ASPECTS OF ESTIMATION IN THE LINEAR MODEL WITH SPECIAL REFERENCE TO COLLINEARITY

by

Christien Thiart

Thesis Presented for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Statistical Sciences

UNIVERSITY OF CAPE TOWN

April 1994

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Assoc. Prof. Tim Dunne and co-supervisor Prof. Cas Troskie. Without their guidance, patience and expertise this thesis would not have been feasible. I record special thanks to Assoc. Prof. Dunne who read and corrected draft after draft.

Shirley Breed shared her expertise on T^3.

Finally I would like to thank the FRD for their financial support over the first two years and the staff of the Department of Statistical Sciences for their support.

ASPECTS OF ESTIMATION IN THE LINEAR MODEL WITH SPECIAL REFERENCE TO COLLINEARITY

ABSTRACT

Violations in the assumptions of the linear regression model lead to problems that make the ordinary least square estimator inappropiate. When the error terms are non-normal, the inference procedure that usually follows ordinary least squares estimation is no longer strictly valid. Collinearity, error terms that are non-normal and outliers are usually factors present to some degree in combination in a regression data set. Their effects may be confounded. In examining the factors separately the whole story is not told. We therefore set up a factorial design to investigate these problems under various scenarios.

The four factor (8 x 5 x 4 x 2) design consists of simulated data for 8 distributions, 5 collinearity levels, 4 variance levels and 2 orientations of the beta vectors in the model. The complete design was replicated 100 times, for a 30×5 regressor matrix X in the regression model. For a full 200×5 X regressor matrix a reduced design involving 8 distributions, 2 collinearity levels, 1 variance level and 2 orientations was used.

To compensate for collinearity in the model we propose biased estimators. The issue of non-normal distributions is addressed by the use of L_p-norm estimators. In our search for a better, robust estimator, ordinary least square estimators were compared to 13 biased, and 26 L_p norm estimators in the factorial design. The programs for the various estimators were obtained by using various algorithms in the literature, by adapting some of them, and by writing new algorithms. Comparisons were made between the different algorithms, and estimators. Generally, biased estimators are influenced by variance, orientation and collinearity but are impervious to distribution changes for the regular distributions of this study. Overall the performances of the L_p-norm estimators were disappointing. L_p-norm estimators are not influenced by variance, orientation or collinearity. It was only in long-tail distributions, like the Slash, that L_p-norm estimators seem to perform better than ordinary least squares estimators. Evidence was

obtained that the so-called robust (L_p-norm) estimators, were robust only to a point, but that very large outliers may influence the estimates substantially. Problems in estimating the moment ratio parameter were reported, and a possible improved estimator proposed. However in certain circumstances there are no reasonable estimates for the usual moment ratio parameter, in which case we suggest that an L_1-norm estimator and an appropriate median-like moment ratio estimator should be used.

A thorough overview of the theory of biased and L_p-norm estimators is presented.

A mixed general model is proposed. By changing our view of the parameters, the model can be adapted to address any of the scenarios in the linear regression model. Biased estimators, L_p-norm estimators, heteroscedasticity, transformations, misspecifications (overfitting and underfitting) and outliers can all be viewed as special cases of the mixed general model.

Christien Thiart
Department of Statistical Sciences
University of Cape Town
Private Bag
7700 Rondebosch

April 1994

CONTENTS

INTRODUCTION

In the linear regression model

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

the β is a vector of unknown coefficients, X is a n×r matrix of fixed regressors whose rank is r (n > r) and the ϵ is a n×1 vector of independent random errors with identical distributions. In estimation, we find a value of β that will minimize the residuals in some sense. The choice of function to be minimized will usually be an L_p-norm

$$\|Y - X\hat{\beta}_{L_p}\|_p = \left(\sum_{i=1}^{n} |\hat{\epsilon}_i|^p\right)^{\frac{1}{p}} \qquad (0.1)$$

where $\hat{\beta}_{L_p}$ is the L_p-norm estimator of β and the term $\hat{\epsilon}_i$ denotes the i-th element in the vector of residuals $\hat{\epsilon}$.

Classically, minimizing (0.1), with p = 2, gives the ordinary least squares (OLS) solution, which is appropriate for the case of normally (Gaussian) distributed errors.

This thesis was constructed to address two problems:

1. Collinearity: the assumption that X has full rank, r(X) = r, may be tenuous, leading to difficulties known as collinearity.

2. Error terms that are non-normal and outliers.

Under these problem conditions, OLS estimators may by inappropriate, because collinearity leads to "nonsense estimators" (see Chapter 2) and because a single outlier can perturb the $\hat{\beta}$ of OLS. When the error terms are not normal, the inference procedure that usually follows OLS is no longer strictly valid.

We seek estimators that are robust, ie estimators that are not effected by problems (1) and (2). Problem (1) is addressed by fitting biased estimators and problem (2) by fitting L_p-norm estimators. However since both problems may occur in the same data set, their effects may be confounded. In examining them separately the whole story is not told. We therefore set up a factorial design to investigate these problems under various scenarios.

The four factor (8 x 5 x 4 x 2) design consists of simulated data for 8 distributions, 5 collinearity levels, 4 variance levels and 2 orientations of the beta vectors in the model. The design was replicated 100 times, for a 30×5 matrix X in the regression model.

In our search for a robust estimator, OLS is compared to 13 biased estimators, and 26 L_p norm estimators.

The thesis structure is as follows:

The statistical and mathematical theory and background required to explore collinearity and L_p-norms is presented in Chapter 1. Chapter 2 discusses collinearity *per se* and gives a summary of some properties of biased estimators, and is in fact a summary of previously examined work, by the author. Chapter 3 presents some theory of L_p-norm estimators. In Chapter 4 the layout of the factorial design, and the simulation structure is discussed. Chapter 5 is a discussion of the results of the factorial simulation, together with various recommendations. Chapter 6 is a discussion of the general form of the linear model.

Appendices A - G are summaries of the results of the simulation study.

Chapter 1

THE LINEAR REGRESSION MODEL

1.1 The Model

The linear regression model is given by

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

where Y is a $n\times 1$ observed response vector,
ϵ is a $n\times 1$ vector of uncorrelated random error variables with expectation $E(\epsilon) = 0$, and variance matrix $Var(\epsilon) = V(\epsilon) = \sigma^2 I$,
β is a $r\times 1$ vector of regression coefficients that must be estimated, and
X is a $n\times r$ matrix of fixed regressors or independent variables, whose rank is r (we will assume that $n>r$).

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

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

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

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

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

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

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

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

1.2 Ordinary Least Square estimation

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

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

and the minimum sum of squares of residuals is

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

Properties:

1. $E(\hat{\beta}) = \beta$ (unbiasedness) and the bias matrix
   $B = (E(\hat{\beta}) - \beta)(E(\hat{\beta}) - \beta)' = bb' = 0$

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

3. $MSE(\hat{\beta}) = var(\hat{\beta}) + B = \sigma^2(X'X)^{-1}$

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

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

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

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

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

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

7. If ϵ is distributed Normally then

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

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

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

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

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

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

1.3 Other criteria of estimation

There are basically two models of interest when estimating the Beta coefficients. Consider a set of variables $(Y, X_1 \ldots X_p)$ where Y is a random variable depending on controlled or fixed explanatory variables $(X_1, \ldots, X_p)$. The LRM (1.1) written

$$Y = \beta_o + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon \qquad (1.1)$$

where assumptions are then made on random variable ϵ. If ϵ is normally distributed the maximum likelihood can be used to estimate the β_i. If the distribution of ϵ is not normal but reasonably well behaved (symmetric with tails not too long) the ordinary least squares estimators can be used. If the distribution is still well behaved but skew then some transformation can be used on Y (usually log y or $\sqrt{y}$).

A serious problem arises when the distribution of ϵ has long tails (large kurtosis). None of the above procedures yields satisfactory estimates for the β_i. The presence of even a single outlying observation in the dependent variable Y can give very unrealistic estimates for the β_i. In the above case a robust procedure is usually proposed. Detection of outlying observations has been a major research effort in statistics. The problem of a single outlier has been resolved (Doornbos (1981)). The detection of multiple outliers has also received attention and many solutions have been suggested, depending mainly on graphical plots (Hawkins (1980, 1984), Chalton and Troskie (1990)).

Another series problem is the presence of collinearities between the explanatory variables (X_1, .., X_p). A biased estimation procedure is then proposed essentially as a trade-off between bias and precision objectives. A mixture of long tails for the distribution of ϵ and collinearities between the explanatory variables exacerbates the problem quite considerably.

In the second model of interest the random variables ($Y, X_1, \ldots, X_p$) follow some multivariate distribution. In most cases a multivariate Normal is assumed. One is then interested in the conditional distribution of Y given that the random variables ($X_1, \ldots X_p$) take on some fixed values. The quantities of interest are the conditional means $E(Y|X_1, \ldots, X_p)$ and conditional variance. For the Normal distribution we have the linear conditional relationship

$$E(Y|X_1, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

which is of the same form as for model (1.1). Also the variance of (1.1) and conditional variance of the conditional model are assumed to be of the same form. For all practical purposes, ie estimation, tests of hypothesis, confidence intervals, tests for outliers, the two models can be considered to be identical. The central distributions are all the same. However the non-central distributions differ (Troskie (1971)).

The difficulty with the conditional model arises when the underlying model is not multivariate normal but has longer tails than the normal. The stable Pareto distributions are common alternatives (Press (1972)). This topic will not be discussed in this theses.

The problem of collinearity between the random variables $(X_1,\ldots,X_p)$ can be tackled in exactly the same way as for model (1.1) by using biased estimation procedures for the β_i. In both cases exact distribution results may not be available but asymptotic results are readily available.

There are countless examples of model (1.1) and the conditional model in theoretical and applied research. The LRM (1.1) often appears in contexts relating to science, engineering and, to a lesser extent, in medicine. A typical example is the Analysis of Variance. The conditional model often appears when human and economic variables are present which are not controllable, for instance in the human, medical, social and economic sciences. A typical example is the estimation of risk factors (betas) in finance which is becoming a huge growth area of research.

In all cases reliable estimators for β_i and their standard errors are sought.

The $\hat{Y}$ vector is a good estimator of Y, but $\hat{\beta}$ need not be a good estimator of β. Yet it is of interest to know something about $\hat{\beta}$. We therefore explore other criteria than OLSE, viz. generalized least squares, maximum likelihood and robustness, in estimating β.

1.3.1 Generalized least squares (GLS)

Consider model 1.1, but with a more general error covariance matrix:

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

where ϵ, as in model 1.1 is a $n \times 1$ vector of uncorrelated random error variable with $\epsilon \sim (\tau, \Sigma_\epsilon)$, $\mathrm{Var}(\epsilon) = \Sigma_\epsilon$ and $E(\epsilon\epsilon') = \Sigma_\epsilon + \tau\tau'$. Then the generalized least squares estimator (GLSE), denoted by $\hat{\beta}_G$, sometimes known as the Aitken (1935) estimator is obtained by minimizing

$$(Y - X\beta)'\Sigma_\epsilon^{-1}(Y - X\beta)$$

with respect to β. Thus

$$\hat{\beta}_G = \{X'\Sigma_\epsilon^{-1}X\}^{-1}X'\Sigma_\epsilon^{-1}Y \qquad (1.3.1)$$

Properties:

1. $$E(\hat{\beta}_G) = \beta + \{X'\Sigma_\epsilon^{-1}X\}^{-1}X'\Sigma_\epsilon^{-1}\tau \qquad (1.3.2)$$

   Thus $\hat{\beta}_G$ is biased for β, and we denote the bias of $\hat{\beta}_G$ by $\theta = \{X'\Sigma_\epsilon^{-1}X\}^{-1}X'\Sigma_\epsilon^{-1}\tau$. When $X'\Sigma_\epsilon^{-1}\tau = 0$, eg if $\tau = 0$, the GLS estimator will be unbiased.

2. $$\begin{aligned}\mathrm{Var}[\hat{\beta}_G] &= \{X'\Sigma_\epsilon^{-1}X\}^{-1}X'\Sigma_\epsilon^{-1}\mathrm{Var}(Y)\Sigma_\epsilon^{-1}X\{X'\Sigma_\epsilon^{-1}X\}^{-1} \\ &= \{X'\Sigma_\epsilon^{-1}X\}^{-1}X'\Sigma_\epsilon^{-1}\Sigma_\epsilon\Sigma_\epsilon^{-1}X\{X'\Sigma_\epsilon^{-1}X\}^{-1} \\ &= \{X'\Sigma_\epsilon^{-1}X\}^{-1} \qquad (1.3.3)\end{aligned}$$

3. $$\begin{aligned}\mathrm{MSE}[\hat{\beta}_G] &= \mathrm{Var}[\hat{\beta}_G] + \{X'\Sigma_\epsilon^{-1}X\}^{-1}X'\Sigma_\epsilon^{-1}\tau\tau'\Sigma_\epsilon^{-1}X\{X'\Sigma_\epsilon^{-1}X\}^{-1} \\ &= \{X'\Sigma_\epsilon^{-1}X\}^{-1} \quad (\text{when } X'\Sigma_\epsilon^{-1}\tau = 0, \text{ eg if } \tau = 0) \qquad (1.3.4)\end{aligned}$$

4. $$\begin{aligned} \mathrm{TMSE}[\hat{\beta}_G] &= \mathrm{tr}[\{X'\Sigma_\epsilon^{-1}X\}^{-1} + \{X'\Sigma_\epsilon^{-1}X\}^{-1}X'\Sigma_\epsilon^{-1}\tau\tau'\Sigma_\epsilon^{-1}X\{X'\Sigma_\epsilon^{-1}X\}^{-1}] \\ &= \mathrm{tr}[\{X'\Sigma_\epsilon^{-1}X\}^{-1}] + \mathrm{tr}\{\{X'\Sigma_\epsilon^{-1}X\}^{-1}X'\Sigma_\epsilon^{-1}\tau\tau'\Sigma_\epsilon^{-1}X\{X'\Sigma_\epsilon^{-1}X\}^{-1}\} \\ &= \mathrm{tr}[\{X'\Sigma_\epsilon^{-1}X\}^{-1}] + \tau'\Sigma_\epsilon^{-1}X\{X'\Sigma_\epsilon^{-1}X\}^{-1}\{X'\Sigma_\epsilon^{-1}X\}^{-1}X'\Sigma_\epsilon^{-1}\tau \\ &= \mathrm{tr}[\{X'\Sigma_\epsilon^{-1}X\}^{-1}] \quad (\text{when } X'\Sigma_\epsilon^{-1}\tau = 0,\ \tau = 0) \end{aligned} \tag{1.3.5}$$

5. $\hat{\beta}_G$ is the BLUE of β (for a proof see Schmidt (1967)).

6. The residuals

$$\begin{aligned} \hat{\epsilon} &= Y - X\hat{\beta}_G \\ &= X\beta + \epsilon - X\{X'\Sigma_\epsilon^{-1}X\}^{-1}X'\Sigma_\epsilon^{-1}(X\beta + \epsilon) \\ &= [I - X\{X'\Sigma_\epsilon^{-1}X\}^{-1}X'\Sigma_\epsilon^{-1}]\epsilon \\ &= [I - M]\epsilon \end{aligned} \tag{1.3.6}$$

where $M = X\{X'\Sigma_\epsilon^{-1}X\}^{-1}X'\Sigma_\epsilon^{-1}$,

$$\begin{aligned} \mathrm{tr}(M) &= \mathrm{tr}(X\{X'\Sigma_\epsilon^{-1}X\}^{-1}X'\Sigma_\epsilon^{-1}) \\ &= \mathrm{tr}(X'\Sigma_\epsilon^{-1}X\{X'\Sigma_\epsilon^{-1}X\}^{-1}) \\ &= \mathrm{rank}(X) \\ &= r \quad (\text{when X is of full column rank}) \end{aligned}$$

and

$$\begin{aligned} M'\Sigma_\epsilon^{-1}M &= \Sigma_\epsilon^{-1}X\{X'\Sigma_\epsilon^{-1}X\}^{-1}X'\Sigma_\epsilon^{-1}X\{X'\Sigma_\epsilon^{-1}X\}^{-1}X'\Sigma_\epsilon^{-1} \\ &= \Sigma_\epsilon^{-1}X\{X'\Sigma_\epsilon^{-1}X\}^{-1}X'\Sigma_\epsilon^{-1} \\ &= \Sigma_\epsilon^{-1}M \\ &= M'\Sigma_\epsilon^{-1} \end{aligned} \tag{1.3.7}$$

Suppose $\Sigma_\epsilon = \sigma^2 W_\epsilon$, where , σ^2 is unknown, but W_ϵ is known, then

$$M = X\{X'W_\epsilon^{-1}X\}^{-1}X'W_\epsilon^{-1}$$

$$\begin{aligned} M'W_\epsilon^{-1}M &= W_\epsilon^{-1}X\{X'W_\epsilon^{-1}X\}^{-1}X'W_\epsilon^{-1}X\{X'W_\epsilon^{-1}X\}^{-1}X'W_\epsilon^{-1} \\ &= W_\epsilon^{-1}M \end{aligned} \tag{1.3.8}$$

thus

$$E[\hat{\epsilon}'W_\epsilon^{-1}\hat{\epsilon}] = E[\epsilon'(I - M')W_\epsilon^{-1}(I - M)\epsilon]$$

$$= E\{\epsilon'(W_\epsilon^{-1} - W_\epsilon^{-1}M - M'W_\epsilon^{-1} + M'W_\epsilon^{-1}M)\epsilon\}$$

$$= E\{\epsilon'(W_\epsilon^{-1} - M'W_\epsilon^{-1})\epsilon\})$$

$$= E\{tr(\epsilon\epsilon'(I - M')W_\epsilon^{-1})\}$$

$$= tr\{\Sigma_\epsilon + \tau\tau'\}(I - M')W_\epsilon^{-1}))$$

$$= tr\{\sigma^2 W_\epsilon(I - M')W_\epsilon^{-1}\} + \tau'(I - M')W_\epsilon^{-1}\tau$$

$$= \sigma^2 tr((I - M')W_\epsilon^{-1}W_\epsilon) + \tau'(W_\epsilon^{-1} - M'W_\epsilon^{-1}M)\tau$$

$$= \sigma^2(tr(I_n) - tr(M)) + \tau'(W_\epsilon^{-1} - M'W_\epsilon^{-1}M)\tau$$

$$= \sigma^2(tr(I_n) - tr(X\{X'W_\epsilon^{-1}X\}^{-1}X'W_\epsilon^{-1})) + \tau'(W_\epsilon^{-1} - M'W_\epsilon^{-1}M)\tau$$

$$= \sigma^2(n - tr(I_r)) + \tau'(W_\epsilon^{-1} - M'W_\epsilon^{-1}M)\tau$$

assuming X has full column rank r

$$= \sigma^2(n-r) + \tau'(W_\epsilon^{-1} - M'W_\epsilon^{-1}M)\tau \qquad (1.3.9)$$

Thus if $\Sigma_\epsilon = \sigma^2 W_\epsilon^{-1}$, with W_ϵ^{-1} known, and $X'W_\epsilon^{-1}\tau = 0$, we still require $\tau = 0$ to have an unbiased estimator of σ^2 given by

$$\hat{\sigma}^2 = \hat{\epsilon}'W_\epsilon^{-1}\hat{\epsilon}/(n-r) \qquad (1.3.10)$$

7. When $\Sigma_\epsilon = \sigma^2 I$, the GLSE and the OLS estimator coincide.

1.3.2 Maximum Likelihood (ML)

The likelihood function of n random variable X_1, X_2,...,X_n is defined to be the joint density of the n random variables, say

$$f_{X_1 \ldots X_n}(x_1, \ldots x_n; \theta)$$

which is viewed to be a function of θ given the data. The likelihood function will be denoted by

$$L(\theta;x_1,\ldots x_n) \quad \text{or} \quad L(.;x_1,\ldots x_n)$$

where θ may be a single element or a vector of parameters. The likelihood function is presumed to describe the relative likelihood of each parameter value, as a function of the observed data. The maximum likelihood estimator (MLE) of θ is that value of θ that maximizes the likelihood.

Many likelihood functions satisfy regularity conditions; so that the MLE is the solution of the equation

$$\frac{dL(\theta)}{d\theta} = 0$$

Sometimes it is easier to maximize the log $L(\theta)$, which by monotonicity of the log function attains its maximum at the same value of θ as $L(\theta)$.

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

1.3.3 Robustness

In any model (such as the linear regression model (1.1)) one usually makes assumptions about the underlying situation eg error terms are independent, X matrix is fixed, and so on.

Statistical inferences while based partly upon an estimation criterion and the available data, also inherently involve the underlying assumptions of the model, particularly in the choice of the type of estimator that will be used. For instance under a ML criterion if we assume error terms are random and Gaussian (normal) in distribution, the MLE's are adopted, with distributional properties arising from model assumptions.

If assumptions are not exactly true, apparently minor violations might lead to estimators whose claimed properties under the assumptions are no longer valid due to the violations. An estimator or statistical procedure may be excessively sensitive to such model violations. We therefore seek estimators or statistical procedures that are "robust".

A robust estimator is an estimator that performs well under a variety of underlying conditions. Robustness signifies insensitivity to small deviations from the commonly imposed assumptions (Huber (1981)).

When an estimation procedure adapts itself to an underlying distribution it is known as an adaptive procedure (Hogg(1974)).

Huber (1981) consider three basic types of (robust) estimators namely maximum likelihood type (M) estimators, linear combinations of order statistics (L) and estimators observed from rank tests (R). In this thesis we will only be interested in one type of estimator namely the Maximum likelihood type (M) for which L_p, L_1 and L_∞ are special cases.

1.4 Singular Value Decomposition

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

Let X be an n×r matrix of rank r(X) = m, $m \leq r \leq n$. Then there is an n×n orthogonal matrix U, an r×r orthogonal matrix V, and an n×r matrix Δ such that

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

where $$\Delta : n\times r = \begin{bmatrix} D_a & 0 \\ 0 & 0 \end{bmatrix}$$

$$D_a = \mathrm{Diag}(\sqrt{\lambda}_1, \sqrt{\lambda}_2 \ldots \sqrt{\lambda}_{r(X)})$$

and $$\sqrt{\lambda}_1 \geq \sqrt{\lambda}_2 \geq \ldots . \sqrt{\lambda}_{r(X)} > 0$$

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

Unless the converse is stated explicitly, we will assume that X has full column rank, ie the rank of X, r(X), is r. Observe that for $n \geq r \geq m$ we may also write the SVD (also called the basic structure of X) as

$$X = U\Delta V' \quad (X:n\times r,\ U:n\times r,\ \Delta:r\times r,\ V:r\times r) \tag{1.4.1}$$
$$\Delta = D_\alpha$$
$$V = [\ _1 \quad _r \quad _i \quad 1$$
$$U = [\ldots u_r],\quad u_i:n\times 1$$
$$V'V = U'U = I_r$$

Note that $U'U = I_r$ but that $UU' \neq I_n$. Using r = m the SVD of X, the following equations will be useful in the subsequent Chapters.:

1.
$$X = [u_1\ldots u_r]\ \mathrm{Diag}(\sqrt{\lambda_1},\sqrt{\lambda_2}\ldots\sqrt{\lambda_r})[v_1\ldots v_r]'$$
$$= \sum_{i=1}^{r} \sqrt{\lambda_i} u_i v_i' \tag{1.4.2}$$

2.
$$X'X = V\ \Delta\ U'U\ \Delta\ V'$$
$$= V\ \Delta^2 V'$$
$$= \sum_{i=1}^{r} v_i v_i' \lambda_i \tag{1.4.3}$$

3.
$$(X'X)^{-1} = V\Delta^{-2}V'$$
$$= \sum_{i=1}^{r} v_i v_i'/\lambda_i \tag{1.4.4}$$

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

where $\Delta:n\times r = \begin{bmatrix} D_\alpha & 0 \\ 0 & 0 \end{bmatrix}$

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

and $\sqrt{\lambda_1} \geq \sqrt{\lambda_2} \geq \ldots \sqrt{\lambda_{r(X)}} > 0$

where $\Delta: n \times r = \begin{bmatrix} D_\alpha & 0 \\ 0 & 0 \end{bmatrix}$

$$D_\alpha = \text{Diag}(\sqrt{\lambda}_1, \sqrt{\lambda}_2 \ldots \sqrt{\lambda}_{r(X)})$$

and $\sqrt{\lambda}_1 \geq \sqrt{\lambda}_2 \geq \ldots . \sqrt{\lambda}_{r(X)} > 0$

4.
$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'Y \\ &= V \Delta^{-2} V' V \Delta U'Y \\ &= V \Delta^{-1} U'Y \\ &= \sum_{i=1}^{r} v_i u_i' Y / \sqrt{\lambda}_i \end{aligned} \qquad (1.4.5)$$

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

$$\hat{\beta} = \sum_{i=1}^{r} v_i \, c_i / \lambda_i \qquad (1.4.6)$$

5. $$V(\hat{\beta}) = \sigma^2 \sum_{i=1}^{r} v_i v_i' / \lambda_i \quad \text{(from 1.4.4)} \qquad (1.4.7)$$

1.5 Variance Inflation Factors

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

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

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

We refer to VIF's and their use in diagnosing collinearity in Chapter 2.

1.6.1 Analysis of variance (ANOVA)

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

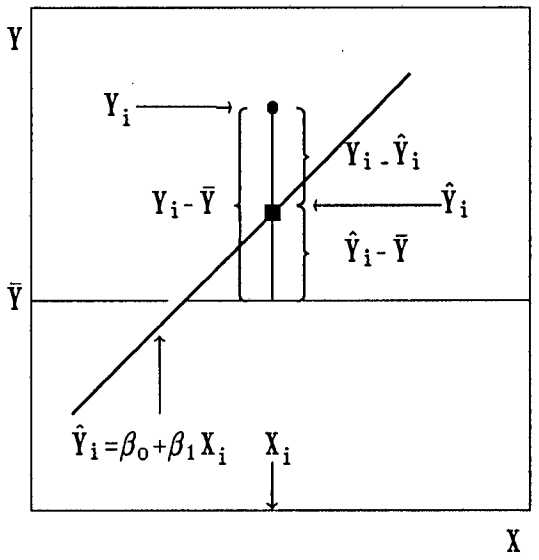
$$\hat{\epsilon}_i = Y_i - \hat{Y}_i$$

and $\hat{\epsilon}$ will be an n×1 column vector of all the n residual values.

The deviation $Y_i - \bar{Y}$, a quantity measuring the variation of the observations Y_i, can be decomposed as follows

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

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



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

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

or SSTO = SSR + SSE

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

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

$$SSTO = Y'Y - n\bar{Y}^2 \qquad (1.6.1)$$

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

$$SSE = Y'Y - \hat{\beta}'X'Y \qquad (1.6.3)$$

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

ANOVA Table

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

Sometimes the random variable SSE will also be specified as SSE $(X_1, X_2, \ldots, X_r)$, where the bracket denotes the subset of the independent

regressor variables that are included in the model. Where this notation is not explicitly used, the set of regressor variables in the model will be clear in the context.

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

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

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \qquad (1.6.4)$$

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

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

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

$$R = \sqrt{R^2} \qquad (1.6.5)$$

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

The coefficient of partial determination is defined as

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

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

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

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

1.6.2 Subset selection of regressor variables

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

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

1.6.2.1 All possible regressions

In the all possible regressions search procedure, all possible regression equations are computed and selection of an optimal equation is performed under some criterion (R^2, Adjusted R^2, MSE and C_p). If there are $(r-1) = k$ independent variables and one intercept term there will be 2^k possible

equations. For example if r = 3 (constant, X_1, X_2) the following 2^2 models are possible:

$E(Y) = \beta_0$; $E(Y) = \beta_0 + \beta_1 X_1$; $E(Y) = \beta_0 + \beta_2 X_2$; $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$; where the meaning (and the values) of the coefficients β_0, β_1, β_2 differ in each model.

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

(i) R^2 Criterion

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

(ii) Adjusted R^2

Adding more independent variables to the model can only increase R^2 and never reduce it. A modified measure that recognises the number of independent variables was introduced by Theil (1971) and Kennard (1971). The adjusted coefficient of multiple determination, denoted $R_a^2(\varphi)$, (where as before φ denotes the set of independent variables) is defined for $q \leq r$ ($n(\varphi) = q$) as:

$$R_a^2(\varphi) = 1 - \frac{n-1}{n-q}\frac{SSE(\varphi)}{SSTO} = 1 - \frac{n-1}{n-q}(1 - R_{y\varphi}^2)$$

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

(iii) MSE Criterion

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

(iv) C_p Criterion

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

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

where $\nu_i = \nu(X_{i1}, X_{i2}, \ldots)$ is the expected value of the i-th observation, arising from true equation for the conditional expectation of $(Y_i | X_{i1} \ldots X_{ir})$,

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

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

r = k+1 when $\beta_0 \neq 0$

= k when $\beta_0 = 0$

$\sigma^2 = \mathrm{var}(\epsilon_i)$ in the true model

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned} C_p &= SSE_p/\hat{\sigma}^2 - (n - 2p) \\ &= \frac{SSE_p}{SSE_r}(n-r-1) - (n - 2p) \quad \text{when } \beta_0 \neq 0 \\ &= \frac{SSE_p}{SSE_r}(n-r) - (n - 2p) \quad \text{when } \beta_0 = 0 \end{aligned} \tag{1.6.12}$$

When there is no bias in the p-variable regression equation

$$E[C_p|\nu_i=\eta_i] = (n-p)\sigma^2/\sigma^2 - (n - 2p)$$
$$= p$$

Thus, when the C_p values for all possible regressions are plotted against p, those regressions with little bias will tend to cluster near the line $C_p = p$, while those for equations with substantial bias will fall above this line. With this criterion we identify the sets of independent variables that lead to smallest C_p for each p and we would prefer those sets that have little bias (ie those near the line $C_p = p < r$). The method will detect such subsets with high probability whenever the effective contribution of some r-p variables in the full model is small or negligible in comparison with variance. By construction C_p focuses on the fitted values $\hat{Y}_i$ and the total bias and error variance associated with alternative submodels. As such it is impervious to collinearity problems.

1.6.2.2 Stepwise regression

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

(i) Forward Selection Procedure

In the forward selection procedure, the emphasis is on finding the best single predictor, then the best two predictors (which include the best single predictor) then the best three predictors (which include the best two

predictors, and in turn the best single predictor), and so forth. The procedure as outlined by Graybill (1976) is as follows:

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

2. Compute all squared multiple correlation coefficients of Y with all pairs of independent variables involving X_1, that is compute R_{y12}^2, R_{y13}^2, $R_{y14}^2, \ldots, R_{y1r}^2$, and select the largest. Suppose it is R_{y12}^2, then X_1 and X_2 are the best two predictors of Y which include the best single predictor X_1.

3. Compute all squared multiple correlation coefficients of Y with all sets of three variables that include X_1 and X_2, that is compute R_{y123}^2, $R_{y124}^2, \ldots, R_{y12r}^2$ and select the largest.

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

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

$$H_o : \rho_{y12\ldots q+1}^2 = \rho_{y12\ldots q}^2 \qquad (1.6.13)$$

which is true if and only if

$$H_o : \rho_{yq+1\cdot 12\ldots q}^2 = 0 \qquad (1.6.14)$$

Compute the test statistic W, where

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

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

The test statistic W is used as a diagnostic, and does not have a F-distribution. This complication arises because at each stage the variable included is the one with the largest F value. Hence the true distribution of the F-statistic is that of the maximum of a number of correlated F statistics. The distribution is unknown and depends upon the correlation structure of the explanatory variable for the particular problem at hand. The hypothesis H_0 is rejected (for a size α test) if and only if w the computed value of W satisfies $w \geq F(\alpha:1,n-q-2)$, the critical value of the $F(1,n-q-2)$ distribution. So the forward selection procedure is terminated at the step where H_0 is accepted. In some contexts α is chosen to be quite large, or almost equivalently, the tabulated F-value criterion is replaced by a suitable constant (eg F_{in} = 2.00 by default in BMDP)

(ii) Backward Elimination Procedure.

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

1. A regression equation containing all variables is computed.

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

3. The lowest partial F-test value, F_L, is compared with a preselected significance level and the corresponding critical value F_0. Then if $F_L < F_0$ we remove the variable X_L from the equation, and recompute the regression

equation without X_L, then re-enter step 2 again. If $F_L > F_0$, we adopt the regression equation.

(iii) Forward Selection with a View Back

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

1.7 Case Diagnostics

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

1.7.1 Outliers

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

1.7.2 Influence

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

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

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

The matrix H is called the Hat matrix, because it maps Y into $\hat{Y} = HY$ (Hoaglin and Welsh (1978)). The matrix H is symmetric ($H' = H$), idempotent ($HH = H$), and a projection matrix (into the column space of X). The diagonal elements of the Hat matrix, whose role as a diagnostic measure is discussed in, inter alia, Thiart, (1990) are

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

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

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

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

where $\hat{\sigma}$ is the residual mean square. It does not follow a t-distribution because $\hat{\epsilon}_i$ and $\hat{\sigma}$ are not independent. The distribution of $r_i^2/(n-r)$ is beta with parameters $\frac{1}{2}$ and $(n-r)/2$ (Atkinson (1987)). When σ is estimated by $\hat{\sigma}(i)$, the estimated error variance when the i-th row of X and Y have been deleted, the result is a studentised residual

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

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

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

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

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

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

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

The distance is regarded as large when $D^2 > F(1-\alpha,r,n-r)$, where $F(1-\alpha,r,n-r)$ is the $1-\alpha$ probability point of the central F-distribution with r and n-r degrees of freedom. While it is known that D^2 does not follow an F-distribution, comparison with tabulated F is an effective measure of relative change in estimated coefficients. All of these case diagnostics focus upon expectations of the individual cases, under alternative omission schemes. Thus these methods too are impervious to collinearity problems.

1.8 Bias and Jackknifing

1.8.1 Biased estimation

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

these biased estimators will perform better than LSE's in the presence of collinearity, in the sense of reduced mean square error (MSE).

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

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

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

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

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

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

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

1.8.2 Jackknifing

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

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

Define

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

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

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

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

$$\begin{aligned}\hat{\beta}_{-i} &= (X_{-i}'X_{-i})^{-1}X_{-i}'Y_{-i} \qquad (1.8.6)\\ &= [X'X - x_i x_i']^{-1}[X'Y - x_i Y_i] \\ &= [(X'X)^{-1} + (X'X)^{-1}x_i(I - x_i'(X'X)^{-1}x_i)^{-1}x_i'(X'X)^{-1}][X'Y - x_i Y_i] \\ &= \hat{\beta} - (X'X)^{-1}x_i[Y_i - Y_i h_i - x_i'\hat{\beta} + h_i Y_i](1-h_i)^{-1} \\ &= \hat{\beta} - (X'X)^{-1}x_i[Y_i - x_i'\hat{\beta}](1-h_i)^{-1} \\ &= \hat{\beta} - (X'X)^{-1}x_i[\hat{\epsilon}_i](1-h_i)^{-1} \qquad (1.8.7)\end{aligned}$$

for $i = 1,2,\ldots,n$.

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

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

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

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

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

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

and the variance estimator is

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

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

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

1.9 Vector and Matrix Norms, and Decompositions

1.9.1 Vector norms

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

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

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

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

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

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

where y_i is the i-th element of the vector $y: n \times 1$.

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

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

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

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

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

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

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

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

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

1.9.2 Matrix norms

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

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

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

The Frobenius norm of a matrix A is defined as

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

The Frobenius norm can also be shown to satisfy

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

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

A unitarily invariant matrix norm is a norm that satisfies

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

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

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

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

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

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

Using the notation of Horn and Johnson (1987) for a matrix norm $|||\cdot|||$ then

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

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

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

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

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

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

1.9.3 Decomposition

The theory of norms (given in the previous two sections) is used in the development and proof of the singular value and QR decomposition.

1.9.3.1 SVD

If the SVD of X is given by (1.4.1) and

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

then

$$r(X) = m \leq r \tag{1.9.16}$$

$$N(X) = \text{span}\{v_{m+1}, \ldots, v_r\} \tag{1.9.17}$$

$$R(X) = \text{span}\{u_1, \ldots, u_m\} \tag{1.9.18}$$

where $r(X) = m$ is the rank of X, $N(X)$ is the null space of X, and $R(X)$ is the range (or column space) of X.

Then the Frobenius norm of X can be written as

$$\|X\|_F^2 = \lambda_1 + \lambda_2 + \cdots + \lambda_m \tag{1.9.19}$$

and the matrix 2-norm of X is

$$\|X\|_2 = \sqrt{\lambda}_1 \tag{1.9.20}$$

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

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

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

1.9.3.2 QR

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

Let X:n×r and Y:n×1 be given and suppose that an orthogonal matrix Q:n×n exists and is computable, with the property that

$$Q'X = R = \begin{bmatrix} R_1 : rxr \\ 0 : (n-r)xr \end{bmatrix} \tag{1.9.21}$$

is upper triangular. Clearly X = QR.

