=1}^{n} Z_{i2}+\sum_{i=1}^{n} Z_{i1}Z_{i2}\right)$$

$$n_4 = \tfrac{1}{4}\left(m-\sum_{i=1}^{n} Z_{i1}-\sum_{u=1}^{n} Z_{i2}-\sum_{i=1}^{n} Z_{i1}Z_{i2}\right)$$

(b) Three independent variables

Here, we may add point in 2^3 corners subject to the conditions

$$(3.6.11)\quad \sum_{i=1}^{N} Z_{i1} = 0$$

$$\sum_{i=1}^{N} Z_{i2} = 0$$

$$\sum_{i=1}^{N} Z_{i3} = 0$$

$$\sum_{i=1}^{N} Z_{i1} Z_{i2} = 0$$

$$\sum_{i=1}^{N} Z_{i1} Z_{i3} = 0$$

$$\sum_{i=1}^{N} Z_{i2} Z_{i3} = 0$$

The number of points to be added to each of the eight corners are the unknowns n_k $k = 1,\ldots,8$.

The equation (3.6.11) with the $\sum_{i=1}^{8} n_k = m$ constitute a system of 7 equations in 8 unknowns. Theoretically there are an infinite number of solutions. Add the condition $\sum_{i=1}^{N} Z_{i1} Z_{i2} Z_{i3} = 0$ because this does not affect the properties and this condition leads to a simple solution.

The kth corner, at which n_k new points are added, is defined by the coordinates given in Table B. For the theoretical solution of the number of points to be added at the first corner

$$n_1 = \frac{1}{8}(m+\sum_{i=1}^{n} Z_{i1}+\sum_{i=1}^{n} Z_{i2}+ \sum_{i=1}^{n} Z_{i3}- \sum_{i=1}^{n} Z_{i1} Z_{i2}$$

$$- \sum_{i=1}^{n} Z_{i1} Z_{i3}- \sum_{i=1}^{n} Z_{i2} Z_{i3}+ \Sigma Z_{i1} Z_{i2} Z_{i3}$$

TABLE B

	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8
Z_1	-1	1	-1	1	-1	1	-1	1
Z_2	-1	-1	1	1	-1	-1	1	1
Z_3	-1	-1	-1	-1	1	1	1	1
m	+	+	+	+	+	+	+	+
ΣZ_{i1}	+	-	+	-	+	-	+	-
ΣZ_{i2}	+	+	-	-	+	+	-	-
ΣZ_{i3}	+	+	+	+	-	-	-	-
$\Sigma Z_{i1}Z_{i2}$	-	+	+	-	-	+	+	-
$\Sigma Z_{i1}Z_{i3}$	-	+	-	+	+	-	+	-
$\Sigma Z_{i2}Z_{i3}$	-	-	+	+	+	+	-	-
$\Sigma Z_{i1}Z_{i2}Z_{i3}$	+	-	-	+	-	+	+	-

(c) P-independent variables

Table analogous to Table B can be constructed and used to obtain the equations for the n_k in p-dimension. In general, the values of n_k will not be integers and some values may be negative or smaller.

Rounding rule B

Set the negative numbers equal to zero and round the remaining numbers to integers such that they sum to m.

Note: The additional points can be added one at a time such that $|X'X|$ is maximized at each step. This is accomplished by adding one point to a corner and evaluating $|X'X|$. This is repeated for each of the 2^p corners. The point that is selected is the one that maximizes $|X'X|$. This process is repeated until m-points are added.

C H A P T E R 4

RIDGE REGRESSION AND EXTENSIONS

4.1 Introduction

The method of ridge regression was proposed by Hoerl and Kennard (1970) as an alternative to least squares estimation of the coefficients of a linear model. This concept has generated considerable interest in the literature.

Marquardt (1970) noted the relation between ridge estimators and the generalised inverse estimator, and in 1974 noted the relation to robust regression. Mayer and Wilke (1973) considered a general class of estimators based on linear transforms of least squares estimators, which included ridge and shruken estimators as special cases. MacDonald and Schwing (1973) and Marquardt and Snee (1975) provided applications to real data sets.

Confine and Stone (1973) examined the concept of ridge regression and were generally critical.

Much of the discussion of ridge regression centres around the choice of the constant k which will be defined below.

The concept of shrinkage estimation is also discussed.

4.2 Properties of ridge regression

Hoerl and Kennard (1970) have proposed the following estimator

$$\hat{\beta}_R = (X^{*\prime}X^{*}+kI)X^{*\prime}Y^{*} \tag{4.2.1}$$

such that $k \geq o$ instead of the least squares estimator

$$\hat{\beta}^{*} = (X^{*\prime}X^{*})^{-1}X^{*\prime}Y^{*} \tag{4.2.2}$$

We note that we are working with the standardized model A_s (see (1.2.5)).

(i) The relation between $\hat{\beta}_R$ and $\hat{\beta}^{*}$ is given by

$$\begin{aligned}\hat{\beta}_R &= (X^{*\prime}X^{*}+kI)^{-1}X^{*\prime}Y^{*} \\ &= WX^{*\prime}Y^{*} \\ &= (X^{*\prime}X^{*}+kI)^{-1}X^{*\prime}X^{*}\hat{\beta}^{*} \\ &= [I+k(X^{*\prime}X^{*})^{-1}]^{-1}\hat{\beta}^{*} \\ &= Z\hat{\beta}^{*}\end{aligned} \tag{4.2.3}$$

Note

$$W = (X^{*\prime}X^{*}+kI)^{-1} \quad \text{and} \quad Z = (X^{*\prime}X^{*}+kI)^{-1}X^{*\prime}X^{*} \tag{4.2.4}$$

(ii) If $\lambda_{max} = \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p = \lambda_{min} > 0$ are the eigenvalues of $X^{*\prime}X^{*}$, $\lambda_i(W)$ and $\lambda_i(Z)$ the eigenvalues of W and Z, respectively then

$$\begin{aligned}\lambda_i(W) &= \frac{1}{\lambda_i+k} \\ \lambda_i(Z) &= \frac{\lambda_i}{\lambda_i+k}\end{aligned} \tag{4.2.5}$$

(iii) The following relation holds between Z and W:

$$Z = Z(k) = I-k[X^{*\prime}X^{*}+kI]^{-1} = I-kW. \tag{4.2.6}$$

We have

$$(4.2.7) \quad Z = (X^{*\prime}X^{*}+kI)^{-1}X^{*\prime}X^{*} = WX^{*\prime}X^{*}$$

$$(4.2.8) \quad W^{-1}Z = W^{-1}WX^{*\prime}X^{*}$$

$$(4.2.9) \quad W^{-1}(I-kW) = W^{-1} - W^{-1}kW$$
$$= X^{*\prime}X^{*}+kI - kI$$
$$= X^{*\prime}X^{*}$$

From (4.2.8) and (4.2.9) we have (4.2.6).

(iv) $\hat{\beta}_R$ for $k \neq 0$ is shorter than $\hat{\beta}^*$, i.e.

$$(4.2.10) \quad \hat{\beta}_R'\hat{\beta}_R < \hat{\beta}^{*\prime}\hat{\beta}^* \quad \text{(Euclideian norm)}.$$

Since $\hat{\beta}_R = Z\hat{\beta}^*$, we have that

$$(4.2.11) \quad \hat{\beta}_R'\hat{\beta}_R = \hat{\beta}^{*\prime}Z'Z\hat{\beta}^*$$
$$\leq (\lambda_i(Z))_{max}\hat{\beta}^{*\prime}\hat{\beta}^*$$
$$< \hat{\beta}^{*\prime}\hat{\beta}^* .$$

We have to use the fact that $(\lambda_i(Z))_{max} = \frac{\lambda_1}{\lambda_1+k} < 1$, and the following proposition.

Proposition

If λ_1 and λ_2 are the max and min characteristic roots of a symmetric matrix A respectively, then $x'x\lambda_2 \leq x'Ax \leq x'x\lambda_1$. (Refer to Press, T. "Applied Multivariate Analysis" p.36.)

(v) For $\hat{\beta}_R$ the residual sum of squares is given by

$$(4.2.12) \quad R.S.S.(\hat{\beta}_R) = (Y^*-X^*\hat{\beta}_R)'(Y^*-X^*\hat{\beta}_R)$$

$$= (Y^*-X^*\hat{\beta}^*)'(Y^*-X^*\hat{\beta}^*) + (\hat{\beta}_R-\hat{\beta}^*)'X^{*\prime}X^*(\hat{\beta}_R-\hat{\beta}^*)$$

$$= R.S.S.(\hat{\beta}^*) + \phi(\hat{\beta}_R).$$

(4.2.13) (vi) $E(\hat{\beta}_R) = E(Z\hat{\beta}^*)$

$$= ZE(\hat{\beta}^*)$$

$$= Z\beta^*$$

and $\hat{\beta}_R$ is therefore a biased estimator.

(4.2.14) $cov(\hat{\beta}_R) = cov(Z\hat{\beta}^*)$

$$= \sigma^2 Z(X^{*\prime}X^*)^{-1}Z'$$

4.3 Ridge trace

Ridge regression has two aspects. The first is the ridge trace which is a two dimensional plot of $\hat{\beta}_{R,i}(k)$ and $R.S.S.(\hat{\beta}_R(k))$ for various values of $k \in [0,1]$. The second is the determination of the value of k that gives a better estimate of β^*. As we will see later on, using the ridge estimate we allow a little bias but we reduce the variance considerably.

Let B be any estimate of β^* then

(4.3.1) $R.S.S.(B) = R.S.S.(\hat{\beta}^*) + \phi(B)$

where $\phi(B) = (B-\hat{\beta}^*)'X^{*\prime}X^*(B-\hat{\beta}^*)$.

Contours of constant $R.S.S.(B)$ are the surfaces of hyperellipsoids centered at $\hat{\beta}^*$. There is a continum of values of B that will satisfy the $R.S.S.(B) = R.S.S.(\hat{\beta}^*) + \phi_0$ where

$\phi_0 > 0$ is a fixed increment.

The distance from $\hat{\beta}^*$ to β^* is

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

and

(4.3.3) $E(L_1^2) = \sigma^2 \sum_{i=1}^{p} (1/\lambda_i)$. (For further details see §4.5).

Now if $X^{*\prime}X^*$ becomes ill-conditioned (4.3.3) will tend to be large. The value of $\phi(B)$ will be large if B is far from $\hat{\beta}^*$.

In particular, the worse the conditioning of $X^{*\prime}X^*$, the more $\hat{\beta}^*$ can be expected to be large; but the worse the conditioning, the further one can move from $\hat{\beta}^*$ without an appreciable increase in R.S.S., i.e. $(B-\hat{\beta}^*)'X^{*\prime}X^*(B-\hat{\beta}^*)$ will not be inflated.

Thus if B moves away from the $\hat{\beta}^*$, the movement should be in the direction which will shorten the length of the regression vector B. The ridge trace can be shown to be following a path through the sum of squares surface so that for fixed value of R.S.S.(B) a single value of B is chosen, that is the one with minimum length.

Now the problem can be stated as follows:

Problem "minimize $B'B$ subject to $(B-\hat{\beta}^*)'X^{*\prime}X^*(B-\hat{\beta}^*) = \phi_0$." We use Lagrange multipliers to solve the problem. We have to minimize $F = B'B + \frac{1}{k}\{(B-\hat{\beta}^*)'X^{*\prime}X^*(B-\hat{\beta}^*) - \phi_0\}$.

Taking the first derivative of F with respect to B and setting it equal to zero, we obtain

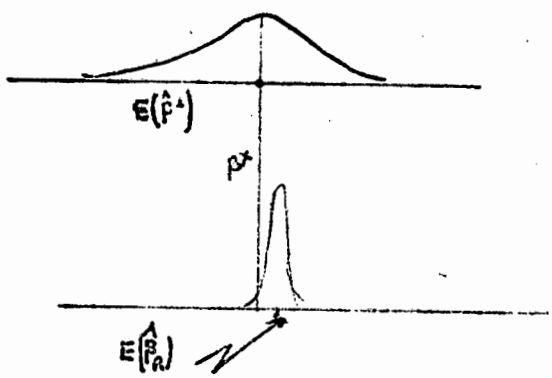
$$\frac{\partial F}{\partial B} = 2B + \frac{1}{k}(2X^{*\prime}X^*B - 2X^{*\prime}X^*\hat{\beta}^*) = 0$$

or $B = \hat{\beta}_R = (X^{*\prime}X^* + kI)^{-1}X^{*\prime}Y^*$, where k is chosen to satisfy the constraint.

In practice it is easier to chose $k \geq 0$ and then to compute ϕ_0. Note that the ridge estimates give the smallest regression coefficients consistent with a given degree of increase in R.S.S.

4.4 Geometric picture of ridge regression

(a)



Zero Bias $E(\hat{\beta}^*) = \beta^*$
Large variance

Non-zero Bias $E(\hat{\beta}_R^*) \neq \beta^*$
small variance

Figure 1

Figure 1 above illustrates the situation where an estimator $\hat{\beta}^*$ is unbiased, but is plagued by large variance. Typical con-

fidence limits for this estimator would be nearly half the width of the figure.

At the bottom is the corresponding frequency function for a biased estimator with much smaller variance.

(b)

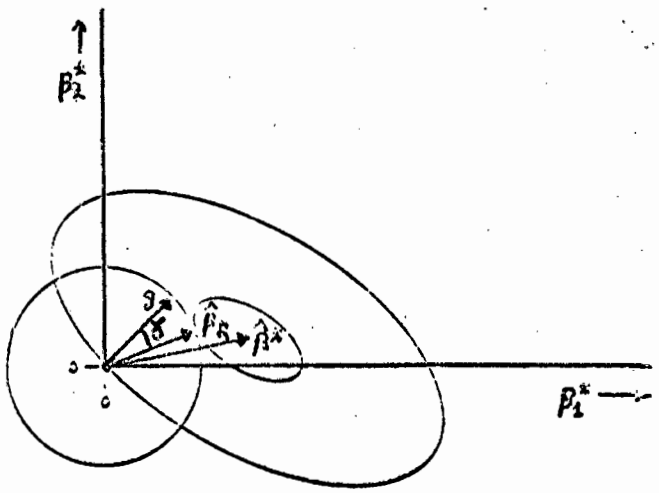


Figure 2

Suppose that we want to estimate the parameter vector $\beta^* = (\beta_1^*, \beta_2^*)$. The point $\hat{\beta}^*$ at the center of the ellipses is the least-square solution. At $\hat{\beta}^*$ the R.S.S. say ϕ achieves its minimum.

The small ellipse is the locus of points in the β_1^*, β_2^* - plane, where the R.S.S. is constant at some value, say ϕ_0, larger than the minimum value. The circle about the origin is tangent to the small ellipse at $\hat{\beta}_R$.

We note that $\hat{\beta}_R$ is the shortest vector that will give a R.S.S. as small as the ϕ value anywhere on the small ellipse.

The gradient $g = X^{*\prime}Y^*$ is perpendicular to the ϕ contour through the origin. The ridge estimator $\hat{\beta}_R$ always lies between $\hat{\beta}^*$ and g.

4.5 Mean square properties of ridge regression

Definition 4.5.1

If $\tilde{\beta}$ is any estimate of β^* then the square distance from $\tilde{\beta}$ to β^* is given by $L^2=(\tilde{\beta}-\beta^*)'(\tilde{\beta}-\beta^*)$.

Let $L_1^2(k)$ and L_1^2 be the squared distances of $\hat{\beta}_R$ and $\hat{\beta}^*$ from β^*, respectively, then

(4.5.2) (i) $$\begin{aligned} E(L_1^2) &= E\{tr(\hat{\beta}^*-\beta^*)'(\hat{\beta}^*-\beta^*)\} \\ &= E\{tr(\hat{\beta}^*-\beta^*)(\hat{\beta}^*-\beta^*)'\} \\ &= E\{tr(\sigma^2(X^{*\prime}X^*)^{-1}\} \\ &= \sigma^2 tr(X^{*\prime}X^*)^{-1} \end{aligned}$$

or equivalently

(4.5.3) $$E(\hat{\beta}^{*\prime}\hat{\beta}^*) = \beta^{*\prime}\beta^* + \sigma^2 tr(X^{*\prime}X^*)^{-1}$$

(ii) $var(L_1^2) = 2\sigma^4 tr(X^{*\prime}X^*)^{-2}$ if $e \sim N(0,\sigma^2 I)$

(iii) It is easy to express the $E(L_1^2)$ and $var(L_1^2)$ in terms of the eigenvalues of the $X^{*\prime}X^*$ matrix.

(4.5.4) $$E(L_1^2) = \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i}$$

(4.5.5) $$var(L_1^2) = 2\sigma^4 \sum_{i=1}^{p} (1/\lambda_i)^2$$

Hint: We have $(\hat{\beta}_i-\beta_i) \sim N(0,\sigma^2/\lambda_i)$, this implies that $(\hat{\beta}_i^*-\hat{\beta}_i^*)^2 \sim (\sigma^2/\lambda_i)\chi_1^2$, so $\text{var}(L_1^2) = \text{var}(\sum_{i=1}^p (\frac{\sigma^2}{\lambda_i})\chi_1^2)$

$$= 2\sigma^4 \sum_{i=1}^p (1/\lambda_i)^2$$

Lemma 4.5.2

$E(L_1^2(k)) = \gamma_1(k) + \gamma_2(k)$ where $\gamma_1(k)$ is the sum of the variances of the $\hat{\beta}_{R,i}$ (total variance) and $\gamma_2(k)$ is the bias introduced when $\hat{\beta}_R$ is used instead of the $\hat{\beta}^*$.

Proof

(4.5.6) $E(L_1^2(k)) = E(\hat{\beta}_R-\beta^*)'(\hat{\beta}_R-\beta^*)$

$= E(Z\hat{\beta}^*-Z\beta^*+Z\beta^*-\beta^*)'(Z\hat{\beta}^*-Z\beta^*+Z\beta^*-\beta^*)$

$= E[(Z\hat{\beta}^*-Z\beta^*)'(Z\hat{\beta}^*-Z\beta^*)+(Z\beta^*-\beta^*)'(Z\beta^*-\beta^*)]$

$= E[(\hat{\beta}^*-\beta^*)'Z'Z(\hat{\beta}^*-\beta^*)+\beta^{*\prime}(Z-I)'(Z-I)\beta^*]$

$= E[\text{tr}(\hat{\beta}^*-\beta^*)'Z'Z(\hat{\beta}^*-\beta^*)+\beta^{*\prime}(Z-I)'(Z-I)\beta^*]$

$= \text{tr}[Z'ZE(\hat{\beta}^*-\beta^*)(\hat{\beta}^*-\beta^*)']+\beta^{*\prime}(Z-I)'(Z-I)\beta^*$

$= \sigma^2\text{tr}(Z'Z(X^{*\prime}X^*)^{-1})+\beta^{*\prime}(-kW)'(-kW)\beta^*$ see (4.2.6)

$= \sigma^2\text{tr}(W'WX^{*\prime}X^*)+k^2\beta^{*\prime}W'W\beta^*$

$= \sigma^2\text{tr}(W'Z)+k^2\beta^{*\prime}(X^{*\prime}X^*+kI)^{-2}\beta^*$

$= \sigma^2\text{tr}(X^{*\prime}X^*+kI)^{-1}-k\sigma^2\text{tr}(X^{*\prime}X^*+kI)^{-2}+k^2\beta^{*\prime}$

$(X^{*\prime}X^*+kI)^{-2}\beta^*$

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

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

$= \gamma_1(k) + \gamma_2(k)$

Now we will show that $\gamma_1(k)$ is the total variance and $\gamma_2(k)$ is the bias.

From (4.2.14) we have

(4.5.7) $\text{var}(\hat{\beta}_{R.i}) = \sigma^2 g_{ii}$ where g_{ii} is the ith diagonal element of the

(4.5.8) $G = Z(X^{*\prime}X^*)^{-1}Z'$

(4.5.9)
$$\begin{aligned}\sum_{i=1}^{p} \text{var}(\hat{\beta}_{R\,i}) &= \sigma^2 \sum_{i=1}^{p} g_{ii} \\ &= \sigma^2 \text{tr}(G) \\ &= \sigma^2 \text{tr}(Z(X^{*\prime}X^*)^{-1}Z') \\ &= \sigma^2 \text{tr}(Z'Z(X^{*\prime}X^*)^{-1}) \\ &= \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)^2} \\ &= \gamma_1(k)\end{aligned}$$

$\gamma_2(k) = (Z\beta^*-\beta^*)'(Z\beta^*-\beta^*)$ is the squared distance from $Z\beta^*$ to β^* and $\gamma_2(k) = 0$ if $k = 0$. Since $Z(k) = I$ when $k = 0$, thus $\gamma_2(k)$ is considered as the square of the bias introduced when $\hat{\beta}_R$ is used instead of $\hat{\beta}^*$.

Theorem 4.5.3

The total variance $\gamma_1(k)$ is a continuous monotonic decreasing function of k.

Proof

We have

$$\gamma_1(k) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)^2}$$

(4.5.10) $\frac{d\gamma_1(k)}{dk} = -2\sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)^3} < 0$ thus $\gamma_1(k)$ is monotonous decreasing function and continuous.

Now we examine the value of the derivative (4.5.10) in the neighbourhood of the origin.

(4.5.11) $\lim_{k\to o^+} \frac{d\gamma_1(k)}{dk} = -2\sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i^2}$

$= -2p\sigma^2$ i.e. $X^{*\prime}X^*$ orthogonal

$= -\infty$ as $\lambda_p \to 0$ i.e. $X^{*\prime}X^*$ becomes singular

Theorem 4.5.4

The squared bias $\gamma_2(k)$ is a continuous monotonical increasing function of k.

Proof

We have

$$\gamma_2(k) = k^2\beta^{*\prime}(X^{*\prime}X^*+kI)^{-2}\beta^*$$

We know that there exist an orthogonal matrix P such that

(4.5.12) $X^{*\prime}X^* = P'\Lambda P$, $P'P = I$ and $\Lambda = \mathrm{diag}\{\lambda_1,..,\lambda_p\}$ contains the eigenvalues of the $X^{*\prime}X^*$.

(4.5.13)
$$\begin{aligned}\gamma_2(k) &= k^2\beta^{*\prime}(P'\Lambda P+kI)^{-2}\beta^* \\ &= k^2\beta^{*\prime}(P'\Lambda P+P'kIP)^{-2}\beta^* \\ &= k^2\beta^{*\prime}(P'(\Lambda P+kIP))^{-2}\beta^* \\ &= k^2(P\beta^*)'(\Lambda+kI)^{-2}(P\beta^*) \quad | \text{ we set } P\beta^* = a\end{aligned}$$

$$= k^2 a'(\Lambda+kI)^{-2} a$$

$$= k^2 \sum_{i=1}^{p} \frac{a_i^2}{(\lambda_i+k)^2}$$

We observe $\gamma_2(k) > 0$ because $\lambda_1 + k > 0$ and there are not singularities in the sum.

Clearly $\lim_{k\to o} \gamma_2(k) = \gamma_2(0) = 0$, that is, $\gamma_2(k)$ is a continuous function for $k \geq 0$.

For $k > 0$ (4.5.13) can be written as

$$(4.5.14) \quad \gamma_2(k) = \sum_{i=1}^{p} \frac{a_i^2}{(1+\lambda_i/k)^2}$$

since $\lambda_i > 0$, $i = 1,\ldots,p$, that is λ_i/k are monotone decreasing as k increases, $\left(1 + \frac{\lambda_i}{k}\right)^2$ are monotone decreasing as k increases, that is, $\frac{1}{(1+\lambda_i/k)^2}$ are increasing as k increases. So $\gamma_2(k)$ as a sum of monotonous increasing terms is monotonous increasing function of k. We note that $\lim_{k\to\infty} \gamma_2(k) = \beta^{*\prime}\beta^*$.

Corollary 4.5.5

$$\lim_{k\to o^+} \frac{d\gamma_2(k)}{dk} = 0.$$

Proof

$$\gamma_2(k) = k^2 \sum_{i=1}^{p} \frac{a_i^2}{(\lambda_i+k)^2}$$

$$(4.5.15)\qquad \frac{d\gamma_2(k)}{dk} = 2k\sum_{i=1}^{p} \frac{\lambda_i a_i^2}{(\lambda_i+k)^3}$$

Each $\dfrac{\lambda_i a_i^2}{(\lambda_i+k)^3}$ is a continuous function of k and

and $\lim\limits_{k\to o^+} \dfrac{\lambda_i a_i^2}{(\lambda_i+k)^3} = 0$. $\dfrac{d\gamma_2(k)}{dk}$ is the sum of continuous functions, and the limit of each term as $k\to o^+$ is 0, this implies

$$\lim_{k\to o^+} \frac{d\gamma_2(k)}{dk} = 0.$$

In Figure 3 the relationship between the variances, the squared bias and the parameter k is shown. The total variance decreases as k increases, while the squares bias increases with k. As indicated by the dotted line, which is $\gamma_1(k) + \gamma_2(k)$, the possibility exists that there are values of k for which the M.S.E. is less for $\hat{\beta}_R$ than it is for the $\hat{\beta}^*$.

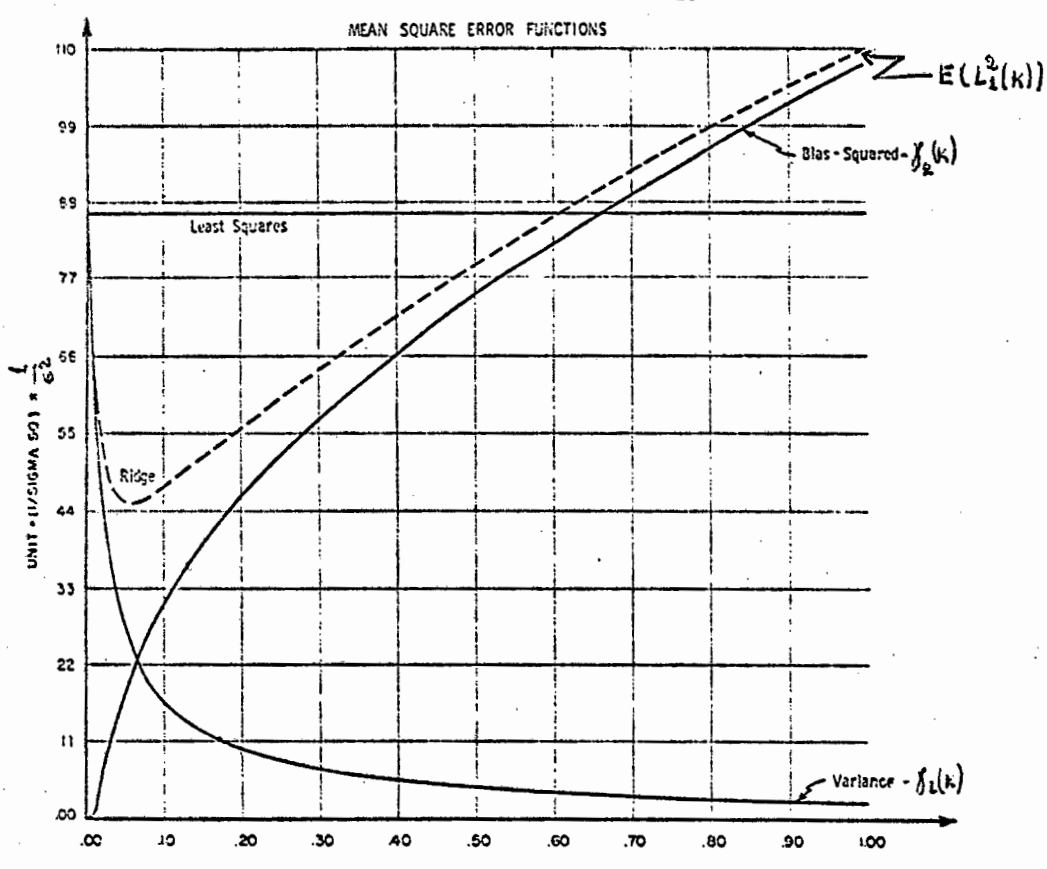


Figure 3

Theorem 4.5.6

There always exists a $k > 0$ such that $E(L_1^2(k)) < E(L_1^2(0))$, i.e. there exist $k > 0$ such that the mean square error is smaller than the total variance of the corresponding least squares estimator.

Proof

From (4.5.6) we have that

$$E(L_1^2(k)) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)^2} + k^2 \beta^{*\prime}(X^{*\prime}X^*+kI)^{-2}\beta^*$$

taking the first derivative

$$\frac{dE(L_1^2(k))}{dk} = -2\sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)^3} + 2k \sum_{i=1}^{p} \frac{\lambda_i a_i^2}{(\lambda_i+k)^3} . \tag{4.5.16}$$

We note that $\gamma_1(0) = \sigma^2 \sum_{i=1}^{p} 1/\lambda_i$, $\gamma_2(0) = 0$.

From Theorem 4.5.3 and Theorem 4.5.4 we have $\gamma_1'(k) < 0$ and $\gamma_2'(k) > 0$ as k increases.

The statement there exist $(k > 0$ such that $E(L_1^2(k)) < E(L_1^2(0))$ is equivalent to

(4.5.17) that there exist $(k > 0$ such that $(E(L_1^2(k)) - E(L_1^2(0)))/(k-0) < 0$ or equivalent to

(4.5.18) there exist $(k > 0$ such that $\frac{dE(L_1^2(k))}{dk} < 0$.

We will prove (4.5.18).

Consider

$$(4.5.19)\quad \frac{d}{dk} E(L_1^2(k)) = -2\sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)^3} + 2k \sum_{i=1}^{p} \frac{\lambda_i a_i^2}{(\lambda_i+k)^3}$$

$$< -2\sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)^3} + 2k\, a_{max}^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)^3}$$

$$= 2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)^3} (k a_{max}^2 - \sigma^2)$$

From the last relation of (4.5.19) it is easy to see that

$$(4.5.20)\quad \frac{d}{dk} E(L_1^2(k)) < 0 \quad \text{if} \quad k\, a_{max}^2 - \sigma^2 < 0$$

or

$$(4.5.21)\quad k < \frac{\sigma^2}{a_{max}^2} .$$

We note that a_{max}^2 is the maximum component of $a = P\beta^*$ (see (4.5.13)).

Corollary 4.5.7

There always exists $k > 0$ such that $E(L_1^2(k)) < \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i+k}$

Proof

From (4.5.6) we have

$$E(L_1^2(k)) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)^2} + k^2 \sum_{i=1}^{p} \frac{a_i^2}{(\lambda_i+k)^2}$$

$$= \gamma_1(k) + \gamma_2(k)$$

or

$$(4.5.22)\quad E(L_1^2(k)) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_{i+k-k}}{(\lambda_i+k)^2} + k^2 \sum_{i=1}^{p} \frac{a_i^2}{(\lambda_i+k)^2}$$

$$= \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i+k} + \sum_{i=1}^{p} k \frac{(a_i^2 k-\sigma^2)}{(\lambda_i+k)^2}$$

So if we take

(4.5.23) $k < \dfrac{\sigma^2}{a^2_{max}}$ we have

(4.5.24) $E(L_i^2(k)) < \sigma^2 \sum_{i=1}^{p} \dfrac{1}{\lambda_i+k}$

Theorem 4.5.7

The ridge estimator (4.2.1) is equivalent to a least squares estimator when the X^* $(n\times p)$ matrix is augmented by an orthogonal matrix H_p $(p\times p)$ of fictious set of data and the components of Y^* corresponding to the rows of H_p are set equal to zero.

Proof

The model has the form

(4.5.25)
$$\begin{pmatrix} Y_1^* \\ Y_2^* \\ \cdot \\ \cdot \\ \cdot \\ Y_n^* \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{pmatrix} = \begin{pmatrix} X_{11}^* & X_{12}^* & \cdots & X_{1p}^* \\ X_{21}^* & X_{22}^* & \cdots & X_{2p}^* \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ X_{n1}^* & X_{n2}^* & \cdots & X_{np}^* \\ H_{11} & H_{12} & \cdots & H_{1p} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ H_{p1} & H_{p2} & \cdots & H_{pp} \end{pmatrix} \begin{pmatrix} \beta_1^* \\ \beta_2^* \\ \cdot \\ \cdot \\ \cdot \\ \beta_p^* \end{pmatrix} + \begin{pmatrix} e_1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ e_{n+p} \end{pmatrix}$$

or

(4.5.26) $\left(\dfrac{Y^*}{0}\right) = \left(\dfrac{X^*}{H_p}\right) \beta^* + e$ where $e \sim N(0,\sigma^2 I)$.

The normal equations have the form

$$(4.5.27)\quad (X^{*\prime}|H_p')\begin{pmatrix}X^*\\ H_p\end{pmatrix}\hat{\beta}^* = (X^{*\prime}|H_p')\begin{pmatrix}Y^*\\ O\end{pmatrix}$$

or

$$(4.5.28)\quad (X^{*\prime}X^*+H_p'H_p)\,\hat{\beta}^* = X^{*\prime}Y^*$$

For any value k the matrix H_p can always be scaled such that

(4.5.29) $H_p'H_p = kI$, so $\hat{\beta}^*$ is actually the ridge estimator proposed by Hoerl and Kennard.

For example one can choose $H_p = k^{\frac{1}{2}}I$. The above estimator can be viewed as a type of weighted average between the actual data and other data (prior information in Bayesian terms), for which the response values are set arbitrarily equal to zero.

Note that for nonstandardized variables, the response values for the fictious data would be set equal to the mean of response of the actual data.

4.6 Generalizations of the Mean Square Error

Let $\tilde{\beta}_1$ and $\tilde{\beta}_2$ be two estimators of the vector parameter β,

$$(4.6.1)\quad M_j = E(\tilde{\beta}_j-\beta)(\tilde{\beta}_j-\beta)'$$

and

(4.6.2) $m_j = E(\tilde{\beta}_j-\beta)'B(\tilde{\beta}_j-\beta)$, $j = 1,2$ were B is a non-negative definite matrix (n.n.d.), the second order moment matrices and mean square errors respectively.

Definition 4.6.1

The $m = E(\tilde{\beta}-\beta)'B(\tilde{\beta}-\beta)$ where B, a n.n.d. matrix, is called generalized mean square error (g.m.s.e.). We say that $\tilde{\beta}_1$ is a better estimator than $\tilde{\beta}_2$ if the g.m.s.e. of $\tilde{\beta}_1$ is less than the g.m.s.e. of $\tilde{\beta}_2$.

Theorem 4.6.2

The following conditions are equivalent

(1) M_1-M_2 is n.n.d.

(2) $m_1-m_2 \geq 0$ for all n.n.d. B

Proof

(1) implies (2).

We observe that m_j is scalar so we have

$$
\begin{aligned}
(4.6.3) \quad m_j &= E(\mathrm{tr}(\tilde{\beta}_j-\beta)'B(\tilde{\beta}_j-\beta)) \\
&= E(\mathrm{tr}(B(\tilde{\beta}_j-\beta)(\tilde{\beta}_j-\beta)')) \\
&= \mathrm{tr}(E(B(\tilde{\beta}_j-\beta)(\tilde{\beta}_j-\beta)')) \\
&= \mathrm{tr}(B(E(\tilde{\beta}_j-\beta)(\tilde{\beta}_j-\beta)')) \\
&= \mathrm{tr}(B\,M_j)
\end{aligned}
$$

Therefore

(4.6.4) $m_1-m_2 = \mathrm{tr}\,B(M_1-M_2)$. So in order to prove that $m_1-m_2 \geq 0$ we must prove $\mathrm{tr}(B(M_1-M_2)) > 0$ for all B n.n.d. iff M_1-M_2 is n.n.d. We note that M_1-M_2 is symmetric. We set $M_1-M_2 = A$. Let $\mu_1,\ldots,\mu_p$ and $u_1,\ldots,u_p$ be the latent roots and latent vectors respectively, of A, then

(4.6.5) $A = \sum_{i=1}^{p} \mu_i u_i u_i'$

Now

(4.6.6) $tr(BA) = \sum_{i=1}^{p} \mu_i u_i' B u_i$, but $tr(BA) \geq 0$ if each $\mu_i \geq 0$ or equivalently A is n.n.d.
(2) implies (1).

If $m_1 - m_2 \geq 0$ for all B n.n.d. implies $tr\, B(M_1 - M_2)$ for all B n.n.d. implies $tr(BA) \geq 0$ implies

(4.6.7) $tr(BA) = \sum_{i=1}^{p} \mu_i u_i' B u_i \geq 0$ for all B n.n.d.

So in order to prove that A is n.n.d. it is necessary to prove that the roots of A, i.e. μ_i, $i = 1,\ldots,p$ are non-negative.

If in (4.6.7) we set $B = u_1 u_1'$ we have that $\mu_1 \geq 0$, so successively setting $B = u_i u_i'$, $i = 2,\ldots,p$ we have $\mu_2,\ldots,\mu_p$ are non-negative, i.e. A n.n.d. If $M(k) = E(\hat{\beta}_R - \beta^*)(\hat{\beta}_R - \beta^*)'$ then Theorem 4.6.3 is analogous to Lemma 4.5.2.

Theorem 4.6.3

Let $M(k) = D(k) + \Gamma_2(k)$ be the second order moment matrix, where $D(k) = \sigma^2 (X^{*\prime}X^* + kI)^{-1} (X^{*\prime}X^*) (X^{*\prime}X^* + kI)^{-1}$ the dispersion matrix, and $\Gamma_2(k) = k\,(X^{*\prime}X^* + kI)^{-1} \beta^* \beta^{*\prime} (X^{*\prime}X^* + kI)^{-1}$. Then there exists $K_1 > 0$ such that $M(0) - M(k)$ is positive definite (p.d.) whenever $0 < k < K_1$.

Proof

(4.6.8) $M(0) - M(k) = \sigma^2 (X^{*\prime}X^*)^{-1} - \sigma^2 (X^{*\prime}X^* + kI)^{-1} (X^{*\prime}X^*) (X^{*\prime}X^* + kI)^{-1}$
$- k^2 (X^{*\prime}X^* + kI)^{-1} \beta^* \beta^{*\prime} (X^{*\prime}X^* + kI)^{-1}$

$$= \sigma^2 (X^{*\prime}X^*+kI)^{-1}(X^{*\prime}X^*+kI)(X^{*\prime}X^*)^{-1}(X^{*\prime}X^*+kI)$$
$$(X^{*\prime}X^*+kI)^{-1}-\sigma^2(X^{*\prime}X^*+kI)^{-1}(X^{*\prime}X^*)$$
$$(X^{*\prime}X^*+kI)^{-1}-k^2(X^{*\prime}X^*+kI)^{-1}\beta^*\beta^{*\prime}(X^{*\prime}X^*+kI)^{-1}$$
$$= k(X^{*\prime}X^*+kI)^{-1}[(\sigma^2/k)\{(X^{*\prime}X^*+kI)(X^{*\prime}X^*)^{-1}$$
$$(X^{*\prime}X^*+kI)-X^{*\prime}X^*\}-k\beta^*\beta^{*\prime}](X^{*\prime}X^*+kI)^{-1}$$
$$= k(X^{*\prime}X^*+kI)^{-1}[(\sigma^2/k)\{(I+k(X^{*\prime}X^*)^{-1})(X^{*\prime}X^*+kI)$$
$$-X^{*\prime}X^*\}-k\beta^*\beta^{*\prime}](X^{*\prime}X^*+kI)^{-1}$$
$$= k(X^{*\prime}X^*+kI)^{-1}[\sigma^2\{2I+k(X^{*\prime}X^*)^{-1}\}-k\beta^*\beta^{*\prime}]$$
$$(X^{*\prime}X^*+kI)^{-1}$$

For $k > 0$ the matrix of the last relation is p.d. if

(4.6.9) $\sigma^2\{2I+k(X^{*\prime}X^*)^{-1}\}-k\beta^*\beta^{*\prime}$ is p.d. or equivalently if

(4.6.10) $2\sigma^2 I-k\beta^*\beta^{*\prime}$ is p.d.

Since the roots of $k\beta^*\beta^{*\prime}$ are zero (with multiplicity p-1) and $k\beta^{*\prime}\beta^*$, we have that the roots of

(4.6.11) $2\sigma^2 I-k\beta^*\beta^{*\prime}$ are the $2\sigma^2$ and $2\sigma^2-k\beta^{*\prime}\beta^*$. Because $2\sigma^2 > 0$ then the sufficient condition is $2\sigma^2-k\beta^{*\prime}\beta^* > 0$ or $k < (2\sigma^2/\beta^{*\prime}\beta^*)$.

Remarks

(1) We are able to invoke this theorem to show that if $k < (2\sigma^2/\beta^{*\prime}\beta^*)$ then according to the g.m.s.e. criterion $\hat{\beta}_R$ performs better that $\hat{\beta}^*$.

(2) If P is a matrix whose rows are the eigenvectors of the $X'X$ and $a = P\beta^*$, then Theorem 4.5.6 shows that for $B = I$

the upper limit on k may be raised to σ^2/a^2_{max}, a proportional increase of at most $p/2$.

(3) If X_0 is our predictor set, then taking $B = X_0X_0'$ gives the mean square error of $X_0'\hat{\beta}_R$ as an estimator of $X_0'\beta^*$.

We will now deal with some methods for the selection of k.

4.7 The method of Hoerl and Kennard (1970)

According to them the best method for achieving a better $\hat{\beta}_R$ is to use $k_i = k$ for all i, and use the ridge trace to select a single value of k and a unique $\hat{\beta}_R$. The following four criteria can be used for recognizing the appropriate value of k.

(1) At a certain value of k the system will stabilize and will then have general characteristics of an orthogonal system.

(2) Coefficients will not have unreasonable absolute values with respect to the factors for which they represent rates of change.

(3) Coefficients with apparently incorrect sign at $k = 0$ will have changed to have the proper sign.

(4) The R.S.S. will not have been inflated to an unreasonable value. It will not be large relative to the minimum residual sum of squares or large relative to what would be a reasonable variance for the process generating the data.

4.8 The method of Marquardt (1970) and Marquardt and Snee (1973)

The authors suggested using the value of k for which the maximum variance inflation factor "VIF" is between one and ten but, closer to one. The VIF associated with each coefficient represents the amount by which the variance of the coefficient is inflated by the correlations between the variables.

Specifically the VIF's are the diagonal elements of $\text{var}(\hat{\beta}_R)/\sigma^2 = (X^{*\prime}X^{*}+kI)^{-1}X^{*\prime}X^{*}(X^{*\prime}X^{*}+kI)^{-1}$.

4.9 The method of Mallows (1973)

The author extends the concept of C_p-plots to C_k-plots which may be used to determine k. He suggested plotting C_k against V_k where:

$$C_k = \frac{R.S.S_K}{\sigma^2} - n + 2 + 2\ \text{tr}(X^{*\prime}L)$$

$$V_k = 1 + \text{tr}(X^{*\prime}X^{*}LL')$$

$$L = (X^{*\prime}X^{*}+kI)^{-1}X^{*\prime}$$

Here $R.S.S_k$ is the residual sum of squares as a function of k. The suggestion is to chose k to minimize C_k.

4.10 The method of McDonald and Galarleau (1975)

The authors have proposed the following rules:

R1. This rule corresponds to least squares estimation and is defined as follows:

Choose $k = 0$

The second and the third rules are somewhat similar to each other. The choice of k is made in such a way that the squared length of the corresponding ridge estimator equals an estimated squared length of β^*. An unbiased estimator of $\beta^{*\prime}\beta^*$ can be obtained as follows:

$$Q = \hat{\beta}^{*\prime}\hat{\beta}^* - \hat{\sigma}^2 \sum_{j=1}^{p} 1/\lambda_j \qquad \text{(see (4.5.3))}.$$

The R2 and R3 rules are defined as follows:

R2. Choose k such that $\hat{\beta}_R'\hat{\beta}_R = Q$ if $Q > 0$; choose $k = 0$ otherwise.

R3. Choose k such that $\hat{\beta}_R'\hat{\beta}_R = Q$ if $Q > 0$; choose $k = \infty$ otherwise.

Remarks

(1) If $Q > 0$ then R2 and R3 define the same type of ridge estimator.

(2) If $Q < 0$ R2 gives the least squares estimate while R3 tives a zero vector as an estimate β^*.

(3) Since $\hat{\beta}_R'\hat{\beta}_R$ is a decreasing function of k and approaches zero as $k \to \infty$, these rules are well defined.

(4) It has been found from simulation studies made by them that:

(i) this method usually chooses a k which is less than the k chosen by the ridge trace examination.

(ii) Negative Q occurs in cases where ridge type estimators have the potential for doing much better than least squares.

So defaulting to the least squares estimate in such cases, or estimating $\hat{\beta}^*$ by the zero vector, is unacceptable. It is more preferable to default to a constant k-value. Thus to summarize : choose k such that $\hat{\beta}_R'\hat{\beta}_R = Q$ if $Q > 0$. Choose $k = k'$ otherwise.

4.11 An explicit solution for generalized ridge regression W. Hemmerle (1975)

The model (1.2.5) can be written as follows:

(4.11.1) $Y^* = Wa + e$

where 1) $a = P\beta^*$

2) $W = X^*P'$

3) P - is orthogonal matrix such that $PX^{*\prime}X^*P' = \text{diag}(\lambda_1,\ldots,\lambda_p) = \Lambda$ contains the eigenvalues of $X^{*\prime}X^*$.

The ridge estimator for a is

(4.11.2) $\hat{a}_R = [W'W+K]^{-1}W'Y^*$

where $K = \begin{pmatrix} k_1 & & 0 \\ & \ddots & \\ 0 & & k_p \end{pmatrix}$ $\quad k_i \geq 0 \quad i = 1,2,\ldots,p.$

We now consider the estimation of $\hat{a}_R$.