If $$Q'Y = \begin{bmatrix} c : rx1 \\ d : (n-r)x1 \end{bmatrix} \tag{1.9.22}$$

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

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

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

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

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

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

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

$$X\Pi = QR \quad \text{where } \underset{n \times r}{R} = \begin{bmatrix} R_{11} & R_{12} \\ 0 & 0 \end{bmatrix} \begin{matrix} m \\ n-m \end{matrix} \qquad (1.9.24)$$

$$\begin{matrix} m & r-m \end{matrix}$$

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

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

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

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

where $$\Pi'\beta = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \begin{matrix} m \\ r-m \end{matrix} \qquad (1.9.26)$$

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

$$\Pi'\beta = \begin{bmatrix} R_{11}^{-1}(c - R_{12}Z_2) \\ Z_2 \end{bmatrix} \begin{matrix} m \\ r-m \end{matrix} \qquad (1.9.27)$$

We may set Z_2 to zero and obtain the basic solution

$$\Pi'\beta_{basic} = \begin{bmatrix} R_{11}^{-1}c \\ 0 \end{bmatrix} \begin{matrix} m \\ r-m \end{matrix} \qquad (1.9.28)$$

If R_{12} is zero the basic solution is the minimal 2-norm solution, since

$$\|\beta\|_2^2 = \min \left\| \begin{bmatrix} R_{11}^{-1}c \\ 0 \end{bmatrix} - \begin{bmatrix} R_{11}^{-1}R_{12}Z_2 \\ Z_2 \end{bmatrix} \right\|_2^2$$

$$= c'R_{11}^{-2}c + Z_2'Z_2$$

$$= c'R_{11}^{-2}c$$

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

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

1.10 Probability limit

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

1.11 Distributions

1.11.1 Uniform

The (continuous) Uniform random variable Y on a general interval [a,b], will be denoted by U(a,b). The density function of Y, for $-\infty < a < b < \infty$, is

$$f(y) = \frac{1}{b-a} I_{[a,b]}(y), \qquad (1.11.1)$$

with mean $\mu = \frac{a+b}{2}$, variance $\sigma^2 = \frac{(b-a)^2}{12}$,

central moments $\mu_r = 0$ for r odd

$$\mu_r = \frac{(b-a)^r}{2^r(r+1)} \quad \text{for r even,}$$

and kurtosis, $\kappa = \mu_4/\mu_2^2 = \frac{(b-a)^4}{2^4(4+1)} \div \frac{(b-a)^4}{12\times 12} = \frac{9}{5} = 1.8.$

1.11.2 Univariate Normal

When the random variable Y has a Normal (Gaussian) distribution with mean μ and variance σ^2, we will write $Y \sim N(\mu,\sigma^2)$. The density function of Y, for $-\infty < y < +\infty$, $-\infty < \mu < \infty$ and $\sigma > 0$ is

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\tfrac{1}{2}(y-\mu)^2/\sigma^2] \qquad (1.11.2)$$

with mean μ, variance σ^2,

central moments $\mu_r = 0$ for r odd

$$\mu_r = \frac{r!}{(r/2)!} \frac{\sigma^r}{2^{r/2}} \quad \text{for r even,}$$

and kurtosis, $\kappa = \mu_4/\mu_2^2 = \frac{4!}{(4/2)!} \frac{\sigma^4}{2^{4/2}} \div \sigma^4 = 3.$

1.11.3 Symmetric contaminated normal

The weighted sum $Y = w_1Y_1 + w_2Y_2$ of two $N(\mu_i,\sigma_i^2)$ variables follows a Contaminated normal distribution. The random variable Y has the density function

$$f(y) = w_1\frac{1}{\sqrt{2\pi}\,\sigma_1} \exp[-(y-\mu_1)^2/2\sigma_1^2] + w_2\frac{1}{\sqrt{2\pi}\,\sigma_2} \exp[-(y-\mu_2)^2/2\sigma_2^2] \qquad (1.11.3)$$

with $-\infty < y < \infty$ and $w_1 + w_2 = 1$. The central moments are

$$\mu_r = 0 \quad \text{for r odd}$$

$$\mu_r = \frac{r!}{(r/2)!2^{r/2}} [w_1\sigma_1^r + w_2\sigma_2^r] \quad \text{for r even.}$$

Thus the variance and kurtosis are respectively given by

$$\mu_2 = [w_1\sigma_1^2 + w_2\sigma_2^2]$$

and

$$\begin{aligned}\kappa &= \mu_4/\mu_2^2 \\ &= \frac{4!}{(4/2)!2^{4/2}} [w_1\sigma_1^4 + w_2\sigma_2^4] \div [w_1\sigma_1^2 + w_2\sigma_2^2]^2 \\ &= 3\,[w_1\sigma_1^4 + w_2\sigma_2^4] \div [w_1\sigma_1^2 + w_2\sigma_2^2]^2\end{aligned}$$

1.11.4 Multivariate normal

When the random variables in $\mathbf{Y}' = [Y_1\ Y_2 \ldots Y_n]$ have a joint multivariate normal distribution with vector of means μ and positive-definite variance-covariance matrix $\mathbf{V}$, we write $\mathbf{Y} \sim N(\mu, \mathbf{V})$. The density function of $\mathbf{Y}$ is then

$$f(Y_1, Y_2, \ldots, Y_n) = \frac{\exp[-\frac{1}{2}(\mathbf{Y}-\mu)'\mathbf{V}^{-1}(\mathbf{Y}-\mu)]}{(2\pi)^{\frac{n}{2}}|\mathbf{V}|^{\frac{1}{2}}} \qquad (1.11.4)$$

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

1.11.5 Slash

The Slash distribution is that distribution associated with the random variable obtained by dividing a N(0,1) deviate by an independent U[0,1] deviate.

$$Y = V/W, \quad \text{where } V \sim N(0,1),\ W \sim U(0,1),$$
$$\text{and } V \text{ and } W \text{ are independent.}$$

Rice (p89, 1988) gives the distribution of $Y = V/W$ as

$$f(y) = \int_0^1 |w| f(w) f_v(wy) dw$$

$$= \int_0^1 w.1.(2\pi)^{-\frac{1}{2}} \exp(-w^2y^2/2) dw.$$

Set $u = w^2$, thus $du = 2w\,dw$ and

$$= (2\pi)^{-\frac{1}{2}}\, 2^{-1} \int_0^{\infty} \exp(-uy^2/2)du$$

$$= (2\pi)^{-\frac{1}{2}}\, 2^{-1}\, (y^2/2)^{-1}\{1-\exp(-y^2/2)\}$$

$$= (2\pi)^{-\frac{1}{2}}\, (y^2)^{-1}\{1-\exp(-y^2/2)\} \text{ for } -\infty < y < \infty$$

$$= \begin{cases} [N(0,1) - N(y,1)]/y^2 & y \neq 0 \\ N(0,1)/2 & y = 0 \end{cases} \qquad (1.11.5)$$

The Slash is similar to the Normal, except that its tails are much heavier, so that it resembles the Cauchy, and has infinite even moments. Hence we preserve the median as zero, and change the scale in Y by the corresponding scale change in the V element.

1.11.6 Exponential

A random variable Y with the Exponential distribution has the density function,

$$f(y) = \lambda\exp\{-\lambda y\}, \quad \lambda > 0 \text{ and } 0 < y < \infty \qquad (1.11.6)$$

with mean $1/\lambda$, variance $1/\lambda^2$,
raw moments $\mu'_r = \dfrac{\Gamma(r+1)}{\lambda^r}$
and kurtosis, $\kappa = \mu_4/\mu_2^2 = 9$

1.11.7 Laplace

A random variable Y with Laplace distribution has the density function,

$$f(y) = \tfrac{1}{2}b\, \exp\{-|y-a|/c\}, \qquad -\infty < y < \infty, \quad -\infty < a < \infty, \quad c > 0 \qquad (1.11.7)$$

with mean a, variance $2c^2$,

central moments $\mu_r = 0$ for r odd

$\mu_r = r!c^r$ for r even

and kurtosis, $\kappa = \mu_4/\mu_2^2 = 4!c^4/4c^4 = 6$.

1.11.8 Central χ^2

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

$$f(u) = \frac{u^{\frac{n-2}{2}} \exp(-u/2)}{(2)^{\frac{n}{2}} \Gamma(n/2)}, \quad u > 0 \qquad (1.11.8)$$

1.11.9 Central $F(n_1,n_2)$

The ratio of two independent variables each having χ^2-distributions, over their degrees of freedom has an F-distribution. Thus if

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

$$f(v) = \frac{\Gamma(\frac{1}{2}(n_1+n_2))\ n_1^{\frac{1}{2}n_1}\ n_2^{\frac{1}{2}n_2}\ v^{\frac{1}{2}n_1-1}}{\Gamma(\frac{1}{2}n_1)\ \Gamma(\frac{1}{2}n_2)\ (n_2+ n_1v)^{\frac{1}{2}(n_1+n_2)}}, \qquad v > 0 \qquad (1.11.9)$$

1.11.10 Central t

The Student's t distribution is that distribution associated with the ratio of a standard normal random variable to the square root of an independently distributed chi-square random variable which has been divided by its degrees of freedom.

$$T = Z/\sqrt{U/k} \sim t_k$$

where $Z \sim N(0,1)$, $U \sim \chi^2_k$ with k degrees of freedom and Z and U are independent.

The density function of T is

$$f(t) = \frac{\Gamma[\frac{1}{2}(k+1)]}{\Gamma[\frac{1}{2}k]} \frac{1}{\sqrt{k\pi}} \frac{1}{(1+t^2/k)^{\frac{1}{2}(k+1)}} \qquad -\infty < t < \infty, \quad k > 0$$

with mean $\mu = 0$ for $k > 1$, variance $\sigma^2 = \frac{k}{k-2}$ for $k > 2$,

central moments $\mu_r = 0$ for $k > r$ and r odd

$$\mu_r = \frac{k^{\frac{1}{2}r}}{} \frac{B(\frac{1}{2}(r+1),\frac{1}{2}(k-r))}{B(\frac{1}{2},\frac{1}{2}k)} \quad \text{for } k > r \text{ and } r \text{ even}$$

and kurtosis, $\kappa = \mu_4/\mu_2^2 = \frac{3(k-2)}{(k-4)}$ $k > 4$.

1.11.11 Non-central χ^2

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

$$\gamma = \mu'\mu/2 \qquad (1.11.11)$$

Reference to the distribution is by means of the symbol $\chi^2(n,\gamma)$. For $n > 0$, the density function of the non-central χ^2-distribution $\chi^2(n,\gamma)$ is

$$f(u) = \exp(-\gamma) \sum_{k=0}^{\infty} \frac{\gamma^k \, u^{\frac{1}{2}(n+2k-1)} \exp(-\frac{1}{2}u)}{k! \; (2)^{\frac{1}{2}n+k} \quad \Gamma(\frac{1}{2}n+k)}, \qquad u > 0 \tag{1.11.12}$$

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

1.11.12 Non-central $F(n_1,n_2;\gamma)$

If U_1 and U_2 are independent and

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

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

the non-central F-distribution with n_1 and n_2 degrees of freedom and non-centrality parameter γ. For $v > 0$, the density function is

$$f(v) = \sum_{k=0}^{\infty} \frac{\exp^{-\gamma}\gamma^k}{k!} \frac{n_1^{\frac{1}{2}n_1+k} \; n_2^{\frac{1}{2}n_2} \; \Gamma(\frac{1}{2}n_1+\frac{1}{2}n_2+k)}{\Gamma(\frac{1}{2}n_1+k) \; \Gamma(\frac{1}{2}n_2)} \frac{v^{\frac{1}{2}n_1+k-1}}{(n_2+n_1v)^{\frac{1}{2}n_1+\frac{1}{2}n_2+k}}, \qquad v > 0 \tag{1.11.13}$$

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

1.11.13 Doubly non-central $F(n_1,n_2;\gamma_1,\gamma_2)$

If U_1 and U_2 are independent and

$$U_1 \sim \chi^2(n_1,\gamma_1) \text{ and } U_2 \sim \chi^2(n_2,\gamma_2)$$

$$\text{then} \qquad V = \frac{U_1/n_1}{U_2/n_2} \text{ is distributed as } F(n_1,n_2;\gamma_1,\gamma_2),$$

the doubly non-central F-distribution with n_1 and n_2 degrees of freedom and non-centrality parameters γ_1 and γ_2. For $v > 0$, the density function is

$$f(v)= \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} \frac{\exp^{-\gamma_1}\gamma_1^j}{j!} \frac{\exp^{-\gamma_2}\gamma_2^k}{k!} \frac{n_1^{\frac{1}{2}n_1+j}\, n_2^{\frac{1}{2}n_2+k}}{B(\frac{1}{2}n_1+j,\frac{1}{2}n_2+k)} \frac{v^{\frac{1}{2}n_1+j-1}}{(n_2+n_1v)^{\frac{1}{2}n_1+\frac{1}{2}n_2+k+j}}, \quad v > 0$$

(1.11.14)

Here too some texts prefer to regard $\mu_1'\mu_1$ and $\mu_2'\mu_2$ as the non-centrality parameters.

1.11.14 Non-central t

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

$$f(t) = \frac{n^{\frac{n}{2}}}{\Gamma(\frac{1}{2}n)} \frac{e^{-\frac{1}{2}\mu^2}}{(n+t^2)^{\frac{1}{2}(n+1)}} \sum_{k=0}^{\infty} \frac{\Gamma(\frac{1}{2}(n+k+1))\ \mu^k\ 2^{\frac{k}{2}}\ t^k}{k!(n+t^2)^{\frac{k}{2}}} \qquad (1.11.15)$$

1.12 Simulation

From pseudo-random Uniform U[0,1] values one can generate various distributions by using specific techniques. In the sections that follow the technique for each distribution is described. For some distributions more than one row of pseudo-random deviates is needed. For the specific simulated data in Chapter four a total of six streams were needed, which we distinguish by labelling them A, B, C, D, E and F. A discussion of the generation of data for the simulation study of this thesis is presented in Chapter 4.

1.12.1 Uniform

The Uniform distribution is defined in §1.11.1. The density function for Y is given in (1.11.1) and the distribution function of Y is

$$F(y) = \frac{y-a}{b-a} \qquad a \leq y \leq b$$

Thus a random deviate, U, with Uniform [0,1] distribution, has mean $\frac{1}{2}$ and variance $\frac{1}{12}$. To generate a random deviate Y from the Uniform distribution with mean 0 and variance equal to σ^2, we use the cumulative distribution function technique. We require $y = F^{-1}(u)$, where u is a Uniform [0,1] random number (stream A). Setting a = -b, where f(y) = 1/2b, $F(y) = \frac{y+b}{2b}$, and

$$\begin{aligned} y = F^{-1}(u) &= 2bu - b \\ &= b(2u - 1) \end{aligned}$$

Thus if Y ~ Uniform[-b,b], then var(Y) = $b^2/3$, and b = $\sqrt{3}\sigma$.

1.12.2 Normal (Gaussian)

A normal (Gaussian) random variable has the density function given in 1.11.2.

To obtain Normal(0,1) deviates from Uniform deviates, we use the Box-Muller transformation. Box and Muller (1958) proposed a method in which two independent Uniform variables U[0,1] generated from separate seeds, eg stream A and B, are used to generate two independent standard normal variables N(0,1). The transformations are

$$\begin{aligned} z_1 &= (-2 \ln u_1)^{\frac{1}{2}} \sin 2\pi u_2, \\ z_2 &= (-2 \ln u_1)^{\frac{1}{2}} \cos 2\pi u_2, \end{aligned}$$

where u_1 and u_2 are the independent Uniform deviates (from stream A and B respectively).

To obtain the desired level of variance, we multiply by the required σ.

For a discussion of the performances of the Box and Muller technique under various generators see §4.2.2.2.

1.12.3 Symmetric contaminated normal

The weighted sum $Y = w_1Y_1 + w_2Y_2$ of two $N(\mu_i,\sigma_i^2)$ variables follows a Contaminated normal distribution. The density function, variance and kurtosis are given in §1.11.3. To obtain a Contaminated random deviate we use the same process as in the normal case: we take two independent N(0,1) variables from the A and B streams, say z_1 and z_2, multiply by the square root of the relevant variance factors for σ_1^2 and σ_2^2, and mix out a single stream randomly under the weights (using the C-stream of random deviates).

For specific choices of weights and variance see §4.2.2.3.

1.12.4 Laplace

The Laplace distribution is defined in §1.11.7. The distribution function is

$$F(y) = \tfrac{1}{2}\exp\{(y-a)/c\}, \qquad \text{when } y \le a$$

and

$$F(y) = 1 - \tfrac{1}{2}\exp\{-(y-a)/c\}, \qquad \text{when } y > a$$

For zero mean we choose a=0. Then by using the cumulative distribution function technique, $y = F^{-1}(u)$, where u is a Uniform [0,1] random number (stream A), and we solve for y when $0 \le u \le 0.5$ in the equation

$$\tfrac{1}{2}\exp[y/c] = u.$$

Hence

$$y/c = \ln(2u)$$

$$y = c\ln(2u)$$

and $y \le 0$.

For $0.5 \leq u \leq 1.0$ we solve the equation

$$1-\tfrac{1}{2}\exp\{-(y)/c\} = u.$$

Hence $$-(y)/c = \ln(2\{1-u\})$$
$$y = -c\ \ln(2\{1-u\})$$
and $y > 0$.

1.12.5 Exponential

The Exponential distribution is defined in §1.11.6 and has the distribution function

$$F(y) = 1 - \exp\{-\lambda y\}$$

Since $\lambda > 0$ the mean is not set to zero. By using the cumulative distribution function technique, $y = F^{-1}(u)$, where u is a Uniform [0,1] random number (stream A) we have

$$\exp[-\lambda y] = 1-u$$
and $$y = -\ln(1-u)/\lambda.$$

Now y is a sample observation from an Exponential distribution with parameter λ, and variance $1/\lambda^2$.

Note that when errors from this distribution are transferred into the simulated data, least squares centering will result in the estimation of the intercept $\beta_0 + \lambda$ as the overall constant term.

1.12.6 Student's t

The Student's t distribution is defined in §1.11.10. To find T from the Uniform deviates we have to form N(0,1) deviates. Two streams of pseudo-random Uniform deviates, will be fed into the Box-Muller transformation, and will result in two z_i (N(0,1)) streams.

$$t = z_0\sqrt{k}/(z_1^2 + z_2^2 + \ldots + z_k^2)^{1/2}.$$

For k = 5, we will need six streams (A, B, C, D, E and F) of Uniform deviates to yield six streams (0,1,2,3,4,5) of Normal deviates.

1.12.7 Slash

The Slash distribution is defined in §1.11.5, and is found by dividing a N(0,1) deviate by an independent U[0,1] deviate. In this thesis the N(0,1) deviate is found as described in §1.13.2 (stream A and B) and the U[0,1] deviate is from the C stream.

Chapter 2

COLLINEARITY AND BIASED ESTIMATORS

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

Collinearity can not be described in simple terms as being present or absent. Rather, what is important is the degree of collinearity and what effect this degree can have on the regression model. For the statistician, near-collinearities inflate the variances of regression coefficients and magnify the effects of error in the regression response variable. For the numerical analyst, collinearities combine with rounding errors to introduce inaccuracies in computations. In the opinion of the writer collinearity is much a problem as an inherent part of the data set and model.

Inherently coefficient estimation *per se* is an extrapolation from the available data. Primarily it is this desired extrapolation that is at issue in any collinearity discussion.

2.1 Defining and detecting Collinearity

For a complete discussion on definitions and ways of detecting collinearity the reader is referred to Thiart (1990, §2.2 and 2.3). One of these definitions has been used by several writers, Johnston (1963), Silvey (1969), Mason *et al.* (1975) and others, and is in terms of the linear dependence of a set of column vectors, X_j of the matrix X.

Definition 2.1.1

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

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

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

The distinction between defining and detecting collinearity is elusive. Some authors use ways of detecting collinearity as an implicit method of defining collinearity. For example, (i) if at least one singular value of X is small, then X is collinear; (ii) if the determinant of X'X approaches zero, then X is near-singular; (iii) if X has large VIF's or large condition numbers these values are taken as indicators of collinearity. The main issue is that the user of OLS must be aware of what is happening in the space of the regressor variables. What is important is the awareness of the inherent inadequacy of the model for coefficient estimation.

Collinearity can be detected by various indicators and methods, some more adequate and correct, and others merely helpful but not conclusive. Some of these indicators and methods are: (i) sensitive estimators, (ii) correlation matrix of scaled regressors, (iii) determinant of X'X, (iv) departure from orthogonality, (v) smallest singular value, (vi) condition number and condition index (vii) regression coefficient variance decomposition (viii) mixed condition index (ix) variance inflation factors (x) signal-to-noise tests and (xi) use of prior information.

2.2 Collinearity in practice

In practice collinearity becomes harmful to inferences when estimation or hypothesis testing is influenced more by the relationship between the regressor variables than by the relationship between the response and the

regressor variables. Such an influence can result in poor parameter estimates and restrictions on the applicability and generality of the model in use.

2.2.1 Sources and origins

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

1. An over-defined model is one in which there are more regressor variables than observations. This type of model arises frequently in medical research where many elements of information are recorded on each individual in a study.
2. Any in-built mathematical constraint in variables that forces them to add to a constant for each case will generate a collinearity in a centred model. Generating new variables as transformations of other variables can produce a collinearity among the set of variables involved eg ratios or powers of variables frequently may be nearly collinear with the original variables.
3. Component variables of a system may show near linear dependencies because of biological or physical constraints of the system (eg various measures of size of an organism will show dependencies). Such correlation structures are properties of the system and can be expected to be present in all observations obtained from the system. Gunst (1983) referred to this type of collinearity as 'population-inherent collinearities'.
4. Inadequate sampling occurs when the experimenter unknowingly samples only from a subspace of the space of the regressor variables. Collinearities due to sampling deficiencies are a property of the particular data set which has been collected and would not be expected to occur in data sets arising from alternative sampling.
5. Poor experimental design may give rise to collinearities. When possible the levels of the experimental factors are generally chosen in such a way so that the different treatment factors are statistically orthogonal to each other.

6. Outliers in the design space can induce artificial collinearities among the predictor variables, and will be discussed in Chapter 6 or see Thiart (Chapter 9, 1990).

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

2.2.2 Effects of collinearity

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

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

The variance covariance matrix of the OLS estimator is given by

$$V(\hat{\beta}) = \sigma^2 \sum_{i=1}^{r} v_i v_i'/\lambda_i . \qquad (2.2.1)$$

Clearly $V(\hat{\beta}_i)$ can also be written as (Marquardt (1970), Stewart (1987) or Thiart (1990), 2.2.24)

$$V(\hat{\beta}_i) = \frac{\sigma^2}{(X_i' X_i)^{-1}} VIF_i \qquad (2.2.2)$$

Because the bias is zero

$$MSE(\hat{\beta}) = \sigma^2 \sum_{i=1}^{r} v_i v_i'/\lambda_i \qquad (2.2.3)$$

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

$$\text{TMSE}(\hat{\beta}) = \sigma^2 \sum_{i=1}^{r} 1/\lambda_i > \sigma^2/\lambda_r \qquad (2.2.4)$$

2.2.2.1 Inflation of variance

In the presence of near collinearity $\lambda_r \to 0$ so that the $\text{Var}(\hat{\beta})$ is inflated and $\text{TMSE}(\hat{\beta}) \to \infty$. From (2.2.2) the individual variance of the i-th element of $\hat{\beta}$ is

$$V(\hat{\beta}_i) = \frac{\sigma^2}{(X_i'X_i)^{-1}} \text{VIF}_i \qquad \text{for } i = 1,2,\ldots r \qquad (2.2.5)$$

Thus by the definition of the variance inflation factor, the variance of the estimator of the i-th regression coefficient is a function of the VIF_i's. If VIF_i is large (indicating collinearity) then the variance will be inflated as well. In the case of a nearly orthogonal design, for each i $\text{VIF}_i \approx 1$, and there is no effect on the variance.

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

2.2.2.2 Unexpected coefficient values and signs

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

computational error. To these we may add (iv) outliers in the response variable and (v) influential cases.

We illustrate the effect of collinearity by considering the OLSE which, from (1.4.5), can be written as

$$\hat{\beta} = \sum_{i=1}^{r} v_i u_i' Y/\sqrt{\lambda_i}$$

$$= \sum_{i=1}^{r} v_i c_i/\lambda_i \quad \text{where } c_i = u_i' Y\sqrt{\lambda_i} \qquad (2.2.6)$$

Assume in the SVD of X that the eigenvalues are ordered, eg

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{r-1} \geq \lambda_r$$

Suppose that λ_r is much smaller than λ_{r-1} (implying a single very strong collinearity), so much smaller that the summation in (2.2.6) is completely dominated by λ_r, eg

$$\hat{\beta} \simeq v_r c_r/\lambda_r .$$

Then we may infer

$$\hat{\beta}_j \simeq v_{rj} c_r/\lambda_r \qquad (2.2.7)$$

Gunst and Mason (1980) claim two characteristics of expression (2.2.7) when a strong collinearity occurs in X and $v_{rj} \neq 0$:

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

(ii) the signs of the estimates tend to be determined more by the collinearity associated with v_r than by relationship of the predictor

variables with the response: ie if $c_r > 0$, the sign of $\hat{\beta}_j$ is the same as that of v_{rj}; if $c_r < 0$, the sign of $\hat{\beta}_j$ is opposite of that of v_{rj}.

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

2.2.2.3 Unstable regression coefficients

In the presence of collinearity a small perturbation in X or Y can result in a relatively unstable regression coefficient $\hat{\beta}_i$. However the $\hat{\beta}_i$ are estimated in a linearly interdependent way so that some estimable functions may in fact be very precisely determined despite the collinearity.

2.2.2.4 Linear combinations of regression variables

Poor precision in the estimation of individual parameters does not imply that the estimated model is a poor predictor of linear functions of the $\hat{\beta}_i$. Although some individual parameters may be estimated poorly, the Y values may be estimated adequately for each data case, as the whole vector of $\hat{\beta}_i$'s is used. When the collinearity persists into the prediction area, the collinearity is not harmful. Use of the model outside the data-defined region of observed estimable functions (extrapolation) will result in poor prediction.

2.3 Centering and standardisation of the X matrix

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

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

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

The condition number has its own scaling problem. Stewart (1987) show that by scaling down any column of X, the condition number can be made arbitrary large and cause 'artificial ill-conditioning'. Therefore it is recommended that before computing the condition number, the columns should be standardised to have unit column length (Belsley *et al.* (1980, Appendix 3B and §3.3))

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

The effect of centering on VIF's is discussed by Schall and Dunne (1987b), it is invariant over a mean shift in the variables $[X_{(i)}\ X_i]$.

Mean-centering does not impose an arbitrary (data dependent) mean on the model. It removes arbitrariness from collinearity measures, by making them invariant over the choice of an origin. The origin of measurement of the variables in a regression model is usually determined by the experimenter or the measuring instrument. Thus collinearity measures based on data without centering, will depend on the specific implicit choice of origin.

In the case of a model with a constant term and one covariate, one would not speak of collinearity as collinearity occurs due to the presence of more than one covariate. The VIF due to having the covariates $X_{(i)}$ in addition to the constant term, is measured by the partial variance inflation factor (thus the VIF obtained from centered data).

Belsley (1984,1986) holds the view that centering can mask elements of ill-conditioning and produce meaningless and misleading collinearity diagnostics. He argues that the intercept term should not be dealt with as a 'nuisance parameter' as it could play a vital role in the collinearities involved. His view was demonstrated theoretically and practically in Simon and Lesage (1988). However Schall and Dunne (1987b) do not regard the mean (intercept) as a nuisance parameter so much as an essential parameter accounting for an interpretation of the origin.

Stewart (1987) and Gunst (1983) concluded that one should mean-center and only examine the conditioning of the uncentred data if the estimate of the

intercept is of interest. Computationally it may be advantageoûs to center and standarise as more accurate estimations may occur.

Randall and Rayner (1987) show that it could have computational advantages to decenter, ie first center to obtain computational accuracy and then go back to the original model by decentering. They warned that decentering is not easily mechanized.

2.4 Detecting and handling collinearity

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

In §2.1 various ways of detecting collinearities have been discussed. It is of utmost importance that collinearity should be detected. Any method may be used, but it may even be advisable to use several of them. If one were to use for instance VIF's, it is also good practice to look at other methods (e.g condition number, condition indices, variance-decomposition, and small eigenvalues), to get a multi-faceted insight into the problem. What is important here is the identification of collinearity and not which particular method one uses to detect it. The user may have to directly calculate some of the measures, as many currently released regression programs are not designed to warn automatically of the presence of near-collinearities.

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

Subset selection of variables to remove the collinearity should be applied with great care, as this approach may result in removing some of the important regression variables. Hoerl *et al.* (1986) recommend against subset selection as a general strategy to combat collinearity. In the face of severe collinearity one of the best alternatives is to use those biased estimators that are not so severely effected by collinearity. An array of biased estimators will be described in §2.5. A choice for one of them will depend on the circumstances of the problem, and estimators may perform differently in different situations.

2.5 Biased estimators

In the presence of collinearity one of the alternatives to OLS is biased estimation. In this section we present tabular summaries of the expectation and expected square error properties of biased estimators used in the simulation study. Thiart (1990) gives a detailed discussion of these estimators. Table 2.1 consists of a summary of estimators and abbreviations. The tables 2.2 and 2.3 are separated largely for convenience of presentation, as the properties to which they refer are essentially interrelated. Table 2.4 presents some choices (mostly those used in Chapter 4) of k, K, and d for substitution in the estimators appearing in table 2.2 and 2.3.

Table 2.1 Family and estimator abbreviations and references

Family	Estimator	Description
OLS:		Ordinary least squares estimator (Rawlings; 1988)
R:	RHK	Ridge regression, estimated k (Hoerl, Kennard and Baldwin; 1975)
	RLW	Ridge regression, estimated k (Lawless and Wang; 1976)
GR:	GRHK	Generalized ridge regression, estimated K (Hoerl, Kennard and Baldwin; 1975)
	GRT	Generalized ridge regression, estimated K (Troskie; 1990)
JGR:	JGRS	Almost unbiased (jackknife) generalized ridge regression (Singh, Chaubey and Dwivedi; 1986) and (Nomura; 1988)
JR:	JRO	Almost unbiased (jackknife) operational ridge regression (Ohtani; 1986) and (Kadiyala; 1984)
PC:	PC1	Principal component regression, delete the smallest singular value (Rawlings; 1988)
	PC2	Principal component regression, delete the two smallest singular values (Rawlings; 1988)
FPC:	FPCG1	One step version of FPCI estimator (iterative fractional principal component estimator, via the generalized ridge method), where the OLSE is replaced by the PC1 estimator (Lee and Birch; 1988)
	FPCG2	One step version of FPCI estimator, where the OLSE is replaced by the PC2 estimator (Lee and Birch; 1988)
	FPCR1	One step version of FPCV estimator (iterative fractional principal component estimator, via the RR method), where the OLSE is replaced by the PC1 estimator (Lee and Birch; 1988)
	FPCR2	One step version of FPCV estimator, where the OLSE is replaced by the PC2 estimator (Lee and Birch; 1988)
SH:		Shrinkage estimator (Mayer and Willke; 1973)

Table 2.2 Expectation properties of estimators

Family	Definition	Bias	Variance matrix
OLS	$\hat{\beta} = \sum_{i=1}^{r} v_i c_i/\lambda_i$,	0	$\sigma^2 \sum_{i=1}^{r} v_i v_i'/\lambda_i$
	$\hat{\delta} = (Z'Z)^{-1}Z'Y$	0	$\sigma^2 \sum_{i=1}^{r} 1/\lambda_i$
	$= V'\hat{\beta}$		
PC	$\hat{\beta}_{PC} = \sum_{i=1}^{r-m} v_i c_i/\lambda_i$,	$-V_2V_2'\beta$	$\sigma^2 \sum_{i=1}^{r-m} v_i v_i'/\lambda_i$
R	$\hat{\beta}_R = \sum_{i=1}^{r} (\lambda_i+k)^{-1} c_i v_i$	$-k(X'X+kI)^{-1}\beta$	$\sigma^2 V[\Delta^2+kI]^{-2}\Delta^2 V'$
GR	$\hat{\delta}_K = (\Delta^2 + K)^{-1}\Delta^2\hat{\delta}$	$-(\Delta^2 + K)^{-1}K\delta$	$\sigma^2(\Delta^2+K)^{-2}\Delta^2$
JR	$\hat{\delta}_{JW} = [I-[kA^{-1}]^2]\hat{\delta}$ where $A = Z'Z + K$	$-[kA^{-1}]^2\delta$	$\sigma^2[I-[kA^{-1}]^2]^2\Delta^{-2}$
JGR	$[\hat{\delta}_0]_i = \dfrac{(\lambda_i\hat{\delta}_i^2+2\hat{\sigma}^2)\lambda_i\hat{\delta}_i^3}{(\lambda_i\hat{\delta}_i^2+\hat{\sigma}^2)^2}$	see II.1	see II.2
SH	$\hat{\beta}_{SH} = d \sum_{i=1}^{r} v_i c_i/\lambda_i$	$-(1-d)\beta$	$\sigma^2 d^2 \sum_{i=1}^{r} v_i v_i'/\lambda_i$
FPC	$\hat{\delta}_{FPC} = F\hat{\delta}^*$	$-[I-F]\delta$	$\sigma^2 F\Delta^{-2}F$
	$\hat{\beta}_{FPC} = VFV'\hat{\beta}$	$-[I-VFV']\beta$	$\sigma^2 VF\Delta^{-2}FV'$

*see II.4 and II.5

Table 2.3 Mean square errors of estimators

Family	MSE matrix	TMSE = tr(MSE)
OLS	$\sigma^2 \sum_{i=1}^{r} v_i v_i'/\lambda_i$	$\sigma^2 \sum_{i=1}^{r} 1/\lambda_i$
PC	$\sigma^2 \sum_{i=1}^{r-m} v_i v_i'/\lambda_i + V_2 V_2' \beta\beta' V_2 V_2'$	$\sigma^2 \sum_{i=1}^{r-m} 1/\lambda_i + \beta' V_2 V_2' \beta$
R	$\sigma^2 V[\Delta^2+kI]^{-2}\Delta^2 V' + k^2(X'X+kI)^{-1}\beta\beta'(X'X+kI)^{-1}$	$\sum_{i=1}^{r} \dfrac{\sigma^2\lambda_i + k^2\delta_i^2}{(\lambda_i + k)^2}$
GR	$\sigma^2(\Delta^2+K)^{-2}\Delta^2 + (\Delta^2+K)^{-1}K\delta\delta'K(\Delta^2+K)^{-1}$	$\sum_{i=1}^{r} \dfrac{\sigma^2\lambda_i + k_i^2\delta_i^2}{(\lambda_i + k_i)^2}$
JR	$\sigma^2[I-[kA^{-1}]^2]^2\Delta^{-2} + [kA^{-1}]^2\delta\delta'[kA^{-1}]^2$	$\sum_{i=1}^{r} \left\{ \dfrac{\sigma^2\lambda_i(\lambda_i+2k_i)^2 + k_i^4\delta_i^2}{(\lambda_i+k_i)^4} \right\}$
JGR	use II.1 and II.2	
SH	$\sigma^2 d^2 \sum_{i=1}^{r} v_i v_i'/\lambda_i + (1-d)^2\beta\beta'$	$\sigma^2 d^2 \sum_{i=1}^{r} 1/\lambda_i + (1-d)^2\beta'\beta$
FPC	$\sigma^2 F\Delta^{-2}F + [I-F]\delta\delta'[I-F]$	$\sigma^2 \sum_{i=1}^{r} f_i^2/\lambda_i + \sum_{i=1}^{r} (1-f_i)^2\delta_i^2$

Table 2.4 Choices of k, K, and d

Estimator	Estimated value
RHK	$k = r\hat{\sigma}^2/\hat{\beta}'\hat{\beta}$
RLW	$k = r\hat{\sigma}^2/\sum_{i=1}^{r}\lambda_i\hat{\delta}_i^2$
GRHK	$\hat{k}_i = \hat{\sigma}^2/\delta_i^2$
GRT	$\hat{k}_i = \lambda_i/(F_i+1), \quad F_i = {}'\lambda_i\hat{\delta}_i^2/\hat{\sigma}^2$
JR	$k = p\hat{\sigma}^2/(\sum_{i=1}^{r}[\hat{\delta}_i^2/\{1+\sqrt{1+\lambda_i(\hat{\delta}_i^2/\hat{\sigma}^2)}\}$
SH*	$d = \hat{\beta}'\hat{\beta}/\{tr([var(\hat{\beta})] + \hat{\beta}'\hat{\beta}\}$

*an estimated value that minimizes the TMSE of the SH (see II.3)

2.5 Summary

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

Chapter 3

L_p ESTIMATION IN LINEAR REGRESSION

OLS performs badly in the presence of error terms that are not distributed as Normal (Gaussian) and where massive tails are present in the distribution of the error terms (Sposito and Tveite (1986)). Poor performance also arises even when the long-tailed distribution is assumed to be symmetric about its mean. We seek a robust estimator (eg an estimator that is relatively insensitive to departure from normality and massive tails in the distribution).

Alternatives to OLSE include the family of L_p-norm estimators, in which we have OLSE when $p = 2$. We examine other values of p in the hope that they will be more robust than either L_2 (OLS) or the biased estimators introduced in Chapter 2, in the presence of collinearity.

This chapter consists of a survey of L_p estimation. In §3.1 the general L_p-estimator is defined, with particular attention to L_1 estimation, L_2 estimation, and L_∞ estimation. It is of interest to note that these estimation procedures are each two centuries old. Finally we examine the case when p is an arbitrary value other than 1, 2 or ∞. A general introduction to the statistical properties of X-linear L_p-norm estimators is presented in §3.2.

Sensitivity to outliers increases as p moves upwards from unity through 2 towards ∞. Large values of p may be appropriate for distributions in which the tails are highly attenuated.

The device of estimating p from the data amounts to an exploration of alterations to the tails of the error distribution in order to downweight the influence of residuals that are large in some sense.

3.1 Definitions

Consider model (1.1) of Chapter 1:

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

In general, the coefficients (β) may be estimated by minimizing the sum of the p-th powers of the absolute deviation of the estimated values from the observed values of the dependent variables. The X-linear L_p-norm estimation problem is to:

Find an estimate of β, denoted by $\hat{\beta}_{L_p}$ which minimizes

$$\sum_{i=1}^{n} |Y_i - X_i\hat{\beta}_{L_p}|^p = \sum_{i=1}^{n} |\hat{\epsilon}_i|^p \tag{3.1.1}$$

with Y_i the i-th element (row) of Y:nx1 and X_i is the i-th row of X:n×r. The term $\hat{\epsilon}_i$ denotes the i-th residual in the vector of residuals $\hat{\epsilon}$. Note that the notation $\hat{\epsilon}$ will be used for any value of p, because the value of p will be clear from the context.

Let $\hat{\epsilon}_i$ be written as $\hat{\epsilon}_i = u_i - v_i$, where $u_i, v_i \geq 0$, represent the positive and negative deviations respectively. Then the general L_p-norm problem (3.1.1) can be formulated as a mathematical programming problem, namely

$$\text{Minimize} \quad \sum_{i=1}^{n} (u_i^p + v_i^p) \tag{3.1.2}$$

$$\left.\begin{array}{l} \text{subject to} \quad X_i\hat{\beta}_{L_p} + (u_i - v_i) = Y_i \\ \qquad\qquad\quad u_i, v_i \geq 0 \end{array}\right\} \qquad i = 1,\ldots n$$

$$\hat{\beta}_{L_p} \quad \text{unconstrained}$$

The justification for studying L_p-norm estimation lies in a generalized theory of errors that encompasses least square error theory. The distribution of the error terms is presumed to determine an optimal value of p and thus the effectiveness of L_p-norm estimation.

The following theorem (Kiountouzis (1971), Barr (1981)) shows a connection between L_p-norm estimation and MLE-estimation in a subfamily of the Exponential family of distributions.

Theorem 3.1.1 In the general model (1.1), if the error terms satisfy the following conditions:

(i) the errors ϵ_i are contained only in the measurement of Y_i (ie X_i measured without error),

(ii) ϵ_i and ϵ_j $(i \neq j)$ are mutually independent

(iii) the p.d.f. of ϵ_i, for i = 1,2,...,n is

$$f(\epsilon_i) = c \exp\{-h|\epsilon_i|^p\} \qquad (3.1.3)$$

where c, h are constants and $1 \leq p \leq \infty$, and

(iv) no other information is available concerning the coefficients β_j

then the optimal (maximum likelihood) estimate of β is obtained when (3.1.1) is a minimum.

Solutions $\check{\beta}$ of the MLE equations for the vector β in the r-dimensional Euclidean space R^r are then given by

$$\{\check{\beta}: \sum_{i=1}^{n} |Y_i - X_i\check{\beta}|^p \leq \sum_{i=1}^{n} |Y_i - X_i\check{\beta}|^p, \quad \forall \ \check{\beta} \in R^r \}$$

It is clear that each $\check{\beta}$ is a L_p-norm estimator and under appropriate conditions on X, this set is a singleton. When p = 1, (3.1.3) reduces to a Laplace distribution. Thus the L_1-norm estimator is optimal when the underlying distribution of the error terms is Laplace. The least squares

(L_2) estimator is optimal when the errors are distributed as Gaussian normals (p = 2, in (3.1.3)). In §3.4 it will be shown that the L_∞-norm estimator is optimal (maximum likelihood) when the errors are distributed uniformly.

3.1.1 L_1 estimation

The 1-norm (see Chapter one) of a vector ϵ (the vector of residuals), is

$$\|\epsilon\|_1 = \sum_{i=1}^{n} |\epsilon_i| .$$

If $\hat{\beta}_{L_1}$ is the L_1-norm estimator of β, (3.1.1) can be written as

$$\sum_{i=1}^{n} |Y_i - X_i\hat{\beta}_{L_1}| = \sum_{i=1}^{n} |\hat{\epsilon}_i| \qquad (3.1.4)$$

We minimize the sum of the absolute error terms to develop an estimator (MSAE estimator). Other names for MSAE criterion are least absolute value (LAV), minimum absolute deviation (MAD), minimum absolute error (MAE), least absolute deviation (LAD), least sum of absolute errors (LSAE), and L_1-norm estimation. In this thesis we will use the terms L_1-norm or MSAE estimation.

The first known reference to MSAE estimation dates back some two hundred and fifty years. Sometime between 1755 and 1757, Boscovich formulated and applied the minimum sum of absolute errors criterion for obtaining an optimal fitting line. In general, to be computationally practicable, this L_1-norm approach requires a technology that became available 200 years later. Linear programming was a development of the late 1940's and early 1950's. Statisticians ignored MSAE until Charnes *et al.* (1955) introduced linear programming as a procedure to produce a MSAE regression solution.

Nearly 50 years after Boscovich's minimum sum of absolute errors, in 1805, least squares was introduced by Legendre (and by Gauss in 1806).

In subsequent years, and as statistical theory developed, the relative computational ease of least squares, its uniqueness when applied to models of full rank and its attractive properties (eg known distribution for $\hat{\beta}$) made hypothesis testing simple and convenient under Gaussian distribution assumptions.

3.1.1.1 Primal and Dual

When p = 1 the objective function for a linear programming (LP) problem is to minimize

$$\sum_{i=1}^{n} (u_i + v_i) = 1_n'u + 1_n'v$$

$$\text{subject to} \quad X\hat{\beta}_{L_1} + I(u - v) = Y$$

$$u_i, v_i \geq 0 \quad i = 1,\ldots,n$$

$$\hat{\beta}_{L_1} \quad \text{unconstrained}$$

where $u = [u_1,\ldots u_n]'$ and $v = [v_1,\ldots v_n]'$. Thus the LP problem has n equations in r + 2n unknowns $(\hat{\beta}_{L_1}, u, v)$, with 2n non-negativity constraints.

Any well formulated LP problem possesses an equivalent dual problem. Denote the dual vector by $d = [d_1\ldots d_n]'$, where d_i is the i-th dual variable, then the dual of (3.1.4), when the matrix X involves a vector of ones for the intercept in the model, is given by Wagner (1959) as

$$\text{Maximize} \quad \sum_{i=1}^{n} d_i y_i = Y'd$$

$$\text{subject to} \quad -1 \leq d_i \leq 1, \quad i = 1,\ldots,n \qquad (3.1.5)$$

$$\sum_{i=1}^{n} d_i = 0$$

$$\sum_{i=1}^{n} d_i x_{ij} = 0 \quad j = 2,\ldots,r$$

Wagner (1959) set $w_i = d_i + 1$, to obtain dual variables that are non-negative. Thus (3.1.5) then becomes

$$\text{Maximize} \quad \sum_{i=i}^{n} w_i y_i$$

$$\text{subject to} \quad 0 \leq w_i \leq 2, \quad i = 1,\ldots,n \qquad (3.1.6)$$

$$\sum_{i=1}^{n} w_i = n$$

$$\sum_{i=1}^{n} w_i x_{ij} = \sum_{i=1}^{n} x_{ij} \quad j = 2,\ldots,r$$

3.1.1.2 Algorithms

Since Charnes *et al.* (1955) a great deal of work has been done in developing specialized LP algorithms for the L_1-norm estimation problem. For reference to the development of LP algorithms from 1960 to 1970 the interested reader is referred to Lawrence (1979) or the useful annotated bibliography by Dielman (1984).

The most successful and widely used algorithms are those of Barrodale and Roberts (1973, 1974). Their algorithms are modifications of the simplex method applied to the primal formulation of the LP problem.

Armstrong *et al.* (1979) modified the Barrodale and Roberts approaches, providing an algorithm that uses the revised simplex method with the principle of LU decomposition (Golub and Van Loan (1983)) in maintaining the current basis.

According to Bloomfield and Steiger (1983), the three best algorithms available are those of Barrodale and Roberts (1974), Bartels *et al.* (1978) and their own algorithm. Bloomfield and Steiger (1983) give an extensive discussion of the algorithms and compared their CPU times and the iterations necessary to obtain an MSAE estimate, in a simulation study performed over a variety of n and r values. As sample size increases, the algorithm of

Bloomfield and Steiger gains relative advantage over the Barrodale and Roberts algorithms.

LP formulation is not the only technique for finding an L_1-norm estimator: Schlossmacher (1973) proposed an iterative weighted least squares algorithm to obtain L_1-norm estimators.

In this study the algorithm of Barrodale and Roberts (1974) will be used to obtain the L_1-norm estimator, because the algorithm is widely used and because the FORTRAN code is published. For more discussion and development on algorithms the reader is referred to Bloomfield and Steiger (1983) and Gonin and Money (1989, p25-26).

3.1.1.3 Geometric properties

Some properties of L_1-estimation arising from the LP formulation are given by Appa and Smith (1973), Gentle *et al.* (1977) and Kiountouzis (1971). A few of these properties are:

1. There exists at least one L_1 hyperplane passing through m (where $m = r(X)$) of the n observations. When X is of full column rank r then the hyperplane will pass through r of the observations.

2. Let n^+ and n^- be the number of observations above and below the hyperplane respectively, and n^* the maximum number of observations that lie on any hyperplane. Then

$$|n^+ - n^-| \leq n^*$$

3. Multiple optimal solutions can occur, ie two or more different hyperplanes may give the same MSAE for β.

4. Variations in Y do not change the optimal values of the coefficients as long as no change in an observation causes it to cross the optimal hyperplane. Thus theoretically the L_1-estimator is resistant to wild points.

5. Linear dependence among the independent regressor variables will not cause any failures in the estimation procedure.

Statistical properties of the L_1-norm estimator will be discussed in §3.2

3.1.2 L_2 estimation

When p = 2 the L_2-norm estimator is the OLS estimator, $\hat{\beta}$, namely

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

The computation of $\hat{\beta}$ is in principle easy (eg using SVD or Gaussian elimination), and gives the best linear unbiased estimator. Under normality it gives the MLE (Gauss-Markoff theorem). In practice numerical problems arise when $(X'X)^{-1}$ is not well approximated computationally. Its statistical tractability has made least squares the most common method of estimation. The OLS estimator and its properties were discussed in Chapter one.

3.1.3 L_∞ estimation

The ∞-norm of a vector ϵ is defined as

$$\|\epsilon\|_\infty = \max\{|\epsilon_i|,\ i=1,2,\ldots,n\}$$

Let $\hat{\beta}_{L_\infty}$ denote the L_∞-norm estimator of β, then (3.1.1) can be written as

$$\hat{\beta}_{L_\infty} = \min_\beta\{\max_i|Y_i - X_i\beta|,\ i=1,2,\ldots,n\} = \min_\beta \|Y_i - X_i\beta\|_\infty \qquad (3.1.7)$$

Thus we minimize the maximum absolute error. Other names for L_∞-norm estimation are max-norm, Chebychev approximation (norm) or uniform norm estimation.

Laplace proposed the procedure (3.1.7) in 1799 and it was studied in detail by P.L. Chebychev (whence the alternative name). An account of the theory

of Chebychev approximation can be found in Rice (1964), Hand (1978) and Watson (1980).

In §3.1 it was shown that the L_1-norm estimator is optimal in the sense of MLE when the error terms are from a Laplace distribution, and OLS is optimal in the sense of MLE when the error terms are normally distributed. The Chebychev estimator is optimal in the sense of MLE when the error terms are from a Uniform distribution, as proved by the following theorem (Hand (1978)):

Theorem 3.1.2: In the model $Y = X\beta + \epsilon$ assume the errors ϵ_i are contained only in the measurement of Y_i, and the ϵ_i's are distributed uniformly on some symmetric interval $[-a,a]$ where a is fixed but unknown, and ϵ_i and ϵ_j are mutually independent for $i \neq j$. Then the L_∞-norm estimator $\hat{\beta}_{L_\infty}$ of the unknown parameter vector β is the maximum likelihood estimator of β.

Proof: The complete proof is given in Hand (1978) but may be summarised as follows:

By assumption, the p.d.f of ϵ_i is

$$f(\epsilon_i) = \frac{1}{2a} \qquad \forall\ i = 1,2,\ldots,n,\ \epsilon_i \in [-a,a]$$
$$= 0 \qquad \text{otherwise}$$

The quantity $(X\beta)_i = \sum_{j=1}^{r} x_{ij}\beta_j$ will be constant with respect to ϵ_i, thus

$$Y_i \sim U[(X\beta)_i - a, (X\beta)_i + a] \quad \forall\ i$$

$$f(Y_i; (X\beta)_i, a) = \frac{1}{2a} \quad \text{for } ((X\beta)_i - a) \leq Y_i \leq ((X\beta)_i + a)$$

$$= \frac{1}{2a} \qquad |Y_i - (X\beta)_i| \leq a \text{ or } |\epsilon_i| \leq a$$
$$= 0 \qquad \text{otherwise.}$$

Define $\varphi(\epsilon_i, a) = 1$ if $|\epsilon_i| \leq a$
$= 0$ otherwise

then the likelihood function of the Y_i is

$$L((Y_i;(X\beta)_i, a) = (2a)^{-n} \prod_{i=1}^{n} \varphi(\epsilon_i, a)$$

The likelihood function can be maximized by minimizing a subject to the constraint that the maximum absolute residual is less than or equal to a, which is the Chebychev criterion for minimizing the absolute error. Thus the maximum likelihood estimator is the Chebychev estimator.

By a single application of the CHEB algorithm, the Chebychev estimator, as the solution to an LP problem, is found iteratively inside the CHEB algorithm. The LP solution of the CHEB estimator and all other L_p-norm estimators presents computational difficulties and also produces severe obstacles to the development of a distribution theory for all values of p, except p = 2. Farebrother (1985) proves that the CHEB estimator is unbiased under certain conditions, see section 3.2.1.

3.1.3.1 Primal and Dual

The Chebychev estimator can be formulated as an LP problem, in the manner of Wagner (1959), Appa and Smith (1973) and Sposito (1976). If the maximum absolute residual is denoted by D (so that $D \leq a$ in theorem 3.1.2) then the primal of the LP problem is

Minimize D

$$\text{subject to } D \geq Y_i - \sum_{j=1}^{r} x_{ij}\beta_j \quad \forall\ i = 1,\ldots,n \qquad (3.1.8)$$

$$D \geq -Y_i + \sum_{j=1}^{r} x_{ij}\beta_j \quad \forall\ i = 1,\ldots,n$$

where $D \geq 0$ and β_j unconstrained $j = 1,\ldots,r$

The primal (3.1.8) involves 2n constraints in r variables. By defining the following matrices

$$k = [0 \ldots 0\ 1]' : (r+1)x1 \text{ vector}$$
$$\theta = [\beta_1 \ldots \beta_r\ D]' : (r+1)x1$$
$$c = [-Y_1\ -Y_2 \ldots -Y_n, Y_1 \ldots Y_n]' : 2nx1$$
$$1_n = [1 \ldots 1]' : nx1$$
$$A = \begin{bmatrix} -X & 1 \\ X & 1 \end{bmatrix} : 2nx(r+1)$$

the matrix notation of Hand (1978) for (3.1.8) is:

Minimize $k'\theta$

subject to $A'\theta \geq c$

θ unrestricted

In this notation the dual can be formed easily by using the rules of Sposito (1975). The rules are: transpose the constraint matrix A′, change minimization to maximization, interchange k and c, reverse the inequality sign in the constraints, and if the i-th variable (θ_i) is unrestricted in sign then the dual constraint is an equality. Thus the dual problem is

Maximize $c'f$

subject to $Af = k$

$f \geq 0$

where f:2nx1 is a vector containing the dual variables. If f is partitioned as $d' = [z'\ w'] = [z_1\ z_2 \ldots z_n\ w_1 \ldots w_n]$ then a more descriptive notation for the dual is

$$\text{Maximize } \sum_{i=1}^{n} Y_i(w_i - z_i)$$

$$\text{subject to } \sum_{i=1}^{n} (w_i + z_i) = 1$$

$$\sum_{i=1}^{n} (w_i - z_i) = 0$$

$$\sum_{i=1}^{n} x_{ij}(w_i - z_i) = 0 \qquad j = 2, \ldots . r$$

$$w_i, z_i \geq 0 \qquad i = 1, \ldots, n$$

3.1.3.2 Algorithms

Various algorithms for Chebychev estimation are available, and most of them work on the simplex algorithm of LP. Barrodale and Phillips (1974 and 1975) proposed a three-stage algorithm for the solution of the dual in Chebychev estimation and they published a FORTRAN subroutine CHEB (Barrodale and Phillips (1975)). This code (CHEB) will be used in this thesis.

Hand (1978) developed a gradient projection algorithm for the Chebychev approximation. This method is built around the algorithm of Barrodale and Phillips with the following changes: the OLS estimator is used as a starting value in the algorithm, gradient projection pivot selection in stage one and maximum increase pivot selection in stage three. Hand's (1978) results were not totally conclusive but his algorithm is worth mentioning and the examples given by him, showing the tableaux for both methods at each iteration, are quite informative.

Further algorithms can be found in Armstrong and Kung (1980), and a FORTRAN subroutine RLMV is also available in the IMSL library.

3.1.3.3 Geometric properties

From the LP formulation of the Chebychev estimation, various geometric properties have been derived by authors such as Appa and Smith (1973), Kiountouzis (1971) and Hand (1978):

1. There exists at least one optimal hyperplane which is vertically equidistant, at a distance D, from at least r+1 observations.

2. The r+1 observations that determine the optimal Chebychev hyperplane must lie in the convex hull of the n rows of X in R^r.

An important consequence of this last property is that the Chebychev estimator is determined by the convex hull of the observations, ie the estimator is largely determined by the most extremal observations. The

Chebychev estimator is therefore extremely sensitive to outliers, and more sensitive than L_2 estimators. Residuals should be checked for approximate uniformity before accepting a Chebychev estimator, if an MLE criterion is the basis of its choice.

3.1.4 L_p estimation

In the previous sections specific values of p were investigated. There is no theoretical reason why values of p other than 1, 2, and ∞ should not be considered. In the case p = 1,2 and ∞ exact solutions exist, while other values of p each give rise to a non-linear programming problem whose solution can only be found to a given level of convergence.

3.1.4.1 Primal and dual

In (3.1.2) the L_p-norm problem was formulated as a NLP problem. The primal can be written as (3.1.2) and an equivalent formulation is:

$$\begin{aligned} &\text{Min} \sum_{i=1}^{n} (u_i^p + v_i^p) \\ &\text{subject to } X\hat{\beta}_{L_p} + Iu - Iv = Y \qquad (3.1.9) \\ &\qquad u,v \geq 0 \\ &\qquad \hat{\beta}_{L_p} \text{ unconstrained} \end{aligned}$$

The dual of (3.1.2) and (3.1.9) is:

$$\begin{aligned} &\text{Max } Y'd \\ &\text{subject to } -1 \leq d_i \leq 1, \quad i = 1,\ldots,n \qquad (3.1.10) \\ &\qquad \sum_{i=1}^{n} d_i = 0 \\ &\qquad X'd = 0 \end{aligned}$$

3.1.4.2 BFGS and other algorithms

The L_p-norm estimation problem can be formulated (Barrodale and Roberts (1970)) as a non-linear programming (NLP) problem. The objective function is concave for $0 < p < 1$ and convex for $1 \leq p < \infty$ with the constraints being linear. They suggested for $p > 1$ the convex simplex or Newton's method, and for $p < 1$ a modification of the simplex method for LP.

Ekblom (1974) suggested the following algorithms for various p values:

1. For $1 < p < 2$ use the damped Newton method. A problem with this method is that zero residuals are encountered. In the second derivative of the L_p function a zero residual is raised to a negative power and will cause the iteration to terminate prematurely. Therefore Abdelmalek (1971) and Forsythe (1972) applied the Davidson-Fletcher-Powell method, while Ekblom (1973) introduced a perturbation of the problem to avoid zero residuals. Ekblom (1974) rewrites (3.1.1) as

$$\sum_{i=1}^{n} [\hat{\epsilon}_i^2 + c^2]^{\frac{1}{2}p} \qquad (3.1.11)$$

   This form ensures that the second order derivative remains positive definite as long as $c \neq 0$, and that the (damped) Newton method to solve (3.1.11) does not terminate prematurely. Ekblom (1973) then showed that as $c \to 0$, the solution of (3.1.11) is the solution to the original problem (3.1.1). Algorithms for this technique can be found in Ekblom (1973).

2. For $0 < p < 1$ Ekblom suggested the algorithms of Barrodale and Roberts (1970) and Henriksson (1972).

Barr (1981) used the algorithms described by Fletcher and Powell (1963) and Forsythe (1972), which form part of the IBM Scientific Subroutine Package (1968).

Gonin and Money (1989) described a large selection of the algorithms available:

1. Fisher (1981) transformed (3.1.1) to a "linearly constrained" problem and use a constrained version of Newton's method.

2. Merle and Späth (1973) used an iteratively reweighted least-squares algorithm, which set zero residuals equal to a small positive constant.

3. Schlossmacher (1973) also used the iteratively reweighted least squares technique. His method deletes observations with zero residuals and in following iterations reintroduces them as the residuals become larger.

4. Sposito *et al.* (1977) extended the Schlossmacher method for $1 \leq p \leq 2$. The authors consider minimizing

$$I = \sum_{i=1}^{n} W_i R_i^2$$

where the R_i are the residuals and W_i are weighting factors. In the (m+1)-th iteration the above can be written as

$$I(m+1) = \sum_{i=1}^{n} (1/|R(m)_i|^{2-p}) \{R(m+1)_i\}^2$$

5. Barr (1981) extended Sposito *et al.* (1977) method to the r dimensional case (to handle more than one X variable) and compared this method (weighted least squares (WLS)) with that of Fletcher and Powell. Barr suggested that WLS should be used for $1.0 < p \leq 2.6$, and that when p has values greater than 2.6, the Fletcher and Powell (sometimes referred to as Davidson, Fletcher and Powell (DFP)) algorithm should be used. For $p \geq 3$ the WLS method will not necessarily converge.

6. **Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm:**

The basic idea of the quasi-Newton (or variable metric) method is explained in Jacobs (1977, p 237) and Press et. al (1987). The aim is to build up, iteratively, an approximation to the inverse Hessian matrix.

Notation: The function to be minimized is $F(\mathbf{x})$, where $\mathbf{x}$:nx1 is a vector with elements x_i, i = 1..n. The gradient vector $\mathbf{g}$:nx1 is defined as

$$g_i = \frac{\partial F}{\partial x_i}$$

and the Hessian matrix G:nxn as

$$G_{ij} = \frac{\partial^2 F}{\partial x_i \partial x_j}.$$

When the second derivatives are not available the quasi-Newton method builds an iterative approximation to the inverse Hessian matrix G^{-1}. This approximation is denoted by H^i, where the superscript i denotes the i-th iteration, with the property that

$$\lim_{i \to \infty} H^i = G^{-1}$$

Let $\quad g^{i+1} - g^i = \gamma^i$:nx1 and $x^{i+1} - x^i = \delta^i$:nx1

The first quasi-Newton method was suggested by Davidson (1959) and later modified by Fletcher and Powell (1963). They suggested the Davidson-Fletcher-Powell (commonly known as the DFP) updating formula

$$H^{i+1} = H^i + \frac{\delta^i \delta^{i\prime}}{\delta^{i\prime} \gamma^i} - \frac{H^i \gamma^i \gamma^{i\prime} H^i}{\gamma^{i\prime} H^i \gamma^i}$$

The DFP formula has the property that the matrices H^i are all positive definite, provided that H^0 (the starting matrix) is positive definite. Thus the search direction is downhill.

Broyden (1967) was the first to point out that the DFP formula is one of a family of formulae which share the same properties. This one-parameter family is (Jacobs (1977))

$$H^{i+1} = H^i + \frac{\delta^i \delta^{i\prime}}{\delta^{i\prime} \gamma^i} - \frac{H^i \gamma^i \gamma^{i\prime} H^i}{\gamma^{i\prime} H^i \gamma^i} + \phi^i [\bullet] \qquad (3.1.12)$$

$$[\bullet] = (\gamma^{i\prime} H^i \gamma^i) \left\{ \frac{\delta^i}{\delta^{i\prime} \gamma^i} - \frac{H^i \gamma^i}{\gamma^{i\prime} H^i \gamma^i} \right\} \left\{ \frac{\delta^i}{\delta^{i\prime} \gamma^i} - \frac{H^i \gamma^i}{\gamma^{i\prime} H^i \gamma^i} \right\}'$$

The conditions to ensure that the H-matrix remains positive definite can be found in Jacobs (p240, 1977).

When $\phi^i = 0$, (3.1.12) reduces to the DFP formula. The choice $\phi^i = 1$ was suggested independently by Broyden (1970), Fletcher (1970), Goldfarb (1970) and Shanno (1970) and is therefore known as the BFGS algorithm. In practice the BFGS algorithm seemed preferable to the DFP algorithm.

However Dixon (1972) shows that each member of the family of updating formulae (3.1.12) generates the same sequence of points (x^i) when minimizing a general function F. This result implies that any differences (where the difference is picked up by the constant ϕ^i) could be attributed to inaccuracy in the line search and to rounding errors.

In view of the remarks of Dixon (1972), in this thesis the BFGS updating formula will be used because it is more tolerant of inexactitude in the line minimization than the DFP updating formula, and the user friendly FORTRAN subroutines can be found in Press *et al.*(1987).

Brodlie (1977, in Jacobs(1977)) notes: "The trend away from exact line searches has shown serious deficiencies in the DFP formula. It is now universally recognised that the BFGS formula is superior. Indeed it seems hard to find any formula in the Broyden family which performs better. To my mind, the reasons for this are not fully understood. For example, most explanations of the superiority of the BFGS formula to the DFP formula have been in the context of exact line searches (Powell (1971)). Yet it is as the line searches become less accurate that the supremacy of the BFGS formula is more noticeable."

In this thesis we will use the Barr (1981) WLS technique as well as BFGS algorithms. The effect of both techniques will be reported in appendices C and E, and will be discussed in Chapter 5.

3.1.4.3 Choice of p

When fitting a regression plane one seeks the optimal fit, where optimal will be defined in terms of one or other criterion (eg smallest MSE). In minimizing the p-norm the L_p-norm estimator is found. The OLS estimator (p=2) is the BLUE, but p≠2 yields estimators that are not linear and hence can have lower variance than those obtained using OLS. How do we find a value of p that will yield the optimal plane or a plane near the optimal plane (meaning that there could be a set (range) of optimal planes)?

One approach would be to fit the whole range of p values in order to find the unique p that satisfies a more general set of optimality criteria. In practice this approach would not be feasible and one should rather examine the guidelines given in the literature.

Forsythe (1972) suggested that in the case where the errors have a Gaussian distribution, one should use p = 2 (OLS). For Contaminated normal or skewly distributed error distributions he suggest, p = 1.5, as a "good compromise".

Ekblom (1974) found that for the Contaminated normal distribution the L_p-norm is inferior to the Huber (1964, 1972) estimator which, according to Ekblom, is a "mixed L_p-estimate", since it uses L_2 in the middle and L_1 in the tails.

For the Laplace and the Cauchy he suggested p = 1.25. For error densities with very "long" tails and skewly distributed (such as χ^2) he suggests $p \leq 1$. In contradiction, Rice (1964) remarks that $p < 1$ is not of interest.

Harter (1977) proposed an adaptive procedure which depends on the kurtosis, κ, of the regression error distribution. He suggested

$$\begin{array}{lll} p = 1 & \text{for} & \kappa > 3.8 \\ p = 2 & \text{for} & 2.2 \leq \kappa \leq 3.8 \\ p = \infty & \text{for} & \kappa < 2.2 \end{array} \qquad (3.1.13)$$

Barr (1981) showed that the L_p estimates are unbiased for all $p \geq 1$ (see §3.2.1), He therefore based the choice of p on the empirical generalized variance. He defined the empirical generalized variance of the regression coefficient estimates as the determinant of the empirical covariance matrix of their estimator and seeks that p which yields the smallest possible generalized variance.

In a simulation study conducted over a wide range of distributions with kurtosis ranging from 1.8 to ∞, Barr found the optimal p (eg the p that give the minimum generalized variance) and plotted this p against the theoretical kurtosis values. From this plot an approximate functional relationship between p and κ (kurtosis of error distributions) emerges:

$$p = 9/\kappa^2 + 1 \qquad (3.1.14)$$

Formula (3.1.14) has the advantage that it precludes any ambiguity in the choice of the L_p-norm.

In a second simulation study Barr showed that the population kurtosis κ could be estimated by the kurtosis based on the sample data, $\hat{\kappa}$ (see the notes at the end of this section on kurtosis). Then (3.1.14) becomes

$$p = 9/\hat{\kappa}^2 + 1 \qquad (3.1.15)$$

Examining the empirical generalized variance of the corresponding estimates, Barr concluded that the results obtained using this formula (3.1.15) are generally superior to the adaptive procedure (3.1.13) of Harter (1977).

On the basis of the study Barr proposed the following steps in fitting an L_p-norm estimator:

> Obtain a set of residuals by fitting OLS, use these residuals, and compute $\hat{\kappa}$, and hence p. Use this p, fit L_p, then from the resulting residuals compute $\hat{\kappa}$, obtain a new p, fit L_p.

Sposito *et al.* (1983) give the following guidelines with respect to p:

> For small sample size, the formula (3.1.14) yields a reasonable value of p for distributions with a finite range. For long-tailed distributions a large sample ($n \geq 200$) is needed to identify an optimal $p \in [1,2]$. A value for p which is reasonably close to the optimal p value is
>
> $$p = 6/\kappa \qquad (3.1.16)$$

Finally Sposito *et al.* suggest the following modification to the proposals of Harter (1977):

$$\begin{array}{lll} p = 1.0 & \text{for} \quad \kappa \geq 6 & \text{(implicitly)} \\ p = 1.5 & \text{for} \quad 3 < \kappa < 6 & \text{(Forsythe's compromise rule)} \\ p = 2 & \text{for} \quad 2.2 \leq \kappa \leq 3 & \\ p = \infty & \text{for} \quad \kappa < 2.2 & \end{array} \qquad (3.1.17)$$

Gonin and Money (Chapter 5, 1989) consider the choice of p in the non-linear case (also applicable in the linear case). They consider formulae (3.1.15), (3.1.16) and a theoretically sound but possibly impractical approach. Gonin and Money (1985) based their choice of estimator on the p-th order Exponential distribution with density function

$$f(y) = \frac{2^{-(1+1/p)}}{\phi\,\Gamma(1+1/p)} \exp\{-\tfrac{1}{2}|(y-\theta)/\phi|^p\} \qquad -\infty < y < \infty \qquad (3.1.18)$$

where ϕ (> 0) is a scale parameter and θ ($0 < \theta < \infty$) a location parameter. If the residual distribution belongs to this class of Exponential distributions, then the maximum likelihood estimate can be obtained by simply minimizing the sum of the p-th power of the absolute residuals or equivalently by maximizing the likelihood function over p. This fact was already stated in Theorem 1.1.1 where the distribution (3.1.18) was given as

$$f(\epsilon_i) = c \exp\{-h|(\epsilon_i)|^p\}$$

and it was shown that, for p=1, the distribution reduces to the Laplace and the L_1-norm estimator is the MLE; for p = 2, the distribution reduces to

the normal with the OLS estimator as the MLE. In theorem 1.1.2 it was proved that for the Uniform distribution the MLE is the Chebychev estimator.

Gonin and Money (1985) give the r-th central moment of (3.1.18) as

$$\mu_r = 2^{r/p}\phi^r\Gamma\{(r+1)/p\}/\Gamma(1/p) \text{ for r even}$$

The kurtosis κ is then given by

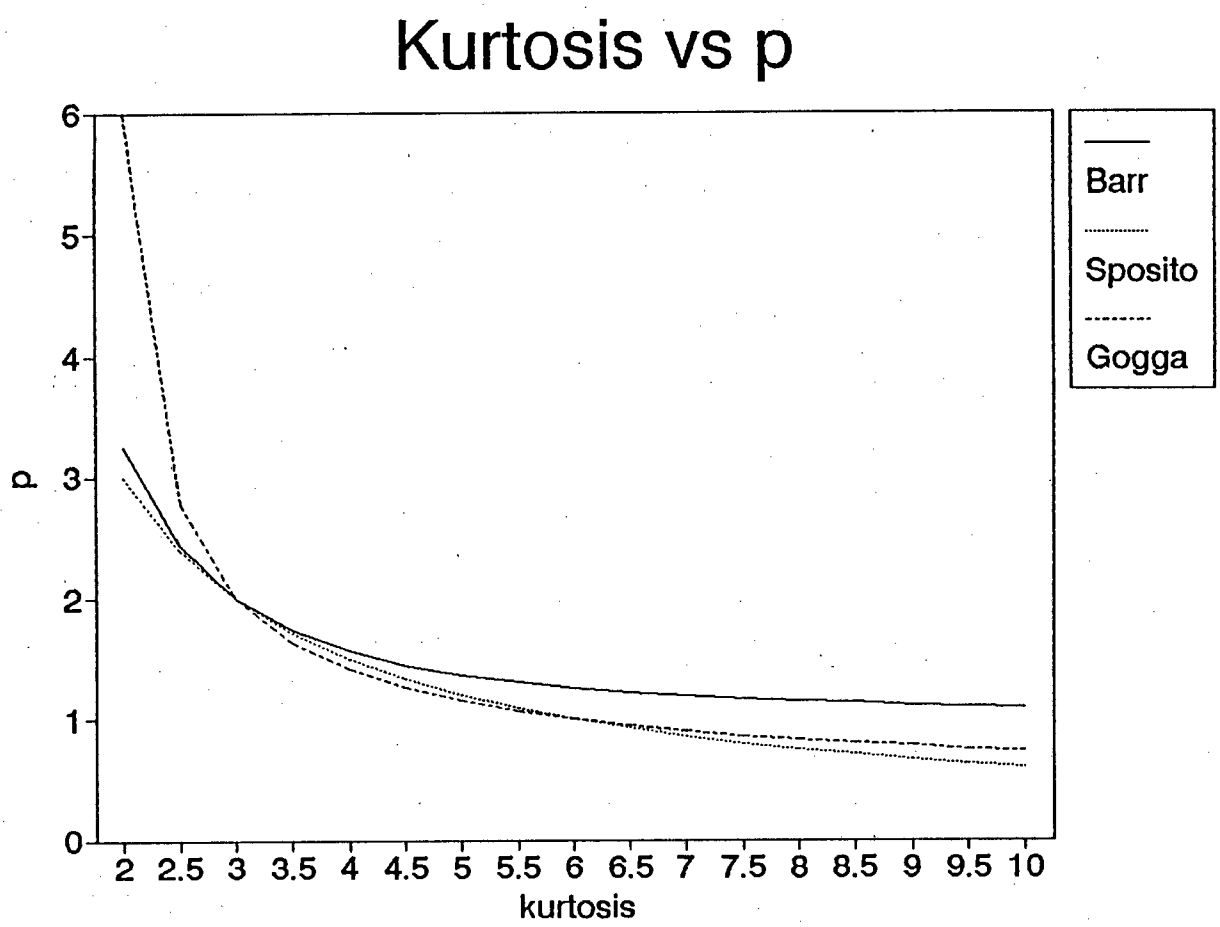
$$\kappa = \frac{\Gamma(5/p)\Gamma(1/p)}{\{\Gamma(3/p)\}^2} \qquad (3.1.19)$$

Use of (3.1.19) is theoretically sound as it yields a p that would insure a MLE of the betas, under these distributional assumptions. If one uses the sampling kurtosis in the equation as an estimate of κ, a method could be found to estimate p as a root of (3.1.19). Gonin and Money (1985) suggested the algorithm of of Steffenson (Dahlquist and Bjorck (1974:230)), but we use the algorithm known as Brent's method. This method was developed by van Wijngaarden, Dekker and others at the Mathematical Center in Amsterdam, and later improved by Brent (1973). Brent's method combines root bracketting, bisection, and inverse quadratic interpolation to converge from the neighbourhood of a zero crossing. The method is guaranteed to converge, so long as the function can be evaluated within the interval known to contain a root. With trial runs we set the interval for the root p between 0.1 and 30. When the kurtosis approaches 1.8 the root seems to move to infinity ($p > 30$) and the algorithm was unstable, so p was set to infinity, and the Chebychev-algorithm was called. The FORTRAN code for Brent's method can be found in Press et. al (1986). Finding the p in terms of (3.1.19) is referred to as the Gogga method.

Gonin and Money (1985) suggest the following adaptive procedure: Fit a curve using least squares (or any finite value of $p \geq 1$). Compute the sample kurtosis of the resulting residuals and make a prediction of the optimal exponent p using either formula (3.1.15), (3.1.16) or (3.1.19). Use this estimated value of p and fit a new curve to the data. Subsequently compute the sample kurtosis of the resulting residuals and make a new

prediction of the true exponent p. Repeat the process until no further change in the values of p is detected.

This method will be used in our simulation study, and we will use all three methods of estimating p (ie Barr, Sposito or Gogga), investigating the "goodness" of the resulting betas plus the rate of convergence. Gonin and Money found that the values of p converged in about 4 iterations. Theoretical convergence has not been proven and we suspect that the p value will not necessarily converge in expectation to a global solution (the true value of p for those distributional assumptions).



It is interesting to note, that if one plots the values (p, κ) of the three methods (Barr, Sposito and Gogga) the shapes of the curves are similar. By construction, they all intersect at the point p = 2, κ = 3. Furthermore it seems that for

$\kappa > 3$ Sposito will outperform Barr
$\kappa < 3$ Barr will be better than Sposito

on the basis of the distributional assumptions. In the following table we try to summarise the theoretical values suggested by various authors for the distribution(D): Uniform, Normal, Contaminated normals, Laplace, Student's t distribution with 5 degrees of freedom and Exponential. The values in the last column were found by a trail and error method using the subroutine GAMMLN from Press *et al.* (1986), and could be subject to a slight error margin.

Table 3.1.1: Theoretical values for p

D	κ	Barr	Sposito	Harter	Harter (modified)	Forsythe	Ekblom	Gonin and Money
U	1.8	3.78	3.33	∞	∞			30±
N	3.0	2.00	2.00	2	2	2		1.99<p<2.00
CN	4.0	1.56	1.50	1	1.5	1.5	Huber	1.40<p<1.50
CN	5.0	1.36	1.20	1	1.5	1.5	Huber	1.10<p<1.20
L	6.0	1.25	1.00	1	1	1.5	1.25	1
t_5	9.0	1.11	0.66	1	1	1.5		0.75<p<0.78
E	9.0	1.11	0.66̇	1	1	1.5		0.75<p<0.78

3.1.4.4 Kurtosis

3.1.4.4.1 Estimation and variance

All three formulae suggested for estimating p (Barr, Sposito, Gonin) are functions of the kurtosis ($\kappa = \mu_4/\mu_2^2$).

Barr (1981) gives an estimator $\hat{\kappa}$ of the kurtosis, applicable for observable error terms with a common mean as

$$\hat{\kappa} = 3 + k_4/k_2^2 \qquad (3.1.20)$$

Here k_r is the r-th k-statistic, an unbiased estimator of the r-th cumulant κ_r. The r-th cumulant is defined as

$$\kappa_r = \left|\left(\frac{\partial}{\partial t}\right)^r [\log M_Y(t)]\right|_{t=0}$$

where $M_Y(t)$ is the moment generating function. The values of k_r for r = 2 and 4 are given by Kendall and Stuart (1966) as

$$k_2 = \frac{n}{(n-1)} m_2$$

$$k_4 = \frac{n^2}{(n-1)(n-2)(n-3)} \{(n+1)m_4 - 3(n-1)m_2^2\}$$

where $m_r = \sum_{i=1}^{n} (y_i - \bar{Y})^r/n$ is the sample r-th central moment. When Y has a Normal distribution, k_2 and k_4 are independent, and $E(k_4/k_2^2) = 0$. In general, k_2 and k_4 are unbiased but are not independent, and $\hat{\kappa}$ will not be an unbiased estimator of κ.

An alternative estimator for κ is the ratio of sample moments

$$\hat{\kappa} = m_4/m_2^2 \qquad (3.1.21)$$

which is also in general biased.

Barr (1981) conducted a simulation study in which the average MSE of the estimators (3.1.20) and (3.1.21) was calculated for a set of variables with known mean zero. The study ranges over 4 distributions (Uniform, Normal, Contaminated normal and Laplace) and three sample sizes (10, 30 and 50). Results indicated that over the distributions studied the estimator (3.1.20) usually generates a smaller MSE. Thus we may prefer (3.1.20) over (3.1.21).

Gonin and Money (1989) use the unbiased estimates of the second and fourth order moments (Cramer (1946))

$$\hat{\mu}_2 = \sum_{i=1}^{n} (\hat{\epsilon}_i - \bar{\epsilon})^2/(n-1)$$

$$\hat{\mu}_4 = \frac{(n^2-2n+3)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} (\hat{\epsilon}_i - \bar{\epsilon})^4 - \frac{3(n-1)(2n-3)}{n(n-2)(n-3)} \hat{\mu}_2^2$$

to obtain in (3.1.21) an estimate of the kurtosis as $\hat{\mu}_4/\hat{\mu}_2^2$. In the simulation study that follows we adopt this method.

Sposito *et al.* (1983) use the analogue of (3.1.21)

$$\hat{\kappa} = \frac{n\Sigma(\hat{\epsilon}_i - \bar{\epsilon})^4}{\{\Sigma(\hat{\epsilon}_i - \bar{\epsilon})^2\}^2}$$

It must therefore be conceded that both these approaches effectively give the fitted residuals $\hat{\epsilon}_i$ the status of independent observations from a common distribution. The effect of correcting n for degrees of freedom lost in fitting the r explanatory variable in X is an open question.

Since the kurtosis plays such an important role in the estimation of p we examine the variance of the kurtosis. A large error margin in $\hat{\kappa}$, may lead to estimates of p that are absurd.

Cramer (1946, §27.7) finds that the variance of the estimated kurtosis, (3.1.21), is given by

$$\text{var}(m_4/m_2^2) = \frac{\mu_2^2\mu_8 - 4\mu_2\mu_4\mu_6 - 8\mu_2^2\mu_3\mu_5 + 4\mu_4^3 - \mu_2^2\mu_4^2 + 16\mu_2\mu_3^2\mu_4 + 16\mu_2^3\mu_3^2}{n\mu_2^6} \qquad (3.1.22)$$

where μ_r is the r-th central moment.

From Chapter 1 (§1.11) we may obtain the higher order central moments of the particular distributions:

(a) Uniform distribution

The central moment is $\mu_r = 0$ for r odd

$$\mu_r = \frac{(2b)^r}{2^r(r+1)} \quad \text{for r even, and } b = \sqrt{3}\sigma$$

$$= \frac{(\sqrt{3}\sigma)^r}{(r+1)}$$

thus $\mu_2 = \sigma^2$, $\mu_3 = 0$, $\mu_4 = 9\sigma^4/5$, $\mu_5 = 0$, $\mu_6 = 27\sigma^6/7$ and $\mu_8 = 81\sigma^8/9$

and $$\text{var}(m_4/m_2^2) = \frac{\mu_2^2\mu_8 - 4\mu_2\mu_4\mu_6 + 4\mu_4^3 - \mu_2^2\mu_4^2}{n\mu_2^6}$$

$$= \frac{\left[\frac{81\sigma^{12}}{9} - \frac{972\sigma^{12}}{35} + \frac{2916\sigma^{12}}{125} - \frac{81\sigma^{12}}{25}\right]}{n\sigma^{12}}$$

$$= \frac{1.3165714}{n}$$

(b) Normal distribution

The central moment is $\mu_r = 0$ for r odd

$$\mu_r = \frac{r!}{(\frac{1}{2}r)!} \frac{\sigma^r}{2^{\frac{1}{2}r}} \quad \text{for r even}$$

thus $\mu_2 = \sigma^2$, $\mu_3 = 0$, $\mu_4 = 3\sigma^4$, $\mu_5 = 0$, $\mu_6 = 15\sigma^6$, and $\mu_8 = 105\sigma^8$

and $$\text{var}(m_4/m_2^2) = \frac{\mu_2^2\mu_8 - 4\mu_2\mu_4\mu_6 + 4\mu_4^3 - \mu_2^2\mu_4^2}{n\mu_2^6}$$

$$= \frac{105\sigma^{12} - 180\sigma^{12} + 108\sigma^{12} - 9\sigma^{12}}{n\sigma^{12}}$$

$$= \frac{24}{n}$$

(c) Symmetric contaminated normal distribution

The central moment is $\mu_r = 0$ for r odd

$$\mu_r = \frac{r!}{(\frac{1}{2}r)!2^{\frac{1}{2}r}} [w_1\sigma_1^r + w_2\sigma_2^r] \quad \text{for r even}$$

$$= \frac{r!}{(\frac{1}{2}r)!2^{\frac{1}{2}r}} [\sigma_1^r + 9\sigma_2^r]/10$$

where the weights are chosen as 1/10 and 9/10 (§4.2.2.3).

thus $\mu_2 = [\sigma_1^2 + 9\sigma_2^2]/10$, $\mu_3 = 0$, $\mu_4 = 3[(\sigma_1^2)^2 + 9(\sigma_2^2)^2]/10$, $\mu_5 = 0$, $\mu_6 = 15[(\sigma_1^2)^3 + 9(\sigma_2^2)^3]/10$, and $\mu_8 = 105[(\sigma_1^2)^4 + 9(\sigma_2^2)^4]/10$.