1. Iterative estimation of optimal $\hat{a}_R$

Optimal values for the k_i's in (4.11.2) can be considered to be those k_i's that minimize

$$\Omega = E[(\hat{a}_R-a)'(\hat{a}_R-a)]$$

or equivalently those k_i's that minimize

$$(4.11.3) \quad \Omega = \sum_{i=1}^{p}(\sigma^2\lambda_i+a_i^2k_i^2)/(\lambda_i+k_i)^2.$$

Differentiation of (4.11.3) with respect to the k_i's yields the minimization equations

$$\frac{\partial\Omega}{\partial k_i} = (2a_i^2k_i)(\lambda_i+k_i)^2 - 2(\sigma^2\lambda_i+a_i^2\cdot k_i^2)(\lambda_i+k_i)/(\lambda_i+k_i)^4$$

$$= 2(\lambda_i+k_i)\lambda_i(k_ia_i^2-\sigma^2)/(\lambda_i+k_i)^4$$

$$= 0$$

for $i = 1,2,\ldots,p.$

The $X^*'X^*$ is a full rank matrix so $\lambda_i > 0$ for all $i = 1,2,\ldots,p$ with the restriction $k_i > 0$, $i = 1,\ldots,p$ we get $k_i = \frac{\sigma^2}{a_i^2}$ $i = 1,\ldots,p$. The iterative procedure is then

$$(4.11.4) \quad k_i(j) = \frac{\hat{\sigma}^2}{(\hat{a}_{R.i}(j))^2} \qquad \begin{matrix} i = 1,\ldots,p \\ j = 1,\ldots \end{matrix}$$

with initial value $\hat{a}_{R.i}(0) = \hat{a}_i$, $i = 1,\ldots,p.$

These obtained values of $k_i(j)$ are then used for the calculation of $\hat{a}_R(j+1)$ from (4.11.2).

The new values of $\hat{a}^2_{R.i}(j+1)$ are substituted into (4.11.4)

and the $k_i(j+1)$ is obtained. The procedure is then repeated. In what follows we will give an explicit formulation using the matrix notation, and the criteria for divergence and convergence are given.

If we represent the p-vectors $W'Y^*$ and $\hat{a}_R(j)$ as diagonal matrices we have

$$B = \begin{pmatrix} [W'Y^*]_1 & & 0 \\ & \ddots & \\ 0 & & [W'Y^*]_p \end{pmatrix} \quad \text{and} \quad A_j = \begin{pmatrix} \hat{a}_{R.1}(j) & & 0 \\ & \ddots & \\ 0 & & \hat{a}_{R.p}(j) \end{pmatrix}$$

where j indicates the iteration. For $j = 0$ the initial value is given as $A_0 = \Lambda^{-1}B$. The above iterative procedure is given by

$$\begin{aligned} (4.11.5) \quad A_{j+1} &= [\Lambda+\hat{\sigma}^2A_j^{-2}]^{-1}B \\ &= [\Lambda+\hat{\sigma}^2A_j^{-2}]^{-1}\Lambda A_0 \\ &= [\Lambda(I+\hat{\sigma}^2\Lambda^{-1}A_j^{-2})]^{-1}\Lambda A_0 \\ &= [I+\hat{\sigma}^2\Lambda^{-1}A_j^{-2}]^{-1}IA_0 \end{aligned}$$

If we set $D = \Lambda/\hat{\sigma}^2$ we have

$$(4.11.6) \quad A_{j+1} = [I+D^{-1}A_j^{-2}]^{-1}A_0$$

and an expression for A_{j+1}^{-2} is given by

$$\begin{aligned} (4.11.7) \quad A_{j+1}^{-2} &= A_0^{-1}(I+D^{-1}A_j^{-2})A_0^{-1}(I+D^{-1}A_j^{-2}) \\ &= A_0^{-2}(I+D^{-1}A_j^{-2})^2 \end{aligned}$$

because A_0^{-1} and $(I+D^{-1}A_j^{-2})$ are diagonal matrices and can be commuted.

Multiplying (4.11.7) by D^{-1} (both sides) we have

(4.11.8) $D^{-1}A_{j+1}^{-2} = D^{-1}A_0^{-2}(I+D^{-1}A_j^{-2})^2$;

if we set $E_j = D^{-1}A_j^{-2}$ then (4.11.8) becomes

(4.11.9) $E_{j+1} = E_0(I+E_j)^2$

If we assume that $\hat{a}_i \neq 0$ for all $i = 1,\ldots,p$ and that there exist

(4.11.10) $$\lim_{j\to\infty} E_{j+1} = E^* = E_0[I+E^*]^2$$

because $\lim_{j\to\infty} E_j = \lim_{j\to\infty} E_{j+1}$.

Now from (4.11.10) we have

(4.11.11) $E^* = E_0[I+E^*]^2$

or

(4.11.12) $(E^*)^2 + (2I-E_0^{-1})E^* + I = 0$.

The equation (4.11.12) consists of p-equations of the form

$$(e^*)^2 + (2-1/e_0)e^* + 1 = 0$$

or $$e^* = (1-2e_0 \pm \sqrt{1-4e_0})/2e_0 .$$

We wish to note the following:

(1) $\lambda_i > 0$ for all $i = 1,\ldots,p$ because $r(X^{*\prime}X^*) = p$ so if $\hat{a}_i = 0$ implies $(W'Y^*)_i = 0$ implies $\hat{a}_{R.i} = 0$, for $k_i > 0$ consequently we can exclude the ith equation.

(2) $E_{j+1} = E_0(I+E_j)^2$ consists of p-equations of the form

(4.11.13) $e_i(j+1) = e_i(0)\ (1+e_i(j))^2$ $i = 1,\ldots,p$, where $e_i(0)$, $e_i(j)$, $e_i(j+1)$ are scalars, for all $i = 1,\ldots,p$, and j denotes the iteration.

(3) Using the following lemma 4.11.1 we can prove for all $i = 1,2,\ldots,p$ that $e_i(j+1)$ converges if $0 < e_i(0) \leq \frac{1}{4}$ and diverges if $e_i(0) > \frac{1}{4}$ as j goes to infinity.

Lemma 4.11.1

There exist $\lim_{j\to\infty} e_i(j+1) = \lim_{j\to\infty} (e_i(0)(1+e_i(j))^2)$ and it is a finite number if $e_i(0) = \frac{1}{4}$ and is equal to infinity if $e_i(0) > \frac{1}{4}$. (A generalization of this lemma is given by Theorem 5.6.1)

Proof

See Technometrics vol. 17, No. 3, 1975 p.311.

2. Solution for optimal $\hat{a}_R$

Let $e_i(j)$ denote the ith equation and the jth iteration, then $\lim_{j\to\infty} e_i(j) = e_i^*$, $i = 1,2,\ldots,p$.

We have that

$$(4.11.13)\quad e_i(j) = \frac{\hat{\sigma}^2}{\lambda_i \hat{a}_{R.i}^2(j)},\ \hat{a}_{R.i}(j) \underset{j\to\infty}{\to} 0 \quad \text{if} \quad e_i(0) > \tfrac{1}{4}.$$

Let $\hat{a}_{R.i} = \lim_{j\to\infty} \hat{a}_{R.i}(j)$ then we set

$$(4.11.14)\quad \hat{a}_{R.i} \begin{cases} = 0 & \text{if } e_i(0) > \frac{1}{4} \\ = \dfrac{\hat{a}_i}{1+e_i^*} & \text{if } 0 < e_i(0) \leq \frac{1}{4} \end{cases}$$

where (1) $e_i^* = (1-2e_i(0)-(1-4e_i(0))^{\frac{1}{2}})/2e_i(0)$

(2) $e_i(0) = \hat{\sigma}^2/\lambda_i \hat{a}_i^2 \qquad i = 1,\ldots,p$.

We note the following

(1) When $a_i = 0$, we minimize the ith component of (4.11.3) by letting k_i approach infinity. We have

$$\begin{aligned} pr\{\hat{a}_{R.i} = 0\} &= pr\{e_i(0) > \tfrac{1}{4}\} \\ &= pr\{\frac{1}{e_i(0)} \leq 4\} \\ &= pr\{\frac{\lambda_i \hat{a}_i^2}{\hat{\sigma}^2} \leq 4\} \end{aligned}$$

But $var(\lambda_i^{\frac{1}{2}}\hat{a}_i) = \frac{\lambda_i \sigma^2}{\lambda_i} = \sigma^2$, so we have that under $H_0 : a_i = 0$ $\frac{1}{e_i(0)} \sim t^2$ where t^2 is the square of the Student t-statistic, which is used to test for zero regression coefficients.

$$\begin{aligned} pr\{\hat{a}_{R.i} = 0 | H_0\} &= pr\{t^2 \leq 4\} \\ &= pr\{F_{1,n-p} \leq 4\} \end{aligned}$$

This probability increase with $n-p$.

(2) If we allow the k_i's to become infinite and set $\hat{a}_{R.i} = 0$ if $e_i(0) > \frac{1}{4}$, we may produce a significant increase in the R.S.S.

<u>The following methods can be used in order to constrain the increase in the R.S.S.</u>.

(A) Using the relation (4.2.12) we have

$$\begin{aligned} (4.11.15) \quad \Delta^* &= R.S.S.(\hat{a}_R) - R.S.S.(\hat{a}) \\ &= (\hat{a}_R - \hat{a})' \Lambda (\hat{a}_R - \hat{a}) \geq 0 . \end{aligned}$$

Let $\Delta_i^*(j)$ be the ith component of the jth iteration then

$$\begin{aligned} (4.11.16) \quad \Delta_i^*(j) &= (\hat{a}_{R.i}(j) - \hat{a}_i)^2 \lambda_i \\ &= \hat{\sigma}^2 \left(\frac{1}{\sqrt{e_i(j)}} - \frac{1}{\sqrt{e_i(0)}} \right)^2 \end{aligned}$$

$$= \hat{\sigma}^2 \; e_i^2(j-1)/e_i(j)$$

$$= (\hat{\sigma}^2/e_i(0)) \cdot e_i^2(j-1)/(1+e_i^2(j-1))^2.$$

It is easy to show that $\Delta_i^*(j)$ is an increasing function of j because $e_i(j) > e_i(j-1)$.

We have

$$\lim_{j\to\infty} \Delta_i^*(j) = \hat{\sigma}^2 e_i^* \quad \text{if} \quad e_i(0) \leq \tfrac{1}{4} \tag{4.11.17}$$

$$\lim_{j\to\infty} \Delta_i^*(j) = \hat{\sigma}^2/e_i(0) \quad \text{if} \quad e_i(0) > \tfrac{1}{4} .$$

The Δ^* can be written in the jth iteration as

$$\Delta^*(j) = \textstyle\sum_{i=1}^{p} \Delta_i^*(j) = \hat{\sigma}^2 \sum_{i=1}^{p} e_i^2(j-1)/e_i(j) \tag{4.11.18}$$

so

$$\lim_{j\to\infty} \Delta_j^* = \sigma^2 \left(\sum_{e_i(0)\leq\frac{1}{4}} e_i^* + \sum_{e_i(0)>\frac{1}{4}} \frac{1}{e_i(0)} \right) \tag{4.11.19}$$

Now suppose that we want the R.S.S. to be increased by no more than 100M% so that $\Delta^* \leq M(n-p)\hat{\sigma}^2 = M^*$.

We distinguish two cases

(i) $\lim_{j\to\infty} \Delta_j^* < M^*$

then the solution in (4.11.14) satisfies the above constraint.

(ii) if $M^* < \lim_{j\to\infty} \Delta_j^*$

In this case we constrain each component Δ_i^* as follows

$$\Delta_i^* = (\hat{a}_{R.i} - \hat{a}_i)^2 \lambda_i \leq M_i^*$$

and $$\textstyle\sum_{i=1}^{p} M_i^* = M^*$$

The advantage is due to the fact that no iteration is required.

We can assume that

(4.11.20) $M_i^* < \hat{\sigma}^2/e_i(0)$ because the

(4.11.21)
$$\begin{aligned}\lim_{k_i\to\infty} \Delta_i^* &= \lim_{k_i\to\infty} (\hat{a}_{R.i}-\hat{a}_i)^2\lambda_i \\ &= \lim_{k_i\to\infty} k_i^2([W'Y]_i)^2/\lambda_i(\lambda_i+k_i)^2 \\ &= \lim_{k_i\to\infty} k_i^2\lambda_i\hat{a}_i^2/(\lambda_i+k_i)^2 \\ &= \lim_{k_i\to\infty} k_i^2\hat{\sigma}^2/e_i(0)(\lambda_i+k_i)^2 \\ &= \hat{\sigma}^2/e_i(0)\end{aligned}$$

Then whenever $\hat{\sigma}^2 e_i^* \leq M_i^*$ we use (4.11.14) to obtain $\hat{a}_{R.i}$; however, whenever $\hat{\sigma}^2 e_i^* > M_i^*$ including the cases when $e_i^* = \infty$ we set $\Delta_i^* = M_i^*$ and solve the

$$k_i^2\hat{\sigma}^2 = M_i^* e_i(0)(\lambda_i+k_i)^2$$

for k_i. Then for $k_i > 0$ we have

$$k_i = \lambda_i(M_i^* e_i(0))^{\frac{1}{2}}/(\hat{\sigma} - (M_i^* e_i(0))^{\frac{1}{2}}).$$

One way to apportion M^* is to make its components M_i^* proportional to the individual components of (4.11.19).

We make M_i^* proportional to ℓ_i where

$$\ell_i = \begin{cases} e_i^* & \text{if } e_i(0) \leq \frac{1}{4} \\ 1/e_i(0) & \text{if } e_i(0) > \frac{1}{4} \end{cases}$$

Following this procedure we are assured that (4.11.20) will hold

for all $i = 1,2,\ldots,p$ since $e_i^* < 1/e_i(0)$ for $e_i(0) \leq \frac{1}{4}$ and $M^* \leq \lim_{j\to\infty} \Delta_j^*$.

(B) If one is reluctant to allocate M^* and constrain the individual component Δ_i^* of Δ^*, then the alternative is to perform the recursion given by (4.11.9) evaluating $\Delta_{(j)}^*$ given by $\hat{\sigma}^2 \operatorname{trace}(E_j^{-1}E_{j-1}^2)$ at each step.

We would stop the recursion whenever $\Delta_{(j)}^* > M^*$. Then E_{j-1} would be used for limiting E^* matrix in evaluating $\hat{a}_{R.i}$ for all i by (4.11.14), irrespective of the value of $e_i(0)$.

4.12 The method of Guilkey and Murphy (1975)

The authors suggested the following two algorithms for the selection of k. Their methods are called directed ridge regression techniques (D.R.T.T.).

D.R.T.T.(1)

Step 1: Determine the eigenvalues and eigenvectors of the $X^{*\prime}X^*$ and form the following matrices

$$\Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix}$$

$$P = \begin{pmatrix} v_1 \\ \vdots \\ v_p \end{pmatrix}$$

$$W = X^*P'$$

where $\lambda_1,\ldots,\lambda_p$ are the eigenvalues and P is the matrix whose rows are the eigenvectors of $X^{*\prime}X^*$.

Step 2: Find $\hat{a} = \Lambda^{-1}W'Y^* = P\hat{\beta}^*$ and σ^2

Step 3: Find $k_i^{*(0)} = \frac{\hat{\sigma}^2}{\hat{a}_i^2}$ and set $k_i^{*(0)} = 0$ for i such that $\lambda_i \leq 10^{-c}\lambda_{max}$, c-constant.

Step 4: Find the directed ridge estimator

$$a^{*(0)} = \Lambda_{k^*}^{-1}W'Y^*$$

where $$\Lambda_{k^*} = \Lambda + k^* = \begin{pmatrix} \lambda_1+k_1^* & 0 \\ 0 & \lambda_p+k_p^* \end{pmatrix}$$

Step 5: Re-estimate $k_i^{*(1)} = \frac{\sigma^2}{\left(a_i^{*(1)}\right)^2}$

Step 6: Repeat Steps 4 and 5 until stabilization is achieved; say, on the mth iteration. By the term stabilization the authors mean that the change in $a^{*\prime}a^*$ is one or five or ten percent from one iteration to the next.

Step 7: Determine $\hat{\beta}_R = P'a^{*(m)}$.

D.R.T.T.(2)

Step 1: Calculate $\hat{\beta}^*$ and σ^2 using O.L.S.

Step 2: Find the eigenvalues and the eigenvectors of $X^{*\prime}X^*$ and form $W = X^*P'$.

Step 3: Let $k_i = k$, and allow k to increase until the unexplained variance has increased from $\hat{\sigma}^2$ to $\hat{\sigma}^2+q\hat{\sigma}^2$.

Step 4: Determine $\hat{\beta}_R = P\hat{a}^*$ where $\hat{a}^* = \Lambda_k^{-1} W'Y^*$.

Notes and Remarks

(1) k is added in D.R.T.T.(2) to the diagonal element of Λ if the corresponding $\lambda_i < 10^{-c}\lambda_{max}$.

(2) q is taken to be 10%.

(3) We have relatively good results if we take as stabilization criterion 1%.

(4) The directed ridge estimation will result in estimates of β^*, less biased than the estimate proposed by Hoerl and Kennard, because we change only the diagonal elements of the Λ-matrix corresponding to small eigenvalues.

(5) It is possible that the estimates which are obtained using the above algorithms to have smaller mean square error than those which are obtained using the ridge trace.

(6) There are some arbitaries involved in the above methods, e.g.

(i) In choosing c (the authors took $c = 1$ or 2 or 3).

(ii) In defining the term stabilization.

(7) We use as an estimate of $\frac{\sigma^2}{a_i^2}$ the $\frac{\hat{\sigma}^2}{\hat{a}_i^2}$ which is very likely a poor estimate in view of the tendency of $\hat{a}_i^2$ to overestimate a_i^2.

4.13 The method of Hoerl and Kennard (1976)

The authors' method can be stated as follows:

Step 1: (i) Find $\hat{\beta}^*$ and $\hat{\sigma}^2$ using the O.L.S.

(ii) Find $k_0 = \dfrac{p\,\hat{\sigma}^2}{\hat{\beta}^{*\prime}\hat{\beta}^*}$

(iii) Find $\hat{\beta}_R = \hat{\beta}_R(k_0) = (X^{*\prime}X^* + k_0 I)^{-1} X^{*\prime}Y^*$.

Step 2: (i) Find $k_i = \dfrac{p\,\hat{\sigma}^2}{\hat{\beta}_R'(k_{i-1})\hat{\beta}_R(k_{i-1})}$

(ii) Find $\hat{\beta}_R(k_i) = (X^{*\prime}X^* + k_i I)^{-1} X^{*\prime}Y^*$.

Repeat (i) and (ii) in Step 2, and as a stopping rule use anyone of the following criteria.

Criterion 1: If $(k_{i+1} - k_i)/k_i > \delta$ continue; otherwise stop and take as an estimate of β^* the ridge estimate $\hat{\beta}_R(k_i)$; δ is taken to be equal to $20 \times (\text{trace}(X^{*\prime}X^*)^{-1}/p)^{-1,30}$.

Criterion 2: If $\hat{\beta}_R'(k_i)\hat{\beta}_R(k_i) < (\frac{1}{1+\delta})\hat{\beta}_R'(k_i - 1)\hat{\beta}_R(k_i - 1)$ continue; otherwise stop and use $\hat{\beta}_R(k_i)$.

Proposition 4.13.1

Criterion 1 is equivalent to Criterion 2.

Proof

(Criterion 1 implies Criterion 2.)

$$(4.13.1)\quad \frac{\hat{\beta}_R'(k_i)\hat{\beta}_R(k_i)}{\hat{\beta}_R'(k_{i-1})\hat{\beta}_R(k_{i-1})} = \frac{P\hat{\sigma}^2/(k_i)}{P\hat{\sigma}^2/(k_{i-1})} = \frac{k_{i-1}}{k_i}$$

But from

$$(4.13.2)\quad (k_i-k_{i-1})/k_{i-1} > \delta \text{ implies } \frac{k_{i-1}}{k_i} < 1/1+\delta.$$

From (4.13.1) and (4.13.2) implies Criterion 2. (Criterion 2 implies Criterion 1).

The proof is similar to the above.

Notes and Remarks

(1) $k_{i+1} \geq k_i$ implies $\delta > 0$

(2) $\delta > 0$ implies $\frac{1}{1+\delta} < 1$

(3) length of $(\hat{\beta}_R(k_i)) < (\frac{1}{1+\delta})^{\frac{1}{2}} \cdot$(length of $(\hat{\beta}_R(k_i-1))$)

(4) The simulation study shows:

(i) There is a significant reduction in M.S.E. using the δ-criterion and making more than one iteration. The improvement in M.S.E. becomes greater as X*'X* becomes less condition.

(ii) More than one iteration is needed in the solution in which X*'X* is moving toward ill-conditioned matrix.

(iii) The δ-criterion produces an error distribution with smaller standard deviation.

(iv) The ridge estimator based on δ-criterion has a probability greater than 0,5 of producing estimates with smaller M.S.E. than the least squares.

(5) The first step of the above algorithm has been proposed as a method for selecting k by Hoerl, Kennard and Baldwin (1975).

4.14 The method of Lawless and Wang (1976)

Consider the model A_s as in (1.2.5). There exists an orthogonal matrix P such that $PX^{*\prime}X^{*}P' = \Lambda$, where $\Lambda = \text{diag}(\lambda_1,\ldots,\lambda_p)$ contains the eigenvalues of $X^{*\prime}X^{*}$. Let $W = X^{*}P'$ and $a = P\beta^{*}$. Then the model in (1.2.5) can be written as

$$Y^{*} = Wa+e,\ e \sim N(0,\sigma^2 I) \tag{4.14.1}$$

The O.L.S. estimator for a is

$$\hat{a} = \Lambda^{-1}W'Y^{*} = P\hat{\beta}^{*} \tag{4.14.12}$$

Consider any particular estimator $\tilde{\beta}^{*}$ of β^{*}, with $\tilde{a} = P\tilde{\beta}^{*}$ as the corresponding estimator of a. Then Lawless and Wang (1976) have proposed as a measure of goodness of an estimator the total mean square error of prediction. This is given by

$$\text{M.S.E.P} = E\{\textstyle\sum_{i=1}^{n}(\hat{Y}_i-Y_i)^2\} \tag{4.14.3}$$

$$= \sigma^2+\textstyle\sum_{i=1}^{p}\lambda_i E(\tilde{a}_i-a_i)^2$$

The authors said if one temporarily adopts a Bayesian approach and assumes that $a = P\beta^{*}$ has a prior distribution that is $N(0,\sigma_a^2 I)$ (this is equivalent to assuming $\beta^{*} \sim N(0,\sigma_a^2 I)$), then the Bayes estimator for a is $\tilde{a}_B$, i.e.

$$(4.14.4)\quad \tilde{a}_{B,i} = \frac{\lambda_i}{\lambda_i+\sigma^2/\sigma_a^2}\,\hat{a}_i \qquad (i = 1,\ldots,p)$$

Since σ^2 and σ_a^2 are unknown to us we must estimate them.

Since $X^{*\prime}X^*$ is in a correlation form $\mathrm{tr}(X^{*\prime}X^*) = \sum_{i=1}^{p}\lambda_i = p$; thus unconditionally

$$(4.14.5)\quad E\{\textstyle\sum_{i=1}^{p}\lambda_i\hat{a}_i/p\sigma^2\}-1 = \sigma_a^2/\sigma^2$$

we might therefore estimate σ_a^2/σ^2 by $\Sigma\lambda_i\hat{a}_i^2/p\hat{\sigma}^2-1$. The authors have chosen to estimate it by $\Sigma\lambda_i\hat{a}_i^2/p\hat{\sigma}^2$; since σ_a^2 will presumably be much larger than σ^2 this should provide a reasonable estimate of σ_a^2/σ^2.