When $\kappa = 4.0$, $\sigma_1^2 = \sigma^2[1 + \sqrt{3}]$ and $\sigma_2^2 = \sigma^2[1 - \sqrt{1/27}]$ and

$\mu_2 = \sigma^2$, $\mu_3 = 0$, $\mu_4 = 4\sigma^4$, $\mu_5 = 0$, $\mu_6 = 37.698004\sigma^6$, and $\mu_8 = 625.17373\sigma^8$

and $$\text{var}(m_4/m_2^2) = \frac{\mu_2^2\mu_8 - 4\mu_2\mu_4\mu_6 + 4\mu_4^3 - \mu_2^2\mu_4^2}{n\mu_2^6}$$

$$= \frac{625.17373\sigma^{12} - 603.16806\sigma^{12} + 256\sigma^{12} - 16\sigma^{12}}{n\sigma^{12}}$$

$$= \frac{262.00567}{n}$$

For $\kappa = 5.0$, $\sigma_1^2 = \sigma^2[1 + \sqrt{6}]$ and $\sigma_2^2 = \sigma^2[1 - \sqrt{2/27}]$,

thus $\mu_2 = \sigma^2$, $\mu_3 = 0$, $\mu_4 = 5\sigma^4$, $\mu_5 = 0$, $\mu_6 = 66.773242\sigma^6$ and $\mu_8 = 1513.1693\sigma^8$

and $$\text{var}(m_4/m_2^2) = \frac{\mu_2^2\mu_8 - 4\mu_2\mu_4\mu_6 + 4\mu_4^3 - \mu_2^2\mu_4^2}{n\mu_2^6}$$

$$= \frac{1513.1693\sigma^{12} - 1335.4648\sigma^{12} + 500\sigma^{12} - 25\sigma^{12}}{n\sigma^{12}}$$

$$= \frac{652.7045}{n}$$

(d) Laplace distribution

The central moments is $\mu_r = 0$ for r odd

$\mu_r = r!c^r$ for r even, and $c = \sigma/\sqrt{2}$

thus $\mu_2 = \sigma^2$, $\mu_3 = 0$, $\mu_4 = 6\sigma^4$, $\mu_5 = 0$, $\mu_6 = 90\sigma^6$, and $\mu_8 = 2520\sigma^8$

and $$\text{var}(m_4/m_2^2) = \frac{\mu_2^2\mu_8 - 4\mu_2\mu_4\mu_6 + 4\mu_4^3 - \mu_2^2\mu_4^2}{n\mu_2^6}$$

$$= \frac{2520\sigma^{12} - 2160\sigma^{12} + 864\sigma^{12} - 36\sigma^{12}}{n\sigma^{12}}$$

$$= \frac{1188}{n}$$

(e) Exponential distribution

The raw moment is $\mu_r' = \frac{\Gamma(r+1)}{\lambda^r}$

thus $\mu_1' = \frac{1}{\lambda}$, $\mu_2' = \frac{2}{\lambda^2}$, $\mu_3' = \frac{6}{\lambda^3}$, $\mu_4' = \frac{24}{\lambda^4}$, $\mu_5' = \frac{120}{\lambda^5}$, $\mu_6' = \frac{720}{\lambda^6}$, $\mu_7' = \frac{5040}{\lambda^7}$ and $\mu_8' = \frac{40320}{\lambda^8}$. So that the central moment is $\mu_r = E[X-E(X)]^r$. Thus $\mu_1 = 0$, $\mu_2 = \frac{1}{\lambda^2}$, $\mu_3 = \frac{2}{\lambda^3}$, $\mu_4 = \frac{9}{\lambda^4}$, $\mu_5 = \frac{44}{\lambda^5}$, $\mu_6 = \frac{265}{\lambda^6}$, $\mu_7 = \frac{1854}{\lambda^7}$ and $\mu_8 = \frac{14833}{\lambda^8}$.

$$\text{The } \operatorname{var}(m_4/m_2^2) = \frac{\mu_2^2\mu_8 - 4\mu_2\mu_4\mu_6 - 8\mu_2^2\mu_3\mu_5 + 4\mu_4^3 - \mu_2^2\mu_4^2 + 16\mu_2\mu_3^2\mu_4 + 16\mu_2^3\mu_3^2}{n\mu_2^6}$$

$$= \frac{14833 - 9540 - 704 + 2916 - 81 + 576 + 64}{n}$$

$$= \frac{8064}{n}$$

(f) Student's t distribution

The central moment is $\mu_r = 0$ for $k > r$ and r odd

$$\mu_r = \frac{k^{\frac{1}{2}r}\, B(\frac{1}{2}(r+1), \frac{1}{2}(k-r))}{B(\frac{1}{2}, \frac{1}{2}k)} \quad \text{for } k > r \text{ and r even}$$

thus μ_6 and μ_8 do not exist (when t has 5 degrees of freedom) and therefore the $\operatorname{var}(m_4/m_2^2)$ is taken as ∞.

(g) Slash distribution

For the Slash distribution the moments are infinite and therefore the variance is undefined.

3.1.4.4.2 Unbiasedness

For the simulation study of Chapter 4 we used the unbiased estimates of the second and fourth order sample moments eg

$$\hat{\mu}_2 = \sum_{i=1}^{n} (\hat{\epsilon}_i - \bar{\epsilon})^2/(n-1) = \frac{n}{n-1}\, m_2 = c_0\, m_2$$

$$\hat{\mu}_4 = \frac{(n^2-2n+3)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} (\hat{\epsilon}_i - \bar{\epsilon})^4 - \frac{3(n-1)(2n-3)}{n(n-2)(n-3)}\hat{\mu}_2^2$$

$$= \frac{n(n^2-2n+3)}{(n-1)(n-2)(n-3)}\, m_4 - \frac{3n(2n-3)}{(n-1)(n-2)(n-3)}\, m_2^2$$

$$= c_1 m_4 - c_2\, m_2^2$$

where $c_0 = \frac{n}{n-1}$, $c_1 = \frac{n(n^2-2n+3)}{(n-1)(n-2)(n-3)}$ and $c_2 = \frac{3n(2n-3)}{(n-1)(n-2)(n-3)}$.

Thus $\text{var}[\hat{\mu}_4/\hat{\mu}_2^2] = \text{var}[(c_1 m_4 - c_2 m_2^2)/(c_0^2 m_2^2)]$

$= \text{var}[c_1 m_4/(c_0^2 m_2^2) - c_2/c_0^2]$

$= (c_1/c_0^2)^2 \text{var}[m_4/m_2^2]$

and $c_1/c_0^2 = \frac{n(n^2-2n+3)}{(n-1)(n-2)(n-3)} \times \frac{(n-1)}{n} \times \frac{(n-1)}{n}$

$= \frac{(n^2-2n+3)(n-1)}{n(n-2)(n-3)}$ (3.1.23)

Now the $\text{var}(m_4/m_2^2)$ as found in (a) through (e) should be multiplied by the square of (3.1.23). Thus the variance as given by Gonin and Money (1989, p225) should also be multiplied by (3.1.23) as should the values they provide in their table.

The above results are summarised in the following Table:

Table 3.1.2: Variance for $\hat{\kappa}$

D	$\text{var}(m_4/m_2^2)$ n in general	$\text{var}(m_4/m_2^2)$ n = 30	$\text{var}[\hat{\mu}_4/\hat{\mu}_2^2]$
U	1.3166/n	0.0438857	0.0509903
N	24/n	0.8	0.929512
CN4	262.01/n	8.7335223	10.147393
CN5	652.70/n	21.756817	25.279029
L	1188/n	39.6	46.010847
E	8064/n	268.8	312.31605

These results suggest caution is appropriate. However if we consider the adaptive algorithm (discussed in §3.1.4.3) we are fitting, the sample kurtosis yields only a starting value for the L_p-norm estimation process, and in this algorithm, we continue until convergence of the p's is reached. Thus we calculate the limit of the sequence $\{p^i\}$ (theoretical convergence, has not yet been proven, are we still going to try this - em algorithm)). Gonin and Money (1989) claimed that using examples where the true p is known, they could find a p with any method that converges to a particular estimate of p in about 4 iterations. Their claim and our findings will be discussed in Chapter 5.

3.2 Properties of L_p -norm estimators

3.2.1 Unbiasedness

Theoretically it can be proved that OLS estimators ($p = 2$) are unbiased (Chapter 1). However when $p \neq 2$ the property of unbiasedness is not so clear. In simulation studies no evidence of bias in L_p estimates could be found by Forsythe (1972), Barr (1981) and Money *et al.* (1982).

Barr (1981) and Money *et al.* (1982) performed a simulation study on nine symmetric error distributions: Uniform, Normal, Contaminated normal (with 5 levels of kurtosis), Laplace and Cauchy. Means of 500 sample estimates of regression coefficients were computed and compared to the true parameter values for six values of p namely (1.00, 1.25, 1.50, 1.75, 2.00 and ∞). Then by using the normal approximation to the binomial distribution the hypothesis of unbiasedness was tested (using the number of estimates falling above the true parameter value as a test statistic). Those authors concluded that the L_p-norm estimators are unbiased for all $p \geq 1$ when the error distribution is symmetric.

Nyquist (1983) (correctly) criticised the above studies by Money *et al.* (1982) and Barr (1981) and claimed that the notions of "symmetric distributions" and "unbiasedness" have been confused. Nyquist (1983) suggested that the correct conclusion should be that when the errors are

symmetrically distributed about $X\beta$, the L_p-norm estimators ($\hat{\beta}_{L_p}$) are symmetrically distributed. In effect Barr (1981) and Money et. al (1982) showed that the distribution of $\hat{\beta}_{L_p}$ is symmetric about a median which is close to β.

Harvey (1978) proved that the L_p-norm estimators of β are unbiased for $1 < p < \infty$ if the regression errors are symmetrically distributed, the first moment exists and X is of full column rank. Thus for the Cauchy and the Slash distribution (whose first absolute moments do not exist) the L_p-norm estimators are not unbiased, but do possibly have a median close to β.

Farebrother (1985) defined the (weighted) L_p-norm estimators of β as those which minimizes

$$\|\hat{\epsilon}\| = \left(\sum_{i=1}^{n} w_i |\hat{\epsilon}_i|^p\right)^{1/p}$$

where $p \geq 1$ is a given constant and $w \geq 0$ is a given $n \times 1$ vector of weights. Farebrother (1985) claimed that it is easy to find L_p-norm estimators of β that are symmetrically distributed about β when ϵ is symmetrically distributed about zero when $1 < p < \infty$. He then shows that the L_1 and CHEB estimators can be obtained by means of any of four linear programming problems. Implementing them carefully will yield unbiased L_1 and CHEB estimators. Usually the L_1 and CHEB estimators may be biased if the L_1 and CHEB norm have multiple minima. Harvey (1978) established the conditions given in a previous paragraph under which the L_p-norm estimator will be unbiased for $1 < p < \infty$. Under these same conditions, plus a further condition that the L_1 and CHEB norm must have a unique minimum for all y, he showed that the L_1 and CHEB estimator are unbiased. For $w_i=1$, conditions for the L_1 norm to have a unique minimum have been established by Kripke and Rivlin (1965), Rivlin (1969) and Nyquist (1980). Farebrother (1985) pointed out that when $w_i = 1$, the L_p norm estimator is a member of Huber's class of M-estimators so that $\sqrt{n}(\hat{\beta}_{L_p} - \beta)$ is asymptotically normally distributed with

mean zero and variance $\omega_p^2 Q^{-1}$, where $Q = \lim(\frac{1}{n}X'X)$ and where ω_p^2 is defined by Nyquist (1980, 1983), for sufficiently small values of $p \geq 1$.

Sielken and Hartley (1973) used other variants of the LP formulation than Farebrother in the developing of an algorithm which ensures unbiased L_1-norm estimators. Sposito (1982) extended the results of Sielken and Hartley (1973) to the general case where $p \geq 1$. Both Sielken and Hartley (1973) and Sposito (1982) fixed $w_i = 1$. Farebrother (1985) pointed out that if X has full column rank then the result of Sposito (1982) is redundant as Harvey has already established that the L_p-norm estimator is unique and unbiased. Furthermore neither Sielken and Hartley nor Sposito have explicitly shown that the unbiased estimators of β are symmetrically distributed about β. The advantage of Sposito's method is that it may be used when X has less than full column rank.

Before one may claim the property of unbiasedness for an L_p-norm estimator, it seems that one should take into account the error distribution (meaning the type of distribution, the symmetry and outliers), the value of p, the rank of X and the algorithm used in applying the estimator. In certain cases there may be possible evidence against unbiasedness, or it may be possible to show MSE convergence to 0. However the explicit demonstration of theoretical unbiasedness is in general an unsolved analytic problem.

3.2.2 Asymptotic distributions

In the case $p = 2$, the OLS estimator of β is a linear function of Y, and any linear function of normally distributed random variables is again normally distributed. Hence by assuming normally distributed errors, the whole spectra of significance testing and of confidence intervals are readily obtained from sums of squares distributed as Chi-square and from ratios distributed as F-distributions. For p strictly between one and infinity, the dual space of the L_p space (ie an n-dimensional vector space with the L_p norm) is (isometrically isomorphic to) the L_r space where $r = p/(p-1)$. Thus the L_2 space of the L_2-norm is self-dual, and it is the only space of the L_p spaces whose norm can be defined as an inner product (ie the only L_p

space which is a Hilbert space). If the inner product of two elements of a vector space is zero, the elements are said to be orthogonal. This result allows us to generalize many geometric concepts including the orthogonal decomposition of Y with respect to the column space of X, and leads to the exact distribution theory for least squares!.

In contrast to OLS, there is no closed-form solution of distributional issues for the general L_p-norm estimators. We must attempt to derive the distribution of the L_p-norm solution, $\hat{\beta}_{L_p}$, to an iterative non-LP problem.

The asymptotic distribution of $\sqrt{n}(\hat{\beta}_{L_p} - \beta)$, $1 \leq p < \infty$ is given by Nyquist (1983) as

$$\sqrt{n}(\hat{\beta}_{L_p} - \beta) \sim N(0, \omega_p^2 Q^{-1}) \qquad (3.2.1)$$

where both

$$\omega_p^2 = \begin{cases} 2F'(0)^{-2} & \text{if } p = 1 \\ E[|\epsilon_i|^{2p-2}]\{(p-1)E[|\epsilon_i|^{p-2}]\}^{-2} & \text{if } 1 < p < \infty \end{cases}$$

and the following conditions are satisfied:

A1: ϵ_i, $i = 1,\ldots,n$ are independent and identically distributed stochastic variables with common (cumulative) distribution function F.

A2: The L_1- and L_∞-norm estimators are unique (for $1 < p < \infty$, L_p-norm estimators are always unique).

A3: $Q = \lim_{n\to\infty}(\frac{1}{n}X'X)$, is a positive-definite matrix.

A4a: When $p = 1$: F is continuous with a continuous positive derivative at the median.

A4b: When $1 < p < \infty$ the following expectations exist:

$$E[|\epsilon_i|^{p-2}],\ E[|\epsilon_i|^{2p-2}],\ \text{and}\ E[|\epsilon_i|^{p-1}] = 0.$$

Basset and Koenker (1978) prove the case $p = 1$, and proofs for $p > 1$ are given in Huber (1973), Ronner (1977) and Nyquist (1980).

The theoretical value for ω_p^2 (the moment ratio parameter), is given by Gonin and Money (1989), and values for several distributions are summarised in the following two tables covering the whole range of distributions used in the simulation study of Chapter 4.

Table 3.2.1: ω_1^2 for several distributions.

Distribution	ω_1^2
Uniform	$3\sigma^2$
Normal	$1.57\sigma^2$
CN4	$1.393\sigma^2$
CN5	$1.278\sigma^2$
Laplace	$0.5\sigma^2$
Student's t_5 (scaled)	$1.04\sigma^2$
Exponential	σ^2

Table 3.2.2: ω_p^2 for several distributions.

Distribution	ω_p^2 $(1 < p < \infty)$
Uniform	$3\sigma^2/(2p-1)$
Normal	$2\sqrt{\pi}\sigma^2\Gamma(p-\frac{1}{2})/\{(p-1)\Gamma[\frac{1}{2}(p-1)]\}^2$
CN	$20\sqrt{\pi}(\sigma_1^{2p-2}+9\sigma_2^{2p-2})\Gamma(p-\frac{1}{2})/\{(p-1)\Gamma[\frac{1}{2}(p-1)](\sigma_1^{p-2}+9\sigma_2^{p-2})\}^2$
Laplace	$\sigma^2\Gamma(2p-1)/(2\{(p-1)\Gamma(p-1)\}^2)$
Student's t_5 (scaled)	$\frac{\sigma^2\sqrt{3^3}\sqrt{5\pi}}{16}\,\frac{\{\Gamma[\frac{1}{2}(2p-1)]\}^2\Gamma[1-\frac{1}{2}(2p-1)]}{\Gamma[\frac{1}{2}(2p-5)]}\,\frac{\{\Gamma[\frac{1}{2}(p-5)]\}^2}{\{\Gamma[\frac{1}{2}(p-1)]\Gamma[1-\frac{1}{2}(p-1)]\Gamma[\frac{1}{2}(p+1)]\}^2}$
Exponential	$\sigma^2\Gamma(2p-1)/\{(p-1)\Gamma(p-1)\}^2$

When p = 1 the quantity ω_p^2 of (3.2.1) is denoted by λ^2 (Dielman and Pfaffenberger (1983) and Gonin and Money (1989)), where λ^2/n is the

asymptotic variance of the sample median of residuals with distribution F. Thus the least absolute error estimator has strictly smaller asymptotic confidence ellipsoids than the least squares estimator in linear models under any cumulative distribution function F for which the sample median is a more efficient estimator of location than the sample mean.

Prior to the proof by Basset and Koenker (1978), of asymptotic convergence of the L_1 estimator to normality, Rosenberg and Carlson (1971) conducted a Monte Carlo simulation experiment on L_1 estimation, for two sample sizes (31, 59) and three error distributions (Normal and two Contaminated normal distributions). They concluded that the Gaussian distribution (3.2.1), provides an acceptable approximation to the distribution of $\hat{\beta}_p$ for modest sample sizes and well-conditioned designs X, when p = 1. The approximation was significantly worse for ill-conditioned designs (by which they mean that the independent variables of the X-matrix in the regression model could be constant, normal, or dispersed with high kurtosis). The results of Rosenberg and Carlson (1971) were unpublished but are partly presented in Rosenberg and Carlson (1977).

Dielman and Pfaffenberger (1983) extended the Monte Carlo simulation experiment of Rosenberg and Carlson (1971), by adding two further distributions namely Laplace and Cauchy, extended the sample sizes (20, 30, 40, 50, 100, 150, 200) and allowed for collinearity amongst the predictor variables. From their extended study, the following recommendations regarding the minimum sample sizes required to support inferences made using L_1 estimators and the asymptotic normal theory for the error, are given:

(i) if the error distribution is normal and $n \geq 20$, approximate normality of the distribution of the estimators is ensured;
(ii) in the case of the Contaminated normal the minimum sample size required is n = 50;
(iii) for the Cauchy the minimum sample size is n = 150;
(iv) for the Laplace distribution the minimum sample size is n = 200;
(v) collinearity appears to have no effect on the rate on convergence to normality.

The theorem (3.2.1) of asymptotic convergence to normality now allows us to construct confidence intervals for the components of $\hat{\beta}_{L_p}$. The $100(1-\alpha)\%$ confidence interval for the i-th element of β_{L_p} is (Gonin and Money (1989)):

$$[\hat{\beta}_{L_p}]_i \pm z_{\alpha/2}\{\omega_p^2(X'X)^{-1}_{ii}\}^{\frac{1}{2}} \qquad (3.2.2)$$

where $z_{\alpha/2}$ is the appropriate percentile of the standard Normal distribution. The quantity ω_p^2 is unknown. Gonin and Money (1989), by conducting a simulation study, suggested the following estimate of ω_p^2 for $1 < p < \infty$:

$$\hat{\omega}_p^2 = \frac{m_{2p-2}}{[(p-1)m_{p-2}]^2} \qquad (3.2.3)$$

where $m_r = \frac{1}{n}\sum_{i=1}^{n} |\hat{\epsilon}_i|^r$, and $\hat{\epsilon}_i$ is the residual from the L_p-fit.

When p = 1, the $100(1-\alpha)\%$ confidence interval for the i-th element of β_{L_1} is (Gonin and Money (1989)):

$$[\hat{\beta}_{L_1}]_i \pm z_{\alpha/2}\{\lambda^2(X'X)^{-1}_{ii}\}^{\frac{1}{2}} \qquad (3.2.4)$$

In a simulation study, Dielman and Pfaffenberger (1983), used the true value of λ. However, in general, the quantity λ is unknown and various methods to estimate the variance of the median of the residuals (λ^2/n) are given by Gonin and Money (1989, p15-16). In the simulation study of Chapter 4 one of these estimators will be used:

$$\hat{\lambda} = [2\hat{f}(m)]^{-1} \qquad (3.2.5)$$

where f(m) is the ordinate of the error distribution at the median. In this thesis we will estimate $\hat{f}(m)$ by the method of Cox and Hinkley (1974) or by the method of McKean and Schrader (1987). Denote the i-th ordered residuals

(using the L_1-estimator) by $\hat{\epsilon}_{(i)}$. Then Cox and Hinkley (1974) estimate $\hat{f}(m)^{-1}$ by

$$\hat{f}(m)^{-1} = \frac{\hat{\epsilon}_{(i)} - \hat{\epsilon}_{(j)}}{(i-j)/n} \qquad (3.2.6)$$

Cox and Hinkley (1974) stress that i and j should be symmetric about the index of the median sample residual and that the difference between i and j should be kept small.

Sposito and Tveite (1986) consider all the residuals and set

$$i = [n/2] + v$$
$$j = [n/2] - v$$

where [.] denotes the greatest integer value of the argument and v is an appropriate positive integer. They conducted a simulation study to investigate how well (3.2.6) estimates $f(m)^{-1}$. They considered four distributions (Laplace, Cauchy, Normal and Uniform) , 4 sample sizes (50, 100, 200 and 300) and v ranging through 1,2,3,4,5, and 6. In their representation of the results they underline entries for which $|\text{estimate} - \text{true}| \leq 0.05$, without taking into account the variance of the distributions (which were not equal); thus they may be giving some distributions an advantage.

The choice of i and j specified by Sposito and Tveite will make i and j unsymmetric for n even and we propose for n even to set $i = [n/2] + 1 + v$.

Sposito and Tveite (1986) suggested that the difference between i and j should be slightly larger than r + 1 (r parameters) since the number of zero residuals under L_1 is at least r (the L_1 algorithm fit at least r points in the plane).

Gonin and Money (1989) proceed as follows to find the value of i and j:

Consider only the number $\tilde{n}$ of non-zero residuals, and instead of n residuals use $\tilde{n}$ non-zero residuals. Then the i and j can be found by

$$i = (\tilde{n}+1)/2 + v \text{ and } j = (\tilde{n}+1)/2 - v \qquad \text{for } \tilde{n} \text{ odd}$$

$$i = (\tilde{n})/2 + 1 + v \text{ and } j = (\tilde{n})/2 - v \qquad \text{for } \tilde{n} \text{ even}$$

where v is a positive small integer.

Cox and Hinkley (1974) show that $\hat{\lambda}$ of (3.2.2) is a consistent estimator of λ.

McKean and Schrader (1987) only consider the non-zero residuals. Their estimator is based on the α-percent non-parametric confidence interval for λ. The estimate of λ is given by

$$\hat{\lambda}_{1-\alpha} = \frac{\tilde{n}^{\frac{1}{2}}(\hat{\epsilon}_{(\tilde{n}-l+1)} - \hat{\epsilon}_{(l)})}{2z_{\alpha/2}}$$

where asymptotically

$$l = \frac{\tilde{n}+1}{2} - z_{\alpha/2}(\tilde{n}/4)^{\frac{1}{2}}$$

and l is usually rounded to the nearest integer.

3.3 Summary

In this chapter the L_1, L_∞, L_2 and L_p-norm estimators were defined, their properties discussed, and various algorithms (programs) to find these norms were introduced. In the case of the L_p-norm various choices of p were investigated. Finally the asymptotic distributions of the L_p-norm estimators were discussed.

Chapter 4

SIMULATION STUDY

4.1 Introduction

The purpose of this simulation study is to compare the performances of 13 different biased estimators, with OLSE (summarised in table 2.1 and 2.4 of Chapter 2), and with L_1, L_p and L_∞ estimators (discussed in Chapter 3, and summarised in Table 5.1 and Table 5.4 of Chapter 5). The generation of the simulated data and the setup of the factorial experiment are discussed in §4.2. In §4.3 we comment on the tail ratios of four distributions.

4.2 Data

The simulation study of this thesis follows the form of McDonald and Galarneau (1975) and Wichern and Churchill (1978). The X matrix of the data sets were obtained from Chalton (1990) who generated them for a simulation study in his PhD. thesis.

Chalton (1990), considers a five parameter model, with a sample size of 30 and the predictor variables generated from the following relationship:

For $j = 1,2,3$ and $i = 1,2,\ldots,30$

$$X_{ij} = (1 - a_1^2)^{\frac{1}{2}} Z_{ij} + a_1 Z_{i6} \qquad (4.2.1)$$

but for $j = 4,5$ and $i = 1,2,\ldots,30$

$$X_{ij} = (1 - a_2^2)^{\frac{1}{2}} Z_{ij} + a_2 Z_{i6} \qquad (4.2.2)$$

where

(i) Z_{ij} are independent N(0,1) variates generated by the SAS-function RANNOR. The seeds were not recorded by Chalton (1990), as the RANNOR function derives seeds from the time clock of the computer.

(ii) The parameters a_1 and a_2 determine the degree of collinearity between the predictor variables: a_1^2 is the theoretical correlation between any pair of the variables X_1, X_2 and X_3, the product $a_1 a_2$ is the theoretical correlation between any variables in the set $\{X_1, X_2, X_3\}$ and a variable in $\{X_4, X_5\}$ and a_2^2 is the theoretical correlation between X_4 and X_5.

Five different vector combinations of a_1 and a_2 were considered, and two choices (orientations) of β, suggested by Newhouse and Oman (1971), namely the eigenvectors (of $X'X$) corresponding to the largest and smallest eigenvalues, denoted by β_L and β_S. For these 5×2=10 combinations (implying 10 distinct estimated response vectors $X\beta$), four different values of σ where considered, namely 0.01, 1.0, 5.0 and 10.0, for the error terms. We expect that some terms will be swamped by error.

For i = 1,2,...,30 we model the response values as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \epsilon_i \qquad (4.2.3)$$

where the X_{ij} are unstandardized, β_0 is zero, and the ϵ_i are independent $N(0,\sigma^2)$ variates. Chalton (1990) generated 100 Y-vectors as 100 sets of 30 data points from the model (4.2.3), for each of the (10×3=30) combinations of orientations, a_1 and a_2 values, and 3 variance values. (Note Chalton (1990) reported the theoretical correlation as a_1^2 and a_2^2. However, in his generation program he used the values of a_1^2 and a_2^2 as input to a_1 and a_2. We use his X-matrix, but correct the minor error as we change his a_1^2 and a_2^2 to the correct a_1 and a_2). The eigenvalues and condition numbers corresponding to the five different combinations of a_1 and a_2 are shown in Table 4.1 and the coefficients of β_L and β_S are shown in Table 4.2.

Table 4.1: Eigenvalues and condition numbers of X'X (X is standardized).

Correlations $a_1:a_2$	eigenvalues of X'X: λ_i (without β_0)	condition number λ_1/λ_5
.99:.99	(4.920,0.026,0.021,0.013,0.011)	435
.99:.10	(3.157,1.128,0.668,0.031,0.016)	201
.90:.90	(4.215,0.430,0.154,0.126,0.075)	56
.90:.10	(2.755,1.215,0.173,0.168,0.148)	19
.70:.30	(2.283,1.049,0.871,0.496,0.301)	8

Table 4.2: β used in generating Y

Correlations $a_1:a_2$	β'	Eigenvectors of X'X				
.99:.99	β'_L	[0.4474	0.4473	0.4481	0.4470	0.4463]
	β'_S	[0.2846	0.4760	-0.8302	0.0548	0.0163]
.99:.10	β'_L	[0.5534	0.5542	0.5510	0.1705	0.2323]
	β'_S	[-0.7755	0.1675	0.6084	0.0190	-0.0094]
.90:.90	β'_L	[0.4125	0.4547	0.4649	0.4383	0.4636]
	β'_S	[0.1821	-0.3973	0.5673	0.3037	-0.6284]
.90:.10	β'_L	[0.5634	0.5489	0.5644	0.2354	0.0862]
	β'_S	[-0.0162	0.6908	-0.7051	0.0655	0.1451]
.70:.30	β'_L	[0.5177	0.5611	0.5226	0.0286	0.3785]
	β'_S	[-0.6600	0.7049	0.0099	0.1956	-0.1708]

For the simulation study in this chapter, β_0=10 because we add 10 to the Y's generated by the author. For each combination of X, β and σ, 100 replications of the (30×1) ϵ-vector were generated independently for (4.2.3), because we want independent repetitions of Chalton's data and we were concerned that there might be serial correlation present in the

Chalton's first set of data. Additionally, Chalton's deviates were no longer available for transformation to alternative distributions.

Chalton only generated data for $N(0,\sigma^2)$, σ = 0.1, 1.0, 5.0. We extend the study to include an extra σ value, σ = 10.0, and generate the error terms from several additional distributions namely Uniform, Normal, Contaminated normal (2 kurtosis levels), Laplace, Student's t, Exponential and Slash, varying the scale over the four σ-values.

Thus in the setting up of the data we set up a factorial experiment of

8	x	4	x	5	x	2
distributions		variance levels		collinearity levels		orientations of the betas

and the total number of estimators in which we are interested 39: 13 biased, L_1, L_∞, and finally 12 L_p-norm estimators each fitted via two algorithms (WLS and BFGS).

In §4.2.1 we discuss the generation of the pseudo-uniform variates, in §4.2.2 we introduce the distributions, and discuss the transformation from uniform U[0,1] variables to variables with the specified distributions and required scale parameters (4 levels).

In Chapter 5, §5.5.1, two full X (200×5) matrices as well as pseudo-random uniform numbers will be generated using the same scenario as described in this chapter.

4.2.1 Generating pseudo-random uniform numbers

The pseudo-random uniform U[0,1] values were generated using the random generator of Wichmann and Hill (1982). We choose this generator because it is relatively simple and completely portable (eg given the seeds, the unique set of numbers generated can be obtained on any machine that can perform integer arithmetic up to 30320). Furthermore those authors claim

that this generator is "reasonably short, reasonably fast, machine-independent, easily programmed in any language, and statistically sound!". The cycle length exceeds 2.78×10^{13}. We did not perform any rigorous testing of the portable random generator, already shown to be satisfactory by its authors. The generator is already in use in various statistical packages (eg GENSTAT (Payne (1988))).

The structure of this random number generator involves three independent uniform random numbers obtained from three multiplicative congruential generators, where each multiplicative generator uses a prime number for its modules and a primitive root for its multiplier. (For a discussion on linear congruential methods the reader is referred to Knuth (Chapter 3, 1969)). The fractional part of the sum of these three uniform random numbers is obtained, and this fraction is again a uniform U[0,1] random number. The cycle length for pseudo-random numbers obtained in this manner is $c_1 c_2 c_3$, where c_i is the cycle length of the i-th random generator.

For each X-matrix, and for each variance level, 100 repetitions of each ϵ:30 x 1 vector were generated for each variance level. Six origins (meaning six independent sets of 3 seeds) where chosen randomly, and labelled A, B, C, D, E and F. The A files store triplets of uniform deviates from which the Wichman and Hill uniform deviates were obtained. Then by transformation the deviate generates data from distributions for which only one random variable is needed. When an extra random deviate is needed (eg Normal, Contaminated normal) stream B is used, and stream C is used for the weighting process of the Contaminated normal distribution and for the division in the Slash distribution. Streams D, E, F are used additionally for generation of a Student's t distribution with 5 degrees of freedom. In each stream 40 ($4 \times 5 \times 2$) independent sets (each of 3000x3 pseudo-observations) of uniform [0,1] values were generated and these sets are then transformed to the relevant distributions. By using the same origen (A) as the base for each distribution, comparability of residuals were ensured.

Each of the 40 (x6) sets of 3000 random values was tested for serial correlation (lag 1, lag 2, lag 3) and a frequency test (χ^2- goodness of fit test) was performed. The seeds for the generation are given in Appendix A. Only the first three random number seeds were chosen, subsequent values were generated and recorded as at the corresponding stage of computing. These values are recorded to facilitate re-use by restarting the generation process at a designated file element, when restarting was required because of lack of output storage. For instance when using the eighth X-matrix, and a σ-level of 5.0, it is not necessary to start at the first element and store all the output but one may start at a designated file element, and perform repeated smaller loops of operations.

The serial correlations, χ^2-values and other statistics are reported in Appendix A. Some of the χ^2-values appear to be too large and some too small, but on the whole the results seem satisfactory.

The serial correlations were calculated for lag 1, lag 2, and lag 3. The serial correlation is expected to be close to zero. Knuth (1969) gives the 95 percent confidence interval for the serial correlation as

$$\mu_n \pm 2\sigma_n$$

where $\mu_n = \frac{-1}{n-1}$, $\sigma_n = \frac{1}{n-1}\sqrt{\frac{n(n-3)}{n+1}}$, $n > 2$

These checks were performed here simply to help explore what goes into our pseudo-data and hence to the estimators, and what comes out after fitting various estimators.

4.2.2 Transformations for distributions

Several symmetric error distributions with $E(\epsilon) = 0$ and $Var(\epsilon)$ as 0.01, 1.0, 25.0 and 100.0 were considered as well as one unsymmetric distribution, the Exponential with mean 100, 1, 1/5 and 0.01 and variance levels as before.

The Slash distribution is a distribution without finite variance (see 4.2.2.8). Properties of the distributions may be found in Johnson and Kotz (1970) or Mood, Graybill and Boes (1974). Although these distributions are defined in Chapter 1, we included them here for completeness and continuity. The moments of the simulated distributions are reported in Appendix B. The first 6 tables in Appendix B contain the moments for the 6 normal streams, and the remainder of Appendix B contains the moments of the various distributions.

In the case of this thesis the random deviates for each distribution were generated once, and stored. However if the random deviates are generated on hand, it should be pointed out that there are more efficient (faster) methods to generate residuals for normal and Student's t distributions.

4.2.2.1 Uniform

The uniform random variable Y has distribution function

$$F(y) = \frac{y-a}{b-a} \qquad a \leq y \leq b$$

with mean $\mu = \frac{a+b}{2}$, variance $\sigma^2 = \frac{(b-a)^2}{12}$,

central moments $\mu_r = 0$ for r odd

$$\mu_r = \frac{(b-a)^r}{2^r(r+1)} \quad \text{for r even,}$$

and kurtosis, $\kappa = \mu_4/\mu_2^2 = \frac{(b-a)^4}{2^4(4+1)} \div \frac{(b-a)^4}{12x12} = \frac{9}{5} = 1.8$

Thus a random deviate, U, with Uniform (0,1) distribution, has mean $\frac{1}{2}$ and variance $\frac{1}{12}$. To generate a random deviate Y from the Uniform distribution with mean 0 and variance equal to σ^2, we use the cumulative distribution function technique. We require $y = F^{-1}(u)$, where u is a Uniform [0,1] random number. Setting a = -b, where $f(y) = \frac{1}{2b}$, $F(y) = \frac{y+b}{2b}$, and

$$y = F^{-1}(u) = 2bu - b$$
$$= b(2u - 1)$$

Thus if $y \sim$ Uniform[-b,b], then $var(y) = \frac{b^2}{3}$, and $b = \sqrt{3}\sigma$

For given variance values, we may obtain the corresponding uniform pseudo-random deviates for u from U[0,1] as

$\sigma^2 = 0.01^2$, $b = \sqrt{3} \times 0.01$ and $y = 0.0173205 \times (2u - 1)$

$\sigma^2 = 1.0$, $b = \sqrt{3} \times 1.0$ and $y = 1.7320508 \times (2u - 1)$

$\sigma^2 = 25$, $b = \sqrt{3} \times 5.0$ and $y = 8.660254 \times (2u - 1)$

$\sigma^2 = 100.0$, $b = \sqrt{3} \times 10.0$ and $y = 17.320508 \times (2u - 1)$

4.2.2.2 Normal (Gaussian)

A normal (Gaussian) random variable has the density function,

$$f(y) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp[-(y-\mu)^2/2\sigma^2], \qquad -\infty < y < \infty$$

$$-\infty < \mu < \infty$$

$$0 < \sigma$$

with mean μ, variance σ^2,

central moments $\mu_r = 0$ for r odd

$$\mu_r = \frac{r!}{(r/2)!}\,\frac{\sigma^r}{2^{r/2}} \quad \text{for r even,}$$

and kurtosis, $\kappa = \mu_4/\mu_2^2 = \frac{4!}{(4/2)!}\,\frac{\sigma^4}{2^{4/2}} \div \sigma^4 = 3$

Box-Muller Transformation

To obtain Normal(0,1) deviates from Uniform deviates, we use the Box-Muller transformation. Box and Muller (1958) proposed a method in which two independent uniform variables U[0,1] generated from separate seeds, eg stream A and B, are used to generate two independent standard normal variables N(0,1). The transformations are

$$z_1 = (-2 \ln u_1)^{\frac{1}{2}} \sin 2\pi u_2,$$

$$z_2 = (-2 \ln u_1)^{\frac{1}{2}} \cos 2\pi u_2,$$

where u_1 and u_2 are the independent Uniform deviates (from stream A and B respectively).

To obtain the desired level of variance, we multiply by the required σ. Note that in this simulation for convenience we generally use only the z_1 variate-values, and the z_2 variate-values are used only when a second normal variable is required.

Neave (1973) found unsatisfactory sampling distributions for the Box-Muller transformation when used with multiplicative Congruential Pseudo-random Number generators:

$$x_{s+1} = (b \times x_s) (\text{mod } M).$$

He reported that the agreement between the observed and the expected frequencies is poor in the two tails of a Normal distribution. However when the multiplier (b) increases the range of z (the range of the normal deviates) becomes more realistic. But no matter how large the multiplier may be, z is still unsymmetric, bounded on at least one side, and very unsmooth, especially in the tails. Golder and Settle (1976) pointed out that the improvement claimed by Neave for increasing b is only true for $b < M/2$.

Chay *et al.* (1975) consider the problem further and overcomes it by interchanging the order of each successive pair of pseudo-random numbers. Golder and Settle (1976) show by an example in their paper that the Chay interchange can also yield samples with poor properties. They classify the conventional method (used by Neave) and the modified generator of Chay under Single Generator methods. Under Two Generator methods they consider:

1. The Shuffle method of Maclaren and Marsaglia (1965): Two (independent) generators A and B are used. The A produces pseudo-random uniform deviates and B produces a set of randomly generated integers in the range 1 to N. The integers generated in B define an ordering (shuffle) for the sequence in which the A deviate stream will be used.

2. The Neave (1972) method: One may use two multiplicative congruential generators with moduli m_1 and m_2 yielding sequences $\{y_i\}$ and $\{y_i'\}$, and form

$$p_i \equiv y_i/m_1 + y_i'/m_2 \text{ (modulo 1)}$$

where m_1 and m_2 are prime.

3. The two-sequence method: Two independent generators produce two separate sets of pseudo-random uniform deviates. Pairs of deviates are then sent directly into the Box-Muller transformation.

By Monte Carlo simulation Golder and Settle (1976) conclude that of the Two Generator methods, both the Neave and the two-sequence methods are acceptable.

In this study we employ a Three-Generator method feeding into a two-sequence form. Multipliers 171, 172 and 170 in congruential multiplicative generators yield 3 deviates which are combined into one (sum of three uniform deviates) and two independently seeded runs are taken as a single paired stream to enter the Box-Muller transformation.

The resulting values we assume to be adequately Gaussian, and the χ^2-results (df = 30-1) in Appendix B appear satisfactory. The tails of the Gaussian streams were checked and also appeared satisfactory. These empirical findings have further confirmation in the Golder and Settle (1976) study.

4.2.2.3 Symmetric contaminated normal

The weighted sum $Y = w_1Y_1 + w_2Y_2$ of two $N(\mu_i, \sigma_i^2)$ variables follows a Contaminated normal distribution. The random variable Y has the density function

$$f(y) = w_1 \frac{1}{\sqrt{2\pi}\,\sigma_1} \exp[-(y-\mu_1)^2/2\sigma_1^2] + w_2 \frac{1}{\sqrt{2\pi}\,\sigma_2} \exp[-(y-\mu_2)^2/2\sigma_2^2]$$

with $-\infty < y < \infty$ and $w_1 + w_2 = 1$. The central moments are

$$\mu_r = 0 \quad \text{for r odd}$$

$$\mu_r = \frac{r!}{(r/2)!2^{r/2}} \left[w_1\sigma_1^r + w_2\sigma_2^r\right] \quad \text{for r even.}$$

Thus the variance and kurtosis are respectively given by

$$\mu_2 = \left[w_1\sigma_1^2 + w_2\sigma_2^2\right]$$

and

$$\kappa = \mu_4/\mu_2^2$$

$$= \frac{4!}{(4/2)!2^{4/2}} \left[w_1\sigma_1^4 + w_2\sigma_2^4\right] \div \left[w_1\sigma_1^2 + w_2\sigma_2^2\right]^2$$

$$= 3\left[w_1\sigma_1^4 + w_2\sigma_2^4\right] \div \left[w_1\sigma_1^2 + w_2\sigma_2^2\right]^2$$

We use two symmetrically Contaminated distributions by choosing μ_1 and μ_2 equal to zero, the kurtosis equal to 4 and 5 and the weights equal to 1/10 and 9/10. In contrast, Barr (1980) and Gonin and Money (1989) choose $w_1 = w_2 = \frac{1}{2}$.

To obtain the corresponding desired levels of variance ($\mu_2 = \sigma^2$) and specified kurtosis (4 and 5) we solve for σ_1^2 and σ_2^2, by solving the quadratic in

$$\sigma_1^2 + 9\sigma_2^2 = 10\sigma^2$$

$$\sigma_1^4 + 9\sigma_2^4 = (10\kappa\sigma^4)/3$$

thus

$$\sigma_1^2 = \sigma^2 + \sigma^2 \times \left[3(\kappa - 3)\right]^{\frac{1}{2}}$$

$$= \sigma^2\left\{1 + \left[3(\kappa - 3)\right]^{\frac{1}{2}}\right\}$$

and

$$\sigma_2^2 = \sigma^2 - \sigma^2 \times \left[(\kappa - 3)/27\right]^{\frac{1}{2}}$$

$$= \sigma^2\left\{1 - \left[(\kappa - 3)/27\right]^{\frac{1}{2}}\right\}$$

Thus, for $\kappa = 4$, the above simplifies to

$$\sigma_1^2 = \sigma^2(1 + [3]^{\frac{1}{2}})$$

and $$\sigma_2^2 = \sigma^2(1 - [1/27]^{\frac{1}{2}})$$

and for $\kappa = 5$ the solution is

$$\sigma_1^2 = \sigma^2(1 + [6]^{\frac{1}{2}})$$

and $$\sigma_2^2 = \sigma^2(1 - [2/27]^{\frac{1}{2}})$$

To obtain a Contaminated random deviate we use the same process as in the normal case: we take two independent N(0,1) variables from the A and B streams, say z_1 and z_2, multiply by the square root of the relevant variance factors for σ_1^2 and σ_2^2, and mix out a single stream randomly under the weights (using the C-stream of random deviates).

4.2.2.4 Laplace

A random variable Y with Laplace distribution has the density function,

$$f(y) = \tfrac{1}{2}c^{-1}\exp\{-|y-a|/c\}, \qquad -\infty < y < \infty, \quad -\infty < a < \infty, \quad c > 0$$

and distribution function

$$F(y) = \tfrac{1}{2}\exp\{(y-a)/c\}, \qquad \text{when } y \leq a$$

and

$$F(y) = 1 - \tfrac{1}{2}\exp\{-(y-a)/c\}, \qquad \text{when } y > a$$

with mean a, variance $2c^2$,

central moments $\mu_r = 0$ for r odd

$\mu_r = r!c^r$ for r even

and kurtosis, $\kappa = \mu_4/\mu_2^2 = 4!c^4/4c^4 = 6$.

For zero mean we choose a = 0. Then by using the cumulative distribution function technique, $y = F^{-1}(u)$, where u is a Uniform [0,1] random number, we solve for y when $0 \leq u \leq 0.5$ in the equation

$$\tfrac{1}{2}\exp[y/c] = u.$$

Hence $$y/c = \ln(2u)$$

$$y = c \ln(2u)$$

and $y \leq 0$.

For $0.5 \leq u \leq 1.0$ we solve the equation

$$1 - \tfrac{1}{2}\exp\{-(y)/c\} = u.$$

Hence $$-(y)/c = \ln(2\{1-u\})$$

$$y = -c \ln(2\{1-u\})$$

and $y > 0$.

The variance ($2c^2$) must be equal to σ^2, thus $2c^2 = \sigma^2$, thus $c = \sigma/\sqrt{2}$

$\sigma^2 = 0.01^2$, $c = 0.01/\sqrt{2}$

$\sigma^2 = 1.0$, $c = 1.0/\sqrt{2}$

$\sigma^2 = 25$, $c = 5.0/\sqrt{2}$

$\sigma^2 = 100.0$, $c = 10.0/\sqrt{2}$

4.2.2.5 Student's t

The Student's t distribution is that distribution associated with the ratio of a standard normal random variable to the square root of an independently distributed chi-square random variable which has been divided by its degrees of freedom.

$T = Z/\sqrt{U/k}$ where $Z \sim N(0,1)$,
$U \sim \chi^2$ with k degrees of freedom
and Z and U are independent.

The density function of T is

$$f(t) = \frac{\Gamma[\frac{1}{2}(k+1)]}{\Gamma[\frac{1}{2}k]} \frac{1}{\sqrt{k\pi}} \frac{1}{(1+t^2/k)^{\frac{1}{2}(k+1)}} \qquad -\infty < t < \infty, \quad k > 0$$

with mean $\mu = 0$ for $k > 1$, variance $\sigma^2 = \frac{k}{k-2}$ for $k > 2$,

central moments $\mu_r = 0$ for $k > r$ and r odd

$$\mu_r = \frac{k^{\frac{1}{2}r} B(\frac{1}{2}(r+1), \frac{1}{2}(k-r))}{B(\frac{1}{2}, \frac{1}{2}k)} \quad \text{for } k > r \text{ and } r \text{ even}$$

and kurtosis, $\kappa = \mu_4/\mu_2^2 = \frac{3(k-2)}{(k-4)}$ $\quad k > 4.$

To find T from the uniform deviates we have to form N(0,1) deviates. Two streams will result in two z_i streams.

$$t = z_0\sqrt{k}/(z_1^2 + z_2^2 + \ldots + z_k^2)^{\frac{1}{2}}.$$

For k = 5, the variance is 5/3 and the kurtosis is 9. We will need six streams (A, B, C, D, E and F) of normal deviates.

4.2.2.6 Exponential

A random variable Y with the Exponential distribution has the density function,

$$f(y) = \lambda \exp\{-\lambda y\}, \quad \lambda > 0 \text{ and } 0 < y < \infty$$

and distribution function

$$F(y) = 1 - \exp\{-\lambda y\}$$

with mean $1/\lambda$, variance $1/\lambda^2$,

raw moments $\mu'_r = \dfrac{\Gamma(r+1)}{\lambda^r}$

and kurtosis, $\kappa = \mu_4/\mu_2^2 = 9$

Since $\lambda > 0$ the mean is not set to zero. By using the cumulative distribution function technique, $y = F^{-1}(u)$, where u is a Uniform [0,1] random number (stream A) we have

$$\exp[-\lambda y] = 1-u$$

and

$$y = -\ln(1-u)/\lambda.$$

Now y is a sample observation from an Exponential distribution with parameter λ, and variance $1/\lambda^2$, and the value of λ will be chosen so that $\sigma^2 = 1/\lambda^2$, which induces a non-zero mean.

Note that when errors from this distribution are transferred into the simulated data, least squares centering will result in the estimation of the intercept $\beta_0 + \lambda$ as the overall constant term.

4.2.2.7 Slash

The Slash distribution is that distribution associated with the random variable obtained by dividing a N(0,1) deviate by an independent U[0,1] deviate.

$$Y = V/W, \quad \text{where } V\sim N(0,1),\ W\sim U(0,1),$$
$$\text{and V and W are independent.}$$

Rice (p89, 1988) gives the distribution of $Y = V/W$ as

$$f(y) = \int_0^1 |w| f(w) f_y(wy)\,dw$$

$$= \int_0^1 w.1.(2\pi)^{-\frac{1}{2}} \exp(-\tfrac{1}{2}w^2y^2)\,dw.$$

Set $u = w^2$, thus $du = 2w\,dw$ and

$$= (2\pi)^{-\frac{1}{2}}\, 2^{-1} \int_0^1 \exp(-\tfrac{1}{2}uy^2)du$$

$$= (2\pi)^{-\frac{1}{2}}\, 2^{-1}\, (\tfrac{1}{2}z^2)^{-1}\{1-\exp(-\tfrac{1}{2}y^2)\}$$

$$= (2\pi)^{-\frac{1}{2}}\, (y^2)^{-1}\{1-\exp(-\tfrac{1}{2}y^2)\} \text{ for } -\infty < y < \infty$$

$$= \begin{cases} [N(0,1)) - N(y,1)]/y^2 & y \neq 0 \\ N(0,1)/2 & y = 0 \end{cases}$$

The Slash is similar to the Normal, except that its tails are much heavier, so that it resembles the Cauchy, and has infinite even moments. Hence we preserve the median as zero, and change the scale in Y by the corresponding scale change in the V element.

4.3 Tail ratios

To illustrate the influence of the heavy tails the density function of the Laplace, Exponential, Student's t and Slash distribution where expressed as a ratio of

density: Normal density.

The following expressions emerge for $\sigma = 1$

Laplace:Normal ($\mu = 0$)

$$l(x) = \sqrt{\pi}\, \exp[x^2/2 - x\sqrt{2}\,]$$

Exponential:Normal ($\mu = 1$, and $\mu = 0$)

$$e(x) = \sqrt{2\pi}\, \exp[x^2/2 - x\,]$$

Slash:Normal ($V \sim N(0,\sigma^2)$, and $\mu = 0$)

$$s(x) = \frac{1}{x^2}\,\frac{[1-\exp(-\frac{1}{2}x^2)]}{\exp(-\frac{1}{2}x^2)}$$

(scaled) Student's t:Normal

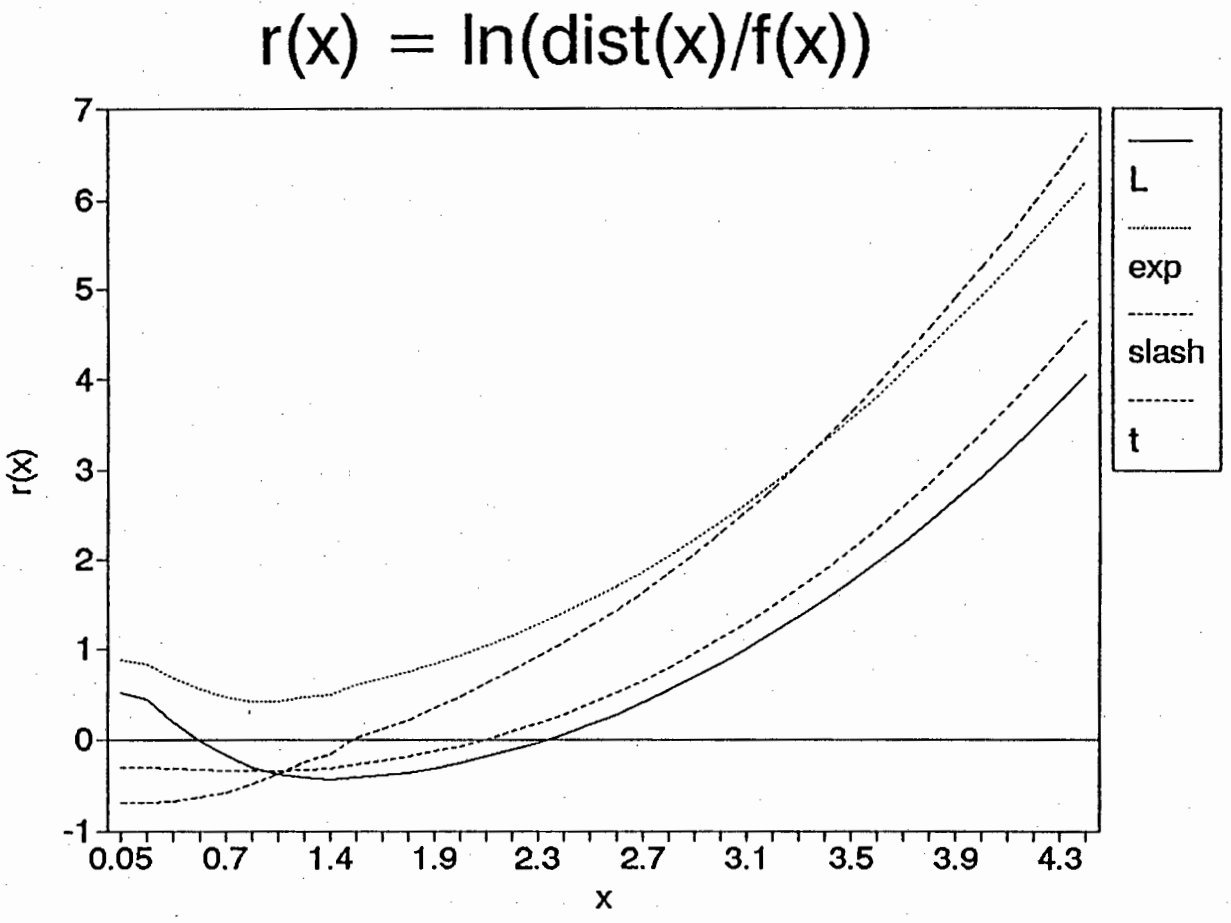
$$t(x) = \frac{\sqrt{k-2}\,\Gamma[\frac{1}{2}(k+1)]}{\sqrt{k}\quad\Gamma[\frac{1}{2}k]}\,\frac{\sqrt{2}}{\sqrt{k}}\,\frac{1}{(1+x^2/k)^{\frac{1}{2}(k+1)}\exp[-\frac{1}{2}x^2]}$$

and when $k = 5$, $t(x)$ reduces to

$$= \frac{\sqrt{3}}{\sqrt{5}}\,\frac{8}{3\sqrt{\pi}}\,\frac{\sqrt{2}}{\sqrt{5}}\,\frac{1}{(1+x^2/5)^3\exp[-\frac{1}{2}x^2]}$$

$$= 0.7370541\,\frac{1}{(1+x^2/5)^3\exp[-\frac{1}{2}x^2]}$$

Plotting the values of $r(x) = \ln(\text{dist}(x)/f(x))$ against x for $\text{dist}(x)$ varying over $l(x)$, $e(x)$, $s(x)$ and $t(x)$, and $f(x)$ the pdf of standard Gaussian distribution, we obtain the following graph:



4.4 Summary

In this chapter a factorial experiment was designed to compare 13 biased and 26 L_p-norm estimators under several factors. The random number generator and the simulation of the X-matrix and of pseudo-random numbers for eight distributions were discussed. The influence of heavy tails was noted.

Chapter 5

DISCUSSION OF SIMULATION STUDY RESULTS

This chapter consists of a discussion of the performances of the 13 biased and the 26 L_p-norm estimators fitted to the simulated data of Chapter 4. In section 5.1 judgement of estimators is discussed. Section 5.2 presents sets of 'best' estimators, a comparison of the WLS vs BFGS algorithm, an examination of an adaptive algorithm and some Bayesian remarks on the issues. Section 5.3 focusses on outliers in L_p-norm estimators, section 5.4 focusses on the moment ratio parameter, section 5.5 and 5.6 investigate what happens in the case of a full X matrix and section 5.7 consists of conclusions.

5.1 Judgement of estimators

5.1.1 Unbiasedness

In the class of unbiased estimators, the OLSE is the best linear estimator (BLUE) in the sense of minimum variance. In the presence of collinearity, the variance of OLSE can be inflated (due to small λ_i's) so that some biased estimators will be more suitable under a changed criterion, eg minimum mean square error. If the collinearity between some of the regressors is however consistently continued in the prediction region (eg in the neighbourhood of the X-observations), the effect of this collinearity on predictions will be less serious. If σ^2 is sufficiently small, β may be estimated by OLS with sufficient accuracy even if strong collinearities exist in X. Thus the choice of whether or not to use the OLSE should be based on the presumed relative magnitudes of the λ_i and unknown σ^2.

5.1.2. MSE criteria

Consider two competing estimators b_1 and b_2. If the matrix difference

$$S = MSE(b_2) - MSE(b_1) \qquad (5.1.1)$$

is positive semi-definite (psd), then b_1 is to be preferred to b_2. S will be psd if $w'Sw \geq 0$ for any non-zero vector w:nx1. The TMSE equivalent to MSE is that b_1 is preferred to b_2 whenever

$$\text{TMSE}(w'b_2) \geq \text{TMSE}(w'b_1), \text{ for every vector } w \qquad (5.1.2)$$

This MSE criterion is the so-called strong MSE criterion, and a weaker criterion for b_1 to be preferred to b_2 is that

$$\text{TMSE}(b_2) \geq \text{TMSE}(b_1) \qquad (5.1.3)$$

Examining these criteria it is worthwhile to point out:

(i) $\text{TMSE}(b_i)$ is the average squared Euclidean distance between b_i and β. One therefore seeks an estimator that minimizes this norm.

(ii) The relations (5.1.2) to (5.1.3) were defined in principal component (PC) estimation as ways to determine which PC's to eliminate.

(iii) Although only some criteria are explicitly stated here, there is a whole range of criteria available. For instance, all those criteria applied in PC estimation to eliminate PC's can be generalized. A detailed discussion of criteria appears in Vinod and Ullah (1981, Chapter 2).

(iv) Some authors perform comparisons based on the simulated relative efficiency (RE) of each estimator to OLSE. There appears to be no statistical analysis of these efficiency ratios, literally only comparisons of the summary values (eg $2 < 3$, or 2.9 is slightly better than 3, and so on)

(v) Empirical comparisons of estimators reveal that no one estimator is always clearly superior to the others. The conditions for superiority depend *inter alia* on the degree of collinearity, the orientation of β, and the value of σ^2. These factors should always be considered when choosing an estimator. Although some rough guidelines can be given, the optimal

estimator for any problem will be unique to that particular problem and no recipe or rule seems practicable at this stage.

In this thesis the performance of each of the (13+12×2+2= 39) estimators listed in Table 2.1, Table 5.1 and Table 5.4, over replications of data-sets of size 30 within (2×4×5×8=320) factor combinations was summarised by computing

$$\sum_{j=1}^{5} \sum_{i=1}^{100} (\tilde{\beta}_{ji} - \beta_j)^2 \qquad (5.1.4)$$

In (5.1.4) $\tilde{\beta}_{ji}$ is the j-th element of $\tilde{\beta}_i$, the estimate (via any method) of β in the i-th replication. The evaluations are based on the relative efficiency (RE) of each estimator compared to $\hat{\beta}$ of OLSE. Thus the tabulated relative efficiency values are

$$\sum_{j=1}^{5} \sum_{i=1}^{100} (\hat{\beta}_{ji} - \beta_j)^2 / \sum_{j=1}^{5} \sum_{i=1}^{100} (\tilde{\beta}_{ji} - \beta_j)^2 \qquad (5.1.5)$$

where $\tilde{\beta}$ is one of the estimators given in §5.2. The RE as defined in (5.1.5) is given in the tables of Appendix C, marked as equal weight RE's. Alternatively

$$\sum_{j=1}^{5} \sum_{i=1}^{100} (\hat{\beta}_{ji} - \beta_j)^2/\sigma^2_{j,LS} \; / \sum_{j=1}^{5} \sum_{i=1}^{100} (\tilde{\beta}_{ji} - \beta_j)^2/\sigma^2_{j,LS} \qquad (5.1.6)$$

The RE as defined in (5.1.6) is given in the tables of Appendix C, marked as diagonal element weight RE's. Lee and Birch (1988), called this RE the standardized empirical mean square error and $\sigma^2_{j,LS}$ is the theoretical variance of $\hat{\beta}_{ji}$.

An adjusted weighted RE measure, defined as

$$\sum_{i=1}^{100} (\hat{\beta}_i - \beta)'(X'X)(\hat{\beta}_i - \beta) / \sum_{i=1}^{100} (\tilde{\beta}_i - \beta)'(X'X)(\tilde{\beta}_i - \beta)$$

is also possible as a measure, but was not used because it focusses essentially on fitted values $X\tilde{\beta}_i$ rather than the individual parameters $\tilde{\beta}_{ij}$ in the vector $\tilde{\beta}_i$, which are the elements of interest here.

5.2 Comparison of Estimators

In this section we compare the estimators on the basis of the measure defined in (5.1.5) as we found that the measure defined in (5.1.6) not much different. We examine biased, WLS, BFGS, L_1 and L_∞ estimators, we make a comparison between these algorithms, we examine the 'adaptive algorithm', and we find the overall 'best' estimator in respective sections.

5.2.1 Biased estimators

The 13 biased estimators applied in the simulation study are given in Table 2.1 of Chapter 2.

5.2.1.1 Program

The program that derives the different estimates is given in Appendix P2. It was written in FORTRAN 5 and ran on a PC. Double precision was used throughout, although we have found in trial runs that it made little difference. The X matrix was standardized before any calculations were performed, then the SVD was computed. After any particular standardized estimate was obtained, we transformed back to unstandardized parameter estimates before calculating the particular statistic of interest, as discussed later.

The SVD and the OLSE's were computed by using the subroutine SVDCMP and SVBKSB of Press *et al.* (1985). To obtain all the biased estimates, SVBKSB was modified for each particular estimation procedure.

To avoid dividing by zero in the FPC estimation procedures, the k_i's were flagged as soon as the delta's (δ's) became smaller than 10^{-10} and in the subroutine that calculates the estimators, the delta's were then set equal to zero.

5.2.1.2 Some apparently best biased estimators

The RE's of the biased estimators are given in , Table C1.1 - C1.64, Appendix C. Tables F.1 to F.4 (Appendix F) present the 'best' three estimators in each category. The relative efficiencies were ranked, from highest to lowest. These rankings might be misleading, as values were used as obtained, but no statistical test was performed. It is difficult to say for instance that 1.7 is better than 1.6. Only RE's greater than 1.10 were taken as better than OLSE. Although we present at most three estimators at any given entry in these tables, the reader should always see the ranking in the context of the whole Appendix C1. Estimators that are of all roughly the same performance as one of higher rank are indicated by the symbol ~ as preceding subscript. Third ranked estimators that are not easily distinguishable from one or more lower-ranked estimators are indicated by the same symbol ~ but as following subscript.

Absence of entries means that there are no estimators that appear really better than OLSE (RE values < 1.10 are interpreted as 1, and not better than OLS). The abbreviations used for the distribution in the tables are U for Uniform, N for Normal, CN4 and CN5 for the Contaminated Normal distributions with kurtosis = 4 and 5 respectively, L for Laplace, t for Student's t, E for Exponential and S for Slash distribution.

From Table F.1 (σ = 0.01) we note that OLSE performs satisfactorily when the collinearity level was modest ($<\sqrt{56}$), except under the Slash distribution. For the first seven distributions (ie U through E) we note that in the

highest collinearity level (99:99) the GR family always is part of the 'best' three. Therefore, for this particular variance level of $\sigma = 0.01$ we can conclude that the GR family is definitely appropriate. Although we only fitted two particular members, GRHK and GRT, in real data one may fit more siblings of this family.

The Slash distribution stands alone and we see in general the FPC family and the JR family become prominent. Note in the Appendix that, under the Slash distribution, the relative efficiency of these families is much higher than for any other distributions. For instance in the U through E distributions a biased estimator was usually better than OLS by RE values from 1.10 to 2.50 whereas RE values that are high (eg highest 49) are not uncommon for the Slash. To classify an estimator better than OLSE for relative efficiencies less than 2, is debatable. The other advantages of OLS seem likely to outweigh small gains in relative efficiency. However in the Slash distribution the relative efficiencies really become obvious and the choice of 'best' estimator (we mean better than OLSE) feasible.

The completeness of Table F.2 is obvious. The variance has increased to 1.0. A glance at the high relative efficiencies in Appendix C1, table C1.17 through to C1.32 indicates how strong the advantages of biased estimators over OLSE become. Here we may note how the FPC family becomes important, explicitly ranked in the best three for 59 out of a total of 80 (8 x 2 x 5) possible blocks. The R family was ranked 30 times, with RLW occurring at 28 of these 30 times.

Table F.3 is also an indication of the increased importance of biased estimators as the variance level increases. The RE's listed in Appendix C1, Table C1.33 through to C1.48 are much higher than in F.2. Here we note the importance of the FPC family, ranked in the first three in all 80 possible blocks. The RLW estimator was ranked 68 times amongst the best three. Furthermore in the Appendix we see how remarkably well these biased estimators perform relative to OLSE, particularly when the collinearity is high.

In Table F.4, where the variance level is now at 100, the RE's listed in Appendix C1, Table C1.49 through to C1.64 are now exploding. Here note the importance of the FPC family, ranked amongst the first three in all 80 possible blocks. The RLW estimator, ranked 66 times amongst the best three. From the Appendix it is clear that, under any distribution, the RE of the estimator classified as 'best' is well above the RE of all the other estimators - meaning that the rank positions of the estimators become much more defined, and with more confidence one can claim that a particular biased estimator is superior (in this simulation) to OLSE. This observation appears true even when we have lower collinearity levels.

From the simulation design we can make the following general comments, about some factors that effect the relative efficiency of the biased estimators:

When the variance is low (σ=0.01): collinearity is not an important factor, and for the first orientation of the betas there seems to be evidence for the superiority of the biased estimators, while in the second orientation the very low RE (marked ** in the Appendix C1) suggest the superiority of OLS over biased estimators. It seems that the distribution of the error terms does not play any role, except for the change of RE's in the case of the Slash (for which technically the variance does not exist).

As variance increases, the relative efficiencies increase over all factors (orientation, collinearity and distribution). The RE for X matrices of high collinearity is much higher than for lower collinearity levels. Furthermore it seems as if the RE is constant as we move from distribution to distribution. It is only in the case of the Slash that the RE seems to change, notably so in the X matrices with the highest collinearity, where the RE is exploding.

In the Slash distribution the variance level of the N(0,1) deviate was taken as the required variance level (scale), in the programs, ie for biased, WLS, BFGS, L_1 and L_∞ estimation. In the case of the Slash distribution we can conclude that scale and collinearity level play an important role in the relative efficiencies. We suggest that only in

distributions with wild outliers will RE values explode. We conclude that biased estimators are influenced only by variance, orientation, and collinearity but are impervious to distribution changes - at least to the regular distributions of this study. From the Slash RE's we believe a biased estimator is likely to be robust against outliers.

5.2.2 L_p norm estimators

5.2.2.1 L_p-norm algorithms in general

The L_p-norm solutions were found via two algorithms WLS and BFGS. We will discuss the adaptive algorithm for determining p and explain the difference between the four main programs. However we first need to examine the 12 L_p-norm estimators.

The 12 L_p-norm estimators fitted via WLS and BFGS are defined in Tables 5.1 and 5.4. We summarise these specific estimators as follows:

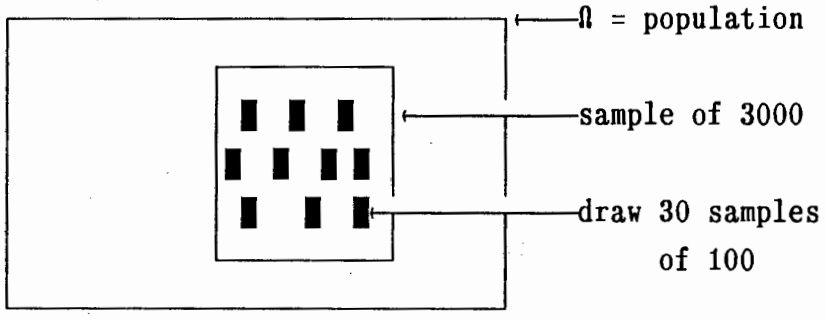
Summary: L_p-norm estimators

Choice of p		Algorithm
Kurtosis based		
(samples of 30)	Barr, Sposito, Gogga	adaptive via WLS; BFGS
	Harter-Sposito	one call WLS/BFGS
(sample of 3000)	Barr, Sposito, Gogga	one call WLS/BFGS
(population)	Barr, Sposito, Gogga	one call WLS/BFGS
Literature choice	Ekblom (p=1.25)	one call WLS/BFGS
	Forsythe (p=1.50)	

The choice of the above L_p-norm estimators stems from Table 3.1.1, the so-called theoretical values of p. Because the variance of the kurtosis is so large (see Table 3.1.2), the kurtosis based on the whole sample of deviates (30 × 100) should be an estimate nearer to the true kurtosis (ie an estimate whose confidence interval is much narrower than an estimate based on 30 deviates). Furthermore the population kurtosis was known for 7 of the 8 distributions, and provides a useful contrast with the large sample estimated kurtosis, in applications to parameter estimator and relative efficiency. These values were intended to provide some objective standards of comparison for the sample size 30 study.

Bayesians would not estimate parameters as we do, but derive posteriors. One could imagine that the underlying theoretical distribution of error terms is precisely known, and thus a choice of p could be based on one of the theoretical choices of p. In practice, knowing the true underlying distribution is doubtful, but the idea that a particular small sample is drawn from a larger sample, is feasible. After a researcher has worked with a certain kind of data, one might use such knowledge to build a model of how the underlying 'large' sample behaved, and thus might base the prior for p, or for the moments of current small samples, on the observed p or observed moments of past larger samples.

The following Venn diagram illustrates the drawing of small samples from a pool of larger samples, within the population framework:



Thus, in expectation,

variance of sample kurtosis based on 30 $\gg$ variance of sample kurtosis based on 3000 $>$ population variance of sampling kurtosis

We fitted the p found from the population kurtosis, to compare the results for the p from the adaptive estimators, where we suspect that the large variance of the sample kurtosis should be dampened by the adaptive procedure. This issue is taken up again in the later discussion of the apparent success of the adaptive algorithm.

The adaptive algorithm: was suggested by Gonin and Money (1985), as mentioned in §3.1.4.3. The programming of the algorithm for this simulation study can be styled as follows:

1. Read next Y(30), (X(30,6) is already in memory)

2 Fit OLS, find residuals, compute kurtosis and estimate p

3 Fit L_p-norm algorithm, find residuals, compute kurtosis and estimate p

4 Test p for convergence: If No: go to step 3,
Yes: store p and betas go to step 1.

As before the estimate of p can arise from Barr, Sposito or Gogga (Chapter 3) methods and the L_p-norm algorithm can be WLS or BFGS. For this study notice that the maximum number of loops between 4 and 3 was set at 30 (the outside loop), the maximum iterations (within WLS or BFGS) was set at 50. The convergence tolerance for p was set at 0.001.

The four main programs:

We applied all the estimators in one program. This program is so large that the size of the FORTRAN version exceeds the allowed 64K segments of MICROSOFT FORTRAN 5. The gt option was applied at compilation, but still there were problems. Past experience also warned that when we allow the

size of the program to approach 64000 (bytes), the behaviour of the MICROSOFT FORTRAN compiler is not what it should be. (The opinion of Digby (1992) is similar). Instead of creating a stack overflow, if the storage space of a matrix cannot fit into the 64K segments, we have found that the last rows and columns just overflow into the first rows of the matrix.

The large program problem was overcome by avoiding the use of unnecessary matrices or vectors. Whenever a value is calculated, and will be needed later, we write it to an appropriate file and then reuse the storage space again. Of course using a number of files causes other problems. The number of files attached to a program is limited by the configuration of the computer and the buffer size. Furthermore by writing to files the program is slowed, as these files are on the hard drive and not in the RAM area.

To cope with the above problems we divide the estimators into 4 main programs, DRELP1 through DRELP4. These main programs find the following estimators:

DRELP1	DRELP2	DRELP3	DRELP4
WB-C	WSAMB	FB-C	FSAMB
WS-C	WSAMS	FS-C	FSAMS
WG-C	WSAMG	FG-C	FSAMG
WHART	W125	FHART	F125
WPOPB	W15	FPOPB	F15
WPOPS		FPOPS	L_1
WPOPG		FPOPG	CHEB

The explanation of the abbreviations can be found in Tables 5.1 and 5.4. Notice that the L_1 and the Chebychev estimators are fitted in program 4, together with the BFGS algorithm. The programs were sometimes changed (usually at format statements) to accommodate the large outliers in the Slash distribution, and the estimator found using the population kurtosis was absent in the case of the Slash.

The way the data are read and the OLSE found, is the same in all four programs and is in fact the first part of the biased estimators program.

One drawback of all four main programs is that the testing for convergence (in the adaptive algorithm) is done within the main program, and all the bookkeeping, writing to files, computation of relative efficiencies and so on are carried out in the main program. These programs would look much neater or more readable if the bookkeeping of residuals, betas, error codes, norms, iterations, RE and so on could have been managed inside a subroutine. For example part of a main program would consist of

```
c     Method Barr, adaptive algorithm, BFGS
      IER = 0  }
               } initialize bookkeeping  and find p from residuals of OLS
      .....    }
c     start of outside loop
100   call BFGS(......)
c     was BFGS fit succesful?
      If (IER=...) then do
c     find residuals
      call SIGMA(....)
      call MOMENT(....)
      p =
c     test for convergence of p
      If(     )

c     p has converged, now copy all relevant statistics to files, and
c     restore all destroyed matrices
```

For further detail see the program codes, in Appendixes P4 trough to P5..

When all the estimators inside a program had been applied in all 100 repetitions, the files containing the betas, relevant bookkeeping and residuals were saved, and the RE was calculated and written to another file.

5.2.2.2 L_p-norm estimators via WLS

The 12 L_p-norm estimators fitted in the simulation study are given in Table 5.1. In general the leading W in a name, indicates that the WLS program was used, the characters B, S and G are the methods to find p, viz Barr, Sposito or Gogga (explain in §3.1.4.3); the -C indicates that the adaptive algorithm was used; POP stands for population, and SAM means the sample, hence WSAMG and so on. The fitted RE of the WLS estimators are given in Appendix C2, while the summary statistics of the programs are given in Appendices E1A - E1D (summary statistics of DRELP1) through to E2A - E2D (summary statistics of DRELP2), each variance level is labelled as A, B, C and D, where E1A contains summary statistics of estimators fitted via DREGLP1, σ = 0.01, through to E2D, summary statistics of estimators fitted via DREGLP2, σ = 10.0.

Table 5.1. L_p-norm estimators using WLS

Estimator	Description
WB-C	WLS program, p calculated via Barr, adaptive algorithm
WS-C	WLS program, p calculated via Sposito, adaptive algorithm
WG-C	WLS program, p calculated via Barr, adaptive algorithm
WHART	WLS program, p calculated via Harter, kurtosis form 30 OLS residuals, one call
WPOPB	WLS program, p calculated via Barr, kurtosis from specific population (distribution), one call
WPOPS	WLS program, p calculated via Sposito, kurtosis from specific population (distribution), one call
WPOPG	WLS program, p calculated via Gogga, kurtosis from specific population (distribution), one call
WSAMB	WLS program, p calculated via Barr, kurtosis from sample of 3000 generated deviates, one call
WSAMS	WLS program, p calculated via Sposito, kurtosis from sample of 3000 generated deviates, one call
WSAMG	WLS program, p calculated via Gogga, kurtosis from sample of 3000 generated deviates, one call
W15	WLS program, p = 1.5, one call
W125	WLS program, p = 1.25, one call

5.2.2.2.1 Program

The WLS algorithm is an iteratively reweighted least square technique (see points 3, 4 and 5 on p3.15). It was developed by Schlossmacher (1973), extended by Sposito *et al.* (1977) for $1 \leq p \leq 2$ and in 1981 Barr extended the one regressor variable case of Sposito to an X matrix with more than one independent variable. Barr suggested that WLS should be used for $1 < p \leq 2.6$.

Though convergence has not been proved, Sposito claimed that the WLS routine has converged for every one of the problems attempted by his group and that the rate of convergence was sufficiently rapid.

We found in many cases there was no convergence (see §5.2.2.6), and that compared to BFGS, WLS was slow. This conclusion might be due to the fact that all values of p were considered, and not only those that fall into the interval suggested by Barr or the interval of Sposito *et al.* See also the comments on the values of p in the discussion of the performance of the WLS technique.

The documentation for the WLS program can be found in Barr (1981), and additional comments on the algorithm can be found in Sposito *et al.* (1977). The following additional comments on the use of the WLS algorithm may be useful:

1. The X matrix sent to WLS as Z is an X matrix augmented with an extra column Y, the response variable, ie Z = [X Y] where X is the X matrix without the column of ones (Z(30,6): 5 independent variables and one dependent).

2. The minimum norm returned from WLS is stored in SD, iterations in IT and the failure indications in IFAULT, where

   IFAULT = 0 if the routine converged
   = 1 if the return was due to an increase in the norm
   = 2 maximum iterations exceeded
   = 3 X moment matrix is non-singular (all though in program code this was never called.

An error code of 0 or 1 is taken as a success.

The error code reported as a 3 in Appendix E1 is in fact the number of repetitions in which the p loop did not converge.

3. The betas are stored in the matrix BWLS(10,1). The 10 is just the maximum number of the regressor variables plus a constant, chosen by Barr, and can be changed within WLS if necessary. Note that in this study only the first 6 positions are used. The estimated betas are stored in the first 5 positions and the constant (an estimate of β_0) is stored in the (6,1) position.

4. When we return to the main program we check for norm increase (ie IFAULT = 1). We have found with IFAULT = 1 and IT = 2 that the betas returned from WLS are usually OLSE, and hence do not correspond to the p sent to WLS. Therefore in the main program we check to see if the betas are OLSE, and if they are, we change p to be equal to 2, and the resulting betas (OLSE) are incorporated into the RE calculation. As our basic concern is the RE, we reason that adding an OLSE is not contributing to the difference in the RE. Cases like this are counted and reported in the Appendix E1, and E2 under the column heading ignr.

5. The limiting parameter vector (LPV) inside WLS consists of two elements ie EPS = 10^{-6} and maximum iterations set at 50.

The constant EPS has three functions in the WLS algorithm

(i) EPS is a constant used for assigning zero weight to observations ie if in the k-th iteration, the i-th residual is less or equal that EPS, then in the (k+1)-th iteration the weight of the i-th observation is zero.

(ii) The variable EPS2 is set equal to twice the value of EPS and is taken as a convergence criterion. If the maximum absolute difference in residuals from the k-th to the (k+1)-th iteration is less than

2×EPS, the routine is considered to have converged. The program code for this convergence criterion is:

```
      ......
      ISW = 0
      DO 4 I = 1,N
          RES =       /*i-th residual at (k+1)-th iteration
          ABSRI = ABS(RES)
         IF (ABS(ABSRI-ABS(R(I))) .GT. EPS2) ISW = 1
                     /* R(I) is i-th residual at k-th iteration
          ......
4     CONTINUE
      .....
      IF (ISW .EQ. 0) RETURN
```

Inside the DO loop, the program compares each residual of the (k+1)-th iteration with that of the k-th iteration. When this absolute difference is greater than EPS2, ISW is set to 1 (flag to continue). If (N-1) residual differences were zero and only one by chance is not, the flag is set to 1 and more iterations will be carried out until all N differences in residuals are zero.