4.15 New Ridge regression - H. Vinod (1976)

The disadvantages of using the ridge trace as it has been proposed by Hoerl and Kennard (see 4.7) are stated as follows:

(1) The k on the horizontal axis cannot be used for plotting generalized ridge regression defined by $\hat{\beta}_R = P'\mathrm{diag}(\lambda_1/\lambda_1+k_1, \lambda_2/\lambda_2+k_2, \ldots, \lambda_p/\lambda_p+k_p)P\hat{\beta}^*$ where $k_i > 0$ for all $i = 1,\ldots,p$.

(2) The k scale has the unfortunate property that the ridge trace may appear to be more stable for larger k even for completely orthogonal data, because of the

$$\left|\frac{d\hat{\beta}_R}{dk}\right| = \frac{\hat{\beta}^*}{(1+k)^2} \quad \text{for} \quad X^{*\prime}X^* = I.$$

Definition 4.15.1

The multicollinearity allowance m is given by the

$$m = p - \sum_{i=1}^{p} \lambda_i/(\lambda_i+k_i) = p - \sum_{i=1}^{p} \delta_i$$

Proposition 4.15.2

Having m on the horizontal axis of the ridge trace will not give an appearance of greater stability at larger m.

Proof

From Definition 4.15.1, if we take $k_i = k$, we have

$$(4.15.1) \quad \frac{dm}{dk} = \sum_{i=1}^{p} \lambda_i/(\lambda_i+k)^2 = S$$

Now

$$(4.15.2) \quad \frac{d\hat{\beta}_R}{dk} = -P'\mathrm{diag}(\lambda_1/(\lambda_1+k)^2, \lambda_2/(\lambda_2+k)^2, \ldots, \lambda_p/(\lambda_p+k)^2)P\cdot\hat{\beta}^*$$
$$= -G\hat{\beta}^*$$

So

$$(4.15.3) \quad \frac{d\hat{\beta}_R}{dm} = \frac{d\hat{\beta}_R}{dk} \cdot \frac{dk}{dm}$$
$$= -G\hat{\beta}^*/S$$

For completely orthogonal data S becomes $p/(1+k)^2$, $(p/S)G = I$, and $\frac{d\hat{\beta}_R}{dm} = -\hat{\beta}^*/p$ which does not change with m.

The condition $(p/S)G = I$ previously stated will not be satisfied for non-orthogonal data, and the absolute values of the elements of $(p/S)G - I$ will be large. This suggests the following definition.

Definition 4.15.3

The index of stability of relative magnitudes (ISRM) of $\hat{\beta}_{R,i}$, defined for $m < p$ by

$$ISRM = \sum_{i=1}^{p}[(p\delta_i^2/S\lambda_i)-1]^2$$

Note that ISRM = 0 for $X^{*\prime}X^* = I$.

An important advantage of ISRM as a quantification of Hoerl and Kennard's concept of stable region is that in most cases ISRM yields a considerably narrow range of desirable values for m. Moreover the theoretical advantage of ISRM is that it is not stochastic. The $\hat{\beta}_{R,i}$ plotted in a ridge trace are stochastic, hence their visual inspection leads to a stochastic determination of k(or m).

Choice of m, consequences and trade offs

The monotonic behaviour of $var(\hat{\beta}_{R,i})$ and $\sum_{i=1}^{p}var(\hat{\beta}_{R,i}) = \gamma_1(k)$ (see Theorem 4.5.3) make it difficult to use them for choosing m because they seem to always favour a larger m while ignoring the bias completely.

Any linear transformation does not change the heuristic ratio $t_i = \hat{\beta}_{R,i}/\hat{\sigma}(g_{ii})^{\frac{1}{2}}$ whose distribution unfortunately involves unknown β_i $(i = 1,\ldots,p)$ except when $k = 0 = m$. For orthogonal data $X^{*\prime}X^* = I$ and $g_{ii} = g_{jj}$.

Define $\hat{\beta}_{R,i}$ to be more significant than $\hat{\beta}_{R,j}$ when

$|t_i| > |t_j|$ which may be assumed to imply that $\Pr\{|\beta_i^*| > 0\} > \Pr\{|\beta_j^*| > 0\}$ and hence the true $|\beta_i^*| > |\beta_j^*|$ holds with high probability. Therefore we may require the estimates $|\hat{\beta}_{R,i}|$ to be numerically larger than $|\hat{\beta}_{R,j}|$.

In general the off diagonal elements of the $X^{*\prime}X^*$ can be large and g_{ii} and g_{jj} can be so unequal that more significant $\hat{\beta}_{R,i}$ may not be numerically larger than $\hat{\beta}_{R,j}$.

So we suggest a scatter plot of $|\hat{\beta}_{R,i}|$ against $|t_i|$ having p-points for different values of m. Then we compute the ordinary correlation coefficient (R_m^2) between the $(|\hat{\beta}_{R,i}|, |t_i|)$ $i = 1,\ldots,p$ for each different m. In this way the monotonicity of $\text{var}(\hat{\beta}_{R,i})$ can be avoided by considering only the relative magnitude of $|\hat{\beta}_{R,i}|$ and $|t_i|$ for a given m.

As m increases, there is a degradation of the fit. To avoid those choices of m that lead to serious degradation, we assess the fit of the model in the original units with coefficients as they are given by (1.2.6). But even so both R.S.S. and R^2 are monotonic and tend to favour a smaller m. These are being the trade off against $\sum_{i=1}^{p}\text{var}(\hat{\beta}_{R,i})$, which is also monotonic and tends to favour a larger m.

How to improve the goodness-of-fit

We know that as m increases, the actual magnitudes of $\hat{\beta}_{R,i}$ shrink toward zero. This may not be appropriate for

some applications.

Let μ denote a scale factor for rescaling the elements of $\hat{\beta}_{R,i}$, so each $\hat{\beta}_{R,i}$ becomes $\mu\hat{\beta}_{R,i}$ and hence $\hat{Y}^*$ becomes μY^* keeping the relative magnitude of $\hat{\beta}_{R,i}$ unchanged. The μ can be found as follows

minimize the R.S.S. $= (Y^*-\mu\hat{Y}^*)(Y-\mu\hat{Y}^*)$, where $\hat{Y}^* = X^*\hat{\beta}_R$. So taking the partial derivative with respect to μ and setting the result equal to zero, we find

$$\mu = (\hat{\beta}_R'X^{*\prime}Y^*)/(\hat{\beta}_R'X^{*\prime}X^*\hat{\beta}_R)$$

The work that we must do when we use the new ridge method is summarised as follows:

(a) We construct the following table for different values of m:

m	k	R^2	ISRM	R_m^2	μ
m_1	k_1	R_1^2	$ISRM_1$	$R_{m_1}^2$	μ_1
m_2	k_2	R_2^2	$ISRM_2$	$R_{m_2}^2$	μ_2
m_3	k_3	R_3^2	$ISRM_3$	$R_{m_3}^2$	μ_3
m_4	k_4	R_4^2	$ISRM_4$	$R_{m_4}^2$	μ_4

(b) We choose the m from the above table which satisfies simultaneously the following:

(1) ISRM is minimum

(2) R_m^2 is maximum

(3) R^2 is maximum

(c) From (b) we compute $\hat{\beta}_R$ and then we rescale the elements of $\hat{\beta}_R$ using the corresponding μ.

4.16 A critical view of Ridge regression

(i) Hoerl and Kennard establish only the existence of a k leading to smaller mean square error than the variance of least squares estimator. There is no guarnatee that any particular choice of k improves in the least squares estimator.

(ii) Since $k \in [0,1]$ chosen by their method is estimated from the data, it is not a constant but rather a variable, so the moments of $\hat{\beta}_R(k)$ for fixed k are not the moments of the estimator which is used in practice.

(iii) Stability of the ridge trace is a trivial property of $\hat{\beta}_R = (X^{*\prime}X^* + kI)X^{*\prime}Y$ as k increases. It is easy to show that

$$\frac{d\hat{\beta}_R}{dk} = \begin{cases} -Z(X^{*\prime}X^*)^{-1}Z\hat{\beta}^* & \text{non-orthogonal system} \\ & Z = I + k(X^{*\prime}X^*)^{-1} \\ -\frac{1}{(1+k)^2}\hat{\beta}^* & \text{the system is orthogonal} \end{cases}$$

So even if the matrix is perfectly conditioned $\hat{\beta}_R$ values would change more slowly with increasing k.

(iv) In any case, any procedure based on inspecting the slopes of the ridge trace plots could have a high associated error variance if $X^{*\prime}X^*$ is ill-conditioned.

The average variance of the slopes is given by the formula

$$(4.16.1) \quad \frac{\sigma^2}{p} \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^4}$$

The quantity in (4.16.1) will be large if k is small. For example if $\lambda_{min} = 0,05$ and $k = 0,1$ then that term of the sum will be $\simeq 100$. This term will be small if k is very large, but the idea of ridge regression is that small value of k will achieve optimum mean square error.

(v) In Model (4.14.1) the ridge estimator of the a_i-component is given by

$$(4.16.2) \quad \hat{a}_{R,i} = \frac{\lambda_i}{\lambda_i + k} \hat{a}_i$$

Because $k \geq 0$ $\frac{\lambda_i}{\lambda_i + k} \leq 1$ so $\hat{a}_{R,i} \leq \hat{a}_i$. So if λ_i is small the least squares estimator has large variance so could over-estimate or under-estimate a_i. If we had an idea of the value of a_i we could judge whether $\hat{a}_i$ was an over-estimate or under-estimate. If an over-estimate, we would like to reduce it, and if an under-estimate to increase it. But ridge estimator reduces it in both cases.

Since the β_i are functions of the a_i, the ridge procedure could "worsen" some estimates of the β_i. An estimate that had the "wrong" sign because an a_i was over-estimated could change to the right sign, but if the sign were "wrong" because an a_i was under-estimated, the ridge procedure could worsen the results. Changes from the right to the wrong sign are also possible.

(vi) The ridge estimator of β in the linear model can be expected to be better than the least square estimate when the orientation of the true regression vector coincides with eigenvector associated with the largest eigenvalue of $X^{*\prime}X^*$.

4.17 Shrinkage-estimators

The estimators of this type are divided in two classes.

(i) Deterministic shrunken estimators, i.e. $C_\lambda = \lambda\hat{\beta}$ $\lambda \in [0,\infty)$, λ is a fixed scalar.

(ii) Stochastic shrunken estimator, i.e. $\tilde{a}_{A_2} = \lambda\hat{\beta}$ where $\lambda = P(\hat{\beta}'\hat{\beta})$ is a scalar function of $\hat{\beta}'\hat{\beta}$.

In this section we examine the estimators of Stein (1960), of Sclove (1968) as well as the estimators proposed by Goldstein and Smith (1973). The ridge estimator is derived as a special case of the Goldstein and Smith estimator. Some generalizations are given. Other results concerning the shrinkage estimator will be given in Section 4.18, where we consider linear transformations of $\hat{\beta}$.

The first estimator that we consider is the Stein estimator.

(a) Stein shrinkage estimator (1960)

Consider the Model A with X_i-orthogonal and independent variables. It is assumed that the "regression plane" passes through the origin. Let $\beta_A^{(2)} = X'Y$ be the maximum likelihood estimator of $\beta_A^{(2)}$. The following general theorem due to Banachik (1970) can be used to obtain the estimator of Stein.

Theorem 4.17.1

Consider the Model A and the following assumptions

(i) $\hat{\beta} \sim (\beta, \sigma^2 I)$ and $p \geq 3$

(ii) The loss function $L(\tilde{\beta};\beta,\sigma^2) = (\tilde{\beta}-\beta)'(\tilde{\beta}-\beta)/\sigma^2$ where $\tilde{\beta}$ is an estimator of β

(iii) $F = \hat{\beta}'\hat{\beta}/S$, where $S \sim \sigma^2\chi^2_{n-p}$ and independent of $\hat{\beta} = \hat{\beta}^{(2)}_A$

(iv) $r(.)$ is monotone, nondecreasing

(v) $0 < r(.) < 2(p-2)/(n-p+2)$

Then relative to the loss function in (ii) an estimator of the form:

$\phi(\hat{\beta},S) = (1-r(F)/F)\hat{\beta}$ is better than $\hat{\beta}$.

Before we give the proof we give some definitions.

We define <u>risk</u> of an estimator to be the expected value of the loss function. So $\phi(\hat{\beta},S)$ will be a better estimator than $\hat{\beta}$ if the risk of $\phi(\hat{\beta},S)$ is less than the risk of $\hat{\beta}$.

Proof

Since the $\hat{\beta}$ is the maximum likelihood estimator of β it will be sufficient to show that

(4.17.1) $E\|\phi(\hat{\beta},S) - \beta\|^2 - E\|\hat{\beta}-\beta\|^2$ is not positive for all parameter values of (β,σ^2). Setting $g(F) = 1-r(F)/F$, (4.17.1) becomes

(4.17.2) $E(\hat{\beta}'\hat{\beta}g^2(F)) - 2\beta'E(g(F)\hat{\beta}) + \|\beta\|^2 - p\sigma^2$

Computing conditionally given $S = s$, we obtain the conditional expectations (4.17.3) - (4.17.8).

$$(4.17.3)\quad E[\hat{\beta}'\hat{\beta}g^2(\hat{\beta}'\hat{\beta}/s)] = e^{-\|\beta\|^2/2\sigma^2}\sum_{k=0}^{\infty}\frac{(\|\beta\|^2/2\sigma^2)^k}{k!}$$

$$\cdot E[\sigma^2\chi^2_{p+2k}g^2(\sigma^2\chi^2_{p+2k}1/s)]$$

where χ^2_{p+2k} is chi-square random variable with $p+2k$ degrees of freedom. To compute

$$(4.17.4)\quad \beta'E(g(F)\hat{\beta}) = \beta'E[g(\hat{\beta}'\hat{\beta}/s)\hat{\beta}],$$

we choose coordinate system so that the first coordinate axis lies along β. This does not affect the values of σ^2 and s. Then (4.17.4) is equal to

$$(4.17.5)\quad \|\beta\|E(g(\hat{\beta}'\hat{\beta}/s)\hat{\beta}_1)$$

where $\hat{\beta}_1$ is the first coordinate of $\hat{\beta}$.
Writing out (4.17.5) in terms of the distribution of $\hat{\beta}$ it becomes

$$\frac{\sigma^2\|\beta\|e^{-\|\beta\|^2/2\sigma^2}}{(2\pi\sigma^2)^{p/2}}\ \frac{d}{d\|\beta\|}\left[\int\cdots\int g(\Sigma b_i^2/s)\right.$$

$$\left.\cdot\exp(-\textstyle\sum_{i=1}^{p}b_i^2-2\|\beta\|b_1)/2\sigma^2)\,\pi_{i=1}^{p}db_i\right]$$

or

$$(4.17.6)\quad \sigma^2\|\beta\|e^{-\|\beta\|^2/2\sigma^2}\frac{d}{d\|\beta\|}\,e^{\|\beta\|^2/2\sigma^2}E(g(\sigma^2\chi^2_{p+2k_1}1/s))$$

where k_1 is Poisson random variable with mean $\|\beta\|^2/2\sigma^2$.
Thus (4.17.5) equals

$$(4.17.7)\quad 2\sigma^2\sum_{k=0}^{\infty} e^{-\|\beta\|^2/2\sigma^2}\left(\frac{\|\beta\|^2}{\sigma^2}\right)^k k\, E[g(\sigma^2\chi^2_{p+2k}{}^{1/s})/k!]$$

Combining (4.17.3) and (4.17.7) and noting $E(2k_1) = \frac{\|\theta\|^2}{\sigma^2}$, (4.17.2) (conditional on $S = s$) becomes

$$(4.17.8)\quad \sigma^2 e^{-\|\beta\|^2/2\sigma^2}\sum_{k=0}^{\infty}\frac{(\|\beta\|^2/2\sigma^2)^k}{k!}$$

$$\cdot\{E[\chi^2_{p+2k}g^2\left(\frac{\sigma^2\chi^2_{p+2k}}{s}\right)]-4kE[g\left(\frac{\sigma^2\chi^2_{p+2k}}{s}\right)]-p+2k\}$$

Averaging (4.17.8) over S and writing $S = \sigma^2\chi^2_{n-p}$, we see that the theorem will be proved if we show that

$$(4.17.9)\quad E\{\chi^2_{p+2k}g^2(\chi^2_{p+2k}/\chi^2_{n-p})-4kg(\chi^2_{p+2k}/\chi^2_n)-p+2k\}$$

is not positive for each value $k = 0,1,\ldots$

In the computation which follows we write $U = \chi^2_{p+2k}/\chi^2_{n-p}$ and will use the notation

$$(4.17.10)\quad r(U) = (1-g(U))U$$

and the fact that

$$(4.17.11)\quad g(U) \geq 1 - \frac{2(p-2)}{n+2-p}U^{-1}$$

It follows from (4.17.10) and the fact that $E(\chi^2_{p+2k}) = p+2k$ that (4.17.9) is equal to

$$E\{-2r(U)\chi^2_{n-p}+r(U)(1-g(U))\chi^2_{n-p}+4kr(U)/U\}$$

which is

$$(4.17.12)\quad E\{r(U)\chi^2_{n-p}(-1-g(U)+4k/\chi^2_{p+2k})\}$$

Using (4.17.11) we see that (4.17.12) is bound above by

(4.17.13) $E\{r(U)Z\} = E_{\chi^2_{n-p}}\{E\{r(\chi^2_{p+2k}/\chi^2_{n-p})Z|\chi^2_{n-p}\}\}$ where

$$Z = \chi^2_{n-p}\{-2+(4k+2\,\frac{p-2}{n-p+2}\,\chi^2_{n-p})/\chi^2_{p+2k}\}$$

Fixing χ^2_{n-p}, we define the constant α by

(4.17.14) $$-2+(4k+2\,\frac{p-2}{n-p+2}\,\chi^2_{n-p})/\alpha = 0$$

From condition (iv) we have the inequality

$$E\{r(\chi^2_{p+2k}/\chi^2_{n-p})Z|\chi^2_{n-p}\} \leq r(\alpha/\chi^2_{n-p})E\{Z|\chi^2_{n-p};\chi^2_{p+2k}\leq\alpha\}\cdot P[\chi^2_{p+2k} \leq \alpha]$$

$$+ r(\alpha/\chi^2_{n-p})E\{Z|\chi^2_{n-p};\chi^2_{p+2k}>\alpha\}\ P\{\chi^2_{p+2k} > \alpha\}$$

$$= r(\alpha/\chi^2_{n-p})E(Z|\chi^2_{n-p})$$

$$= r(\alpha/\chi^2_{n-p})\chi^2_n\{-2+(4k+2\,\frac{p-2}{n+2}\chi^2_{n-p})/(p-2+2k\}$$

Multiplying through by $(p-2+2k)/2(p-2)$ and using (4.17.13) and (4.17.14), we see that $\phi(\hat{\beta},S)$ will have minimum risk if

(4.17.15) $$E\{r\left(\frac{2k}{\chi^2_{n-p}} + \frac{p-2}{n-p+2}\right)\chi^2_{n-p}\{-1+\chi^2_{n-p}/(n+2-p)\}\}$$

is less than or equal to 0. But, (4.17.15) is bounded above by

$$r\left(\frac{2k+p-2}{n-p+2}\right)E\{\chi^2_{n-p}\{-1+\chi^2_{n-p}/(n-p+2)\}|\chi^2_{n-p}<n-p+2\}P\{\chi^2_{n-p}<n-p+2\}$$

$$+ r\left(\frac{2k+p-2}{n-p+2}\right)E\{\chi^2_{n-p}[-1+\chi^2_{n-p}/(n-p+2)]|\chi^2_{n-p}\geq n-p+2\}P\{\chi^2_{n-p}\geq n-p+2\}$$

$$= r\left(\frac{2k+p-2}{n-p+2}\right)E\{\chi^2_{n-p}[-1+\chi^2_{n-p}/(n-p+2)]\} = 0$$

Application 4.17.2

The above theorem will be used to obtain the estimator of Stein and of Alam and Thompson.

(1) Setting r equal to a constant c we obtain the Stein estimator for $0 \leq c \leq \frac{2(p-2)}{n-p+2}$. These estimators may be improved by replacing $(1-c/F)$ by $\max\{0,1-c/F\}$. It is worth noting that the "improved" estimators also satisfy the conditions of the theorem (here we take $r(F) = c$ if $c < F$, and equal to F, otherwise).

(2) Setting $r(F) = c/(1+cF^{-1})$ where $(0 \leq c \leq ((p-2)/(n-p+2))$ we have $\left(\frac{\hat{\beta}'\hat{\beta}}{\hat{\beta}'\hat{\beta}+cS}\right)\hat{\beta}$

The estimator given by Alam and Thompson.

(3) Define $f(F) = \begin{cases} c \; ; \; 0 \leq c \leq (p-2)/(n-p+2) & \text{if } F > c \\ 0 & \text{otherwise} \end{cases}$

Then the estimator is given by

$$\phi(\hat{\beta}) = \begin{cases} (1-c/F)\hat{\beta} & \text{if } F > c \\ \hat{\beta} & \text{if } F \leq c \end{cases}$$

(b) Sclove shrinkage estimator (1968)

Consider the model in (4.14.1) $Y^* = Wa + e$, $e \sim N(0,\sigma^2 I)$. Suppose that we partition a as (a_1,a_2), where a_1 is a m-component vector and a_2 is a q-component vector $(m+q = p)$,

in such a way so that the components of a_2 are corresponding to the q-variables with smallest eigenvalues (i.e. less than 0,03). We also partition $P = (P_1, P_2)$ in order to conform to $a = (a_1, a_2)$.

If we want to test the hypothesis

$$H_a : a_2 = 0 \quad \text{against} \quad H_1 : a_2 \neq 0$$

we use the following statistic

$$F^* = \hat{a}_2' \hat{a}_2 / q / \frac{R.S.S.(\hat{a})}{n-p} \sim F_{q,n-p}$$

If $F^* < \frac{n-p}{q} c$ then $\hat{a}_2 = 0$, where c is a constant which will be determined later on.

Use of the Sclove estimator $\hat{\beta}_s$, where

$$(4.17.16)\quad \hat{\beta}_s = \begin{cases} P_1 \begin{pmatrix} a_1 \\ 0 \end{pmatrix} & \text{if } F^* \leq \frac{n-p}{q} c \\ P_1 \hat{a}_1 + \left(1 - c\left(\frac{R.S.S.(\hat{a})}{\hat{a}_2' \hat{a}_2}\right)\right) P_2 \hat{a}_2 & \text{if } F^* > \frac{n-p}{q} c \end{cases}$$

and $c \in (0, 2(q-2)/(n-p+2))$, corresponds to making a preliminary test of the hypothesis $a_2 = 0$, at a level of significance a_c dictated by the value c such that $a_c = \Pr\{F_{q,n-p} > \frac{(n-p)}{q} c\}$. By taking $c = c_0 = \frac{q-2}{n-p-2}$ the values of $a_c = \Pr\{F_{q,n-p} > (n-p)(q-2)/q(n-p+2)\}$ are given in Table (4.17.A), for $n-p = q$ and for $n-p = \infty$.