From an L_p-norm viewpoint, where the objective is to minimize the p-norm (ie $\Sigma|\hat{\epsilon}_i|^p$) a more logical test would be to test norm convergence. Due to less sensitivity to rounding, a test based on a suitable definition of norm convergence would converge faster than one that is based on whether N residual differences have converged.

(iii) The third situation where EPS is used as a criterion, is in a test for norm increase. For a discussion on this test see Porter and Winstanley (1979). Sposito *et al.* (1977) claim empirical results strongly suggest that an increase in the norms only occurs when the process has converged to a solution within tolerance for rounding errors.

In the test for norm increase, a variable SD3 is set equal to the difference between the value of the L_p-norm from the (current) (k+1)-th iteration and the k-th iteration. If this difference is greater than EPS, then the algorithm returns to the main program.

Thus, to summarise, convergence in WLS can be obtained via (ii), EPS2 = 2 x 10^{-6} and via (iii) EPS = 10^{-6}.

As convergence criteria in the other programs (L_1, CHEB and BFGS) are set to 10^{-6}, we wonder how one would put these two criteria of WLS (EPS2 and EPS) on the same level as those of the other programs. To explore the effect on the convergence when both EPS2 and EPS are set equal to 10^{-6} in the WLS routine, we ran some trial runs with a program WLS2 (EPS = EPS2 = 10^{-6}) and compared this with Program WLS1 (EPS = 10^{-6}, and EPS2 = $2x10^{-6}$). We make the following observations:

(a) The number of repetitions converging in WLS2 is usually lower than that of WLS1. This property is clear from the convergence criteria:

absolute difference in residual > $2x10^{-6}$ > 10^{-6}

Table 5.2: An empirical comparison on convergence for WLS1 and WLS2

Datafile	B99101C5		B90102E		B99991T		C70302E		D99992C5		D70301N	
Routine	WLS1	WLS2	WLS1	WLS2	WLS1	WLS2	WLS1	WLS2	WLS1	WLS2	WLS1	WLS2
Estimator												
WB-C	76	76	65	63	67	66	61	58	69	67	74	74
WS-C	69	67	65	64	61	60	59	56	67	67	65	65
WG-C	62	61	61	60	56	55	55	53	57	56	74	72
WHART	99	99	93	93	98	97	94	94	99	99	100	100
WPOPB	99	99	21	21	18	18	8	8	100	100	100	100
WPOPS	77	39	100	100	100	100	99	99	1	1	100	100
WPOPG	10	8	99	99	97	97	98	98	4	4	100	100

Table 5.3: A comparison on convergence for WLS1 and WLS2 - Slash distribution

Datafile	A90101S		B99992S		D99102S	
Routine	WLS1	WLS2	WLS1	WLS2	WLS1	WLS2
Estimator						
WB-C	53	49	29	27	22	21
WS-C	56	55	62	62	62	63
WG-C	75	74	63	63	64	63
WHART	85	85	73	73	72	72

Thus a difference in absolute residual would be found quicker in WLS1 than in WLS2. The numbers of converging repetitions are given in the Tables 5.2 and 5.3.

The separate table for the Slash distribution is given because in §5.2.2.6 we will show that, for distributions U through E, the BFGS program was much more stable than WLS. However for the Slash distribution the choice between the WLS and the BFGS algorithm is not so clear.

(b) The results in WLS1 and WLS2 were identical when return to the main program was due to a norm increase (ifault = 1).

(c) When no convergence was reach in WLS2, but convergence was found in WLS1, the number of iterations in WLS1 is near 50 (the maximum iterations) and thus when WLS2 is run, and then more iterations are needed, maxit is exceeded and non-convergence was declared. (This phenomenon is particularly apparent in file B99101C5, for estimator WPOPS, where the number of convergence WLS1 is 77, compared to 39 for WLS2.

(d) Usually when IFAULT = 0 (convergence) was reported we found that WLS2 needed between 0 and 3 more iterations for convergence (In some cases WLS2 did not converge within the outer p loop but ended in non-convergence, or singularity).

(e) The norms and betas found using WLS2 correspond for roughly at least the first 4 decimals to WLS1.

Our conclusion was to use algorithm WLS1 (EPS = 10^{-6}, EPS2 = $2x10^{-6}$) for the following reasons:

In the original study reported by Sposito *et al.* (1977), they chose EPS2 = $2x10^{-6}$.

As previously reported we find the results with WLS1 to have more apparently converging repetitions than program WLS2.

It is an open question which of the two tests, norm increase or maximum absolute residual, should be the criterion for convergence.

Results reported in Appendix C2 are for the WLS1 (labelled as WLS program) routine with EPS = 10^{-6} and EPS2 = 2 x 10^{-6}. In future research, the two reported problems with WLS should be investigated: The fact that WLS sometimes returns OLSE for a p that is not 2 is a serious drawback, and should be checked in WLS and not in the main program, as in this study. For an iterative weighted least square algorithm the convergence criterion of maximum absolute difference in residuals is much stricter than a criterion based on norm convergence. It is strongly recommended that in using WLS, a norm convergence test (such as the one used in BFGS) should be included instead of the maximum absolute residual test. It is suspected that this change would speed up the WLS algorithm considerably.

The programs (DRELP1 and DRELP2) to obtain the WLS L_p-norm estimators is given in Appendices P2 and P3, and the summary statistics on these programs in Appendix E (E1A through to E2D).

5.2.2.2.2 Some optimal WLS estimators

The relative efficiencies (RE) of the L_p-norm estimators using the WLS program are given in Appendix C2, Table C2.1 - C2.64. Tables F.5 through to F.8 present the apparently best three estimators in each category. The remarks in the first paragraph of section 5.2.1.2 are applicable here, except, that because of space limitations we will no longer have a symbol indicating that the third ranked estimators are distinguishable from lower ranked estimators. An extra symbol will be used in these tables, namely *, which when preceding any ranked value indicates that less than 50 out of a 100 repetitions converged. Thus a * symbol is an indication of a failure of the particular program. We include these values, when by rights one should ignore these values in this study, because with real data, or in a future simulation study, the failure for a particular program can be investigated, the program changed and then fitted to the particular data set. We suspect the failures are usually due to the restriction of the ALPV.

In the case of the L_p-norm estimators (via WLS) we immediately see that the variance levels, collinearity and orientation of the β did not influence the fit. Therefore the presentation of the discussion of the ranked tables for WLS, will be different from that in the case of the biased estimators. Rather, we discuss the results in the framework of each distribution. This approach will also complement the theoretical background of L_p-norm estimators (ie the optimal estimators for certain distributions are L_p-norm estimators (page 3.3))

Uniform distribution: Theoretically the Chebychev estimator is optimal for the Uniform distribution. If we examine the rank entries in table F.5 through F.8 we notice that the estimator chosen is usually an estimator for which p is calculated by the Gogga function. This predominance arises from the structure of the program. Although the subroutine for the Gogga function is stable for p approaching 30 ($p<30$), we find that when this subroutine is used within the WLS program, function values calculated in the FFUNC subroutine were unstable (ie in the region of 10^{20}), and math overflow occurs. We therefore set $p = \infty$ whenever the kurtosis approaches

1.84 (which corresponds with a p of roughly 14). Thus the estimators in Tables F.5 through F.8 are actually Chebychev estimators. Because Chebychev is optimal, they rank among as apparently best of all estimators, compared to OLSE.

All three adaptive estimators found under the Uniform distribution over all variance levels where unstable (less than 50 converge). The average p values reported (see Appendix E1A/B/C/D) in these adaptive estimators are 2.09 and higher. This finding support those of Barr and of Sposito's (ie Barr suggests $p \leq 2.6$ and Sposito claimed WLS is stable for $p \leq 2.$). In this study we have found WLS unstable when $p \geq 2.09$.

Normal and CN4 distributions: The missing cells in Table F.5 through F.8 reveal immediately where the OLSE appeared to be the best estimator. This superiority of OLSE is expected, since for $p = 2$ the L_p estimator under the Normal distribution is the MLE (Chapter 3, p3.3). The kurtosis of the Contaminated Normal is 4, and CN4 is the nearest (in terms of kurtosis) to the Normal amongst all the distributions under consideration. Notice that in Table 3.1.1 values for p for CN4 were in the range 1.40 - 1.56.

CN5 distribution: The number of ranked estimators increases as the variance increases (the tails get heavier) but this increase might be due to chance. Note that of the 10 ranked estimators in table F.8, 9 where unstable, the error code return in these cases are IFAULT = 3 (to many iterations in the p loop). Because only 20 estimators were ranked, of which a large number were unstable, we concluded that for this distribution an L_p-norm estimator with $p = 2$ is best.

Laplace distribution: Over all variance levels L_p-norm estimators are chosen for all orientations and collinearity levels. From the theoretical background we know that the optimal estimator for the Laplace distribution is the L_1-norm estimator. Then if we examine at the theoretical values for p, (Table 3.1.1) we expect that the estimators in which p was calculated by the methods due to Sposito, Harter or Gogga, should be ranked amongst the best. However, on the contrary we found the W1.25 ($p = 1.25$) estimator ranked amongst the best in 32 out of the 40 cases. Perhaps then $p = 1.25$ is

the appropriate choice. Note that this value of p is the choice of Ekblom for the Laplace distribution. The fact that p = 1.25 emerges as best may also explain why the methods of Sposito and Gogga are in the shadow of the Barr method. Given the population kurtosis, the method of Barr yields a p of 1.25 (the optimal p), Sposito p = 1.0 and Gogga a p of 1.0.

These findings apply to the WLS algorithm, which we suspect is not always stable. Later in this chapter we will discuss the optimal p, p = 1, fitted with the stable L_1 algorithm. We expect this algorithm (L_1) to be superior to the WLS algorithm, and for the Laplace distribution, it should yield an estimator that is superior to all other estimators (optimal p = 1.0, the MLE).

Student's t distribution: The number of ranked blocks (where a block denotes an specific X-matrix) and estimators found as σ increases can be summarised as follows:

σ	0.01	1.0	5.0	10.0
blocks	4	8	7	10
estimators	10	18	15	22

It seems that as σ increases the L_p-norm estimators become more prominent over OLSE. However if we look at the actual RE's there is no evidence to support this impression. The RE's fluctuate around fixed numbers and do not change from one variance level to another. Thus although the RE does not increase with the variance levels, there is an increase in the number of L_p-norm estimators ranked better than OLSE.

A value p = 1.5 was found to be amongst the three best in 21 out of a possible 28 blocks with entries. A value p = 1.5 was also the choice of Forsythe (Table 3.1.1). There was 19 missing cells, implying that no estimator was better than OLSE (p = 2) in 19 blocks. Thus the choice of p could be 1.5 or 2.

Exponential distribution: This distribution is the only non-symmetric distribution used in this study. The L_p-norm estimators were ranked better

than OLSE in 39 out of the 40 blocks. The W1.5 and W1.25 estimators appear most often; p = 1.25 was found in 32 out of the 39 blocks, and p = 1.5 was ranked in 30 of the 39 blocks. Thus a choice of p = 1.5 or p = 1.25 seems appropriate. Notice that p = 1.5 was also Forsythe's choice.

Slash distribution: If we examine the value of the RE over the variance (scale), orientation and collinearity levels there seems to be no pattern. The highest RE in tables were:

σ	0.01	1	5	10
value	9217	17947	8474412	226522
X matrix	99:10:2	99:99:2	70:30:1	99:99:2

Here the first four digits in the X matrix indicator are the value of $a_1:a_2$ and the fifth digit is the orientation of the β, 1 for largest and 2 for smallest eigenvector. Thus we conclude that collinearity, orientation and the variance (in the sense of scale) play no role. Also see the comments in section 5.3, were the values in the above table will be investigated further, and the comments under the Slash distribution when we fit the BFGS algorithm.

In the 40 blocks, the adaptive estimators were ranked (in the best three) 40 times and the WSAMB estimator was ranked 27 times out of 40. The frequencies of some L_p-norm estimators ranking amongst the three best are summarised in the following table:

	σ				
	0.01	1.0	5.0	10.0	tot
WB-C	7	3	6	3	19
WS-C	3	5	6	3	18
WG-C	6	9	7	9	30
WSAMB	8	7	4	8	27

From the above table it seems that the Gogga method is superior to the others. The WB-C estimator seems to be unstable (see the entries marked with a * in Tables F.5 through F.8). The question of adaptive estimator

choice will be discussed later, with the claim of and Gonin and Money (1989) that the 'method' (Barr, Sposito or Gogga) does not matter.

The reason why WSAMB achieved high RE's might be suggested by the method of Barr,

$$p = \frac{9}{(\text{sample kurtosis})^2} + 1$$

$$\cong 0 + 1 = 1$$

The WSAMB estimator is in effect the L_1-norm estimator. We will refer to this equivalence later in this chapter. (Recall that in the case of the Slash distribution the population kurtosis is unknown and we could not fit 3 of the estimators).

In conclusion if we examine only the 12 L_p norm estimators that use WLS, then collinearity, variance, and orientation play no role. The most important issue is the distribution. We have found that for the distributions U through to E there is a slight improvement over OLS in using L_p estimation. In the case of the Slash the improvement is excellent and the superiority and the adaptivity of the L_p-norm estimator emerge. In most cases where the Gogga function are used, the WLS program seems to be unstable. As instability is also found under other criteria there is not a clear cut case, against the WLS program, as WLS and Gogga are confounded. On its own we have found the Gogga function stable, as long as the bracketting of the minimum and maximum is done correctly.

Roughly, the choice of p and the L_p estimator fitted (under WLS) can be summarised as

	Distribution							
	U	N	CN4	CN5	L	E	T	S
p	∞	2	2	no clear	1.25	1.5/1.25	1.5	<1
	L_∞	L_2	L_2	choice	$L_{1.25}$	$L_{1.25}L_{1.5}$	$L_{1.5}$	WG-C

It should be pointed out, that only in the case of the Slash distribution, did one of the adaptive estimators emerge superior. In all the other distributions, the optimal p was fitted, by a fixed value estimator, that was found with one fit (or call to WLS), under the adaptive estimator. This issue will be discussed further in §5.2.2.5.

5.2.2.3 L_p norm estimators via the BFGS program.

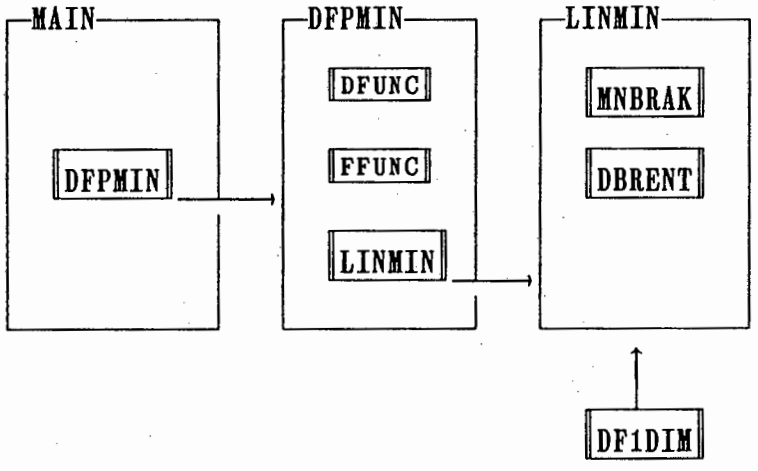
The 12 L_p-norm estimators fitted in the simulation study are given in Table 5.4. In general the F indicates that the BFGS program was used, B, S and G are the methods to find p via Barr, Sposito or Gogga (§3.1.4.3), POP stands for population, and SAM means the sample. Table 5.4 contrasts onto Table 5.1.

Table 5.4 L_p-norm estimators using BFGS

Estimator	Description
FB-C	BFGS program, p calculated via Barr, adaptive algorithm
FS-C	BFGS program, p calculated via Sposito, adaptive algorithm
FG-C	BFGS program, p calculated via Barr, adaptive algorithm
FHART	BFGS program, p calculated via Harter, kurtosis form 30 OLS residuals, one call
FPOPB	BFGS program, p calculated via Barr, kurtosis from specific population (distribution), one call
FPOPS	BFGS program, p calculated via Sposito, kurtosis from specific population (distribution), one call
FPOPG	BFGS program, p calculated via Gogga, kurtosis from specific population (distribution), one call
FSAMB	BFGS program, p calculated via Barr, kurtosis from sample of 3000 generated deviates, one call
FSAMS	BFGS program, p calculated via Sposito, kurtosis from sample of 3000 generated deviates, one call
FSAMG	BFGS program, p calculated via Gogga, kurtosis from sample of 3000 generated deviates, one call
F15	BFGS program, p = 1.5, one call
F125	BFGS program, p = 1.25, one call

5.2.2.3.1 Program

The BFGS algorithm: The Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm was discussed in section 3.1.4.2. We used the program code for the BFGS algorithm as given by Press *et al.* (1985). It consists of various subroutines linked together, and can be illustrated by the following diagram:



The function of each subroutine can be summarised as:

DFPMIN: performs BFGS minimization on a function FFUNC, using its gradient as calculated by a routine DFUNC;
DFUNC: calculates the gradient of the FFUNC;
FFUNC: calculates the function value;
LINMIN: implements line minimization (one-dimensional, through the function DF1DIM) and is described by Press *et al.* (1985) as a kind of bookkeeping swindle, and 'slightly dirty FORTRAN';
MNBRAK: searches in the downhill direction and brackets a minimum of the function;
DBRENT: given a function and its derivative, and given the bracketing of the function, this routine isolates the minimum to a fractional precision by a modification of Brent's method that uses derivatives;
DF1DIM: is an external function that must accompany LINMIN. It is used as an interface between a multidimensional minimization strategy and a one-dimensional minimization strategy.

The above description is short and if the reader wants to implement the algorithm, as a whole, it is crucial that Press *et al.* (1985), should be consulted. Note that the subroutines have been changed to fit the needs of this study. In particular, in finding the function value we have to send more program parameters (ie X, Y and p) through to DFPMIN, which in turn sends them through to FFUNC, DFUNC, LINMIN, MNBRAK and DF1DIM. This change could have been avoided by declaring X, Y and p as common, but had they been in the common area more problems would have occurred.

The following notes apply on the use of the BFGS algorithm:

1. The subroutines of Press as modified for this study, were all used in double precision.

2. The only element of LPV that is passed to DFPMIN is the convergence criteria for the norm FTOL = 10^{-6}. All other tolerance limits can be found at the beginning of each subroutine, and the interested reader is referred to these. In both subroutines DFPMIN and DBRENT the maximum iterations is set at 50.

3. DFPMIN tested for convergence in the following manner

```
IF (2*ABS(FRET-FP).LE.FTOL*(ABS(FRET)+ABS(FP)+EPS)) RETURN
```

where FRET is the minimum norm at the (k+1)-th iteration and FP is the minimum norm at the k-th iteration. FTOL is the convergence criterion and EPS = 10^{-10} is taken as a very small value to ensure that ABS(FRET)+ABS(FP)+EPS > 0.

4. In the original DFPMIN of Press *et al.*(1985) the code includes:

```
      DO 17  I = 1,N
         FAC = .....
         FAE = .....
17    CONTINUE
      FAC = 1/FAC
      FAE = 1/FAE
```

These last two steps caused problems when FAC and FAE are very small, ie virtual division by zero. To obviate this situation we add a test for small values for FAC and FAC before division:

```
17   CONTINUE
     IF (ABS(FAC).LE.SMALL.OR.ABS(FAE).LE.SMALL) THEN
         IER = 4
```

where SMALL = 10^{-25}, and IER = 4 is the error code describing this situation. When an error code of 4 is encountered we suspect multiple solutions and test if p = 1, in which case L1 is called and then control from DFPMIN is returned to the main program. Further discussion on the error code 4 will be given in section 5.2.2.3.2

5. The following error codes are possible within the BFGS algorithm:

```
IER = 0   successful BFGS fit
    = 1   only when L1 or the CHEB estimator is called
    = 2   maximum iterations (50) exceeded in DFPMIN
    = 3   maximum iterations (50) exceeded in DBRENT
    = 4   divide by zero in DFPMIN
```

An IER > 0 causes DFPMIN to return to the main program.

When an IER = 4, was encountered in DFPMIN, and we have found p = 1, the L_1 algorithm was called the following error codes were possible:

```
IER = 40: optimal, non-unique solution
    = 41: unique solution
    = 42: calculations terminated prematurely due to rounding error
```

(note in the programs, error code 40 and 41 are indication of a failure in DFPMIN but a successful fit via L_1, and thus in program code they are flagged as -40 or -41, to indicate a successful call (IER $\leq$ 1))

When p is found to be equal to infinity in estimator algorithms which then call the CHEB estimator, the following error codes are allocated:

An error code 0 or 1, returned from CHEB is set equal to 0;
an error code 2 returned from CHEB is set equal to 20 (failure in the CHEB routine).

In the case of program 4 (dreglp4) the error codes used within the L1 or the CHEB estimators are as described in algorithms elsewhere (§ 5.2.2.4). The error codes and other statistics of the programs are summarised in Appendix E3- and E4- for each variance level.

5.2.2.3.2 Optimal BFGS estimators

The relative efficiencies (RE) of the L_p-norm estimators via the BFGS program are given in Appendix C3, Table C3.1 - C3.64. Tables F.9 through to F.12 present the apparently best three estimators in each category. The remarks given previously in section 5.2.2.2.2 are applicable here.

Basically the conclusion from BFGS is the same as that from the WLS program, in section 5.2.2.2.2. We will discuss the stability of the WLS and the BFGS program in the following section.

In the case of the BFGS L_p-norm estimators we see immediately that the variance levels, collinearity and orientation of the β did not influence the RE's. As before (under WLS), the presentation of the discussion of the ranked tables will be within the framework of each distribution. This structure will suit the theoretical background of L_p-norm estimators (ie the optimal estimators for certain distributions are L_p-norm estimators (page 3.3))

Uniform distribution: Although theoretically $p = \infty$ is expected the CHEB was never one of the 3 best in all 40 instances. If we examine the ranked entries in Tables F.9 through F.12, we observe that the estimator recorded is usually an estimator involving the population or sample variance, and where the p is calculated by the Barr method. FPOPB and FSAMB were ranked

39 times in a total of 40 blocks. The adaptive algorithm was only ranked 11 times, which we interpret as failure of the adaptive algorithm. The p fitted in FPOPB is 3.78, and for FSAMB the p was in the interval [3.68, 3.94]. (Note that math overflow can occur in some uniform files, to overcome this, INFCUR should be increased, to route the program to the CHEB algorithm)

Normal and CN4 distributions: As in the case of WLS, the missing cells in Tables F.9 through F.12 reveal immediately that the OLSE is apparently the best estimator. This phenomenon is expected since for p=2 and the Normal distribution, L_p is the MLE (Chapter 3, p3.3). The kurtosis of the Contaminated Normal is 4, and thus it is the nearest to the Normal in all the distributions under consideration. It is therefore not too surprising that the 0 E is best (p = 2). Notice that in Table 3.1.1 under CN4 values for p were in the interval [1.40, 1.56].

CN5 distribution: Only three out of a possible 40 blocks had any rankings, indicating that p = 2 is optimal, and in effect, we could group this distribution with the Normal and CN4 distribution.

Laplace distribution: Over all variance levels L_p-norm estimators are chosen for all orientations and collinearity levels. From the theoretical background we know that the optimal estimator for the Laplace distribution is the L_1-norm estimator, and we expect that some of the best empirical estimators will have p = 1.0. However, we found the F1.25 (p =1.25) estimator ranked amongst the best in 39 out of the 40 cases. The FPOPB estimator always followed the F1.25 estimator, because the p via FPOPB is 1.25 and in effect is equivalent. The FSAMB estimator was ranked 30 out of 40 times as best, and the observed p fitted for this estimator fall in the interval [1.22, 1.38]. Notice that p = 1.25 is in this interval. The adaptive estimators were never ranked amongst the apparently best.

Thus the optimal p for this distribution is 1.25. This value is the choice of Ekblom for the Laplace distribution, and the same result was found under the WLS criteria. Fitting the L_1 algorithm, we found the RE to be lower than that of p = 1.25 (see also remarks in 5.2.6).

Student's t distribution: We find p = 1.5 to be amongst the three best in 26 out of 27 blocks with entries. This finding corresponds with the result in the WLS case and a p value of 1.5 was also the choice of Forsythe (Table 3.1.1).

Exponential distribution: The L_p-norm estimators were ranked better than OLSE in 39 out of the 40 blocks. The W1.5 and W1.25 estimators appear most often, p = 1.25 was found in 35 out of the 39 blocks, and p = 1.5 was ranked in 24 of the 39 blocks. Thus a choice of p = 1.5 or p = 1.25 seems appropriate. Notice that p = 1.5 was also Forsythe's choice and these results correspond with the WLS results.

Slash distribution: As for WLS, the RE's are exploding, usually because the basis for comparison is the OLS estimator, which in the case of long tail distribution is a disaster.

If we examine the value of the RE over the variance, orientation and collinearity level there seems to be no pattern. The highest RE's in tables are:

σ	0.01	1	5	10
WLS	9217	17947	8474412	226522
BFGS	6447	8055	16687	71796
L_1	7250	22382	2518622	130063
X matrix	99:10:2	99:99:2	70:30:1	99:99:2

where the first 4 digits in the X matrix are the value of $a_1:a_2$ and the 5 digit is the orientation of the β, 1 for largest and 2 for smallest eigenvector. Thus we conclude that collinearity, orientation and the variance (in the sense of scale) play no role. For comparison, notice the similarity of pattern for the WLS and L_1-norm estimators. We will refer to this phenomenon again in section 5.3.

In the 40 blocks the adaptive estimators were ranked 35 times and specifically FB-C, 34 times. The number of times L_p-norm estimators were

ranked amongst the best is summarised in the following table:

	σ				
	0.01	1.0	5.0	10.0	total
FB-C	10	9	6	9	34
FHART	9	6	4	3	22
F1.25	2	4	8	7	21
FSAMB	7	9	1	2	19

From the above table it seems that the Barr method is superior to the others. This conclusion is no surprise since for Barr $p > 1.0$, where as for Sposito and Gogga $p < 1$ is likely. This study supports the Rice (1964) claim that $p \geq 1$, and contradicts on an Ekblom (1974) suggestion that for long tail distributions $p \leq 1$. Later we will discuss the superiority of the L_1-norm estimator.

It should be pointed out, that in the RE given for the Slash distribution there are some peculiar results. Sometimes an RE of a 1 is reported, when the RE should be in the range near the RE of L1. This phenomenon will be discussed in the section on outliers.

Overall conclusion: if we only examine the 12 BFGS L_p norm estimators, we find: collinearity, variance, and orientation play no role; the most important issue is the distribution. We have found that for the U through to E a slight improvement in L_p-norm estimators over OLS. In the case of the Slash the improvement is excellent and the superiority and the adaptivity of the L_p-norm estimator emerges. In section 5.3 the effect of outliers in L_p-norm estimators will be discussed.

Roughly, the invariant choice of p for an optimal estimator can be summarised as

	Distribution							
	U	N	CN4	CN5	L	E	T	S
p	±3.78	2	2	2	1.25	1.5/1.25	1.5/2.0	1.11 - 1.24
Estimator	$L_{3.78}$	L_2	L_2	L_2	$L_{1.25}$	$L_{1.25}/L_{1.5}$	$L_2/L_{1.5}$	FB-C

In the DRELP3 and the DREGP4 programs, an error code of 4 (that is division by a near-zero in DFPMIN) was reported in 116 of the cases. When an error code of 4 was reported, and p = 1 was found, L_1 was called, and when fitted successfully, an error code of 41 was recorded. The error code 41 was recorded in 710 cases. If we examine what happens in DFPMIN the algorithm can be described as

1. Call DFPMIN with p = 1
2. inside DFPMIN minimize the function, test for division by zero,
3. if divide by zero, and when p = 1, fit L_1 and return to main program

Earlier we had suspected multiple solutions. What is really happening is that when BFGS cannot minimize a function with p = 1, it fails. But when the same function is sent to L_1 algorithm, this subroutine could minimize the function, and find a solution for the betas.

The 710 occurrences of error codes of 41, is a sure indication of failure in DFPMIN. However there were cases of p = 1 successfully fitted.

Interesting to note that in the estimators FS-C, and FG-C where $p < 1$, there were 57 cases of instability (that is less than 50 repetitions converge). The error code of 3 for non-convergence, indicates that the number of iterations in DBRENT exceeds the maximum iterations. Thus there is some evidence of instability of the BFGS algorithm for $p <$ one.

The reader should note that in the BFGS algorithm, the number of iterations can be exceeded in two ways. Firstly in the number of iterations that the algorithm is called inside DBRENT (ier = 3) and the number of iterations needed in DFPMIN.

The instability of BFGS for $p < 1$, the failure of DFPMIN, (division by zero), and its failure of DFPMIN to minimize a function when p = 1 (in contrast to L_1) have not previously been recorded.

5.2.2.4 L_1 and L_∞ estimators

The RE's of these two estimators are presented for convenience as the last two rows of the BFGS tables: Appendix C3, Tables C3.1 through to C3.64.

5.2.2.4.1 Program

L_1-algorithm: In this study we have used the L_1-algorithm of Barrodale and Roberts (1974), who provide a complete discussion and documentation of the algorithm. Also see the discussion in section 3.1.1.2.

The following notes apply on the use of the L_1-algorithm:

1. The X-matrix sent to L_1 is destroyed by the L_1-subroutine. The dimensions of this matrix must be (n+2) and (r+2), where n is the number of rows of the X-matrix and r the number of columns. In the program we refer to this augmented matrix as XAUG(32,8).

2. On return of the L_1-subroutine to the main program, the error code is stored in the XAUG(32,7) element. The number of iterations is contained in the XAUG(32,8) element and the value of the L_1-norm is found in the XAUG(31,7) element.

3. An error code (variable IER) of less than 2 is taken as a successful fit. When IER is equal to

   0 - fit successful but not unique
   1 - fit successful and unique
   2 - calculations are terminated prematurely

   In all our 320 (5 × 2 × 8 × 4) scenarios we found no error code of either 2 or 0.

4. The LPV sent to the L_1-algorithm consists of one element. This element is a tolerance limit equal to 10^{-6} (a small positive tolerance). Inside the L_1-algorithm the variable BIG is initialized as 10^{37} (double

precision), as compared to the original value of 10^{75} (single precision) in the L_1-algorithm.

5. Note we change all the variables in the original subroutine to double precision. This change was an attempt to get more precision as well as to make the L_1-algorithm compatible within the main program.

CHEB algorithm: In this study we used the CHEB algorithm discussed and documented by Barrodale and Phillips (1975). Also see the discussion in section 3.1.3.2.

1. The X-matrix sent to the CHEB routine must be a (n+1) x (r+3) matrix. The transpose of the X matrix must be sent over from the main program to the CHEB routine. In our program the transpose of the X-matrix is called AT(9,31). The Y vector sent to CHEB must be a 31×1 vector, and is denoted by YCHEB(31). Note both AT and YCHEB are destroyed when sent over to CHEB. For this reason every time we return from CHEB to the main program, the original matrix of X and Y is copied back into AT and YCHEB.

2. The LPV vector sent to CHEB consist of 2 elements, TOL and RELERR. TOL was set to 10^{-6} and RELERR = 0.0 (assuming a Chebychev solution). Inside CHEB the variable BIG is initialized as 10^{35} (double precision).

3. On return of the CHEB subroutine to the main program the exit code (error code) is contained in the variable OC, where OC is equal to

   0 - optimal solution, which is not unique
   1 - unique optimal solution
   2 - calculations terminated prematurely due to rounding errors.

In all the runs we found no error code other than one.

4. The maximum residual is stored in the variable RESMAX, and the number of iterations in IT. BCHEB contain the betas that were fitted when the Chebyshev norm was found.

5. Note that the CHEB routine was changed to double precision.

5.2.2.4.2 Performance

Chebychev estimator: The CHEB estimator was never ranked as one of the best. Even for the Uniform distribution, the distribution for which it is theoretically optimal, the CHEB was only ranked better (RE $\geq$ 1.10) than OLSE in 22 (6, 7, 4, 5) blocks out of a total of 40. Given this finding and the lack of asymptotic theory, we suggest this estimator can be ignored.

The program for the CHEB estimator never failed and usually converged in 12 iterations.

L_1- norm estimator: The program for the L_1-norm estimator was found to be very stable and quick. It never failed and usually converged within 9 to 14 iterations. The Laplace distribution theoretically has the optimal L_p estimator for p = 1. We found that within the Laplace distribution this estimator was ranked better (RE $\geq$ 1.10) than OLSE in only (2+4+4+2) 12 blocks, and compared to other L_p-norm estimators, it was never ranked as one of the best for the Laplace. The ranking within the collinearity, orientation and variance levels was random and thus there is no evidence that this poor performance is associated with any of the controlled factors.

In the Slash distribution the RE's of the L_1-estimator explode. If one compared the RE's of L_1 with those of the BFGS L_p-norm estimators, one finds it is ranked amongst the best three in all 40 of the blocks.

Comparing the RE's of L_1 in the Slash distribution with those of the WLS L_p-norm estimators, one finds it is ranked amongst the best three in 35 out of 40 blocks.

5.2.2.5 The adaptive algorithm

The adaptive algorithm was introduced in Chapter 3 and described in this chapter in §5.2.2.1. Gonin and Money (1985) claimed that any one of the methods (Barr, Sposito or Gogga) can be used to estimate p, and that the adaptive estimator will converge to the optimal p.

Under WLS we have seen that the L_p-norm estimator is not so prominent for the U through to E distribution. In fact it was only in the Slash that the adaptive estimator seems to be the best, so we focussed our attention on the Slash distribution. We have found that the adaptive estimator usually converges between 1 and 4 iterations for p.

In the case of BFGS as in the case of WLS, it is really only in the Slash that the adaptive estimator is ranked often enough to be of consequence, and usually converged to some p after between 1 and 3 iterations.

In both the case of WLS and BFGS mainly one but sometimes two adaptive estimators emerge, and in no cases could we found similarity between the p fitted in the three methods. The value of p via Barr is always above one, Sposito and Gogga are usually below one and nearer to each other than to the p found in Barr's method.

5.2.2.6 Comparison between WLS, BFGS, L_1 and CHEB programs

1. Comparing the p found in WLS and BFGS it is interesting to note that the p via BFGS is usually higher that the p found in WLS. Only in the case of FB-C estimator (over all variance levels) is the p found to be below the corresponding p found in WLS.

2. The number of iterations needed in WLS (meaning the iterations for the algorithm) is much higher than the iterations needed in BFGS. WLS did not stop quickly , the iterations before convergence were many and often no convergence was reached. The high frequency of error code 2 in the Appendices E1 and E2 are evidence of this claim.

3. If comparing the summaries in Appendices E, of the two HART estimators one may note that they are virtually mirror images of each other.

4. For the outer loop (the p loop in the adaptive estimator) note that the frequency of error code 3 in Appendix E1 the number of times the WLS program was terminated due to exceeding the maximum iterations for the adaptive algorithms. In BFGS the corresponding error code in Appendix E3 is marked as c.

5. Both programs had their failures. As mentioned before WLS seems to be unstable when $p \geq 2.09$. Furthermore the subroutine WLS returns to the main, with a value of p in the memory, and betas that are in fact OLSE betas, without any explicit indication (see comments earlier). On the other hand BFGS was sometimes unstable, usually when $p \leq 1$. We have note the division by zero in DFPMIN.

Table 5.5 was drawn up to explore whether WLS or BFGS are unstable. Instability of the algorithm is arbitrarily defined as convergence of less than 50 out of the 100 replications. The count in the table (under the column heading <50) is the number of 100 replications blocks under a particular distribution for which instability is observed. For the distributions U through to E this count is performed over a total of 120 blocks (12 estimators × 5 × 2), and for the Slash distribution the count is performed over 90 (9 estimators × 5 × 2) blocks. Under the column heading 100 we present the number of 100 replication blocks in which all 100 repetitions converge.

The superiority of BFGS is remarkable for the distributions U through E. In only 8 cases was no convergence reached, compared to 283 cases for the WLS program. When a $p > 2.09$ was fitted, WLS is unstable. In handling real data the p-value could be monitored and when found to be greater than 2.09, the next step would be to use say BFGS, for a stable fit). The number of 100% convergences for BFGS was much higher than that for WLS, eg the lowest percentage for WLS is 7%, compared to the lowest of 54% in the case of BFGS.

Under the Slash distribution we find that for WLS 48 and for BFGS 96 cases did not converge. 100% convergence was reach in 109 cases for WLS and 134 cases for BFGS. From these contrasting results, the two programs seem to be efficient in different contexts. A further frequency table (Table 5.6) was set up to clarify the comparison:

From the table it is clear that we can collapse the classes for σ and we need examine only the total row. It is not clear which of WLS or BFGS emerges as the best. The failure in BFGS usually yields an error code IER3 (failure in DBRENT max itererations exceed 50). In practice the limitation of 50 iterations can be avoided by setting the maximum high, but in a simulation study such as this it seems inappropriate.

Table 5.5 Counts of number of repetitions that converge

	WLS								BFGS							
σ	0.01		1.0		5.0		10.0		0.01		1.0		5.0		10.0	
rep	<50	100	<50	100	<50	100	<50	100	<50	100	<50	100	<50	100	<50	100
U	33	59	0	80	0	79	0	79	0	99	0	100	0	110	0	105
N	8	69	0	75	0	77	0	80	0	98	0	107	0	103	0	100
CN4	6	27	0	59	0	58	0	66	0	99	0	100	0	100	0	99
CN5	5	5	10	23	33	21	28	30	0	98	0	97	0	99	0	96
L	3	7	0	9	9	10	14	12	0	98	0	86	0	75	0	80
t	3	17	21	12	23	13	20	13	0	84	0	81	3	73	3	72
E	3	17	21	12	23	13	20	13	0	72	0	73	0	68	2	65
S	0	26	10	27	18	30	20	26	21	40	26	37	26	29	25	28

6. Surprisingly, the RE's of WLS and BFGS differ quite a lot. Only with W15 and F15, did we find in both 100% convergence, and hence the RE's of these two estimators were equal.

Table 5.6: Number of repetitions converging for the Slash distribution

	WLS					BFGS				
rep	<25	<50	<75	<100	=100	<25	<50	<75	<100	=100
σ										
0.01	0	0	25	39	26	0	20	19	11	40
1.0	0	10	35	18	27	7	18	14	14	37
5.0	1	17	41	1	30	5	21	14	21	29
10.0	12	8	37	7	26	6	18	16	22	28
Tot	13	35	138	65	109	18	77	63	68	134
	48			174		95			202	
%	13%		38%	48%		26%		18%	56%	

7. The large difference in the RE's of WLS and BFGS in the Slash is somewhat disturbing. Sometimes very high RE's are reported in WLS, compared to the RE's of 1 in BFGS. It is an open question if it is one more of the 100 Monte Carlo repetitions that contributes to the low/high RE. The possibility of outliers will be investigated in §5.3.

8. The norms, reported in Appendix E, depend on the value of p. Therefore no direct comparison can be made between WLS and BFGS. In the norm columns in Appendix E, note how within any single table the norms fluctuate. Also note the occasional very high standard errors. In an attempt to bring the norms all to a standardized value, the standardized norm was computed as

$$\text{std-norm} = (\Sigma f_j/n)^{1/p_j}/\sigma)/\text{count}$$

where f_j is the L_p-norm, p_j the p fitted, and σ is square root of the theoretical variance (ie the variance level chosen in the simulation), count is the number of repetitions that converge, and n is the number of rows of the X matrix.

For U through to E std-norm ranges in an interval around one or just below one. Usually the std-norm found in BFGS is slightly higher than the std-norm value found in WLS. In the Slash distribution however, the range of std-norms is from 2 to 15, much wider than in any other distribution. This phenomenon might be due to the fact that σ was used as a scalar and not as a variance level, which in the case of the Slash is infinite.

In §5.3 we will investigate four files that contain at least one outlier, ie files B99992, D99992, C70301 and A99102. It is clear from the high values of the std-norm (and their respective high standard deviations) that they differ from the other files.