Note that $\lim_{n \to \infty} \frac{(n-p)(q-2)}{q(n-p+2)} = q - \frac{2}{q}$.

The asymptotic distribution of $F_{q,n-p}$ as $n \to \infty$ is

χ_q^2/q. Hence $a_c \simeq \text{Prob}\{\chi_q^2/q > (q-2)/q\} = \text{Prob}\{\chi_q^2 > q-2\}$. Clearly as q increases this probability decreases to $\frac{1}{2}$.

We note that for the value $c = c_o$ it has not been shown if it is optimal or not optimal. It is seen that use of (4.17.16) with $c = c_o$ correspond to testing the hypothesis $a_2 = 0$ at a very large significance level. In this sence, then, good values for a_c need not necessarily be low. However, one has a better chance of coming out with a simple regression equation if a_c is small. Taking c larger makes a_c smaller. In fact we can take $c = 2c_o = 2(q-2)/(n-p+2)$ and still have an estimator that is as good as the ordinary one.

TABLE (4.17.A)

		q						
		3	4	5	10	20	30	∞
n-p=q	c_o	0,89	0,84	0,81	0,73	0,67	0,64	0,50
	$2c_o$	0,76	0,65	0,57	0,33	0,14	0,07	0,00
n-p=∞	c_o	0,80	0,74	0,70	0,63	0,59	0,57	0,50
	$2c_o$	0,57	0,41	0,31	0,10	0,02	0,03	0,00

$c_o = q-2/n-p+2$

(c) Goldstein-Smith (1973)

Consider the Model A_s. We know there exist orthogonal Q, $n\times n$, and P, $p\times p$, such that

$$(4.17.17)\quad PX^{*\prime}X^{*}P' = \Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix}$$

and

$$(4.17.18)\quad QX^{*}P' = D = \begin{pmatrix} \Lambda^{\frac{1}{2}} \\ 0 \end{pmatrix}$$

where $\Lambda^{\frac{1}{2}} = \text{diag}\{\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_p}\}$.

Writing $z = QY^{*}$, $\gamma = P\beta$, $v = Qe$, we obtain

$$(4.17.19)\quad z = D\gamma + v, \quad v \sim N(0,\sigma^2 I) \quad \text{explicity}$$

$$z_i \sim N(\sqrt{\lambda_i}\,\gamma_i, \sigma^2) \qquad i = 1,\ldots,p$$

$$z_i \sim N(0,\sigma^2) \qquad i = p+1,\ldots,n.$$

It is easy to show that $\hat{\gamma}_i = \dfrac{z_i}{\sqrt{\lambda_i}} \qquad i = 1,2,\ldots,p.$

we focus on estimators of the form $\hat{\gamma}_{G,i} = c_i z_i$ with $|c_i z_i| \leq |z_i/\sqrt{\lambda_i}|$. Then we consider $c_i = c_i(\sqrt{\lambda_i}, k)$ such that

$$(4.17.20)\quad \begin{cases} \text{(a)} \quad c(\sqrt{\lambda_i}, 0) = \dfrac{1}{\sqrt{\lambda_i}} \\ \text{(b)} \quad c'(\sqrt{\lambda_i}, k)/\sqrt{\lambda_i} < 0 \quad \text{for all} \quad k \geq 0 \end{cases}$$

Lemma 4.17.2

For any γ, there exist $k \geq 0$ such that

$\hat{\gamma}_{G,i} = c(\sqrt{\lambda_i},k)z_i$ has smaller mean square error than $\hat{\gamma}_i = z_i/\sqrt{\lambda_i}$, for all $i = 1,2,\ldots,p$.

Proof

We set $c(\sqrt{\lambda_i},k) = c_i$ then

$$E(c_i z_i - \gamma_i)^2 = \gamma_i^2(c_i\sqrt{\lambda_i}-1)^2 + c_i^2\sigma^2$$

and

$$E\{(z_i/\sqrt{\lambda_i}) - \gamma_i\}^2 = \sigma^2/\lambda_i$$

In order to have $M.S.E.(\hat{\gamma}_{G,i}) < M.S.E.(\hat{\gamma}_i)$ we must have

$$E(c_i z_i - \gamma_i)^2 < E\{(z_i/\sqrt{\lambda_i}) - \gamma_i\}^2$$

or finally

$$(4.17.21)\quad \gamma_i^2 < \frac{\sigma^2}{\lambda_i}\,\frac{(1/\lambda_i)^{\frac{1}{2}}+c_i}{(1/\lambda_i)^{\frac{1}{2}}-c_i} \qquad i = 1,2,\ldots,p.$$

The properties of c_i in (4.17.20) ensure the existence of k such that (4.17.21) is satisfied.

Now if we define:

$$(4.17.22)\quad \hat{\beta}^*_G = P\hat{\gamma}_G \quad \text{where} \quad \hat{\gamma}_{G,i} = c(\sqrt{\lambda_i},k)z_i \qquad i = 1,2,\ldots,p$$

then we can prove the following theorem.

Theorem 4.17.3

For any β^*, there exists $k > 0$ such that, for $\hat{\beta}^*_G$ defined by (4.17.22), $\hat{\beta}^*_{G,i}$ has smaller mean square error than the corresponding least squares $\hat{\beta}^*_i$, for all $i = 1,2,\ldots,p$.

We wish to note the following:

(1) If we take $c(\sqrt{\lambda_i},k) = \frac{\sqrt{\lambda_i}}{(\lambda_i+k)}$, then the estimator $\hat{\beta}^*_G$ defined in (4.17.22) is the ridge estimator $\hat{\beta}_R$ (see 4.2.1).

(2) If we take $c_i(\sqrt{\lambda_i},k_i) = \frac{\sqrt{\lambda_i}}{(\lambda_i+k_i)}$ $i = 1,2,\ldots p$ then the estimator $\hat{\beta}^*_G$ is the generalized ridge estimator.

(3) We can also alter the manner in which the λ_i enter in the expression for c_i. One such alternative is obtained by taking

$$c(\sqrt{\lambda_i},k) = \lambda_i^{(2m-1)/2}/(\lambda_i^m+k) \qquad i = 1,2,\ldots,p$$

where m is an integer.

We can see that the requirements of (4.17.20) are satisfied, and, for $m \geq 2$, has the effect of making the analysis more sensitive to variation in the eigenvalue spectrum.

4.18 Linear transforms of $\hat{\beta}$.

In this section we consider estimators which belong to the following class.

Definition 4.18.1

Let C denote the class of linear transforms of $\hat{\beta}$. Then

$$C = \{\tilde{a}: \tilde{a} = A\hat{\beta} \quad \text{for some} \quad A \in M_{p\times p}\}.$$

Let $\tilde{a}_A = A\hat{\beta}$ be for fixed A then

(1) $E(\tilde{a}_A) = A\beta$

(2) $Var(\tilde{a}_A) = \sigma^2 A\,(X'X)^{-1}A'$

(3) The mean square error is given by

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

(4) $R.S.S.(\tilde{a}_A) = (Y-X\tilde{a}_A)'(Y-X\tilde{a}_A)$

$$= R.S.S.(\hat{\beta}) + \hat{\beta}'(A-I)'X'X(A-I)\hat{\beta}$$

$$= R.S.S.(\hat{\beta}) + L^*(A)$$

The (4) attains its minimum when $A = I$.

Consider now the following mapping:

(4.18.1) $\gamma : M_{p\times p} \to \mathbb{R}^+ : A \to \gamma(A) :=_{def} L^*(A) = \hat{\beta}'(A-I)'X'X(A-I)\hat{\beta}$

which is not 1-1.

The pre-image of any fixed constant τ is given by

(4.18.2) $\gamma^{-1}(\tau) = \{A \in M_{p\times p} : \hat{\beta}'(A-I)X'X(A-I)\hat{\beta} = \tau\}$

Definition 4.18.2

Let $C(\tau)$ be a subclass of C such that $\tilde{a}_{A_0} \in C(\tau)$ iff $L^*(A_0) = \tau$. Then

$$C(\tau) = \{\tilde{a}_{A_0} : \tilde{a}_{A_0} \in C(\tau) \text{ with } L^*(A_0) = \tau\}.$$

If we take as equivalence relation the R.S.S., then $C(\tau)$ with this relation consists of an equivalence class (E.C.) or orbit within the class C.

As a criterion for selection the optimal estimator can be considered the minimization of the norm of the estimator.

In fact different norms lead to different estimators, and there is no reason for preferring one norm or another.

Thus it is probably much better to choose an estimator which has minimum total variance among all estimators in a given equivalence class.

Suppose the criterion for selecting an estimator from an equivalence class is to choose the estimator, which has minimum Euclidean length (norm). Let

(4.18.3) $\|\tilde{a}_{A_0}\| = \hat{\beta}'AA\hat{\beta}$.

Proposition 4.18.3

If $A_0 = (k(X^{*\prime}X^*)^{-1}+I)^{-1}$ for $k \geq 0$ and $\tilde{a}_{A_0} \in C(\tau)$ then

$$\|\tilde{a}_{A_0}\| = \min_{C(t)} \|\tilde{a}_A\|$$

Hint: We differentiate the Lagrangian expression

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

with respect to A and setting the derivative equal to zero.

Proposition 4.18.4

Let the norm be $\|\tilde{a}_A\|_d = \hat{\beta}'A'(X'X)A\hat{\beta}$. If $A_1 = \lambda I$, $\lambda \in [0,1]$ and $C_\lambda = \lambda(X'X)^{-1}X'Y = \lambda\hat{\beta}$ λ-fixed scalar then

$$\|C_\lambda\|_d = \min_{C(\tau)} \|\tilde{a}_A\|_d, \quad \text{where } \|\ \|_d \text{ is the design dependent norm.}$$

Hint: We form and differentiate the Lagrangian expression

$$F = \hat{\beta}'A'X'XA\hat{\beta}+\lambda[\hat{\beta}'(A-I)'X'X(A-I)\hat{\beta}-C]$$

which yields

$$\frac{\partial F}{\partial A} = 2AX'X\hat{\beta}\hat{\beta}' + \lambda[2AX'X\hat{\beta}\hat{\beta}' - 2X'X\hat{\beta}\hat{\beta}']$$

Setting $\frac{\partial F}{\partial A} = 0$ and solving for A.

Since $\tilde{a}_{A_1} = \lambda\hat{\beta} = C_\lambda$ we have shown that both the ridge estimator and the shrunken estimator are minimum length estimators with respect to the appropriate norms.

Proposition 4.18.5

Let $A_2 = \delta\hat{\beta}\hat{\beta}'(I+\delta\hat{\beta}\hat{\beta}')^{-1}$ for some $\delta \in [0,\infty)$, if $\tilde{a}_{A_2} \in C(\tau)$ then

$$\text{var}(\tilde{a}_{A_2}) = \min_{C(\tau)} \text{var}(\tilde{a}_A).$$

Hint: We minimize the R.S.S.$(\tilde{a}_A)$ subject to $\text{tr}(A'(X'X)^{-1}A)=C$ using the Lagrangian multiplier.

Lemma 4.18.6

$$(I+\delta\hat{\beta}\hat{\beta}')^{-1} = I-(1+\delta\hat{\beta}'\hat{\beta})^{-1}\delta\hat{\beta}\hat{\beta}'$$

Proof

Since $\delta\hat{\beta}\hat{\beta}'$ is a $p\times p$ matrix of rank 1 it has a single eigenvalue $\lambda = \delta\hat{\beta}'\hat{\beta}$. There exists an orthogonal matrix P such that $P\delta\hat{\beta}\hat{\beta}'P = \text{diag}(\lambda,0,\ldots 0) = \Lambda$. Therefore

$$\begin{aligned}(I+\delta\hat{\beta}\hat{\beta}')^{-1} &= P'P(I+\delta\hat{\beta}\hat{\beta}')^{-1}P'P \\ &= P'(I+P\delta\hat{\beta}\hat{\beta}'P')^{-1}P\end{aligned}$$

$$= P'(I+\Lambda)^{-1}P$$

$$= P'[\mathrm{diag}(1+\lambda,1,\ldots,1)]^{-1}P$$

$$= P'\mathrm{diag}((1+\lambda)^{-1},\ 1,\ldots,1)P$$

$$= P'[I-\mathrm{diag}(\lambda(1+\lambda)^{-1},\ 0,\ldots,0]P$$

$$= P'[I-(1+\lambda)^{-1}\mathrm{diag}(\lambda,0,\ldots,0)]P$$

$$= I-(1+\lambda)^{-1}\delta\hat{\beta}\hat{\beta}'$$

$$= I-(1+\delta\hat{\beta}'\hat{\beta})^{-1}\delta\hat{\beta}\hat{\beta}'$$

Proposition 4.18.7

The estimator $\tilde{a}_{A_2}$ of Proposition 4.18.5 can be written as $\tilde{a}_{A_2} = \delta[\hat{\beta}'\hat{\beta}-(1+\delta\hat{\beta}'\hat{\beta})^{-1}\delta(\hat{\beta}'\hat{\beta})^2]\hat{\beta}$ and is a shrunken estimator.

Hint: We use the above lemma.

Proposition 4.18.8

The following results can be shown

(1) $\tilde{a}_{A_2} = 0$ if $\delta = 0$

(2) $\tilde{a}_{A_2} = \hat{\beta}$ if $\delta = -2^{-\frac{1}{2}}(\hat{\beta}'\hat{\beta})^{-1}$

(3) The absolute value of the $\tilde{a}_{A_2}$ is an increasing function of δ.

We notice the following:

The difference between C_λ and $\tilde{a}_{A_2}$ is minimal in practice for the following two reasons:

(i) Suppose the R.S.S. is fixed and $C(\tau)$ is determined, then if $\tilde{a}_{A_2}$ and C_λ belong to $C(\tau)$ they are identical.

(ii) If in practice the shrinkage factor is chosen after observing $\hat{\beta}$ then the shrunken estimator being used is stochastic whether it is of the form C_λ or $\tilde{a}_{A_2}$.

How to choose the optimal δ.

(1) By plotting the elements of $\tilde{a}_{A_2}$ against δ and using the stability criteria as is proposed by Hoerl and Kennard (see 4.7).

(2) Use the "maximum inflation factor" (see 4.8).

C H A P T E R 5

PRINCIPAL COMPONENT REGRESSION

5.1 Introduction

The method of principal components has been known for many years, and has been discussed in many different ways by a variety of authors, for example Kendall (1957), Anderson (1958), Massy (1965), Hawkins (1973) and Greenberg (1975).

In this chapter we derive the principal components from the correlation matrix and we give some methods of selecting the components with emphasis on the generalized inverse method. Finally a method called Latent Root Regression Analysis (LRRA) will be treated.

5.2 Derivation of principal components from the correlation matrix

Consider the standardized matrix of Model A_s (see (1.2.5)). Our aim is to find a lineal transformation of the set of p-variates of X^* into a new set denoted by W, where the new set has certain desirable properties. These properties are:

(1) The elements of W are uncorrelated with each other in the sample.

(2) The first element of W will have the maximum possible variance, the second, the maximum possible variance among

those uncorrelated with the first, and so forth.

Let

(5.2.1) $w_1 = X^*v_1$ denote the first new variable, where w_1 is an n-element vector and v_1 is a p-element vector. The sum of squares of w_1 is

(5.2.2) $w_1'w_1 = v_1'X^{*\prime}X^*v_1$.

We wish to choose v_1 to maximize $w_1'w_1$. In order to avoid making $w_1'w_1$ large we impose the following constraint

(5.2.3) $v_1'v_1 = 1$

The problem now is to maximize (5.2.2) subject to (5.2.3).

We define $F = v_1'X^{*\prime}X^*v_1 - \lambda_1(v_1'v_1-1)$ where λ_1 is a Lagrange multiplier. Setting $\frac{\partial F}{\partial v_1} = 0$, we have

(5.2.4) $(X^{*\prime}X^*)v_1 = \lambda_1 v_1$

Thus v_1 is the latent vector of $X^{*\prime}X^*$ corresponding to the latent root λ_1. From (5.2.2) and (5.2.4) we see that $w_1'w_1 = \lambda_1 v_1'v_1 = \lambda_1$ and so we must choose λ_1 as the largest root of $X^{*\prime}X^*$. When the variables are not collinear, the $X^{*\prime}X^*$ will be positive definite and thus will have positive latent roots.

Now define $w_2 = X^*v_2$. We wish to choose v_2 to maximize $v_2'X^{*\prime}X^*v_2$ subject to $v_2'v_2 = 1$ and $v_1'v_2 = 0$.

The reason for the second condition is that w_2 is to be

uncorrelated with w_1. Define $F = v_2'X^{*\prime}X^*v_2 - \lambda_2(v_2'v_2 - 1) - \mu(v_1'v_2)$ where λ, μ are the Lagrange multipliers. We set $\frac{\partial F}{\partial v_2}$ equal to zero, so

$$(5.2.5) \quad \frac{\partial F}{\partial v_2} = 2X^{*\prime}X^*v_2 - 2\lambda_2 v_2 - \mu v_1 = 0$$

Premultiply by v_1' we have

$$2v_1'X^{*\prime}X^*v_2 - \mu = 0$$

But

$$(X^{*\prime}X^*)v_1 = \lambda_1 v_1$$

or

$$v_2'(X^{*\prime}X^*)v_1 = \lambda_1 v_2'v_1 = 0$$

we conclude that $\mu = 0$.

So (5.2.5) can be written as

$$(5.2.6) \quad (X^{*\prime}X^*)v_2 = \lambda_2 v_2$$

where λ_2 should be chosen as the second largest latent root of $X^{*\prime}X^*$.

We can proceed in this way for each of the p latent roots of $X^{*\prime}X^*$ and thus form an orthogonal matrix

$$(5.2.7) \quad P = [v_1, v_2, \ldots, v_p]$$

Now the p-principal components of X^* are given by the $(n \times p)$ matrix W

$$(5.2.8) \quad W = X^*p$$

Moreover

$$(5.2.9) \quad W'W = P'X^{*\prime}X^{*}P = \Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix}$$

If the rank of X^* were $r < p$, then $p-r$ latent roots would be zero and the variation in the X^*'s could be expressed in terms of r independent variables. Even if X^* is a full rank matrix some of the λ's may be fairly close to zero so that a small number of principal components account for a substantial proportion of the variance of the X^*'s.

The total variation in the X^*'s is given by

$$(5.2.10) \quad \sum_{j=1}^{n} X_{j1}^{*2} + \sum_{j=1}^{n} X_{j2}^{*2} + \ldots + \sum_{j=1}^{n} X_{jp}^{*2} = \operatorname{tr}(X^{*\prime}X^{*}) = p$$

But
$$\operatorname{tr}(P'X^{*\prime}X^{*}P) = \operatorname{tr}(X^{*\prime}X^{*}PP')$$
$$= \operatorname{tr}(X^{*\prime}X^{*})$$

so from (5.2.9) we have

$$(5.2.11) \quad \sum_{i=1}^{p}\sum_{j=1}^{n} X_{ji}^{*2} = \operatorname{tr}(X^{*\prime}X^{*})$$
$$= \sum_{i=1}^{p} \lambda_i$$
$$= w_1'w_1 + \ldots + w_p'w_p$$
$$= p$$

Thus $\lambda_1/p, \lambda_2/p, \ldots, \lambda_p/p$ represent the proportional contribution of each principal component to the total variation of the X^*'s.

Important note: Massy (1965) and some other authors scale the components so that all of them have equal variations.

These scaled principal components will be denoted by W^* rather than W; they are defined by

$$W^* = X^*(\Lambda^{-\frac{1}{2}}P)$$

The coefficient matrix $\Lambda^{-\frac{1}{2}}P$ is the inverse of the one ordinarily obtained in a factor analysis equation

$$X^* = W^*B$$

B is called the principal component loading matrix. Thus

$$B = (\Lambda^{-\frac{1}{2}}P)^{-1} = P^{-1}\Lambda^{+\frac{1}{2}}$$
$$= P'\Lambda^{\frac{1}{2}}$$

The elements of B, i.e. the b_{ij} are the correlations between the ith variables and jth components.

The selection of a subset of principal components

Frequently the intercorrelations between economic and social variables means that a small number of components will account for a large proportion of the total variation, and it is desirable to have a test for judging the number of components to retain for further analysis.

In what follows we discuss some methods for deleting principal components.

5.3 The method of Massy (1965)

The author proposes two alternative criteria

(i) Delete the components that are relatively unimportant as predictors of the original independent variables (X^*), i.e. the components having the smallest latent vectors should be dropped.

(ii) Delete components that are relatively unimportant as predictors of the dependent variable (Y^*) in the problem.
In this case the components having the smallest correlations with Y^* should be dropped.

We wish to note the following:

(1) The correlation between Y^* and the ith principal component is

$$(5.3.1) \quad r_{Y^*,w_i} = \frac{w_i' Y^*}{\lambda_i^{\frac{1}{2}}}$$

$$= \frac{v_i' X^{*\prime} Y^*}{\lambda_i^{\frac{1}{2}}} \quad \Big| \quad v_i'(X^{*\prime}X^*)^{-1} v_i = \frac{1}{\lambda_i} v_i' v_i \quad \Big| \quad \lambda_i v_i'(X^{*\prime}X^*)^{-1} = v_i'$$

$$= \lambda_i^{\frac{1}{2}} v_i'(X^{*\prime}X^*)^{-1} X^{*\prime} Y^*$$

$$= \lambda_i^{\frac{1}{2}} v_i' \hat{\beta}$$

$$= \lambda_i^{\frac{1}{2}} \hat{\delta}_i$$

where $\hat{\delta}_i$ is the estimate of the coefficient of the ith principal component.

(2) The coefficient of determination is given by

$$(5.3.2) \qquad R^2_{Y^*:w_1,\ldots,w_p} = \sum_{i=1}^{p} \lambda_i (\hat{\delta}_i)^2$$

5.4 The method of Kendall (1968)

Suppose that we have computed the roots $\lambda_1, \lambda_2, \ldots, \lambda_p$ and that the first r-roots $\lambda_1, \lambda_2, \ldots, \lambda_r$ $(r < p)$ seem sufficiently large and different to be retained.

The question is whether the remaining $(p-r)$ roots and their associated components are sufficiently alike for one to conclude that the true values are equal. A test based on

$$(5.4.1) \qquad \rho = (\lambda_{r+1}, \lambda_{r+2}, \ldots, \lambda_p)^{-1} \left(\frac{\lambda_{r+1} + \lambda_{r+2} + \ldots + \lambda_p}{p-r} \right)^{p-r}$$

can be conducted.

We test the hypothesis

$H_0 : \lambda_{r+1} = \lambda_{r+2} = \ldots = \lambda_p$ against $H_1 : \lambda_{r+1} \neq \lambda_{r+2} \ldots \neq \lambda_p$

and we use the statistic

$$n \log_e \rho \sim \chi^2_{\frac{1}{2}(p-r-1)(p+2-r)}$$

5.5 The method of Marquardt (1970)

Marquardt (1970) suggested that β^* of Model A_s (see (1.2.5)) should be estimated as

$$(5.5.1) \qquad \hat{\beta}^*_{G.I} = P_r \Lambda_r^{-1} P_r' (X^{*\prime} X^*) \hat{\beta}^*$$
$$= P_r P_r' \hat{\beta}^*$$

$$= P_r \hat{\delta}$$

$$= v_1\hat{\delta}_1 + v_2\hat{\delta}_2 + \ldots + v_r\hat{\delta}_r$$

where

(1) $\Lambda_r^{-1} = \text{diag}(\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \ldots, \frac{1}{\lambda_r})$

(2) $\hat{\delta} = (\hat{\delta}_1, \hat{\delta}_2, \ldots, \hat{\delta}_r)'$ is the vector of estimated principal components regression coefficients, $(P_r')\hat{\beta}^*$.