5.2.2.7 Elusive optimality

Tables F.13 through to F.16 contain the combined ranking of all the estimators. The reader should keep in mind previous remarks about such rankings. For distributions U through E we only compare the BFGS, L_1 and the biased estimators, as we have previously shown that BFGS is much more stable than the WLS algorithm, and that the CHEB estimator can be ignored. Although in previous section we compared estimators for which the number of converges was less than 50, in this section we take them as failures and hence ignore them.

For distributions U through E we can conclude: The biased estimators are superior to the L_p-norm estimators. It is only when $\sigma = 0.01$, and the collinearity level is moderate to low that the L_p-norm estimators emerge as being better. When σ is greater than 0.01, the biased estimators are superior to all other estimators. Thus all the conclusions made under biased estimators are applicable here. Not only did the biased estimators cope with collinearity, variance, and orientation, but also with various distributions, ie the biased estimators are robust under this study's violations of normal assumptions.

For the Slash distribution we decided to compare the WLS, BFGS, L_1 and the biased estimators. As shown previously we are not able to conclude which of

WLS or the BFGS algorithms seem superior. In the Slash, the RE's are exploding for some L_p-norm estimators and surprisingly also for the biased estimators. It seems that not only are biased estimators coping with collinearity, orientation, and variance but also with heavy tails. The symmetry of the distributions might contribute to this phenomenon.

The biased estimators, in particular RLW, are only ranked amongst the best three when the collinearity is high. We have already noted that the L_1 estimator is ranked as one of the best on 37 out of 40 occasions

The RE's of the WLS estimators seem higher than those of BFGS. This phenomenon will be discussed in the next section.

5.3 Outliers in L_p-norm estimators

In section 5.2.2.6 we compared WLS and BFGS, and found that it is only in the case of the Slash distribution that the L_p-norm estimators emerge as being better that OLS in terms of RE. The large difference in the RE's of WLS and BFGS in the Slash distribution is somewhat disturbing. One would have expected to have the RE of BFGS near that of L_1, especially for FB-C (where in fact p ~ 1), and for HART (p found to be mostly one). The failure of BFGS to improve on OLS RE = 1 in Table C3.15, C3.31, C3.47 and C3.63) and the very high RE of WLS in the corresponding cases seem to be atypical.

We expected outliers to be the cause of the breakdown. Examining the moments for the Slash distribution (Appendix B) over the four variance levels, it is clear from the large variance that something strange is going on in files:

A99102, B99992, C70301 and D99992

where A, B, C and D represent a σ value of 0.01, 1.0, 5.0 and 10.0. The next four positions are the value of a_1 and a_2, and the last digit is the orientation of the betas. We expected at least one outlier in each of these files. Further investigation into these four files was conducted using the package TSP.

The four files were imported into TSP where various summary statistics and plots were produced for the four files. In all four cases the histograms were skewed with long tails to at least one side. After investigating the histograms the data was sorted and the tails examined. Using a rule of thumb we determined when a value was an outlier. Basically we look at the extremities of the tails and decide whether there was a significant jump in the data. If a jump is identified, we take the value before the jump for the upper tail $X_{(jump-1)}$, multiply it by 10 and if

$$X_{(n)} > X_{(jump-1)} * 10$$

we declare $X_{(n)}$ an outlier.

After a value was identified as an outlier, we constructed another histogram with the data ignoring the outlier. When the histogram appeared to be normal and symmetric we were satisfied that most of the influential outliers were found. Then the outlier was replaced with the value of its nearest neighbour.

In files A99102, B99992 and C70301 we changed one outlier to its nearest neighbour and in file D99992 three outliers were identified and replaced by that nearest neighbour, which is not an outlier. The L_p-norm programs were run with these altered data sets and the old and the new results were compared. Because in the calculations of the RE's OLS is used as a yardstick, we decided simply to compare the sum of the squared differences between the estimators, as the RE's of the OLSE changed dramatically when an outlier is removed. Thus the values reported in Table 5.7 are

$$SUM = \sum_{j=1}^{5} \sum_{i=1}^{count} (\tilde{\beta}_{j\,i} - \beta_j)^2$$

where count is the number of repetitions that converge and $\tilde{\beta}_{j\,i}$ and β_j are defined as in equation (5.1.4). Note that SUM is in fact the divisor of the RE reported so far. SUM values labelled old are for the original files and SUM values labelled new are for the files after the outlier was changed to

its nearest neighbour. The value in brackets is the number of repetitions contributing to the sum (ie count).

Table 5.7: SUM values

Filename	estimator	BFGS old		BFGS new		WLS old		WLS new	
A99102	B-C	0.5258	(100)	0.5251	(100)	0.3180	(62)	0.3180	(62)
	S-C	1.0556	(55)	1.0557	(56)	0.7226	(70)	0.7229	(70)
	G-C	0.4437	(42)	0.4457	(43)	0.5529	(70)	0.5530	(71)
	HART	1.9732	(100)	1.9717	(100)	0.5064	(84)	0.5064	(84)
	L_1	0.4676	(100)	0.4676	(100)				
B99992	B-C	1.76E+08	(99)	1.22E+04	(99)	4718	(29)	4756	(30)
	S-C	2.02E+08	(49)	1.21E+05	(48)	11489	(62)	10490	(61)
	G-C	2.05E+08	(58)	1.54E+06	(57)	11815	(63)	11973	(63)
	HART	1.37E+05	(100)	1.37E+05	(100)	17046	(73)	17085	(74)
	L_1	9215	(100)	9215	(100)				
C70301	B-C	4.11E+09	(99)	8.02E+03	(99)	1882	(29)	1882	(29)
	S-C	1.57E+10	(43)	1.37E+04	(42)	8737	(61)	8675	(61)
	G-C	1.27E+10	(34)	4.20E+03	(33)	10225	(72)	10402	(72)
	HART	1.20E+10	(99)	8.46E+03	(99)	6390	(76)	6391	(76)
	L_1	6333	(100)	6333	(100)				
D99992	B-C	9.79E+10	(98)	4.04E+06	(97)	2.98E+05	(29)	2.98E+05	(29)
	S-C	6.89E+10	(55)	1.41E+07	(56)	1.39E+06	(69)	1.28E+06	(69)
	G-C	9.93E+10	(54)	3.11E+06	(53)	7.85E+05	(67)	7.61E+05	(67)
	HART	9.84E+10	(99)	1.48E+07	(99)	7.49E+05	(72)	7.45E+05	(71)
	L_1	7.64E+05	(100)	7.64E+05	(100)				

In the case when σ = 0.01 (the smallest) we note that there is no marked difference in the SUM from old to new. Observe how the number of

repetitions increase from old to new under BFGS. For the other variance levels the difference in the sum from old to new under BFGS is dramatic.

Under WLS we note that there are slight differences, with very large differences in file D99992 for the WS-C and the WG-C estimators.

For the L_1 algorithm there are no differences over all variance levels.

These findings show that the L_1 estimators (from the L_1-algorithm) are robust and the inclusion or deletion of any outliers has no effect. However it seems that the BFGS algorithm is sensitive to outliers. As the variance increases the sensitivity of certain L_p-norm estimators under the WLS scenario increases.

In conclusion we have found that so-called robust estimators are not really robust. It appears that only the L_1 estimators are robust. This phenomenon has not been reported before, and should be a field for further research. Also see the recommendations in §5.7.

5.4 Moment ratio parameter, ω_p^2

Nyquist (1983) (see §3.2.2) show that the asymptotic distribution of $\sqrt{n}(\hat{\beta}_{L_p} - \beta)$, $1 \leq p < \infty$ is given by

$$\sqrt{n}(\hat{\beta}_{L_p} - \beta) \sim N(0, \omega_p^2 Q^{-1}) \qquad (3.2.1)$$

The theoretical value for ω_p^2 (the moment ratio parameter), is given by Gonin and Money (1989), and values for the distributions used in this simulation study are summarised in Tables 3.2.1 and 3.2.2. When $p = 1$ the quantity ω_p^2 of (3.2.1) is denoted by λ^2, and various ways to estimate λ^2 were discussed in §3.2.2, and the estimates of λ summarised in Appendix D. Gonin and

Money (1989) suggested, for p > 1, the following estimate

$$\hat{\omega}_p^2 = \frac{m_{2p-2}}{[(p-1)m_{p-2}]^2} \qquad (3.2.3)$$

where $m_r = \frac{1}{n}\sum_{i=1}^{n} |\hat{\epsilon}_i|^r$, and $\hat{\epsilon}_i$ is the residual from the L_p-fit. The estimates, $\hat{\omega}_p$ are reported in Appendix E and is listed under the column heading w_p

5.4.1 Estimation of λ

In Appendix D, 6 estimates of λ are summarised. Each estimated value of λ is the mean value of the 100 repetitions, all based on the residuals of the L_i estimates. Values given in brackets below the average are the standard errors of the estimated λ. The estimates were discussed in §3.2.2, and the abbreviations explained in the introduction to the Appendices. In the heading of each table in Appendix D, the theoretical value of λ is given for easy comparisons. Although simulations studies by other authors (see §3.2.2) show that the size of the X matrix (n = 30) is rather small for non-normal distributions we summarise the best mean ± 2*std error (roughly a 95% CI), includes the theoretical value most frequently over the 10 X matrices) estimates in Table 5.8. Absence of an entry implies that there was no best estimator, ie all the CI's failed, and they were all equally poor.

Table 5.8 Best estimators of λ

σ	Distribution						
	U	N	CN4	CN5	L	t	E
0.01	ch-2	mks	-	-	-	mks	mks
1.0	ch-2	mks	-	-	-	-	-
5.0	ch-2	mks	-	-	-	-	-
10.0	ch-2	mks	-	-	-	-	mks

It is clear that for the Uniform distribution the method ch-2 is appropriate and for the Normal distribution the method of mks.

Overall, examining the values found in Appendix D, we can conclude that the methods st and stsim always calculate values that are to low, whereas ch-1 gives values that are too high. It is the three middle columns (ie estimators ch-2, ch-3 and mks) that seems to be nearest to the theoretical values. The issue of which method seems the best will be further investigated in the case of the full X-matrix, §5.5.

5.4.2 Estimation ω_p^2, $p > 1$

Numerous problems arise in the estimation of ω_p^2. When an error term near zero is encountered, and when $p < 2$, division by zero is possible. In an attempt to avoid division by zero, only residuals greater than 0.000001 where used in the calculation of w_p. It is clear from the calculated values that the estimate of ω_p^2, is a nonsense estimator, especially when $p < 1.5$. One small residuals, taken to a negative power, can inflate the divisor of the estimate ω_p^2, to such a degree that w_p^2 tend to zero. Columns with zero values for ω_p^2 where omitted form Appendix E.

Gonin and Money (1989) reported no such problems, and in the literature we have found no reference to the division by zero in the case of small residuals and $p < 1.5$. Thus we cannot agree with the statement of Gonin and Money (1989): "the approximation of ω_p^2 was found to be adequate for other error distributions as well as for varying values of p". We found the estimator unstable, and usually one of the main reasons of the breakdown of the L_p-norm programs.

As our X matrix only consists of 30 cases, the value of ω_p^2 was based on 30 residuals and in most cases less than 30 as those near zero were ignored. We felt that it was not worth pursuing a better estimate of ω_p^2 as for non-normal distributions we need more degrees of freedom to get an estimate that will be asymptotically near to the true ω_p^2. The issue of an improved and stable estimate of ω_p^2 will be pursued further in §5.6, using the residuals of a full X-matrix.

5.5 The full X-matrix

5.5.1 Generating the full X matrix

In the previous section we note how badly the estimator for omega (w_p) performed. There was a complete breakdown of the estimator as given by Gonin and Money (1989). As our X matrix consisted only of 30 cases, we decided to generate X matrices which might better satisfy the asymptotic results as discussed in Chapter 3 (§3.2.2). It would also be a guideline to see if any results obtained so far change when n is large. Two X matrices, using the same scenario as in Chapter 4, were generated except that n (the sample size) is now 200. The choice of n=200 is based partly on the results of Dielman and Pfaffenberger (1983) for p = 1 (see §3.2.2).

In generating the two new X matrices, (4.2.1) and (4.2.2) will change to:

For $j = 1,2,3$ and $i = 1,2,\ldots,200$

$$X_{ij} = (1 - a_1^2)^{\frac{1}{2}} Z_{ij} + a_1 Z_{i6} \qquad (4.2.1)$$

but for $j = 4,5$ and $i = 1,2,\ldots,200$

$$X_{ij} = (1 - a_2^2)^{\frac{1}{2}} Z_{ij} + a_2 Z_{i6} \qquad (4.2.2)$$

where

(i) Z_{ij} are independent N(0,1) variates obtained from random deviates generated with the generator of Wichmann and Hill (1982) and then transformed to N(0,1) random deviates in the programs DISTR and MOMENT. The seeds for the six streams are recorded in Table G1 (Appendix G).

(ii) Two X matrices where generated, namely those specified by the types 99:99:1 and 70:30:2. With this choice we try to include the two extremes: one with the worst collinearity and one with moderate collinearity. Two choices of the orientation were made. Only one value of σ was considered, σ = 10.0, for the error terms, on the grounds that the variance would be at the largest of comparable scenarios.

For i = 1,2,...,200 we model the response values as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \epsilon_i \qquad (4.2.3)$$

where the X_{ij} are unstandardized, β_0 is ten, and the ϵ_i are independent variates from any one of our usual eight distributions. One hundred Y-vectors of dimension 200, were generated for each of the two X matrices and for each distribution under the model (4.2.3). The eigenvalues and condition numbers corresponding to the two matrices are shown in Table 5.9 and the coefficients of β_L and β_S are shown in Table 5.10.

Table 5.9 Eigenvalues and condition numbers of X'X (X is standardized)

Correlations $a_1^2:a_2^2$	eigenvalues of X'X: λ_i (without β_0)					condition number λ_1/λ_5
.99:.99	(4.9604	0.0117	0.0111	0.0092	0.0076)	649
.70:.30	(3.1579	0.7991	0.5216	0.2703	0.2511)	13

Table 5.10 β used in generating Y

Correlations $a_1^2:a_2^2$	β'	Eigenvectors of X'X				
.99:.99	β_L'	[-0.44705	-0.44701	-0.44740	-0.44710	-0.44751]
.70:.30	β_S'	[-0.70052	-0.69884	-0.00910	0.00387	0.10540]

For each X matrix 100 replications of the (200×1) ϵ-vector were generated independently for each distribution. The same method as set out in Chapter 4 for generating these deviates was followed. The seeds of the random deviates for the six streams are given in Table G2. Summaries of the moments of the deviates for the eight distributions are given in Table G3. Once the data was generated the total number of estimators (39) was fitted, and the results presented in Appendix G and discussed in §5.5.2.2.

5.5.2 Fitting the estimators

5.5.2.1 Program problems

In the full X matrix the 4 programs for obtaining L_p-norm estimators and the program for biased estimators had to be changed, as many problems, such as stack overflow and Dgroup variable too large, were encountered. The format to read in the X matrix in previous (for X:30x1) programs was altered to read in an augmented matrix of 30×105 where the first 100 columns each represent one repetition of the required random deviate. In an attempt to have only one Y in memory at a time (where Y is at this point any one of the random deviates from a specific distribution, and pseudo-observation, y is calculated inside the program)) the Y-matrix was read in repetition by repetition as a vector of 200×1. To accomplish this strategy, the Y vector of 200×100 was transposed outside the estimation programs, inside the estimation programs, one repetition was read in at a time, all the estimators calculated, stored and the next Y vector picked up. In this manner we only have to declare a Y vector of dimension 200 and a X matrix of 200×6. To illustrate

```
      READ in X-matrix
      DO 20 I = 1,100
         READ Y                (pick up y)
        fit estimators, write estimators and error terms to files
20    continue
```

However even with this smaller allocation of memory, there still occur problems and it was decided to change all the variables to reals. Before changing, the programs were run under the Salford compiler (which is capable of addressing all the memory available and is not limited to 640K). The first file we ran under the Salford compiler had different results than those reported by the MICROSOFT compiler. In an attempt to explore this difference, three of the small X matrices, and 2 of the full matrices was submitted to both compilers using only a small section of DREGLP3. In the small matrix we choose σ = 1 and for the full matrix σ = 10.0 as usual. The RE (with the number converging in brackets) are given in the Table 5.11 and

the results for the full matrices are in Table 5.12. The letter after the identification of the file indicates the specific distribution. In the two columns under each filename we contrast results via the MICROSOFT and the Salford compilers, in that order.

Table 5.11 Comparison of the RE under the two compilers (for small X matrix)

Estimator	Filename					
	99991S		99992S		99991N	
FB-C	66	69	1.17 (99)	1.17	0.95	0.95
FS-C	2 (50)	13 (46)	1.02 (49)	1.01 (39)	0.97	0.97
FG-C	16 (56)	26 (59)	1.01 (58)	1.00 (56)	0.69	0.69
FHART	46	46	1510	1510	0.86	0.86

Table 5.12 Comparison of the RE under the two compilers (for full X matrix)

estimator	Filename			
	99991CN4		99991S	
FB-C	0.97	0.97	2866	2810
FS-C	0.92	0.92	1 (88)	1 (89)
FG-C	0.90	0.89	1 (54)	1 (58)
FHART	0.97	0.97	2893	2812

For the small X-matrices we note that the RE values for the two files (X99992S and X99991N) are basically the same. However in the file 99991S note the marked difference in the RE of the estimators FS-C and FG-C. For the full X matrices the RE seems to correspond in most cases, as we only interpret RE as a source of ranking on the basis of OLS and not as numbers as such. We were not concerned about the difference in RE for FB-C and FHART as the rankings of these estimators agree under both compilers.

To further explore the difference found in the file X99991S (X indicates the small matrix) we summarised for the 100 repetitions the number of iterations, norms and p fitted and then compared the two compilers. We find for the FB-C estimator that the number of iterations for both the inside

loop and the outside loop coincide in most cases. The p values fitted usually coincide, up to the fourth decimal. The fourth decimal difference would explain the slight difference in the RE from 66 to 69. For the FG-C and FS-C estimators we found that if there was convergence in both, the compilers the p-values usually correspond. The cases that did not converge under both compilers attribute to the difference in the RE. In the case of FHART the p was the same in both compilers.

In all cases the OLS fit for compilers correspond. Only in the minimizations programs did we find differences. The differences in the two compilers may arise from the way they handle double precision or small numbers. As these programs are both linear programming type programs where a small difference in precision could lead the two paths of the two compilers to fork, it is possible that a small change in numbers generates differing local minima. It is interesting that in both (FS-C and FG-C) the forking usually happens when p is less than one. We also note that the vast difference in the values of the omegas between the compilers. This difference is usually observed when p is near one or less than one. The author believes that the Salford compiler might be able to handle small values taken to a small negative power better than the MICROSOFT compiler. This claim was not investigated. It is suggested that the difference between the two compilers should be investigated in further research. It appears that a knowledge of assembler language is required.

At the time when these differences between compilers were discovered, the whole simulation study for the small X-matrices had been performed and summarised. As MICROSOFT has a respectable reputation, the convenient choice was made for the original compiler. Careful programming seems to sort out most of the observed problems under MICROSOFT, the only drawback being the limitation to 640K.

To implement the full X matrix, the programs were changed to reals. Only in the subroutines ZBRELP, ZBRENT and GAMMLN were the variables declared in double precision, because of the sensitivity of the gamma function. In DREGLP2 and DREGLP4, INFCUR was moved up to avoid math overflow (when p is near 6 or greater) in the WLS programs, the INFCUR was set to 1.9 if

necessary and in the BFGS programs, the INFCUR was set to 2.04 if necessary. This cutoff value was often required in the case of the Uniform error term files.

5.5.2.2 The best estimators

The RE's obtained for each estimator are summarised in Table G4 - G19. Because only two X matrices were generated, the results of all the programs appear on one page and not one page per estimator as in the case of Appendix E. The program statistics are reported in Tables G20 through to G221. The three estimators best under each scenario as well as the best overall are summarised in Appendix G, Table G222. Using the same convention as previously described.

5.5.2.2.1 Best estimator under the biased estimators

In the first full X matrix (99:99:1), where the collinearity is at its highest, the RLW estimator was ranked first followed by the FPCRR2 estimator in the second place and either FPCG2 of FPCR1 in the third position regardless of the distribution. In the case of the second X matrix, where the collinearity is not so severe, the FPCR2 estimator was ranked first, followed by RLW and in the third place one of the FPC families over all distributions.

Thus we may conclude the biased estimators are robust against the influence of the distribution even for long tail distributions like the Slash.

5.5.2.2.2 Best estimator under WLS criteria

For the Uniform distribution a p-value of infinity was fitted for both X files. For the N, CN4 and CN5 distribution there was no estimator better than OLS. In the Laplace distribution we had the interesting situation that for the first full X matrix the set of 'best' estimators all fit p = 1.25 and for the second full X matrix the set of best estimators all fit a p = 1 (expected from a theoretical view).

For the Student's t distribution we obtain p = 1.5. For the HART estimator a p of 1.5 was fitted in 70 percent of cases. In the case of the Exponential distribution there was not much choice between the three best estimators and a p of 1.5 or 1.25 is appropriate.

For all the above distributions, ie Uniform through to Exponential, the improvement over OLS was small. However in the case of the Slash distribution the RE exploded (because the fitted Slash error terms under OLS can be highly distorted by the long tails as well as the collinearity). All three estimators in the best group, fitted p=1, or in the case of WB-C an average p of one.

5.5.2.2.3 Best estimator under BFGS

The results obtained under WLS are confirmed by BFGS. Only in the case of the Uniform is FG-C ranked under BFGS but the corresponding WG-C not ranked. This phenomenon can be explained by the different values of INFCUR sent to these programs. In the case of the FG-C estimator, p was sent to infinity whereas in the case under WLS, p was large (around seven).

5.5.2.2.4 Best estimator overall

For the Uniform through to Exponential distribution the overall best estimators are biased estimators (see the section above on biased estimators). Only in the case of the Uniform distribution, did the CHEB estimator emerge.

For the Slash distribution the L_p-norm estimators were ranked better than biased estimators. Any one of FB-C, WB-C or FHART or WHART estimators will do well. Overall the p fitted via these estimators is in fact the L_1 estimator, where the norm is now minimized by WLS or BFGS. The slight difference in RE between the three algorithms (WLS, BFGS and L_1) could be attributed to less than 100 percent convergence in the case of WLS, or in the case of FB-C, to the fact that p was approximately one, not exactly one. The difference between FHART and L_1 might be due to the error code 41,

picked up 9 and 11 times within FHART, or it might be due to rounding errors.

5.5.3 Comparison with the results from the small matrix

1. Stability: there was no evidence as in the case of the small X matrix of any instability in the WLS method. The lowest number of repetitions converging was 54. In the BFGS algorithm the lowest number of repetitions converging were for the FG-C (47 repetitions) and FSAMG (43 repetitions) estimators. We cannot distinguish which of BFGS or Gogga methods, contributes to the instability, just as with the small X matrix. Overall the stability of the estimators was satisfactory.

2. In the Uniform distribution the WLS algorithm sometimes failed, which is evident in the files being ignored in tables G20 through to G25. Although in the case of the full X matrix we can not determine exactly where WLS will breakdown (in the case of the small X-matrix it was for values of $p \geq 2.09$), as for the N through to S distributions the p-values found in the WLS algorithms were < 2, and not near regions where the breakdown should occur.

3. For the full X matrix there is a dramatic drop in the number of iterations needed to find a solution via BFGS or WLS. The number of iterations needed to find p in the case of the adaptive algorithm is usually lower than that needed for the small X-matrix. This phenomenon may be due to the fact that for the full X matrix the parameter space is much better defined despite the collinearities.

4. In the HART estimators the choice of p for the full X matrix is much more definite than in the case of the small X-matrix for which the allocation to 1, 1.5, 2 and ∞, sometimes appeared random.

5. Both WLS and BFGS converges to the same p-values in the case of the full matrix, and $p \geq 1$, within each estimator. In contrast, for the the small X-matrix there was a vast difference between the p-values found in each algorithm.

We conclude that, for nice distributions any method used to find p is acceptable, as all three (Barr, Sposito and Gogga) converges to more or less the same p-value (confirming the claim of Gonin and Money (1985)). However for the Exponential (non-symmetric) and the Slash (very long tails) the three methods yield p-values that do not correspond.

6. In the full matrix under the Slash distribution the L_1 estimator is no longer amongst the set of best three estimators, as in the case of the small X-matrix. However examining its RE's we find it still ranked in the fourth position.

7. Roughly the invariant choice of p (for the full matrix) for an optimal estimator can be summarised as

	Distribution							
	U	N	CN4	CN5	L	E	T	S
p	∞	2	2	2	1.25	1.5/1.25	1.5/2.0	-1.0
Estimator	CHEB	L_2	L_2	L_2	$L_{1.25}$	$L_{1.25}/L_{1.5}$	$L_2/L_{1.5}$	FB-C/FHART

After comparing this table to the corresponding table for small X-matrix, there appear to be differences differ only for the Uniform and the Slash distribution. In the small X-matrices the p under U was approximately 3.8 and the p for the Slash was in the interval [1.11, 1.24].

8. There seems to be no pattern differences between the sets of RE's for the biased estimators. The biased estimators found amongst the best three are the same for the two scenarios (n=30 and n=200).

9. In the case of the full matrix, BFGS did not break down for $p \sim 1$ as in some cases in the small X-matrix. However for $p < 1$, there seems to be a difference between the RE of WLS and BFGS, (just as in the case of the small X-matrix), suggesting that BFGS is not suitable for $p < 1$.

As before we will ignore $\hat{\omega}_p^2$, until §5.6, and the estimate for λ will also be discussed in §5.6.

5.6 Moment ratio parameter, ω_p^2

5.6.1 Estimation of λ

In Appendix G, tables G223 through to G230, 8 estimators of λ are summarised. As before each estimated value of λ is the mean value of the 100 repetitions, given in the first row. The second row consists of the standard deviations of the means and the third row is the coefficient of variation expressed as a percentage. The estimators are similar to that fitted in the small matrix simulation except for two extra Cox and Hinkley estimators (ch-4 and ch-5). The values of v (discussed in § 3.2.2) are extended to include 4 and 5.

The observed choice of best estimator is bases, as before (§5.4.1), on the CI and are summarised in Table 5.13:

Table 5.13 Best estimators of λ (full X matrix)

σ	Distribution						
	U	N	CN4	CN5	L	t	E
10.0	ch-5/mks	mks	mks	mks	-	mks	ch-4/ch-5/mks

In contrast to the small X-matrix results, best estimators emerge for the CN4, CN5 and Student's t distribution.

In the case of the small X-matrix simulation the ch-2 estimator appeared to be the best whereas in the case of the full X-matrix the ch-5 estimator emerges. This phenomenon can be attributed to the value v (the position of the ordered residuals round the median) and the size n of the sample. Cox and Hinkley only indicate that v should be small. It is beyond the scope of

this thesis to give recommendations of the choice of v, for particular sample sizes. However if we look at the estimators ch-1 through to ch-5, it is clear that for the full X-matrix we could have extended v even further as the mean values of the estimators are approaching the theoretical values. In the opinion of the author, this arbitrary choice of v can influence the estimator to such an extent that the user should be careful to avoid using it simply to find the estimators that suits his needs.

From Table 5.13, the mks estimator appears to be best over most distributions. Furthermore it has the lowest coefficient of variation. It involves no arbitrary choices as in the case of the Cox an Hinkley estimator.

Based on this simulation study we recommend the use of the mks estimator of λ. However we have not exhausted the choices of estimators available in the literature. We only use a small sample (8) of the available estimators, as this explanation was not the main purpose of the thesis. Nonetheless for n large the mks estimator seems to approach the theoretical value of λ, and was the most stable of all the estimators fitted.

5.6.2 Estimation of ω_p^2, $p > 1$

The estimator of ω_p^2 proposed by Gonin and Money (1986) and introduced in Chapter three as

$$\hat{\omega}_p^2 = \frac{m_{2p-2}}{[(p-1)m_{p-2}]^2} \qquad (3.2.3)$$

where $m_r = \frac{1}{n}\sum_{i=1}^{n} |\hat{\epsilon}_i|^r$, and $\hat{\epsilon}_i$ is the residual from the L_p-fit.

As previously mentioned when r is negative and $\hat{\epsilon}_i \to 0$, problems occur as we have division by near-zero, or inflation of the divisor of (3.2.3) such that $\hat{\omega}_p^2$ was near zero. This phenomenon is evident from the absence of the columns for $\hat{\omega}_p^2$ in the Tables of Appendix E. Often for values of $p < 1.25$ the resulting estimator of $\hat{\omega}_p^2$ was near zero.

To avoid situations where one single zero or near zero residual could cause the program to terminate prematurely, or deflate $\hat{\omega}_p^2$ to zero, we propose the following perturbation of the problem:

Originally m_r is defined as

$$m_r = \frac{1}{n} \sum_{i=1}^{n} |\hat{\epsilon}_i|^r$$

$$= \frac{1}{n} \sum_{i=1}^{n} (\hat{\epsilon}_i^2)^{\frac{r}{2}}$$

We define the perturbed m_r, denoted by $\tilde{m}_r$, as

$$\tilde{m}_r = \frac{1}{n} \sum_{i=1}^{n} (\hat{\epsilon}_i^2 + c^2)^{\frac{r}{2}} \qquad (5.6.1)$$

where c is a small fixed quantity, usually positive. As $c \to 0$, the perturbed problem reduces to the original problem. Ekblom (1973) suggested a similar perturbation in the case of zero residuals, in the damped Newton method.

Consider the binomial expansion of (5.6.1). The binomial expansion of $(a + x)^k$ is given by

$$(a + x)^k = a^k + ka^{k-1}x + k(k-1)a^{k-2}x^2/2! + k(k-1)(k-2)a^{k-3}x^3/3! + \ldots.$$

$$= a^k[1 + ka^{-1}x + k(k-1)a^{-2}x^2/2! + k(k-1)(k-2)a^{-3}x^3/3! + \ldots.]$$

thus if $x = \hat{\epsilon}_i^2$, $a = c^2$ and $k = \frac{r}{2}$ then

$$(\hat{\epsilon}_i^2 + c^2)^{\frac{r}{2}} = c^r[1 + \tfrac{r}{2}c^{-2}\hat{\epsilon}_i^2 + \tfrac{r}{2}(\tfrac{r}{2} - 1)c^{-4}\hat{\epsilon}_i^4/2! + \tfrac{r}{2}(\tfrac{r}{2} - 1)(\tfrac{r}{2} - 2)c^{-6}\hat{\epsilon}_i^6/3! + \ldots.]$$

thus

$$\tilde{m}_r = c^r\left[n + \tfrac{r}{2}c^{-2}\sum_{i=1}^{n}\hat{\epsilon}_i^2 + \tfrac{r}{2}\left(\tfrac{r}{2}-1\right)c^{-4}\sum_{i=1}^{n}\hat{\epsilon}_i^4/2! + \tfrac{r}{2}\left(\tfrac{r}{2}-1\right)\left(\tfrac{r}{2}-2\right)c^{-6}\sum_{i=1}^{n}\hat{\epsilon}_i^6/3! + \ldots\right]/n$$

$$= c^r\left[n + \tfrac{r}{2}c^{-2}\sum_{i=1}^{n}\hat{\epsilon}_i^2 + O(c^{-4})\right]/n \qquad (5.6.2)$$

The perturbed estimator of $\tilde{\omega}_p^2$ corresponding to $\tilde{m}_r$ defined in (5.6.2) is

$$\tilde{\omega}_p^2 = \frac{\tilde{m}_{2p-2}}{\{(p-1)\tilde{m}_{p-2}\}^2}$$

$$= \frac{c^{2p-2}\left[n + (p-1)c^{-2}\Sigma\hat{\epsilon}_i^2 + O(c^{-4})\right]/n}{\{(p-1)c^{p-2}\left[n + \frac{p-2}{2}c^{-2}\Sigma\hat{\epsilon}_i^2 + O(c^{-4})\right]^2/n\}^2}$$

$$= \frac{c^2 n}{(p-1)}\,\frac{\left[n + (p-1)c^{-2}\Sigma\hat{\epsilon}_i^2 + O(c^{-4})\right]}{\left[n + \frac{p-2}{2}c^{-2}\Sigma\hat{\epsilon}_i^2 + O(c^{-4})\right]^4} \qquad (5.6.3)$$

Notice that when p = 2 (OLS), most of the terms in the binomial expansion reduce to 0 ($\frac{r}{2} - 1 = 0$), thus (5.6.3) reduces to

$$\tilde{\omega}_p^2 = \frac{\left[nc^2 + \Sigma\hat{\epsilon}_i^2\right]}{n}$$

$$= \frac{n-1}{n}\,\hat{\sigma}^2 \qquad \text{when } c \to 0$$

which is a biased estimator of σ^2, but is asymptotically unbiased.

The estimator proposed in (5.6.3) is stable and is not influenced by any zero or small residuals. It is of course affected by the choice of c and by the current value of p. In the case of values of p near one, problems can

arise with the choice of c. The estimator might be influenced more by, the value of c, than the residuals themselves.

5.6.2.1 Practical experience of $\tilde{\omega}_p^2$

In an attempt to find an iterative choice of c, we start with a small value of c, c_o, and then calculate $\tilde{\omega}_p^2(c_o)$, find the next c_i, calculate $\tilde{\omega}_p^2(c_i)$ and test for convergence. However this algorithm was unsuccessful. The problem observed was that the estimates were grouped together for step c_o to c_i, then a sudden jump at c_{i+1} and again the estimates will stay the same for the next group and so on.

We settle for a c value of 0.0145. In Table 5.14 the original, perturbed and theoretical estimators are summarised, the estimators were calculated using the residuals from the FB-C estimator, and the first X matrix (99:99:1). The Uniform files are not included as $p > 2$. Values in brackets are the standard deviation of the means.

Table 5.14 Comparing estimators of ω_p^2

Distribution	ω_p^2 (original)	$\tilde{\omega}_p^2$ (c = 0.0145)	ω_p^2 theoretical
N	95.57 (0.95)	95.64 (0.95)	101.27 (0.17)
CN4	84.50 (2.76)	93.60 (1.69)	100.91 (0.47)
CN5	68.87 (4.03)	88.38 (2.47)	98.60 (0.61)
L	41.66 (28.75)	68.38 (6.39)	58.61 (6.21)
t	56.12 (43.91)	167.55 (43.53)	113.20 (6.73)
E	33.59 (4.20)	135.31 (34.15)	107.57 (1.09)

We observe that for the Normal distribution (p near 2) there is no difference between the original and the perturbed estimators. For the CN4 and the CN5 distribution the perturbed estimator is much more stable and closer to the theoretical estimator. Note that for CN4 and CN5 the p value is usually > 1.5.

In the case of the Laplace distribution the perturbed estimator appears to be more stable. However p values below 1.25 had a severe effect on the average. For the Exponential and Student's t distribution the perturbed estimator was unsatisfactory as it was inflated by the c values and the p near one.

The results for the Slash ($p \sim 1$) distribution are not reported. In both the original cases and the perturbed method the estimated omegas were inflated (perturbed in the region of 10^7) to such an extent that no estimate for omega (ω_p^2) was determined. See the recommendations in §5.7.

There are still problems estimating the moment ratio parameter ω_p^2. The estimator as given by Gonin and Money (1989) is very unstable for $p < 2$. The perturbed estimator proposed in this section, is more stable than the original, although there are still severe problems. When p is between 1.5 and 2, the perturbed method appears to be better than the original problem. For p between 1.25 and 1.5, it works for some cases. But for p approaching 1, the perturbed estimator is not suitable. In the opinion of the author the negative exponent ($p < 2$), violates the assumptions of Nyquist (1983), in the sense that these moments do not exist.

5.7 Overall conclusions

Overall the performances of the L_p-norm estimators were disappointing. L_p-norm estimators are not influenced by variance, orientation or collinearity. It was only in long tail distributions, like the Slash that L_p-norm estimators seem to perform better than OLSE. For the small X matrices the WLS and BFGS algorithms were more unstable and the choice of p

not well defined whereas in the full matrix, the algorithms were stable, quick and the p more uniform.

When $p > 2$ the WLS algorithm is unstable. For $p < 1$ the BFGS algorithm should be avoided. When p is near 1, the BFGS algorithm should be used with caution, and the results compared with the L_1 algorithm.

With real data the LPV vector should be made as loose as possible, imposing minimal restrictions. Furthermore, after the fit, the residuals should be investigated for outliers, even with L_p-norm estimators. Only the L_1 estimator is robust against outliers.

In estimating β, the user would ultimately like to find confidence intervals for the estimates, and make some inferences based on these estimates. However there are still problems with the estimation of $\tilde{\omega}_p^2$. Based on the experience of this simulation study we propose the following strategy for L_p-norm estimators:

1. For $p = \infty$, there are no estimates for ω_p^2 available, and as the CHEB estimator was only slightly better than OLS, we recommend the OLSE.

2. For $p > 2$, the original estimator of ω_p^2. When p becomes large (ie larger than 3) the possibility of setting $p = \infty$, should be investigated.

3. When $1.5 \leq p \leq 2$, use the perturbed estimator of ω_p^2.

4. For $1 < p < 1.5$, the estimator of ω_p^2 is unstable. Therefore set $p = 1$, or if p is near one, use the resulting betas, and residuals, then use the stable estimator of λ for the moment ratio parameter.

5. For $p = 1$, use the mks estimator of λ, discussed previously.

To find a value of p, and n reasonably large, we recommend the adaptive algorithm, and the method of Barr, to find p.

As previously noted, biased estimators are influenced by variance, orientation and collinearity but are impervious to distribution changes for the regular distributions of this study. The size of the sample (n) appeared to play no role in the choice of the biased estimator, as the same estimators were chosen for n = 30 and n = 200. We recommend even for small to moderate collinearity data sets, the use of biased estimators. Based on the evidence from this simulation study, use the RLW or one on the FPC family estimators.

5.8 Summary

This chapter consisted of the general findings of the simulation study for the small X matrices. Two full X-matrices were simulated, and the findings reported. Comparisons were made between the different algorithms, and between the full and the small matrices. Generally it appears that no estimator consistently outperforms other classes on the criterion of relative efficiency. Specific conditions associated with the optimality of specific estimators, were outlined in this chapter. Areas for further research were described.

Chapter 6

A GENERAL FORM OF THE LINEAR MODEL

6.1 Introduction

In chapter 2 we focussed mainly on the biased estimators of parameters to compensate for collinearity in the design or regressor matrix. Chapter 3 on the other hand addresses the issue of non-normal distributions by the use of the robust L_p-norm estimators. Thus, we have looked at two types of tools in the data analyst's toolbox. In chapter 5 we examined problems that occur when tools are used as a black box through which we pass a data set. In this chapter we set up a mixed model and in terms of this model we seek an overview of all the different tools described in this study. Those tools will transpire to be special cases of the overall framework.

So far we have concentrated on the LRM (1.1), $Y = X\beta + \epsilon$. We have assumed a simple linear function of the X's, we have ignored all prior information, and we accepted the LRM as the appropriate complete and correct model. No attempt was made to adjust this model to misspecifications (such as inclusion of unnecessary predictors or omission of necessary regressors), outliers were ignored (except for the short section in Chapter 5), and collinearity was addressed by the use of biased estimators.

In this chapter we introduce a mixed linear model (or general model), and by changing our view of the parameters we will be able to adapt the model to specify any of the scenarios mentioned in the previous paragraph. This task is accomplished by employing the theory of restricted least squares, where the set of restrictions is in general stochastic, but may be non-stochastic as a special case (sure restricted least squares).