(3) P_r is the matrix, whose columns are the r out of p eigenvectors of the $(X^{*\prime}X^*)$ matrix.

(4) $r = r(X^{*\prime}X^*)$ $(r = \text{rank } (X^{*\prime}X^*))$.

Definition 5.5.1

The estimator (5.5.1) is called the generalized inverse estimator.

The rank of $(X^{*\prime}X^*)$ may either be an integer $1 \leq r \leq p$ or fractional.

The fractional rank as well as the modified fractional rank is discussed in Section 5.6.

Marquardt proposes the following method in the case of integer rank.

"Choose the smallest value r for which

(5.5.2) $\left|\frac{\sum_{j=p}^{p-r} \lambda_j}{\text{trace } (\Lambda)}\right| < \omega$ is satisfied

where $\omega \in (10^{-1}, 10^{-7})$, usually $\omega = 10^{-5}$." Another method of determining the integer rank can be found in Section (5.6).

Now we examine some properties of the generalized inverse estimator. We note that the generalized inverse matrix A_r^+ is given by

(5.5.3) $A_r^+ = P_r \Lambda_r^{-1} P_r'$

where r is the assigned rank. In general, there is an optimum value for r for any problem, but it is desirable to examine the generalized inverse solution for a range of admissible values for r.

Theorem 5.5.1

Let $\hat{\beta}^*_{G.I}$ be the solution to the normal equations $X^{*\prime}X^*\hat{\beta}^* = X^{*\prime}Y^* = g$ obtained by assigning rank r to the matrix $A = X^{*\prime}X^*$, and using the generalized inverse (5.5.3). Then $\hat{\beta}^*_{G.I}$ minimizes the

(5.5.4) $R.S.S.(\hat{\beta}^*) = (Y^*-X^*\hat{\beta}^*)'(Y^*-X^*\hat{\beta}^*)$ for all $\hat{\beta}^*$ within the r-dimensional subspace spanned by P_r.

Proof

The normal equations $A\hat{\beta}^*_{G.I} = g$ may be premultiplied by $P' = P^{-1}$ and written in the form

(5.5.5) $(P'X^{*\prime}X^*P)(P'\hat{\beta}^*_{G.I}) = P'g$

Now

(5.5.6) $W = X^*P$ is the projection of the points of X^* onto rotated axes defined by the eigenvectors which are the columns of P.

In the eigenvector coordinates the normal equations are therefore

(5.5.7) $$(W'W)\hat{\delta}_P = W'Y^*$$

where $$\hat{\delta}_P = P'\hat{\beta}^*_{G.I}$$

and $\hat{\delta}_P$ is the projection of $\hat{\beta}^*_{G.I}$ onto the eigenvectors coordinates.

If, now, only the first r columns are used $(r \leq p)$ in (5.5.6) then

(5.5.8) $$W_r = X^*P_r$$

Consequently

(5.5.9) $$(W_r'W_r)\hat{\delta}_{P_r} = W_r'Y^*$$

yields the least squares solution $\hat{\delta}_{P_r}$ minimizing (5.5.4) within that subspace.

From (5.5.8) and (5.5.9) we have

(5.5.10) $$\hat{\delta}_{P_r} = \Lambda_r^{-1}P_r'X^{*'}Y^*$$

If we express the solution in the original coordinates we have

$$\hat{\beta}^*_{G.I} = P_r\hat{\delta}_{P_r}$$

or

$$\hat{\beta}^*_{G.I} = P_r\Lambda_r^{-1}P_r'g$$

Thus, the generalized inverse method confines the solution to a linear subspace containing the origin, whereas the ridge method confines the regression vector to a sphere about the origin.

Theorem 5.5.2

Let $\hat{\beta}^*_{G.I}$ be defined as in Theorem 5.5.1. Then $\|\hat{\beta}^*_{G.I}\|^2$ is a stepwise increasing function of r.

Proof

For $r \leq p$

$$(5.5.11)\quad \|\hat{\beta}^*_{G.I}\|^2 = \hat{\beta}^{*\prime}_{G.I}\hat{\beta}^*_{G.I}$$

$$= (P_r\Lambda_r^{-1}P_r'g)'(P_r\Lambda_r^{-1}P_r'g)$$

$$= g'P_r\Lambda_r^{-2}P_r'g$$

$$= \sum_{i=1}^{r}\lambda_i^{-2}\left[\sum_{j=1}^{p}g_jP_{ji}\right]^2$$

We observe that the value of the ith term is independent of r; moreover $\|\hat{\beta}^*_{G.I}\|^2$ has been expressed as a sum of squares; therefore $\|\hat{\beta}^*_{G.I}\|^2$ is a stepwise increasing function of r. The ith term in the summation is the increase in the squared length of $\hat{\beta}^*_{G.I}$ due to including the ith eigenvector dimension. It is easy to see that if λ_i is small $1/\lambda_i^2$ is large so the length of $\hat{\beta}^*_{G.I}$ is disproportionately increased by including dimensions for which λ_i is small.

While $\|\hat{\beta}^*_{G.I}\|^2$ is an increasing function of r the $\|\hat{\beta}_R\|^2$ is a decreasing function of k.

Theorem 5.5.3

The estimator $\hat{\beta}^*_{G.I}$ is a linear transform of $\hat{\beta}$, and the transform depends on X^* and r.

Proof

$$\begin{aligned}\hat{\beta}^*_{G.I} &= P_r \Lambda_r^{-1} P_r' X^{*\prime} Y^* \\ &= P_r \Lambda_r^{-1} P_r' (X^{*\prime} X^*) \hat{\beta}^* \\ &= Z_r \hat{\beta}^*\end{aligned}$$

It follows immediately that $\hat{\beta}^*_{G.I}$ is a biased estimator of β^*, if Λ_{p-r} is a non-null matrix. If Λ_{p-r} is the null-matrix then $\hat{\beta}^*_{G.I}$ is conditionally unbiased relative to the constraints implied by the columns of P_{p-r}, $E(\beta^*_{G.I}) = Z_r \beta^*$

Theorem 5.5.4

The variance of $\hat{\beta}^*_{G.I}$ is

$$\text{var}(\hat{\beta}^*_{G.I}) = \sigma^2 \, P_r \Lambda_r^{-1} P_r' \, X^{*\prime} X^*$$

Proof

$$\begin{aligned}\text{var}(\hat{\beta}^*_{G.I}) &= \text{var}(Z_r \hat{\beta}^*) \\ &= \sigma^2 Z_r (X^{*\prime} X^*)^{-1} Z_r' \\ &= \sigma^2 (P_r \Lambda_r^{-1} P_r') (X^* X^*) (P_r \Lambda_r^{-1} P_r') \\ &= \sigma^2 P_r \Lambda_r^{-1} P_r'\end{aligned}$$

Lemma 5.5.5

The mean square error of $\hat{\beta}^*_{G.I}$ is

$$E(L_1^2) = \mathrm{tr}(\mathrm{var}(\hat{\beta}^*_{G.I})) + \beta^{*\prime}(Z_r - I)'(Z_r - I)\beta^*$$

Hint: The proof is analogous to the proof of Lemma 4.5.2.

Theorem 5.5.6

The total variance $\mathrm{tr}(\mathrm{var}(\hat{\beta}^*_{G.I}))$ is an increasing function of r.

Hint:

$$(5.5.12)\quad \mathrm{tr}(\mathrm{var}(\hat{\beta}^*_{G.I})) = \sigma^2 \mathrm{tr}(P_r \Lambda_r^{-1} P_r')$$
$$= \sigma^2 \sum_{j=1}^{r} \frac{1}{\lambda_j} \mathrm{tr}(v_j v_j')$$
$$= \sigma^2 \sum_{j=1}^{r} \frac{1}{\lambda_j}$$

Since $\lambda_j \geq 0$ for all j, this implies that (5.5.12) increases monotonically with r.

We note that in ridge regression the total variance $\gamma_i(k)$ is monotonous decreasing function of k.

Theorem 5.5.7

The bias term in $E(L_1^2)$ of Lemma 5.5.5, i.e. $\beta^{*\prime}(Z_r - I)'(Z_r - I)\beta^*$ is a monotonic decreasing function of r.

Proof

$$(5.5.13)\quad (\text{Bias})^2 = \beta^{*\prime}(Z_r - I)'(Z_r - I)\beta$$

$$= \beta^{*\prime}[P_r \Lambda_r^{-1} P_r'(X^{*\prime}X^*) - I]'[P_r \Lambda_r^{-1} P_r'(X^{*\prime}X^*) - I]\beta^*$$

$$= \beta^{*\prime}[P_r P_r' - I]'[P_r P_r' - I]\beta^*$$

$$= \beta^{*\prime}[-P_{p-r} P_{p-r}']'[-P_{p-r} P_{p-r}']\beta^*$$

$$= \beta^{*\prime}(P_{p-r} P_{p-r}')\beta^*$$

$$= \sum_{j=r+1}^{p} a_j^2$$

where $a_{p-r} = P_{p-r}'\beta^*$ is the $(p-r)$ element vector of projecting β^* onto the subspace spanned by P_{p-r}. Since a_j does not depend on r, we have the result that $(\text{Bias})^2$ is a monotonic decreasing function of r. For $r = p$ and $r = 0$ the $(\text{Bias})^2$ becomes zero and $\beta^{*\prime}\beta^*$ respectively.

We note that in ridge regression the $(\text{Bias})^2$ is a monotonic increasing function of k.

Theorem 5.5.8

A sufficient condition for the mean square error $E(L_1^2)$ to be less than the least squares variance is

$$\sum_{j=r+1}^{p} \frac{1}{\lambda_j} > \frac{1}{\sigma^2}\,(\beta^{*\prime}\beta^*)$$

Proof

From Theorems 5.5.6 and 5.5.7 we have that

$$(5.5.14)\quad E(L_1^2) = \sigma^2 \sum_{j=1}^{r} \frac{1}{\lambda_j} + \sum_{j=r+1}^{p} a_j^2$$

while the least squares variance is

$$\mathrm{var}(\hat{\beta}^*) = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}$$

Thus, a sufficient condition for

$$E(L_1^2) < \mathrm{var}(\hat{\beta}^*)$$

is $\sigma^2\left(\sum_{j=r+1}^{p} \frac{1}{\lambda_j}\right) > \sum_{j=r+1}^{p} a_j^2$

or

$$\sum_{j=r+1}^{p}\left(\frac{\sigma^2}{\lambda_j} - a_j^2\right) > 0$$

or

$$\frac{\sigma^2}{\lambda_j} > a_j^2 \quad \text{for all} \quad j > r$$

Since the last relation would be difficult to apply in practice a more useful sufficient condition can be obtained by noting that $\sum_{j=1}^{p} \beta_j^{*2} \geq \sum_{j=r+1}^{p} a_j^2$ for any $r \leq P$. This theorem corresponds to Theorem 4.5.6 of ridge regression.

Theorem 5.5.9

The generalized inverse estimator is equivalent to a least squares estimator according to one of the following circumstances.

(a) Λ_{p-r} is a null matrix. In this case, the columns of P_{p-r} are imposed constraints and the estimator is immediately seen to be a constrained least squares estimator, and is then said to be conditionally unbiased.

or

(b) Λ_{p-r} is not a null matrix. In this case, the $\hat{\beta}^*_{G.I}$ is equivalent to a least squares estimator when the actual data are supplemented by a fictitious set of data points, i.e. by a $(r\times r)$ matrix $H_r = (P_{p-r}\Lambda^{\frac{1}{2}}_{p-r}(-1)^{\frac{1}{2}})$, the components of Y^* corresponding to the rows of H_r are set equal to zero.

Proof

It is necessary to find H_r such that

$$(5.5.15)\quad (X^{*\prime}X^*+H_r'H_r)^{-1} = P_r\Lambda_r^{-1}P_r'$$

for assigned rank r.

The $r(H_r') = p-r$ so we can invert both sides of (5.5.15).

$$(5.5.16)\quad (X^{*\prime}X^*+H_r'H_r) = P_r\Lambda_rP_r'$$

But $X^{*\prime}X^* = P_r\Lambda_rP_r' + P_{p-r}\Lambda_{p-r}P_{p-r}'$, so that

$$H_r'H_r = -P_{p-r}\Lambda_{p-r}P_{p-r}'$$

$$= -(P_{p-r}\Lambda^{\frac{1}{2}}_{p-r})(P_{p-r}\Lambda^{\frac{1}{2}}_{p-r})'$$

from which it follows that $H_r = P_{p-r}\Lambda^{\frac{1}{2}}_{p-r}(-1)^{\frac{1}{2}}$

In practice the matrix Λ_{p-r} may have some zero eigenvalues, some very small eigenvalues. This will cause the generalized inverse estimator to be a conditionally unbiased least squares estimator with respect to the columns of P_{p-r} that correspond to non-zero eigenvalues. This theorem is analogous to Theorem 4.5.7.

5.6 Fractional and modified fractional rank

Consider the model

(5.6.1) $Y^* = Wa+e$, $e \sim N(0,\sigma^2 I)$ where $W = XP'$, $a = P\beta^*$ and P' is the matrix whose columns are the eigenvectors of $X^{*\prime}X^*$. We define a p-component vector θ_i by

(5.6.2) $\hat{\theta}_i = (\hat{a}_1,\hat{a}_2,\ldots,\hat{a}_i,0,\ldots,0)'$

The Marquardt's fractional rank estimator is given by

(5.6.3) $\hat{a}_F = (1-c)\hat{\theta}_r + c\,\hat{\theta}_{r+1}$

where $c \in [0,1]$ and $r(X) \in [r,r+1]$.

Now we introduce a new class of estimators of a by

(5.6.4) $\tilde{a} = \sum_{i=1}^{p} \delta_i \hat{\theta}_i$

where δ_i are to be determined.

An alternative form to (5.6.4) is

(5.6.5) $\tilde{a} = B\hat{a}$

where $$B = \begin{pmatrix} b_1 = \sum_{j=1}^{p} \delta_j & & 0 \\ & & \\ 0 & & b_p = \sum_{j=p}^{p} \delta_j \end{pmatrix}$$

We note that (5.6.4) is well defined because $\hat{\theta}_i$ $i = 1,2,\ldots,p$ are linearly independent vectors.

Assuming B is known the $E(L_1^2)$ for the estimator $\tilde{a}$ in (5.6.5) is given by

$$(5.6.6)\quad E(L_i^2) = E(\tilde{a}-a)'(\tilde{a}-a)$$

$$= E(B\hat{a}-a)'(B\hat{a}-a)$$

$$= E\{\sum_{i=1}^{p} (b_i\hat{a}_i-a_i)^2\}$$

$$= E\{\sum_{i=1}^{p}[(b_i\hat{a}_i-b_ia_i)+(b_ia_i-a_i)]^2\}$$

$$= \sigma^2\sum_{i=1}^{p}\frac{b_i^2}{\lambda_i} + \sum_{i=1}^{p} a_i^2(b_i-1)^2$$

The fractional rank estimator $\tilde{a}_F$ is obtained from (5.6.4) by adjoining the constraints

$$(5.6.7)\quad \delta_r + \delta_{r+1} = 1$$

$$\delta_r \geq 0, \quad \delta_r + 1 \geq 0$$

$$\delta_i = 0 \quad \text{for} \quad i \neq r, r+1$$

Now we will find an estimate for δ_{r+1}; so for a given rank r we minimize $E(L_i^2)$ in (5.6.6) subject to the constraints of (5.6.7). We have that

$$(5.6.8)\quad E(L_i^2) = \sum_{i=1}^{r}\frac{\sigma^2}{\lambda_i} + \frac{\sigma^2}{\lambda_{r+1}}\delta_{r+1}^2 + a_{r+1}^2(\delta_{r+1}-1)^2 + \sum_{i=r+2}^{p} a_i^2$$

because of (5.6.7)

$$(5.6.9)\quad \frac{\partial E(L_i^2)}{\partial \delta_{r+1}} = \frac{2\sigma^2}{\lambda_{r+1}}\delta_{r+1} + 2\,a_{r+1}^2\,(\delta_{r+1}-1) = 0$$

or

$$(5.6.10)\quad \delta_{r+1} = \frac{a_{r+1}^2}{\frac{\sigma^2}{\lambda_{r+1}} + a_{r+1}^2}$$

$$= \frac{a_{r+1}^2\cdot\lambda_{r+1}}{\sigma^2}$$

$$1 + \frac{a_{r+1}^2 \cdot \lambda_{r+1}}{\sigma^2}$$

$$= \frac{\tau_{r+1}^2}{1+\tau_{r+1}^2}$$

where, in general, we use the symbol

$$(5.6.11) \quad \tau_i^2 = \frac{a_i^2 \lambda_i}{\sigma^2} \qquad i = 1,\ldots,p$$

The choice of r is based on the subsequent minimization of the $E(L_1^2)$ criterion with respect to r. This can be achieved by evaluation of (5.6.6) using (5.6.7). With δ_{r+1} defined by (5.6.10) so (5.6.8) using (5.6.10) becomes

$$(5.6.12) \quad \sum_{i=1}^{r} \frac{\sigma^2}{\lambda_i} = \sum_{i=1}^{r} \frac{\sigma^2}{\lambda_i} + \frac{\sigma^2}{\lambda_{r+1}} \frac{\tau_{r+1}^4}{(1+\tau_{r+1}^2)^2} + a_{r+1}^2$$

$$\left(\frac{\tau_{r+1}^2}{1+\tau_{r+1}^2} - 1\right)^2 + \sum_{i=r+2}^{p} a_i^2$$

$$= \sum_{i=1}^{p} \frac{\sigma^2}{\lambda_i} - \frac{\sigma^2}{\lambda_{r+1}} - \sum_{i=r+2}^{p} \frac{\sigma^2}{\lambda_i} + \frac{\sigma^2}{\lambda_{r+1}} \cdot \frac{\tau_{r+1}^4}{(1+\tau_{r+1}^2)^2}$$

$$+ \frac{a_{r+1}^2}{(1+\tau_{r+1}^2)^2} + \sum_{i=r+2}^{p} a_i^2$$

$$= \sum_{i=1}^{p} \frac{\sigma^2}{\lambda_i} - \sigma^2 \left[\sum_{i=r+2}^{p} (1-\tau_i^2)\lambda_i^{-1}\right] - \frac{\sigma^2}{\lambda_{r+1}}$$

$$+ \frac{\sigma^2}{\lambda_{r+1}} \frac{\tau_{r+1}^2}{1+\tau_{r+1}^2}$$

$$= \sum_{i=1}^{p} \frac{\sigma^2}{\lambda_i} - \sigma^2 \left\{\left[\sum_{i=r+2}^{p} (1-\tau_i^2)\lambda_i^{-1}\right] + (1+\tau_{r+1}^2)^{-1}\lambda_{r+1}^{-1}\right\}$$

Inspection of (5.6.12) reveals that a reduction in rank of more than one is seem to be possible if $\tau_i^2 < 1$ for some i. This is not sufficient since the ordering by magnitude of the τ_i^2 is different from the ordering on the λ_i. We note that a fractional reduction in rank is always possible since for $r = p-1$ the first term in brackets defined to be zero but the second term is assured positive. In this case the rank is given by $p-1+\delta_p$.

Before we give the procedure of estimating the integral part r and fractional part δ_{r+1} we prove the following theorem.

Theorem 5.6.1

The sequence defined by

(5.6.13) $x_{n+1} = x_n^2/(\chi_n^2+L)$

has three points of accumulation depending on L and the initial point x_0 if x^* denotes the limiting point then the following situations are possible.

1. If $L > \frac{1}{4}$ then $x^* = 0$
2. if $L \leq \frac{1}{4}$ and
   (a) $x_0 > x_2$ then $x^* = c_1$
   (b) $x_0 < x_2$ then $x^* = 0$
   (c) $x_0 = x_2$ then $x^* = c_2$

where $c_1 = \frac{1}{2} + (\frac{1}{4}-L)^{\frac{1}{2}}$

$c_2 = \frac{1}{2} - (\frac{1}{4}-L)^{\frac{1}{2}}$

Proof

If (5.6.13) converges to x^* then

$$x^* = \frac{x^{*2}}{x^{*2}+L} \tag{5.6.14}$$

or

$$x^*(x^{*2}-x^*+L) = 0 \tag{5.6.15}$$

From (5.6.15) we conclude that possible accumulation points are

$$x_1^* = 0$$

$$x_2^* = \tfrac{1}{2} + (\tfrac{1}{4}-L)^{\frac{1}{2}}$$

$$x_3^* = \tfrac{1}{2} - (\tfrac{1}{4}-L)^{\frac{1}{2}}$$

Case 1 $L > \frac{1}{4}$

The only possibility is $x^* = 0$ because the sequence is monotone decreasing and bounded below by zero.

Case 2 $L \leq \frac{1}{4}$

(i) If $x_0 \geq \frac{1}{2} + (\frac{1}{4}-L)^{\frac{1}{2}}$ then the sequence is monotone decreasing and bounded below by $\frac{1}{2} + (\frac{1}{4}-L)^{\frac{1}{2}}$ if $\frac{1}{2} - (\frac{1}{4}-L)^{\frac{1}{2}} < x_0 < \frac{1}{2} + (\frac{1}{4}-L)^{\frac{1}{2}}$ then the sequence is monotone increasing and bounded above by $\frac{1}{2} + (\frac{1}{4}-L)^{\frac{1}{2}}$. In either case it converges to $\frac{1}{2} + (\frac{1}{4}-L)^{\frac{1}{2}}$.

(ii) $x_0 < \frac{1}{2} - (\frac{1}{4}-L)^{\frac{1}{2}}$ then the sequence converges to zero.

(iii) If $x_0 = \frac{1}{2} - (\frac{1}{4}-L)^{\frac{1}{2}}$ then $x_n = \frac{1}{2} - (\frac{1}{4}-L)^{\frac{1}{2}}$ for all n.

Estimation of r and δ_{r+1}

The initial estimate of r can be made by inspection of (5.6.12). Specifically, we set $\lambda_i = 0$, $i = r+2,\ldots,p$ where r is determined by the maximum of

(5.6.16) $\sum_{i=r+2}^{p}(1-\tau_i^2)/\lambda_i$ where $\tau_i^2 = \dfrac{\hat{a}_i^2\lambda_i}{\hat{\sigma}^2}$

with respect to r if (5.6.16) is positive. Otherwise $r = p-1$

Using Theorem 5.6.1 with $x_{n+1}=(\delta_{r+1})_{n+1}$ $L = (\hat{\tau}_{r+1}^2)^{-1} = \dfrac{\hat{\sigma}^2}{\hat{a}_{r+1}^2\lambda_{r+1}}$

and $x_0 = (1+L)^{-1}$ we have that

1. $\delta_{r+1}^* = 0$ if $\tau_{r+1}^2 < 4$

2. $\delta_{r+1}^* = \frac{1}{2} + \left(\frac{1}{4} - \dfrac{\hat{\sigma}^2}{\hat{a}_{r+1}^2\lambda_{r+1}}\right)^{\frac{1}{2}}$ if $\tau_{r+1}^2 \geq 4$

For this choice of r, the fractional rank estimator is obtained from (5.6.3) with $c = \delta^*$.

It is possible that a re-evaluation of the criterion with $a = \hat{a}_F$ may suggest a different value for r. So we evaluate (5.6.16) with $\hat{\tau}_i^2$ computed from $\hat{a}_F$ rather than $\hat{a}$. It is possible that a further reduction can be obtained by iterating on this procedure. An increase in rank is not possible.

Modified fractional rank

If the variables are labeled such that

$$\hat{\tau}_1^2 \geq \hat{\tau}_2^2 \geq \cdots \geq \hat{\tau}_p^2$$

The determination of r is simpler. That is, we set to zero all λ_i for which $\tau_i^2 \leq 1$. Note that these need not be the smallest eigenvalues.

The determination of δ_{r+1} and the iterative determination of r proceed as in the ordinary fractional procedure.

5.7 The method of Greenberg (1975)

The author interprets the $\hat{\beta}^*_{G.I}$ of (5.5.1) as an estimate of β^* under the set of constraints.

(5.7.1) $v_j'\beta^* = 0 \quad j = r+1,\ldots,p$ where v_j is the eigenvector of $(X^{*\prime}X^*)$ corresponding to λ_j.

The constraints can be tested using the fact that

$$u = \left(\sum_{i=r+1}^{p} \lambda_i (\hat{\delta}_i)^2\right) \Big/ \left(1 - \sum_{i=1}^{p} \lambda_i (\hat{\delta}_i)^2\right)\left(\frac{n-p-1}{p-r}\right)$$

is distributed as $F_{p-r,n-p-1}$.

Massy's two criteria suggest a tradeoff: Dropping components with small latent roots will reduce variance, but including components is likely to reduce bias. As we have seen in Chapter 3, if we compare restricted and unrestricted estimators based on M.S.E., we will have a tradeoff of bias and variance. For the tests of Table A, page 62, u, as defined above, should be

referred to the tables, which are mentioned there.

5.8 Latent root regression analysis

In what follows we assume Model A_s (see (1.2.5)).

Let $Z = (X^*, Y^*)$ and let

$$S = Z'Z = \begin{pmatrix} X^{*\prime}X^* & X^{*\prime}Y^* \\ Y^{*\prime}X^* & Y^{*\prime}Y^* \end{pmatrix}$$

The latent roots and latent vectors can be found by solving

(5.8.1) $|S-\lambda_j^* I| = 0$ and $(S-\lambda_j^* I)\gamma_j = 0$, $j = 1,2,\ldots,p+1$

respectively. Here we suppose $\lambda_1^* \geq \lambda_2^* \geq \ldots \geq \lambda_p^* \geq \lambda_{p+1}^* \geq 0$.

Denote the elements of the jth latent vector by $\gamma_j = (\gamma_{1j},\gamma_{2j},\ldots,\gamma_{p+1j})'$ and let $\gamma_j^{p+1} = (\gamma_{1j},\gamma_{2j},\ldots,\gamma_{pj})'$ i.e. γ_j^{p+1} contains all the elements of the γ_j except the last one.

Finally let $\Gamma = (\gamma_1,\gamma_2,\ldots,\gamma_{p+1})$ and $\Lambda^* = \mathrm{diag}(\lambda_1^*,\lambda_2^*,\ldots,\lambda_{p+1}^*)$. Then $\Gamma'S\Gamma = \Gamma'Z'Z\Gamma = \Lambda^*$ so $S = \Gamma\Lambda^*\Gamma'$. We note that

$$(5.8.2)\quad Z\gamma_j = (X^*;Y^*)\gamma_j$$

$$= (X^*,Y^*)\begin{pmatrix}\gamma_{1j}\\ \gamma_{2j}\\ \vdots\\ \gamma_{p+1j}\end{pmatrix}$$

$$= \begin{pmatrix}X^*_{11}\gamma_{1j}+\ldots+X^*_{1p}\gamma_{pj}+Y^*_1\gamma_{p+1j}\\ \vdots\\ X^*_{n1}\gamma_{1j}+\ldots+X^*_{np}\gamma_{pj}+Y^*_n\gamma_{p+1j}\end{pmatrix}$$

$$= \begin{pmatrix}\sum_{r=1}^{p}X^*_{1r}\gamma_{rj}+Y^*_1\gamma_{p+1j}\\ \vdots\\ \sum_{r=1}^{p}X^*_{nr}\gamma_{rj}+Y^*_n\gamma_{p+1j}\end{pmatrix}$$

But we have

$$(5.8.3)\quad \lambda^*_j = \gamma_j'S\gamma_j$$

$$= \gamma_j'Z'Z\gamma_j$$

$$= \sum_{i=1}^{n}\left(\sum_{r=1}^{p}X^*_{ir}\gamma_{rj}+Y^*_i\gamma_{p+1j}\right)^2$$

so if

(5.8.4) $\lambda_j^* = 0$

- $\gamma_{p+1j} \neq 0$: Then $\sum_{r=1}^{p} X^*_{ir}\gamma_{rj} + Y^*_i\gamma_{p+1j} = 0$ and $Y^*_i = -\gamma_{p+1j}^{-1}\sum_{r=1}^{p} X^*_{ir}\gamma_{rj}$ for all $i=1,2,\ldots,n$. So we have a perfect predictor for Y^*_i as given above
- $\gamma_{p+1j} = 0$: Then $\sum_{r=1}^{p} X^*_{ir}\gamma_{rj} = 0$, so an exact linear dependence exists between the columns of X^* which implies multicollinearity

From (5.8.2) one can define (providing $\gamma_{p+1j} \neq 0$)

$$(5.8.5)\quad \hat{Y}^*_j = \begin{pmatrix} \hat{Y}^*_{1j} \\ \vdots \\ \hat{Y}^*_{nj} \end{pmatrix} = -\gamma_{p+1j}^{-1} \begin{pmatrix} X^*_{11} & \cdots & X^*_{1p} \\ \vdots & & \\ X^*_{n1} & \cdots & X^*_{np} \end{pmatrix} \begin{pmatrix} \gamma_{1j} \\ \vdots \\ \gamma_{pj} \end{pmatrix}$$

$$= X^*\left(-\gamma_{p+1j}^{-1}\right)\gamma_j^{p+1}$$

$$= X^*\tilde{\gamma}$$

where $\tilde{\gamma} = -\gamma_{p+1j}^{-1}\gamma_j^{p+1}$ for $j = 1,2,\ldots,p+1$

In general none of the roots will be zero but some might be quite small.

If the jth prediction equation is used alone to predict values of the dependent variable, one can see easily that the

R.S.S. is given by

$$(5.8.6)\quad (Y^*-\hat{Y}^*_j)'(Y^*-\hat{Y}^*_j) = \sum_{i=1}^{n}(Y^*_i-\hat{Y}^*_{ij})^2$$

$$= \sum_{i=1}^{n}(Y^*_i+\gamma_{p+1j}^{-1}\sum_{r=1}^{p}X^*_{ir}\gamma_{rj})^2$$

$$= \frac{1}{\gamma^2_{p+1j}}\gamma_j' S\gamma_j \quad \text{| see (5.8.3)}$$

$$= \frac{\lambda^*_j}{\gamma^2_{p+1j}}$$

Normally none of the individual prediction equations in (5.8.5) will by itself be a good predictor. So we take the linear combination of these predictors

$$(5.8.7)\quad Y^* = \sum_{j=1}^{p+1}\alpha_j\gamma_{p+1j}\hat{Y}^*_j$$

and we impose the following restriction

$$\sum_{j=1}^{p+1}\alpha_j\gamma_{p+1j} = 1.$$

We have from (5.8.7)

$$(5.8.8)\quad Y^* = -\sum_{j=1}^{p+1}\alpha_j\gamma_{p+1j}\gamma_{p+1j}^{-1}X^*\gamma_j^{p+1}$$

$$= -X^*\sum_{j=1}^{p+1}\alpha_j\gamma_j^{p+1}$$

$$= X^*\bar{\gamma}$$

where $$\bar{\gamma} = -\sum_{j=1}^{p+1}\alpha_j\gamma_j^{p+1}$$

The R.S.S. using this predictor is given by

$$(5.8.9) \quad \text{R.S.S.} = (Y^*-\hat{Y}^*)'(Y^*-\hat{Y}^*)$$

$$= \sum_{j=1}^{p+1} \alpha_j^2 \lambda_j^*$$

$$= \alpha' \Lambda^* \alpha$$

Least squares and modified least square solution

If α is chosen to minimize the R.S.S. in (5.8.9), then we shall have the O.L.S. predictor. Thus we wish to minimize

$$(5.8.10) \quad F(\alpha) = \sum_{j=1}^{p+1} \alpha_j^2 \lambda_j^* - 2\mu_0\left(\sum_{j=1}^{p+1} \gamma_{p+1j}\, \alpha_j - 1\right)$$

where $-2\mu_0$ is the Lagrangian multiplier.

Now we have

$$(5.8.11) \quad \frac{\partial F(\alpha)}{\partial \alpha_j} = 2\alpha_j \lambda_j^* - 2\mu_0 \gamma_{p+1j} = 0$$

or

$$(5.8.12) \quad \alpha_j = \frac{\mu_0 \gamma_{p+1j}}{\lambda_j^*} \qquad j = 1,2,\ldots,p+1$$

But $\sum_{j=1}^{p+1} \gamma_{p+1j}\, \alpha_j = 1$ so

$$\sum_{j=1}^{p+1} \gamma_{p+1j} \frac{\mu_0 \gamma_{p+1j}}{\lambda_j^*} = 1$$

or

$$(5.8.13) \quad \mu_0 = \frac{1}{\sum_{j=1}^{p+1} (\gamma_{p+1j}^2/\lambda_j^*)}$$

From (5.8.8), (5.8.12) and (5.8.13) we have

$$(5.8.14) \quad \hat{\beta}^* = \hat{\gamma} = -\sum_{j=1}^{p+1} \alpha_j \gamma_j^{p+1}$$

$$= -\sum_{j=1}^{p+1} \frac{\gamma_{p+1j}}{\lambda_j^*} \gamma_j^{p+1} \left(\sum_{j=1}^{p+1} \frac{\gamma_{p+1j}^2}{\lambda_j^*}\right)^{-1}$$

$$= -\frac{1}{\left(\sum_{j=1}^{p+1} \frac{\gamma_{p+1j}^2}{\lambda_j^*}\right)} \cdot \sum_{j=1}^{p+1} \frac{\gamma_{p+1j}}{\lambda_j^*} \begin{pmatrix} \gamma_{1j} \\ \vdots \\ \gamma_{pj} \end{pmatrix}$$

So the ith element of $\hat{\beta}^*$ is given by

$$(5.8.15) \quad \hat{\beta}_i^* = -\frac{1}{\left(\sum_{j=1}^{p+1} \frac{\gamma_{p+1j}^2}{\lambda_j^*}\right)} \sum_{j=1}^{p+1} \left(\frac{\gamma_{p+1j}\gamma_{ij}}{\lambda_j^*}\right)$$

for all $i = 1,2,\ldots,p$.

Now (5.8.9) can be written as follows

$$(5.8.16) \quad R.S.S.(\hat{\beta}^*) = \sum_{j=1}^{p+1} \alpha_j^2 \lambda_j^*$$

$$= \left(\sum_{j=1}^{p+1} \frac{\gamma_{p+1j}^2}{\lambda_j^{*2}} \lambda_j^*\right)\left(\sum_{j=1}^{p+1} \gamma_{p+1j}^2/\lambda_j^*\right)^{-2}$$

$$= \left(\sum_{j=1}^{p+1} \gamma_{p+1j}^2/\lambda_j^*\right)^{-1}$$

$$= \mu_0$$

Suppose that the latent vectors $\gamma_{q+1},\gamma_{q+2},\ldots,\gamma_{p+1}$ correspond to non-predictive near singularities.

The above least squares estimations can be adjusted by setting $\alpha_{q+1},\alpha_{q+2},\ldots,\alpha_{p+1}$ to zero.

Then minimizing (5.8.10) we have

$$(5.8.17)\quad \alpha_j = \gamma_{p+1j}\lambda_j^{*-1}\mu_0 \qquad j = 1,2,\ldots,q$$

where $$\mu_0 = \left(\sum_{j=1}^{q} \gamma_{p+1j}^2/\lambda_j^*\right)^{-1}$$

The modified least squares coefficients are given by

$$(5.8.18)\quad \hat{\beta}_{LR}^* = -\left(\sum_{j=1}^{q} \gamma_{p+1j}^2/\lambda_j^*\right)^{-1}\sum_{j=1}^{q}\frac{\gamma_{p+1j}\gamma_j^{p+1}}{\lambda_j^*}$$

while the R.S.S. is given by

$$(5.8.19)\quad \text{R.S.S.}(\hat{\beta}_{LR}^*) = \mu_0 = \left(\sum_{j=1}^{q} \gamma_{p+1j}^2/\lambda_j^*\right)^{-1}$$

Note that if $X^{*\prime}X^*$ is singular, the same procedure as above can be applied and the minimization of (5.8.10) will yield results identical to (5.8.18) and (5.8.19) above.

This follows from the fact that a singular $X^{*\prime}X^*$ implies some of the λ_j^* and the corresponding γ_{p+1j} of S will be zero, which is equivalent to setting the appropriate α_j in (5.8.10) to zero. Thus X^* need not be of full rank to obtain estimates by this procedure.

The estimates obtained from equations (5.8.14) and (5.8.18) are often very different. This is obvious, since some of the λ_j^*'s are close to zero, which would inflate $\hat{\beta}^*$. Another reason is that the α_j corresponding to vectors of non-predictive near singularities are often very large relative to the remaining

α_j. When this occurs the terms $\alpha_j \gamma_j^{p+1}$ $j = q,q+1,\ldots,p$ can dominate $\hat{\beta}^*$. Removing these dominating terms will yield more accurate estimates of the parameter β^*.

Another measure of the effect of the modified procedure is the R.S.S. By definition (5.8.19) is always larger than (5.8.16), i.e. $\text{R.S.S.}(\hat{\beta}^*_{LR}) > \text{R.S.S.}(\hat{\beta}^*)$.

If one is actually removing non-predictive near singularities from the estimates then (5.8.19) should not differ drastically from (5.8.16).

Alternatively the estimates of σ^2 using (5.8.16) and (5.8.19) should be reasonably close.

Geometric picture

One problem which must not be overlooked when discussing near singularities of $S = Z'Z$ is whether these near singularities contain information about the underlying Model A_s. In order to answer this question, consider $(X^*_{i1},\ldots,X^*_{ip},Y^*_i)$ $i = 1,2,\ldots,n$ as n points in $p+1$-dimensions. Euclidean space defined by a mutually orthogonal axes $X^*_1,\ldots,X^*_p,Y^*$.

The latent vector of $S = Z'Z$ define a second set of mutually orthogonal axes $Z_1,\ldots,Z_p,Z_{p+1}$ where Z_j is the axis defined by γ_j.

The direction axis Z_j relative to original axes is given by the vector sum

(5.8.20) $\sum_{r=1}^{p+1} \gamma_{rj} \vec{e}_r$

where $\vec{e}_1,\ldots,\vec{e}_p,\vec{e}_{p+1}$ are vectors from origin in the directions of axes $X_1^*,\ldots,X_p^*,Y^*$.

The last element of γ_j, γ_{p+1j}, represent the cosine of the angle between Y^* and Z_j, while γ_{rj}, $r = 1,\ldots,p$ represents the cosine of angles between axes X_r^* and Z_j. Assuming the latent vectors are normalized, γ_j is uniquely determined apart from a multiple of -1.

The latent root corresponding to a particular latent vector measures the spread of the n-data points in the direction defined by the latent vector. In other words, λ_j^* is the sums of squares of the projections of the n-data points on the Z_j axis. A small value of λ_j^* indicates that there is little variation in the Z_j direction, i.e. $Z_{ij} = Y_i^*\gamma_{p+1j} + \sum_{r=1}^{p} X_{ir}^*\gamma_{rj}$ is near zero for $i = 1,\ldots,n$.

If γ_{p+1j} is near zero, the axis Z_j is nearly orthogonal to the Y^* axis. Hence if both λ_j^* and γ_{p+1j} are small the latent vector γ_j reveals a non-predictive near singularity.

The least squares estimator is a linear combination of all $p+1$ latent vectors, including latent vectors corresponding to

non-predictive near singularities. The modified least squares estimation utilizes only linear combinations not having both λ_j and γ_{p+1j} small.

We give the following figures for the case in which $p+1 = 3$.

In the Figure 5.1 $\vec{e}_1'$, $\vec{e}_2'$, $\vec{e}_3'$ are the unit vectors from the origin in the direction of axes Z_1, Z_2, Z_3. While $\vec{e}_1, \vec{e}_2, \vec{e}_3$ are also unit vectors, but from the origin in the direction of axes X_1^*, X_2^*, Y^*.

The following relation holds between $\vec{e}_i'$'s and $\vec{e}_i$'s

$$\begin{pmatrix} \vec{e}_1' \\ \vec{e}_2' \\ \vec{e}_3' \end{pmatrix} = \begin{pmatrix} \gamma_{11} & \gamma_{21} & \gamma_{31} \\ \gamma_{12} & \gamma_{22} & \gamma_{32} \\ \gamma_{13} & \gamma_{23} & \gamma_{33} \end{pmatrix} \begin{pmatrix} \vec{e}_1 \\ \vec{e}_2 \\ \vec{e}_3 \end{pmatrix}$$

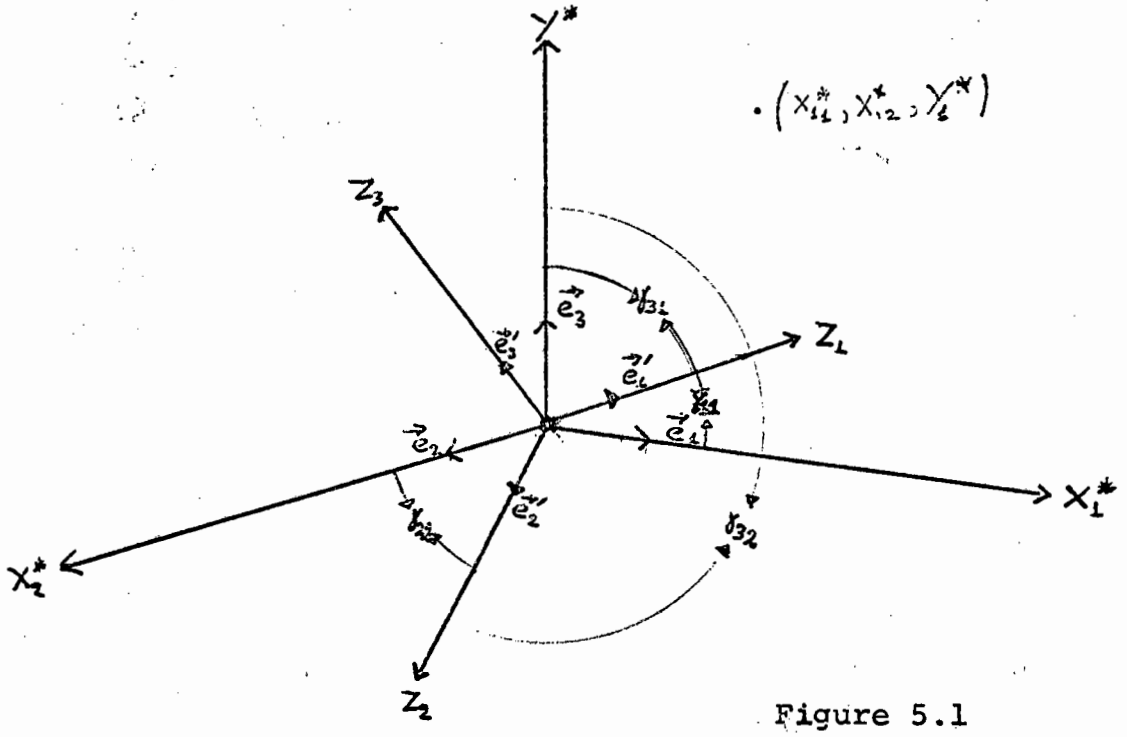


Figure 5.1

C H A P T E R 6

COMPARISONS AND CONCLUSIONS

6.1 Introduction

In this chapter we provide a comparison of the estimators by comparing the optimal value of the criterion used to determine the coefficients. It should be noted that these comparisons are made on the assumption that the optimal values of the coefficients are known as opposed to the more practical situation when they must be estimated from the data. Finally, we give our conclusions, suggestions and problems for further study.

6.2 Comparing the ridge estimator and the generalised ridge estimator

In order to prove that the generalized ridge estimator is better than the ordinary ridge estimator, we must show that

(6.2.1) $E(L_1^2(k)) - \Omega > 0$

where (a) $E(L_1^2(k))$ as in (4.5.6) with $\gamma_2(k)$ as in (4.5.13)

(b) Ω as in (4.11.3)

(c) $k_i = \frac{\sigma^2}{a_i^2}$

From (6.2.1) we have

$$(6.2.2)\quad \sum_{i=1}^{p}\left[\frac{\sigma^2\lambda_i+k^2a_i^2}{(\lambda_i+k)^2}-\frac{\sigma^2\lambda_i+k_i^2a_i^2}{(\lambda_i+k_i)^2}\right]=\sum_{i=1}^{p}\left[\frac{\sigma^2\lambda_i+k^2a_i^2}{(\lambda_i+k)^2}-\frac{\sigma^2a_i^2}{(\lambda_ia_i^2+\sigma^2)}\right]$$

$$=\sum_{i=1}^{p}\frac{(\sigma^2-ka_i^2)^2\lambda_i}{(\lambda_ia_i^2+\sigma^2)(\lambda_i+k_i)^2}$$

We observe that the last relation of (6.2.2) is positive, because $a_i^2,\ \lambda_i > 0$ for all i, $k > 0$, $\sigma^2 > 0$.

6.3 Comparing the O.L.S. estimator and the L.R.R.A.

If a single multicollinearity exists among the columns of X^* there is at least one latent root, λ^*_{p+1}, which is small. It is easy to verify that $\lambda^*_{p+1} \leq \lambda_p$, where λ_p is the smallest root of $X^{*\prime}X^*$.

As we have seen in Section 5.8, if both $|\gamma_{p+1p+1}|$ and λ^*_{p+1} are near zero the multicollinearity is providing little, if any, information about relationships among the response and regressor variables. In this case the elements of γ^{p+1}_{p+1} are very close to the corresponding elements of v_p, the latent vector corresponding to the latent root λ_p. In addition $v_p'v_p \simeq \gamma^{p+1\prime}_{p+1}\gamma^{p+1}_{p+1}$.

We note that if $X^*_{q+1},\ldots,X^*_p$ are such that

$$(6.3.1)\quad \sum_{i=q+1}^{p} c_iX_i^* \simeq 0$$

then the c_i are the appropriate elements of v_p and if $|\gamma_{p+1p+1}|$ and λ^*_{p+1} are both near zero, so do the corresponding

element of γ_{p+1}^{p+1}

The major difference between O.L.S. and L.R.R.A. estimator is the term containing γ_{p+1}^{p+1}. So the O.L.S. estimator tends to be dominated by the v_p term while L.R.R.A. estimator is a linear combination of vectors essentially orthogonal to v_p.

Hence we would expect O.L.S. to estimate β^* more accurately than L.R.R.A. when the $p-q$ coefficients of β^* corresponding to the elements in the multicollinearity (6.3.1) are a constant multiple of the same element in v_p with the constant being large in magnitude, i.e. $\beta_j^* = cv_{jp}$ for the appropriate $p-q$ elements of β^* and v_p. The L.R.R.A. estimator should perform poorly in this case since no linear combination of $\gamma_1^{p+1},\gamma_2^{p+2},\ldots,\gamma_p^{p+1}$ can yield v_p or a constant multiple of v_p.

The coefficient of the variables not involved in multicollinearity can be estimated well by both procedures, because the subspace of these variables is spanned by the vectors $\gamma_1^{p+1},\gamma_2^{p+1},\ldots,\gamma_p^{p+1}$

Let $\beta_{(q+1)}^* = (\beta_{q+1}^*,\beta_{q+2}^*,\ldots,\beta_p^*)'$ be the subvector of β^* whose elements correspond to the variables involved in the multicollinearity in (6.3.1), similarly define $v_{(q)}$ of v_p.

Then if $\beta_{(q)}^* = cv_{(q)}$ and $|c|$ is large, O.L.S. should provide better estimates than L.R.R.A. On the other hand if $\beta_{(q)}^{*\prime}v_{(q)} = 0$ or $|c|$ is small, L.R.R.A. should perform better

than O.L.S. If neither of the above conditions hold L.R.R.A. should still outperform O.L.S. because of the dominance of the v_p term in $\hat{\beta}^*$.

We note that

$$(6.3.2) \quad M.S.E.(\hat{\beta}_{LR}) \simeq \sigma^2 \sum_{i=1}^{p-1} \lambda_i^{-1} + (v_1'\beta^*)^2$$

while

$$(6.3.3) \quad M.S.E.(\hat{\beta}^*) = \sigma^2 \sum_{i=1}^{p} \lambda_i^{-1}$$

We can see that (6.3.3) can become very large if the multicollinearity is strong. On the other hand, the ill effects of multicollinearity are eliminated from the variance term of (6.3.2), but the bias term has been added.

This bias term can be expected to be large if $\beta^*_{(q)} = cv_{(q)}$ and $|c|$ is large; otherwise this bias term should be small and hence $M.S.E.(\hat{\beta}_{LR}) < M.S.E.(\hat{\beta}^*)$.

6.4 Comparing the fractional rank and the generalized ridge estimator

The mean square error of the fractional rank estimator is given by (5.6.8) with

$$\delta_{r+1} = \frac{\lambda_{r+1} a_{r+1}^2}{\sigma^2 + \lambda_{r+1} a_{r+1}^2}$$

The mean square error of the generalized estimator is given by (4.11.3) with $k_i = \frac{\sigma^2}{a_i^2}$. In order to compare the above estimators we take the difference between the mean square errors.

So we have

$$(6.4.1)\qquad (E(L_1^2))_{FR} - (E(L_1^2))_{G.R} = \sum_{i=1}^{r} \frac{\sigma^2}{\lambda_i} + \frac{\sigma^2\lambda_{r+1}a_{r+1}^4}{(\sigma^2+\lambda_{r+1}a_{r+1}^2)^2} +$$

$$+ \frac{a_{r+1}^2\sigma^4}{(\sigma^2+\lambda_{r+1}a_{r+1}^2)^2} + \sum_{i=r+2}^{p} a_i^2$$

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

$$= \sum_{i=1}^{r}\left(\frac{\sigma^2}{\lambda_i} - \frac{\sigma^2 a_i^2}{\lambda_i a_i^2+\sigma^2}\right) + \frac{\sigma^2\lambda_{r+1}a_{r+1}^4+a_{r+1}^2\sigma^4}{(\sigma^2+\lambda_{r+1}a_{r+1}^2)^2}$$

$$- \frac{\sigma^2 a_{r+1}^2}{\lambda_{r+1}a_{r+1}^2+\sigma^2}$$

$$+ \sum_{i=r+2}^{p}\left(a_i^2 - \frac{\sigma^2 a_i^2}{\lambda_i a_i^2+\sigma^2}\right)$$

$$= \sum_{i=1}^{r} \frac{\sigma^4}{\lambda_i(\lambda_i a_i^2+\sigma^2)} + \sum_{i=r+2}^{p} \frac{a_i^4\lambda_i}{\lambda_i a_i^2+\sigma^2} > 0$$

We observe from the last relation that

$$(E(L_1^2))_{GR} < (E(L_1^2))_{FR}$$

6.5 A geometric portrayal of some biased estimators when the predictors are two

Consider the Model A_s as in (1.2.5) with $X^{*\prime}X^* = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$,

and the model (4.11.1) with $\Lambda = \text{diag}(\lambda_1 = 1+\rho, \lambda_2 = 1-\rho)$ and $P' = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$. If we take $\rho = 1$ we have the following figure (6.1).

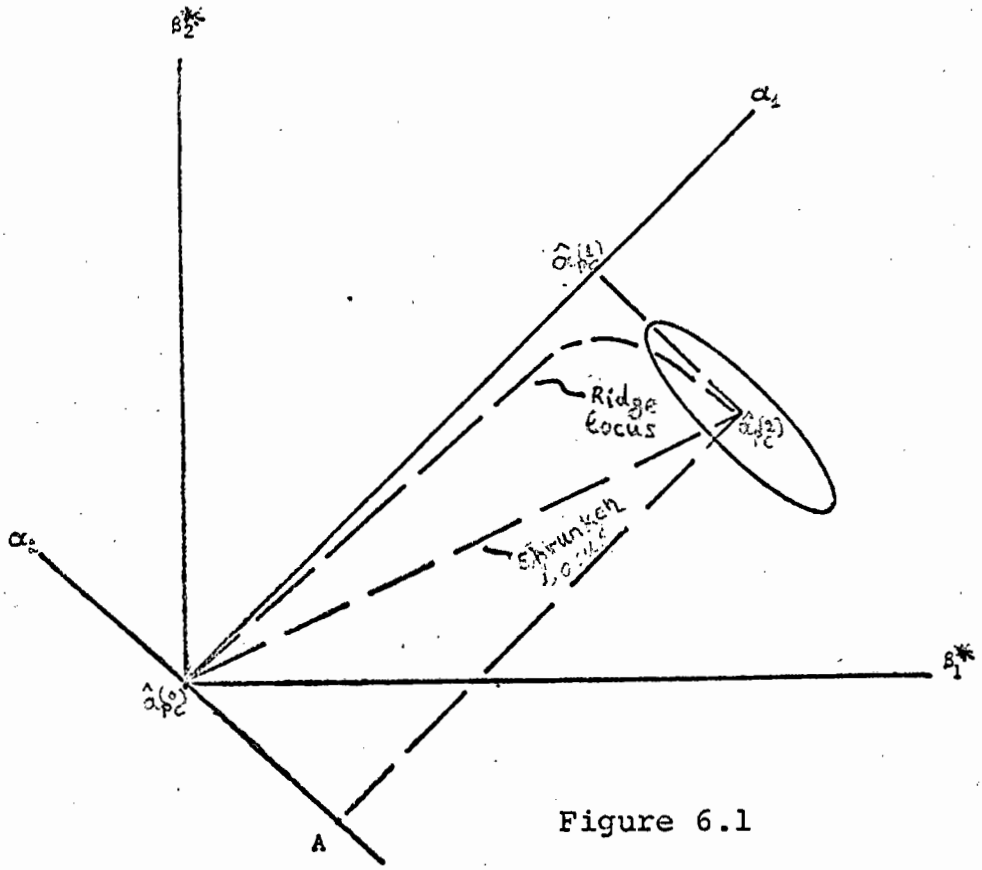


Figure 6.1

(1) The principal component estimates

$\hat{a}_{pc}(0) = (0,0)$, $\hat{a}_{pc}(1) = (\hat{a}_1, 0)$, $\hat{a}_{pc}(2) = (\hat{a}_1, \hat{a}_2)$ are indicated.

(2) The fractional rank estimator for $1 \leq \text{rank} \leq 2$ lie on the line segment $[\hat{a}_{pc}(1), \hat{a}_{pc}(2)]$ and for $0 \leq \text{rank} \leq 1$ it lies on the line segment $[\hat{a}_{pc}(0), \hat{a}_{pc}(1)]$.

(3) The ridge estimator lies on the curved line joining $\hat{a}_{pc}(0)$ and $\hat{a}_{pc}(2)$. In other words the simple ridge estimator will lie in the triangle formed by $\hat{a}_{pc}(0)$, $\hat{a}_{pc}(1)$ and $\hat{a}_{pc}(2)$, with the corresponding result for high dimensions.

(4) The generalized estimator may lie anywhere in the rectangle defined by $\hat{a}_{pc}(i)$, i=0,1,2,and A, we note that $\frac{\lambda_i}{\lambda_i+k_i} \in [0,1]$. If $p > 2$ it will lie in the rectangular prism defined by $\hat{a}_{pc}(i)$ for $i = 0,1,\ldots,p$ and A.

(5) The modified fractional rank may lie along the outer boundary of the $\hat{a}_{pc}(0)$, A, $\hat{a}_{pc}(2)$ triangle.

(6) The shrunken estimator lies on the line segment $[\hat{a}_{pc}(0), \hat{a}_{pc}(2)]$.

This simple picture provides a good indication of the ability of the estimators to adjust to the particular correlation structure. The ridge and fractional rank are more flexible than the principal component estimator. But the generalized ridge estimator is more flexible than the ordinary ridge estimator.

6.6 Conclusions

The question which arises which we would like to answer is: "when correlation, when regression?"

Since regression answers a broader and more interesting set

of questions (and some correlation questions as well), it becomes the preferred technique; correlation is useful primarily as an aid to understanding it, and as an auxiliary tool.

The theory of regression as we know is concerned with the prediction of one or more variables on the basis of information provided by other measurements. Prediction is needed in several practical situations, i.e. a meteorologist wants to forecast weather several hours ahead on the basis of suitable atmospheric measurements taken at a point of time.

But before we will be able to apply regression we must be sure that we have collected "good data". The assumption of "good data" includes the usual linear model assumptions such as homogeneity of variance, etc.

Residual plots may suggest transformations and may also reveal outliers. A serious problem which is included under this heading is that of multicollinearity among the regressors. Here we come to the heart of the problem of multicollinearity.

The problem which arises now is how to define multicollinearty? Some definitions were given in Section 2.1, but other authors define multicollinearity as the situation in which the regressors, although linearly independent, are not necessarily orthogonal.

We do not have a unified definition becuase too little

attention has so far been given to this problem. A considerable amount of work is required in order to cast more light on the problem.

Another problem which faces the analyst is how to overcome the situation in which autocorrelation, erractic data and multicollinearity occur concurrently.

This problem is being investigated and will be reported on at a later stage.

We suggest as a measure of multicollinearity the difference $\sum_{i=1}^{p} \frac{1}{\lambda_i} - \sum_{i=1}^{p} \lambda_i$ where λ_i are the latent roots of the $X^{*\prime}X^{*}$.

We also suggest a procedure for identification of the variables, which are involved in the multicollinearity.

Step 1

Calculate $R^2_{x_1^*.x_2^*}$, $R^2_{x_1^*.x_2^*,x_3^*}$,...,$R^2_{x_1^*.x_2^*,\ldots,x_p^*}$ and find which of the R^2 are statistically significant, and hence which variables are linearly related.

Step 2

If none of the R^2 are significant, repeat Step 1, but now instead of X_1^* use X_2^*, instead of X_2^* use X_3^* etc.

In ridge regression we have a lot of criteria for selecting k but no method has shown to be uniformly better than any other, and none to be uniformly better than least squares. We want to point out that although the theoretical comparisons are in favour of generalized ridge, it does not mean that it will happen in practice.

Now we ask the question: "when should we apply ridge regression or principal components or any other estimator?" To this question there is no answer. We only give a rule which indicates when we should not apply ridge regression.

Rule: If r_{Y^*} denote a p-vector of correlation coefficients r_{Y^*,W_i} between the normalized coordinates of X^* along the ith principal axis and Y^*. We do not do ridge regression whenever $|r_{Y^*,W_p}| \geq |r_{Y^*,W_{p-1}}| \geq \ldots \geq |r_{Y^*,W_1}|$.

In what follows we give some steps which could be useful.

Step 1

Find out if any of the usual linear assumptions are violated.

Step 2

Find out if there exists multicollinearity using the proposed methods in Section 2.6, moreover trace the source of multicollinearity.

Step 3

Based on Step 2 use the proper method for solving the problem.

Generally, before we choose any biased estimator, firstly we must look at the loss or decrease in R^2 of the particular estimator relative to the least squares estimator. Secondly using a subset of the original observations evaluate which biased estimator predicts the Y well.

Finally we note that several programs have been developed for the different methods, but are not included since the whole series of programs will be included in a regression package. We also intend to do a complete simulation study in the near future.

BIBLIOGRAPHY

BOOKS

1. ANDERSON, T.W. (1958): An Introduction to Multivariate Statistical Analysis. Wiley.

2. CRAMER, H. (1946): Mathematical Methods of Statistics. Princeton University Press.

3. DRAPER, N.R. and SMITH, H. (1966): Applied Regression Analysis. Wiley.

4. de WAAL, D.J. (1975): Parametric Multivariate Analysis, Part 1. University of the Orange Free State.

5. GRAYBILL, F.A. (1961): An Introduction to Linear Statistical Models, Vol. 1. McGraw-Hill.

6. JOHNSTON, J. (1973): Econometric Methods, 2nd edition. McGraw-Hill.

7. KSHIRAGAR, A.M. (1972): Multivariate Analysis. Marcel Dekker Inc.

8. MORRISON, D.F. (1967): Multivariate Statistical Methods. McGraw-Hill.

9. PRESS, S.J. (1972): Applied Multivariate Analysis. Holt, Tinehart and Winston.

10. RAO, C.R. (1973): Linear Statistical Inference and its Applications, 2nd edition. Wiley.

11. THEIL, H. (1971): Principles of Econometrics. Wiley.

JOURNALS

1. ALLEN, D.M. (1971): Mean square error criterion as a criterion for selecting variables. Techonometrics, 13, 469-465.

2. ALLEN, D.M. (1974): The relationship between variable selection and data augmentation and a method of prediction. Techometrics, 16, 125-126.

3. ANDERSON, T.W. (1963): Asymptotic theory for principal component analysis. Ann.Math.Statist., 34, 122-148.

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

5. BARTLETT, M.S. (1933): On the theory of Statistical Regression. Proceedings of the Royal Society of Edinburgh, 53, 260-283.

6. BERKSON, J. (1950): Are there two regressions? JASA, 45, 164-179.

7. BIRNBAUM, A. (1962): On the foundations of statistical inference. JASA, 57, 269-306.

8. DYKSTRA, O.J. (1966): The orthogonalization of undersigned experiments. Technometrics, 8, 279-290.

9. EFRON, B. and MORRIS, C. (1975): Data analysis using Stein's estimator and its generalizations. JASA, 70, 311-319.

10. EISENHART, C. (1947): The assumptions underlying the analysis of variance. Biometrica, 3, 1-21.

11. ELLINGSEN, W.R. and LEATHRUM, J.F. (1975): On-line ridge regression; sequential biased estimation for nonorthogonal problems. J.Statist.Comput.Simul. 3, 249-264.

12. FAREBROTHER, R.W. (1972): Principal component estimators and minimum mean square error criteria in regression analysis. The Review of Economics and Statistics, LIV, 332-336.

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

14. FARRAR, D.E. and GLAUBER, R.R. (1967): Multicollinearity in regression analysis : The problem revised. The Review of Economics and Statistics, 49, 92-107.

15. FELDESTEIN, M.S. (1973): Multicollinearity and the mean square error of alternative estimators. Econometrica, 41, 337-345.

16. FISHER, F.M. (1970): Tests equality between sets of coefficients in two linear regressions. Econometrica, 38, 361-366.

17. GAYLOR, D.W. and MERRILL, J.A. (1968): Augmenting existing data in multiple regression. Technometrics, 10, 73-81

18. GOLDSTEIN, M. and SMITH, A.F. (1974): Ridge type estimators for regression analysis. J.R.Statist.Soc.(B), 36, 284-91.

19. GREENBERG, E. (1975): Minimum variance properties of principal component regression. JASA, 70, 194-197.

20. GUILKEY, D.K. and MURPHY, J.L. (1975): Direct ridge regression techniques in cases of multicollinearity. JASA, 70, 769-775.

21. HAITORSKY, Y. (1969): Multicollinearity in regression analysis : comment. The Review of Economics and Statistics LI, 486-489.

22. HAWKINS, D.M. (1973): On the investigation of alternative regressions by principal component analysis. Appl.Statist. 22, 275-286.

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

. HEALY, M.J.R. (1968): Multiple regression with a singular matrix. Appl.Statist., 17, 110-117.

25. HEMMERLE, W.J. (1975): An explicit solution for generalized ridge regression. Technometrics, 17, 309-314.

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

27. HOCKING, R.R., SPEED, F.M. and LYNN, M.J. (1976): A class of biased estimators in linear regression. Technometrics, 18, 425-437.

28. HOERL, A.E. and KENNARD, R.W. (1970): Ridge regression : Biased estimation for nonorthogonal problem. Technometrics,12, 55-67.

29. HOERL, A.E. and KENNARD, R.W. (1970): Ridge regression : Application to nonorthogonal problems. Technometrics, 12, 69-82.

30. HOERL, A.E. and KENNARD, R.W. (1976): Ridge regression. Iterative estimation of biasing parameter. Communication in Statistics, A5(1), 77-88.

31. HOERL, A.E., KENNARD, R.W. and BALDWIN, K.F. (1975): Ridge regression : some simulations. Communications in Statistics 4(2), 105-123.

32. JURITZ, JUNE (1970): The partial multiple correlation coefficient and its application in regression analysis. Mathematics Colloquium, University of Cape Town.

33. KABE, D.G. (1968): On the distribution of the regression coefficient matrix of a normal distribution. Aust.Jour. of Statistics, 21-23.

34. KENDALL, M.G. (1951): Regression structure and functional relationship. Part I. Biometrika, 38, 11-25.

35. KENDALL, M.G. (1952): Regression structure and functional relationship. Part II. Biometrika, 39, 96-108.

36. KRZANOWSKI, W.J. ( ): The algebraic basis of classical multivariate methods. The Statistician, 20, 51.61.

37. KSHIRSAGAR, A.M. (1960): Some extensions of the multivariate t-distribution and the multivariate generalization of the distribution of the regression coefficient. Proc. Camb. Phil. Soc., 56, 80-85.

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

39. LEAMER, E.E. (1975): A result on the sign of restricted least squares estimates. Journal of Econometrics, 3, 387-390.

40. LIN, PI-ERH. (1972): Some characterization of the multivariate t-distribution. Journal of Multivariate Analysis, 2, 339-344.

41. LOTT, W.F. (1973): The optimal set of principal component. Restrictions on a least squares regression. Communications in Statistics, 2(5), 449-464.

42. LOWERRE, J.M. (1974): On the mean square error of parameter estimates for some biased estimators. Technometrics, 16, 461-464.

43. McCALLUM, B.T. (1970): Artificial orthogonalization in regression analysis. The Review of Economics and Statistics LII, 110-113.

44. McDONALD, G.C. and GALARNEAU, D.I. (1975): A Monte Carlo evaluation of some ridge type estimators. JASA, 70, 407-416.

45. MADASKY, A. (1959): The fitting of straight lines when both variables are subject to error. JASA, 54, 173-205.

46. MALLOWS, C.L. (1973): Some comments on C_p. Technometrics, 15, 661-675.

47. MALLOWS, C.L. (1974): Discussion No. 1. Technometrics, 16, 187-188

48. MARQUARDT, D.W. (1970): Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation. Technometrics, 12, 591-612.

49. MARQUARDT, D.W. (1974): Discussion No. 2. Technometrics,16,189-192

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

51. MASON, R.L., GUNST, R.F. and WEBSTER, J.T. (1975): Regression analysis and problems of multicollinearity. Communications in Statistics, 43, 277-292.

52. MASSY, W.F. (1965): Principal components regression in exploratory statistical research. JASA, 60, 234-246.

53. MAYER, L.S. and WILKE, T.A. (1973): One biased estimation in linear models. Technometrics, 15, 497-508.

54. MAYER, L.S. and YOUNGER, M.S. (1976): Estimation of standardized regression coefficients. JASA, 71, 154-157.

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

56. RAO, C.R. (1954): Some problems involving linear hypothesis in multivariate analysis. Biometrika, 46, 49-58.

57. RAO, M.M. and CHIPMAN, J.C. (1964): The treatment of linear restrictions in regression analysis. Econometrica, 32, 198-209.

58. ROLPH, J.E. (1976): Choosing shrinkage estimators for regression problems. Communication in Statistics, A5(9), 789-802.

59. SAMPSON, A.L. (1974): A tale of two regressions. JASA,69, 682-689.

60. SCLOVE, S.L. (1968): Improved estimators for coefficients in linear regression. JASA, 63, 597-606.

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

62. SILVEY, S.D. (1969): Multicollinearity and imprecise estimation. J.R.Statist.Soc. B31, 539-552.

63. SMITH, A.F.M. and GOLDSTEIN, M. (1975): Ridge regression. Some comments on a paper of Stone and Conniffe. The Statistician, 24, 61-66.

64. STONE, J. and CONNIFFE, D. (1973): A critical view of ridge regression. The Statistician, 22, 181-187.

65. STONE, J. and CONNIFFE, D. (1975): A reply to Smith and Goldstein. The Statistician, 24, 67-68.

66. THEOBALD, C.M. (1973): Generalization of mean square error applied to ridge regression. J.R.Statist.Soc. B36, 103-106.

67. TROSKIE, C.G. (1971): Regression and correlation. Proceeding of the Third Symposium on Mathematical Statistics, N.R.I.M.S.

68. TORO-VIZCARRONDO, C. (1968): A test of the mean square error criterion for restrictions in linear regression. JASA, 63, 558-572.

69. VINOD, H.D. (1976): Application of new ridge regression methods to a study of Bell system scale economies. JASA, 71, 835-841.

70. WALLACE, T.D. and TORO-VIZCARRONDO (1969): Tables for the mean square error test for exact linear restriction in regression. JASA, 64, 1649-1663.

71. WALLACE, T.D. (1972): Weaker criteria for linear restriction in regression. Econometrica, 40, 689-698.

72. WALLACE, T.D. and GOODNIGHT, J. (1972): Operational techniques and tables for making weak m.s.e. tests for restrictions in regressions. Econometrica, 40, 699-709.

73. WEBSTER, J.T., GUNST, R.F. and MASON, R.L. (1974): Latent root regression analysis. Technometrics, 16, 513-522.

74. WEBSTER, J.T., GUNST, R.F. and MASON, R.L. (1976): A comparison of least squares and latent root regression estimators. Technometrics, 18, 75-83.