6.2 A general model

The LRM of (1.1), $Y = X\beta + \epsilon$ can be written in the form

$$Y = f_0(X_0)\cdot\beta_0 + f_1(X_1)\cdot\beta_1 + f_2(X_2)\cdot\beta_2 + \ldots + f_{r-1}(X_{r-1})\cdot\beta_{r-1} + \epsilon \quad (6.2.1)$$

where $f_i(X_i)$ is the i-th function operating on the (i)-th column of the X-matrix (the first column of the X-matrix, X_0 is a column of ones for models with a non-zero intercept). There are r functions $f_i(X_i)$. In Chapter 1 (model 1.1), $f_i(X_i) = X_i$ for all i, interpreting $f_0(X_0) = 1$ in the more general form (6.2.1). The nature of the function f_i may be obtained from prior information, or possibly from transformations of the original regressor variables that are suggested by plots of the fitted residuals against the regressor variables. In L_p-estimation when p = 2 (OLS), the use of residual plots is a well-known practice. The use of residual plots for other values of p will only be feasible if p is fixed. But whenever p is obtained from the data, even if r is fixed, the forms $f_i(X_i)$ may change in manners suggested by the fitted residuals as p changes, and the form of the model function will change, with p. A function f_i found under a specific given value of p need not be transferable (though one might explore its use) to another value of p. In the literature (on L_p-norm estimators), the model (and hence each of the functions) is fixed, and the sample values are not used to assist in finding other functions. Relaxing this restriction may be an area for further research, although in the opinion of the author, the formulation of simplistic rules for obtaining the form of the function, for any value of p, (as in the case when p = 2), will constitute a fundamental error. Rather the analyst when confronted with the problem should be open-minded and led by informed collaboration with the subject specialists relevant to the data, and by intuition.

If we specifically allow $X_i \equiv X_j$ and $f_i(X_i) \neq f_j(X_j)$ for $i \neq j$, then the formulation is rich enough to include polynomial and exponential terms in an original subset of regressors.

The fact that the function relating Y to X and β is not always as simple as seen in (3.1.1) leads us to write (3.1.1) in the general form of (6.2.2).

The X-linear L_p-norm estimation problem is then defined as: Find an estimate of β, denoted by $\hat{\beta}_{L_p}$ ($(\hat{\beta}_{L_p})_j$ is the j-th element of $\hat{\beta}_{L_p}$) which

minimizes

$$\sum_{i=1}^{n} |Y_i - \sum_{j=0}^{r-1} f_j(X_j)\cdot(\hat{\beta}_{L_p})_j]|^p = \sum_{i=1}^{n} |\hat{\epsilon}_i|^p \qquad (6.2.2)$$

In Chapter 3, the general L_p-norm problem (3.1.1) was formulated as a mathematical programming problem:

$$\text{Minimize} \quad \sum_{i=1}^{n} (u_i^p + v_i^p) \qquad (3.1.2)$$

$$\left.\begin{array}{l} \text{subject to} \quad X_i'\hat{\beta}_{L_p} + (u_i - v_i) = Y_i \\ u_i, v_i \geq 0 \end{array}\right\} \quad i = 1,\ldots n$$

$$\hat{\beta}_{L_p} \quad \text{unconstrained}$$

where (as in Chapter 3) $\hat{\epsilon}_i = u_i - v_i$. However, we are seeking an overall general model, where the unconstrained (or unrestricted) model is a special case of the constrained problems (with penalized function). If the penalty function or the restrictions are denoted by say a vector $g(\beta,H,\nu)$ then (3.1.1), and the unconstrained (3.1.2), can be formulated as a minimization problem subject to (equality or inequality) constraints, ie

$$\text{minimize} \quad \sum_{i=1}^{n} |Y_i - \sum_{j=0}^{r-1} f_j(X_j)\cdot(\hat{\beta}_{L_p})_j]|^p$$

$$\text{subject to restrictions} \qquad (6.2.3)$$

$$g(\beta,H,\nu) \bullet h$$

where $\bullet$ denotes the relation $=$, $<$,$>$, $\leq$, or $\geq$ between the left and the right hand side of the restrictions. The matrix H, h and ν will be examined in section 6.3.

The advantage of formulating the L_p-norm in this general form (6.2.3) is to emphasize that the blind use of the norm is less likely when the statistician has explicitly formulated functions, and constraints. Linear programming solutions for (6.2.3) are available in the literature (Barrodale and Roberts (1978, p=1), Snyman (1993)). In section 6.3. we investigate the case where p = 2 (OLS).

6.3 The general model (mixed model), when p = 2.

When p = 2, (6.2.3) reduces to

$$\text{minimize} \sum_{i=1}^{n} |Y_i - \sum_{j=0}^{r-1} f_j(X_j)\cdot(\hat{\beta}_{L_p})_j]|^2 \text{ subject to } g(\beta,H,\nu) \bullet h.$$

That is, in addition to having data observations (Y, X) we now assume that we have auxiliary information ($g(\beta,H,\nu)\bullet$ h) on the vector of regression coefficients (β). When $\bullet$ represents inequalities, mathematical optimization techniques can be used to find a numerical solution (Mantel (1969), Judge and Takayama (1966) and Liew (1976))), but the algorithms mentioned in 6.2 for L_p-norm minimization with constraints are also available.

If $\bullet$ represents equalities and if we use a transformation from $f(X,\beta)$ to $X_T\beta_T$ to obtain the simple linear relationship, we apply

$$Y = X_T\beta_T + \epsilon \qquad (6.3.1)$$

where Y, and ϵ as is before, β_T is the vector of regression coefficients from the linearised model, that might be different from those coefficient β in the LRM (1.1). We assume β is estimable (ie X of full column rank) and if not, that we can reparameterise, or eliminate redundant variables, and proceed with a diminished X matrix of full rank. Thus X_T is a n×r matrix, representing a transformed X matrix, and

$$g(\beta_T,H,\nu) = H\beta_T + \nu = h \qquad (6.3.2)$$

where H is an ℓ × r design matrix (fixed) with rank $\ell \leq$ r, h is an ℓ × 1 stochastic vector which is known, and ν is an unobserved ℓ × 1 vector of random error variables with expectation $E(\nu) = \delta$, and variance matrix $\text{Var}(\nu) = \Sigma_\nu$, $E[\nu\nu'] = \Sigma_\nu + \delta\delta'$ and ν is uncorrelated with ϵ. Σ_ν is positive definite (pd), but is possibly unknown. Note that in special cases ℓ can be equal to r. Thus

$$\nu \sim rv(\delta,\Sigma_\nu), \quad \Sigma_\nu \text{ pd} \qquad (6.3.3)$$

For the sake of convenience; we drop the subscript T in the notation that will follow, as the subscript complicates the presentation and development of the model of this chapter. However we interpret the X and β, as those obtained from a transformation of $f(X,\beta)$ to $X\beta$.

In our linear regression model of (1.1) we assume $\epsilon \sim (0,\sigma^2 I)$. To introduce a generalised form, we assume $\epsilon \sim (\tau,\Sigma_\epsilon)$, where $E[\epsilon\epsilon'] = \Sigma_\epsilon + \tau\tau'$ and $Var(\epsilon) = \Sigma_\epsilon$. Furthermore we assume that the auxiliary information (6.3.2) to be independent of the sample and the model (6.3.1), that is

$$E(\nu\epsilon') = 0 : \ell \times n \tag{6.3.4}$$

Toutenberg (1982) augmented (6.3.1), the so-called sample information, with the auxiliary information (6.3.2) to form the following mixed model

$$\begin{bmatrix} Y \\ h \end{bmatrix} = \begin{bmatrix} X \\ H \end{bmatrix}\beta + \begin{bmatrix} \epsilon \\ \nu \end{bmatrix}$$

or

$$\dot{Y} = \dot{X}\beta + \dot{\epsilon} \tag{6.3.5}$$

where $\dot{Y} = \begin{bmatrix} Y \\ h \end{bmatrix} : (n+\ell)\times 1$; $\dot{X} = \begin{bmatrix} X \\ H \end{bmatrix} : (n+\ell)\times r$; and $\dot{\epsilon} = \begin{bmatrix} \epsilon \\ \nu \end{bmatrix} : (n+\ell)\times 1$ are the augmented matrices.

The expectation and covariance of $\dot{\epsilon}$ are then

$$E(\dot{\epsilon}) = \begin{bmatrix} \tau \\ \delta \end{bmatrix} \tag{6.3.6}$$

$$Cov(\dot{\epsilon}\dot{\epsilon}') = \Sigma = \begin{bmatrix} \Sigma_\epsilon & 0 \\ 0 & \Sigma_\nu \end{bmatrix} \tag{6.3.7}$$

using (6.3.4).

The augmented Generalised Least Square Estimator (AGLSE) for (6.3.5), involves minimizing

$$(\dot{Y} - \dot{X}\beta)'\Sigma^{-1}(\dot{Y} - \dot{X}\beta)$$

with respect to β. This estimator will be denoted by $\hat{\beta}_G$ and is obtained as

$$\hat{\beta}_G = \{\dot{X}'\Sigma^{-1}\dot{X}\}^{-1}\dot{X}'\Sigma^{-1}\dot{Y} \tag{6.3.8}$$

Furthermore notice that $\hat{\beta}_G$ is a function of a matrix Σ, which is unknown in the general case. In later sections an operational version of the estimator will be introduced, in which Σ will take specific structural forms involving a small set of parameters to be estimated from the data.

6.3.1 Properties of $\hat{\beta}_G$

1. Relationship to GLSE

Setting $S = [X'\Sigma_\epsilon^{-1}X]^{-1}$ and $D = [\Sigma_\nu + HSH']^{-1}$ in the estimation of $\hat{\beta}_G$ we obtain

$$\begin{aligned}
\hat{\beta}_G &= \{\dot{X}'\Sigma^{-1}\dot{X}\}^{-1}\dot{X}'\Sigma^{-1}\dot{Y} \\
&= \{[X' \; H']\begin{bmatrix} \Sigma_\epsilon^{-1} & 0 \\ 0 & \Sigma_\nu^{-1} \end{bmatrix}\begin{bmatrix} X \\ H \end{bmatrix}\}^{-1}[X'\Sigma_\epsilon^{-1}Y + H'\Sigma_\nu^{-1}h] \\
&= [X'\Sigma_\epsilon^{-1}X + H'\Sigma_\nu^{-1}H]^{-1}[X'\Sigma_\epsilon^{-1}Y + H'\Sigma_\nu^{-1}h] \\
&= \{[X'\Sigma_\epsilon^{-1}X]^{-1} - [X'\Sigma^{-1}X]^{-1}H'[\Sigma_\nu + H[X'\Sigma_\epsilon^{-1}X]^{-1}H']^{-1}H[X'\Sigma_\epsilon^{-1}X]^{-1}\} \\
&\qquad\qquad \times [X'\Sigma_\epsilon^{-1}Y + H'\Sigma_\nu^{-1}h] \\
&= \{S - SH'[\Sigma_\nu + HSH']^{-1}HS\}[X'\Sigma_\epsilon^{-1}Y + H'\Sigma_\nu^{-1}h] \\
&= \{S - SH'DHS\}[X'\Sigma_\epsilon^{-1}Y + H'\Sigma_\nu^{-1}h] \\
&= SX'\Sigma_\epsilon^{-1}Y - SH'DHSX'\Sigma_\epsilon^{-1}Y + SH'\Sigma_\nu^{-1}h - SH'DHSH'\Sigma_\nu^{-1}h \\
&= \hat{\beta}_G - SH'DH\hat{\beta}_G + SH'DD^{-1}\Sigma_\nu^{-1}h - SH'DHSH'\Sigma_\nu^{-1}h \\
&= \hat{\beta}_G - SH'D[H\hat{\beta}_G - D^{-1}\Sigma_\nu^{-1}h + HSH'\Sigma_\nu^{-1}h] \\
&= \hat{\beta}_G - SH'D[H\hat{\beta}_G - \{[\Sigma_\nu + HSH']\Sigma_\nu^{-1} - HSH'\Sigma_\nu^{-1}\}h] \\
&= \hat{\beta}_G - SH'D[H\hat{\beta}_G - \{[I + HSH'\Sigma_\nu^{-1} - HSH'\Sigma_\nu^{-1}\}h] \\
&= \hat{\beta}_G - SH'D[H\hat{\beta}_G - h]
\end{aligned} \tag{6.3.9}$$

Note that

$$\begin{aligned}\{\dot{X}'\Sigma^{-1}\dot{X}\}^{-1} &= [X'\Sigma_{\epsilon}^{-1}X + H'\Sigma_{\nu}^{-1}H]^{-1} \\ &= \{S - SH'[\Sigma_{\nu} + HSH']^{-1}HS\} \\ &= \{S - SH'DHS\} \qquad (6.3.10)\end{aligned}$$

The increase in information is effected in (6.3.10) as the increase in precision and the decrease in variance.

2. Expectation

Since, by assumption $E(h) = H\beta + \delta$, and $E(Y) = X\beta + \tau$, we obtain

$$\begin{aligned}E\{\hat{\beta}_G\} &= E\{\hat{\beta}_G - SH'D[H\hat{\beta}_G - h]\} \\ &= (\beta + SX'\Sigma_{\epsilon}^{-1}\tau) - SH'D[H(\beta + SX'\Sigma_{\epsilon}^{-1}\tau) - H\beta - \delta\} \\ &= \beta + SX'\Sigma_{\epsilon}^{-1}\tau - SH'D[HSX'\Sigma_{\epsilon}^{-1}\tau - \delta\} \\ &= \beta + SX'\Sigma_{\epsilon}^{-1}\tau - SH'DHSX'\Sigma_{\epsilon}^{-1}\tau + SH'D\delta \\ &= \beta + SH'D\delta + S(I - H'DHS)X'\Sigma_{\epsilon}^{-1}\tau \qquad (6.3.11)\end{aligned}$$

Thus $\hat{\beta}_G$ is biased for β, and we denote the bias of $\hat{\beta}_G$ by

$$\theta = SH'D\delta + S(I - H'DHS)X'\Sigma_{\epsilon}^{-1}\tau \qquad (6.3.12)$$

If $X'\Sigma_{\epsilon}^{-1}\tau = 0$, then $\theta = SH'D\delta$, and we require $\delta = 0$ for an unbiased estimator of β.

3. Variance

$$\begin{aligned}\mathrm{var}\{\hat{\beta}_G\} &= \mathrm{var}(\{\dot{X}'\Sigma^{-1}\dot{X}\}^{-1}\dot{X}'\Sigma^{-1}\dot{Y}) \\ &= \{\dot{X}'\Sigma^{-1}\dot{X}\}^{-1}\dot{X}'\Sigma^{-1}\mathrm{var}(\dot{X}\beta + \dot{\epsilon})\Sigma^{-1}\dot{X}\{\dot{X}'\Sigma^{-1}\dot{X}\} \\ &= \{\dot{X}'\Sigma^{-1}\dot{X}\}^{-1}\dot{X}'\Sigma^{-1}\Sigma\Sigma^{-1}\dot{X}\{\dot{X}'\Sigma^{-1}\dot{X}\} \\ &= \{\dot{X}'\Sigma^{-1}\dot{X}\}^{-1} \qquad (6.3.13)\end{aligned}$$

4. Mean mean squared error (MSE) of $\hat{\beta}_G$

$$\begin{aligned} MSE\{\hat{\beta}_G\} &= var\{\hat{\beta}_G\} + \theta\theta' \\ &= \{\dot{X}'\Sigma^{-1}\dot{X}\}^{-1} + \theta\theta' \end{aligned} \tag{6.3.14}$$

where

$$\theta\theta' = S[H'D\delta + (I - H'DHS)X'\Sigma_\epsilon^{-1}\tau][\delta'DH + \tau'\Sigma_\epsilon^{-1}X(I - SH'DH)]S$$

5. Total mean squared error (TMSE) of $\hat{\beta}_G$

$$\begin{aligned} TMSE\{\hat{\beta}_G\} &= tr\{\{\dot{X}'\Sigma^{-1}\dot{X}\}^{-1} + \theta\theta'\} \\ &= tr\{\{\dot{X}'\Sigma^{-1}\dot{X}\}^{-1}\} + tr(\theta\theta') \end{aligned} \tag{6.3.15}$$

$$\begin{aligned} tr(\theta\theta') &= tr([\delta'DH + \tau'\Sigma_\epsilon^{-1}X(I - SH'DH)]SS[H'D\delta + (I - H'DHS)X'\Sigma_\epsilon^{-1}\tau]) \\ &= \delta'DHSSH'D\delta + \tau'\Sigma_\epsilon^{-1}X(I - SH'DH)SS(I - H'DHS)X'\Sigma_\epsilon^{-1}\tau \\ &\quad + 2(\delta'DHSS(I - H'DHS)X'\Sigma_\epsilon^{-1}\tau) \end{aligned}$$

If $X'\Sigma_\epsilon^{-1}\tau = 0$, then $tr(\theta\theta') = \delta'DHSSH'D\delta$

6. Residuals of the AGLSE

$$\begin{aligned} \hat{\dot{\epsilon}} &= \dot{Y} - \dot{X}\hat{\beta}_G \\ &= \dot{Y} - \dot{X}\{\dot{X}'\Sigma^{-1}\dot{X}\}^{-1}\dot{X}'\Sigma^{-1}\dot{Y} \\ &= [I - \dot{X}\{\dot{X}'\Sigma^{-1}\dot{X}\}^{-1}\dot{X}'\Sigma^{-1}][\dot{X}\beta + \dot{\epsilon}] \\ &= \dot{X}\beta + \dot{\epsilon} - \dot{X}\{\dot{X}'\Sigma^{-1}\dot{X}\}^{-1}\dot{X}'\Sigma^{-1}\dot{X}\beta - \dot{X}\{\dot{X}'\Sigma^{-1}\dot{X}\}^{-1}\dot{X}'\Sigma^{-1}\dot{\epsilon} \\ &= \dot{\epsilon} - \dot{X}\{\dot{X}'\Sigma^{-1}\dot{X}\}^{-1}\dot{X}'\Sigma^{-1}\dot{\epsilon} \\ &= [I - \dot{X}\{\dot{X}'\Sigma^{-1}\dot{X}\}^{-1}\dot{X}'\Sigma^{-1}]\dot{\epsilon} \\ &= [I - \dot{M}]\dot{\epsilon} \end{aligned} \tag{6.3.16}$$

where $\dot{M} = \dot{X}\{\dot{X}'\Sigma^{-1}\dot{X}\}^{-1}\dot{X}'\Sigma^{-1}$. Note that $\dot{M}$ is a function of the generally unknown matrix Σ^{-1}. If we assume that

$$\Sigma = \sigma^2\begin{bmatrix} W_\epsilon & 0 \\ 0 & W_\nu \end{bmatrix} = \sigma^2\dot{W} \qquad (6.3.17)$$

where σ^2 is unknown and $\dot{W} = \begin{bmatrix} W_\epsilon & 0 \\ 0 & W_\nu \end{bmatrix}$ is known, then

$$\begin{aligned} E[\hat{\dot{\epsilon}}'\dot{W}^{-1}\hat{\dot{\epsilon}}] &= tr\{[\sigma^2\dot{W}+E(\dot{\epsilon})E(\dot{\epsilon})'](I - \dot{M}')\dot{W}^{-1})\} \quad \text{(from 1.4.9)} \\ &= tr\{[\sigma^2\dot{W}(I - \dot{M}')\dot{W}^{-1}) + E(\dot{\epsilon})E(\dot{\epsilon})'](I - \dot{M}')\dot{W}^{-1})\} \\ &= \sigma^2 tr(I - \dot{M}') + tr\{E(\dot{\epsilon})E(\dot{\epsilon})'](I - \dot{M}')\dot{W}^{-1})\} \\ &= \sigma^2(tr\{I_{n+\ell}\} - tr\{\dot{M}\}) + E(\dot{\epsilon})'(\dot{W}^{-1} - \dot{M}'\dot{W}^{-1})E(\dot{\epsilon}) \\ &= \sigma^2(tr\{I_{n+\ell}\} - tr\{\dot{X}\{\dot{X}'\Sigma^{-1}\dot{X}\}^{-1}\dot{X}'\Sigma^{-1}\} + E(\dot{\epsilon})'(\dot{W}^{-1} - \dot{M}'\dot{W}^{-1})E(\dot{\epsilon}) \\ &= \sigma^2(n+\ell - tr\{I_r\}) + E(\dot{\epsilon})'(\dot{W}^{-1} - \dot{M}'\dot{W}^{-1})E(\dot{\epsilon}) \\ &= \sigma^2(n+\ell-r) + E(\dot{\epsilon})'(\dot{W}^{-1} - \dot{M}'\dot{W}^{-1})E(\dot{\epsilon}) \qquad (6.3.18) \end{aligned}$$

Thus for an unbiased estimator of σ^2, when $\Sigma = \sigma^2\dot{W}$, $\dot{W}$ known and even if $\dot{X}'\dot{W}^{-1}E(\dot{\epsilon}) = 0$, we still require $E(\dot{\epsilon}) = 0$ to obtain

$$\hat{\sigma}^2 = [\hat{\dot{\epsilon}}'\dot{W}^{-1}\hat{\dot{\epsilon}}]/(n-r+\ell) \qquad (6.3.19)$$

6.3.2 Comparison of augmented and GLS models

In this section we will assume that $E(\epsilon) = \tau = 0$. Note that $\hat{\beta}_G$ is a function of the unknown covariance matrix Σ. Special cases of Σ will be discussed at a later stage.

6.3.2.1 Unbiasedness

From (6.3.10) we know that the bias for the augmented estimator is $\theta = -SH'D\delta$, and that for the GLS estimator, the bias is zero. Thus $\hat{\beta}_G$ is unbiased for β only when $H'D\delta = 0$, eg $\delta = 0$ when $\nu \sim (0,\Sigma_\nu)$.

6.3.2.2 Variance

From (6.3.11) the variance of $\hat{\beta}_G$ is

$$\begin{aligned} V\{\hat{\beta}_G\} &= \{\dot{X}'\Sigma^{-1}\dot{X}\}^{-1} \\ &= \{S - SH'DHS\} \end{aligned}$$

and the variance of the GLSE, $V(\hat{\beta}_G) = \{X'\Sigma_\epsilon^{-1}X\}^{-1} = S$.

Thus

$$\begin{aligned} V(\hat{\beta}_G) - V\{\hat{\beta}_G\} &= S - \{S - SH'DHS\} \\ &= SH'DHS \end{aligned} \qquad (6.3.20)$$

Σ_ν, $\Sigma_\epsilon > 0$ i.e positive definite (pd), $S = \{X'\Sigma_\epsilon^{-1}X\}^{-1}$ is positive semi-definite (psd) (see theorem A.9 of Toutenberg (1982)) and HSH' is psd, thus, $D = [\Sigma_\nu + HSH']^{-1}$ is psd (Goldberger (1964) pp 35-37). Thus the difference $V(\hat{\beta}_G) - V\{\hat{\beta}_G\}$ is psd, and the augmented model yields an estimator of β that has smaller variance (better precision) than the normal (unaugmented) model. (See section 6.3.3 for further details, when Σ has to be estimated).

6.3.2.3 MSE criteria

We distinguish between three types of MSE criteria. In this section we will define them, initially with Σ unknown. The criteria are not operational

unless Σ is given or estimated. In later sections special operational conditions are investigated for these three criteria.

6.3.2.3.1 MSE-I Criterion

The first of the MSE criteria is sometimes referred to as MSE criterion I (Toutenberg (1982)) or the Strong MSE criterion. An estimator $\tilde{\beta}$ for β is MSE-I better than estimator $\check{\beta}$ for β if the matrix MSE difference between them (that is $MSE(\check{\beta})$ - $MSE(\tilde{\beta})$) is positive-semi definite.

From (6.3.12) the mean squared error of $\hat{\beta}_G$ is

$$MSE\{\hat{\beta}_G\} = \{\dot{X}'\Sigma^{-1}\dot{X}\}^{-1} + SH'D\delta\delta'DHS$$

and the $MSE\{\hat{\beta}_G\} = \{X'\Sigma_\epsilon^{-1}X\}^{-1} = S$. Thus the matrix difference

$$\begin{aligned} MSE\{\hat{\beta}_G\} - MSE\{\hat{\beta}_G\} &= \{X'\Sigma_\epsilon^{-1}X\}^{-1} - \{\dot{X}'\Sigma^{-1}\dot{X}\}^{-1} - SH'D\delta\delta'DHS \\ &= S - \{S - SH'DHS\} - SH'D\delta\delta'DHS \qquad \text{from (6.3.10)} \\ &= SH'DHS - SH'D\delta\delta'DHS \\ &= SH'D[D^{-1} - \delta\delta']DHS \qquad (6.3.21) \end{aligned}$$

and to ensure that (6.3.21) is psd

$$[D^{-1} - \delta\delta'] \geq 0$$

We have shown earlier that D is pd, so D can be written as

$$D = D^{\frac{1}{2}}D^{\frac{1}{2}} \quad \text{(from Theorem A.2, Toutenberg (1982))}$$

Thus

$$[D^{-1} - \delta\delta'] = D^{-\frac{1}{2}}[I - D^{\frac{1}{2}}\delta\delta'D^{\frac{1}{2}}]\,D^{-\frac{1}{2}}$$

But

$[I - D^{\frac{1}{2}}\delta\delta'D^{\frac{1}{2}}] \geq 0$ if and only if

$\delta'D^{\frac{1}{2}}D^{\frac{1}{2}}\delta \leq 1$ (Toutenberg(1982), Theorem A.17, p 186)

So the MSE-I criterion amounts to requiring

$$\varphi = \delta'D^{1/2}D^{1/2}\delta = \delta'D\delta = \delta'[\Sigma_v + HSH']^{-1}\delta \leq 1 \qquad (6.3.22)$$

Thus by the criterion we have bounded the unknown quantity φ to be smaller than one. The interpretation is that the bias δ is small in relation to the variance. In later sections we will show that under specific conditions $\varphi \rightarrow \lambda$ the non-centrality parameter of the F distribution.

6.3.2.3.2 MSE II Criterion

The TMSE difference between two competing estimators is known as the first weak MSE criterion, MSE-II. Some authors refer to it simply as the weak MSE criterion.

An estimator $\tilde{\beta}$ of β is MSE-II better than $\check{\beta}$ for β if the TMSE difference between the estimators (that is $\text{TMSE}(\check{\beta})$ - $\text{TMSE}(\tilde{\beta})$) is greater than or equal to zero.

From (6.3.14) the $\text{TMSE}(\hat{\beta}_G) = \text{tr}\{\{\dot{X}'\Sigma^{-1}\dot{X}\}^{-1} + \theta\theta'\}$
$= \text{tr}\{S - SH'DHS\} + \text{tr}(\theta\theta')$

$\text{tr}(\theta\theta') = \text{tr}([\delta'DH + \tau'\Sigma_\epsilon^{-1}X(I - SH'DH)]SS[H'D\delta + (I - H'DHS)X'\Sigma_\epsilon^{-1}\tau])$

$= \delta'DHSSH'D\delta + \tau'\Sigma_\epsilon^{-1}X(I - SH'DH)SS(I - H'DHS)X'\Sigma_\epsilon^{-1}\tau$

$+ 2(\delta'DHSS(I - H'DHS)X'\Sigma_\epsilon^{-1}\tau)$

$= \delta'DHSSH'D\delta$ (when $X'\Sigma_\epsilon^{-1}\tau = 0$)]

and from (1.4.5) the

$$\begin{aligned} TMSE(\hat{\beta}_G) &= tr[\{X'\Sigma_\epsilon^{-1}X\}^{-1}] + \tau'\Sigma_\epsilon^{-1}X\{X'\Sigma_\epsilon^{-1}X\}^{-1}\{X'\Sigma_\epsilon^{-1}X\}^{-1}X'\Sigma_\epsilon^{-1}\tau \\ &= tr[S] + tr[\tau'\Sigma_\epsilon^{-1}XSSX'\Sigma_\epsilon^{-1}\tau] \end{aligned}$$

$$TMSE(\hat{\beta}_G) - TMSE(\hat{\beta}_G) = \tau'\Sigma_\epsilon^{-1}XSSX'\Sigma_\epsilon^{-1}\tau + tr\{SH'DHS\} - tr(\theta\theta')$$

$$\begin{aligned} &= tr[SH'DHS] + \tau'\Sigma_\epsilon^{-1}XSSX'\Sigma_\epsilon^{-1}\tau - \delta'DHSSH'D\delta \\ &\quad - \tau'\Sigma_\epsilon^{-1}X(I - SH'DH)SS(I - H'DHS)X'\Sigma_\epsilon^{-1}\tau \\ &\quad - 2(\delta'DHSS(I - H'DHS)X'\Sigma_\epsilon^{-1}\tau \end{aligned}$$

$$\begin{aligned} &= tr[SH'DHS] - \delta'DHSSH'D\delta + \tau'\Sigma_\epsilon^{-1}XSSX'\Sigma_\epsilon^{-1}\tau \\ &\quad - \tau'\Sigma_\epsilon^{-1}X(I - SH'DH)SS(I - H'DHS)X'\Sigma_\epsilon^{-1}\tau \\ &\quad - 2(\delta'DHSS(I - H'DHS)X'\Sigma_\epsilon^{-1}\tau \qquad (6.3.23) \end{aligned}$$

$$= tr[SH'DHS] - \delta'DHSSH'D\delta \quad (\text{when } X'\Sigma_\epsilon^{-1}\tau = 0)$$

We will prefer the $\hat{\beta}_G$ estimator if the right hand side of (6.3.23) is ≥ 0. Thus when $X'\Sigma_\epsilon^{-1}\tau = 0$

$$tr[SH'DHS] \geq \delta'DHSSH'D\delta \qquad (6.3.24)$$

The bound (6.3.24) was also established by Judge and Bock (1978).

The following derivations attempt to obtain bounds for φ in (6.3.24), but applications of these bounds, are not yet clear.

Let $A = SH'[HSSH']^{-1}D^{-1}[HSSH']^{-1}HS$, where $[HSSH']^{-1}$ exist, then $A:r\times r$ is a symmetric matrix and

$$HSASH' = HSSH'[HSSH']^{-1}D^{-1}[HSSH']^{-1}HSSH' = D^{-1}.$$

Denote the eigenvalues of A by $a_1 \geq a_2 \geq \ldots \geq a_r$. Then by using theorem A.13 of Toutenberg (1982), we can say that

$$a_r \leq \frac{[\delta'DHS]A[SH'D\delta]}{[\delta'DHS][SH'D\delta]} \leq a_1$$

$$a_r \leq \frac{[\delta'DD^{-1}D\delta]}{[\delta'DHS][SH'D\delta]} \leq a_1$$

Thus

$$a_r\delta'DHSSH'D\delta \leq \varphi$$

and to ensure that $tr[SH'DHS] \geq \delta'DHSSH'D\delta$, we require

$$\varphi \leq a_r tr[SH'DHS] \qquad (6.3.25)$$

Equation (6.3.25) bounds φ. However the right-hand side still involves the matrix D, an unknown matrix. If we assume that $\Sigma = \sigma^2\begin{bmatrix} W & 0 \\ 0 & W_\nu \end{bmatrix} = \sigma^2\dot{W}$, where σ^2 is unknown and $\dot{W} = \begin{bmatrix} W & 0 \\ 0 & W_\nu \end{bmatrix}$ is known, then

$$S = [X'\Sigma_\epsilon^{-1}X]^{-1} = \sigma^2[X'W^{-1}X]^{-1}$$

$$D = [\sigma^2W_\nu + \sigma^2H[X'W^{-1}X]^{-1}H']^{-1}$$

$$= \sigma^{-2}[W_\nu + H[X'W^{-1}X]^{-1}H']^{-1}$$

$$tr[SH'DHS] = \sigma^2 tr[[X'W^{-1}X]^{-1}H'[W_\nu + H[X'W^{-1}X]^{-1}H']^{-1}H[X'W^{-1}X]^{-1}].$$

It is clear that under these conditions the eigenvalues of A will be a function of σ^{-2}, which will cancel with the σ^2 contained in $tr[SH'DHS]$, thus the right hand side of (6.3.25) involves no unknown matrices and can be

calculated, bounding the unknown LHS. Note that the right hand side of (6.3.25) does not involve the unknown matrix δ as in the bounds given in (6.3.24)

Also see the discussion of Toutenberg (1982), Wallace (1972) and Yancey *et al.* (1973).

6.3.2.3.3 MSE-III Criterion

Toutenberg defined the second weak MSE criterion or MSE-III as:

An unbiased estimator $\tilde{\beta}$ of β is MSE-III better than $\check{\beta}$ for β if

$$E(\check{\beta} - E(\check{\beta}))'W(\check{\beta} - E(\check{\beta})) - E(\tilde{\beta} - E(\tilde{\beta}))'W(\tilde{\beta} - E(\tilde{\beta})) \geq 0 \qquad (6.3.26)$$

where $W = \{X'\Sigma_\epsilon^{-1}X\}$, and $E(\check{\beta}) = \beta = E(\tilde{\beta})$ (ie unbiased) in the notation of Toutenberg (1982). We introduce the general form of (6.3.26), the so-called weighted TMSE or when W is of a certain form, the so-called predictive TMSE criteria.

Thus the MSE-III for $\hat{\beta}_G = \{\dot{X}'\Sigma^{-1}\dot{X}\}^{-1}\dot{X}'\Sigma^{-1}\dot{Y})$, $\hat{\beta}_G = \{X'\Sigma_\epsilon^{-1}X\}^{-1}X'\Sigma_\epsilon^{-1}Y$ and the weighted matrix which we take in this case to be $W = S^{-1} = \{X'\Sigma_\epsilon^{-1}X\}$ is

$$\begin{aligned}
&E(\hat{\beta}_G - E(\hat{\beta}_G))'S^{-1}(\hat{\beta}_G - E(\hat{\beta}_G)) - E(\hat{\dot{\beta}}_G - E(\hat{\dot{\beta}}_G))'S^{-1}(\hat{\dot{\beta}}_G - E(\hat{\dot{\beta}}_G)) \\
&\quad = \mathrm{tr}\{E(\hat{\beta}_G - \beta)'S^{-1}(\hat{\beta}_G - \beta) - E(\hat{\dot{\beta}}_G - E(\hat{\dot{\beta}}_G))'S^{-1}(\hat{\dot{\beta}}_G - E(\hat{\dot{\beta}}_G))\} \\
&\quad = E[\mathrm{tr}\{(\hat{\beta}_G - \beta)'S^{-1}(\hat{\beta}_G - \beta)\} - \mathrm{tr}\{(\hat{\dot{\beta}}_G - E(\hat{\dot{\beta}}_G))'S^{-1}(\hat{\dot{\beta}}_G - E(\hat{\dot{\beta}}_G))\}] \\
&\quad = \mathrm{tr}\{E[S^{-1}(\hat{\beta}_G - \beta)(\hat{\beta}_G - \beta)'] - E[S^{-1}(\hat{\dot{\beta}}_G - E(\hat{\dot{\beta}}_G)(\hat{\dot{\beta}}_G - E(\hat{\dot{\beta}}_G)')]\} \\
&\quad = \mathrm{tr}\{S^{-1}\mathrm{MSE}(\hat{\beta}_G)\} - \mathrm{tr}\{S^{-1}\mathrm{MSE}(\hat{\dot{\beta}}_G)\} \\
&\quad = \mathrm{tr}\{S^{-1}SH'D[D^{-1} - \delta\delta']DHS\} \qquad \text{from (6.3.19)} \\
&\quad = \mathrm{tr}\{DHSH'[I - D\delta\delta']\} \qquad (6.3.27)
\end{aligned}$$

or $\delta'DHSH'D\delta \leq \mathrm{tr}\{H'DHS\}$

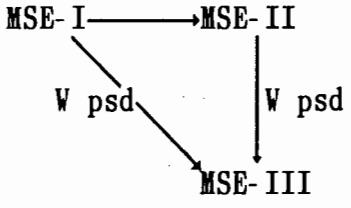
These three criteria can be related as follows: MSE-I implies MSE-II as the trace of a psd matrix is ≥ 0; MSE-I plus the condition that W must be psd implies MSE-III, because MSE-III can be written as

$$\mathrm{tr}[W\{\mathrm{MSE}(\hat{\beta}_G)\}] \geq 0$$

which is only true if W is psd. Similarly MSE-II implies MSE-III, if and only if W is psd:

$$\mathrm{tr}[\mathrm{MSE}(\hat{\beta}_G)] \geq 0 \qquad \text{(MSE-II)}$$

thus $\mathrm{tr}[W\{\mathrm{MSE}(\hat{\beta}_G)\}] \geq 0$ if and only if W is psd. Then to summarise we have



6.3.3 The unknown Σ

The AGLSE $(\hat{\beta}_G)$ defined in (6.3.8) is a function of an unknown positive definite matrix Σ and thus not operational. In practice we specify a structure for Σ (or assumptions on Σ) and then estimate the unknown parameters in Σ with the sample data at hand.

The statistical consequences of incorrectly assuming the wrong structure of Σ will lead to estimators that are inefficient and will lead to incorrect assessment of the reliability of the estimator. For a detailed discussion of this topic and further references, the reader is referred to Judge et.al (1980).

Thus in practice, once a structure for Σ has been specified, a number of questions arise:

(a) Is the structure correct (null hypothesis) and is it possible to envisage specify a specific alternative hypothesis?
(b) How do we estimate Σ?
(c) What are the properties of the AGLSE when a specific estimator of Σ is substituted for Σ?

In the following sections of 6.3.3 we assume different structures for Σ. Estimation of each structure is explored and where applicable some properties of the resulting estimator summarised. It is of course impossible to cover all structures. For more structures see Judge et. al (1980) or the discussion of Vinod and Ullah (1981) on the heteroscedasticity problem, where a wide list of further references are presented.

6.3.3.1 $\Sigma = \sigma^2\dot{W}$, (unknown scalar × known matrix)

Assume that Σ, takes the form defined in (6.3.17):

$$\Sigma = \sigma^2 \begin{bmatrix} W_\epsilon & 0 \\ 0 & W_\nu \end{bmatrix} = \sigma^2 \dot{W}$$

where σ^2 is unknown and $\dot{W} = \begin{bmatrix} W_\epsilon & 0 \\ 0 & W_\nu \end{bmatrix}$ is known, then the GLSE (1.4.1) is

$$\hat{\beta}_G = \{X' W_\epsilon^{-1} X\}^{-1} X' W_\epsilon^{-1} Y \qquad (6.3.28)$$

and the AGLSE of (6.3.8) becomes

$$\hat{\beta}_G = \{\dot{X}' \dot{W}^{-1} \dot{X}\}^{-1} \dot{X}' \dot{W}^{-1} \dot{Y} \qquad (6.3.29)$$

an estimator which is a function of known matrices. The properties of $\hat{\beta}_G$ are summarised in the following table:

Table 6.1 Properties of $\hat{\beta}_G$, for Σ as in (6.3.17)

$\hat{\beta}_G$	$\{\dot{X}'\dot{W}^{-1}\dot{X}\}^{-1}\dot{X}\dot{W}^{-1}\dot{X}\}^{-1}\dot{X}'\dot{W}^{-1}\dot{Y} = \hat{\beta}_G - S_kH'D_kH[H\hat{\beta}_G - h]$
$\mathrm{Var}(\hat{\beta}_G)$	$\sigma^2\{\dot{X}'\dot{W}^{-1}\dot{X}\}^{-1}$
S_k	$[X'W_\epsilon^{-1}X]^{-1}$
D_k	$[W_\nu+H[X'W_\epsilon^{-1}X]^{-1}H']^{-1}$
$(\hat{\beta}_G)$	$\beta + S_kH'D_k\delta + S_k(I - H'D_kHS_k)X'\Sigma_\epsilon^{-1}\tau$
θ	$S_kH'D_k\delta + S_k(I - H'D_kHS_k)X'\Sigma_\epsilon^{-1}\tau$
$\theta\theta'$	$S_k[H'D_k\delta + (I-H'D_kHS_k)X'\Sigma_\epsilon^{-1}\tau][\delta'D_kH + \tau'\Sigma_\epsilon^{-1}X(I-S_kH'D_kH)]S_k$
$\mathrm{MSE}(\hat{\beta}_G)$	$\sigma^2\{\dot{X}'\dot{W}^{-1}\dot{X}\}^{-1} + \theta\theta'$
$\mathrm{TMSE}(\hat{\beta}_G)$	$\sigma^2\mathrm{tr}(\{\dot{X}'\dot{W}^{-1}\dot{X}\}^{-1}) + \mathrm{tr}(\theta\theta')$
$\dot{\epsilon}$	$\hat{\dot{\epsilon}} = [I - \dot{M}_k]\dot{\epsilon};\quad \dot{M}_k = \dot{X}\{\dot{X}'\dot{W}^{-1}\dot{X}\}^{-1}\dot{X}'\dot{W}^{-1}$
$\hat{\sigma}^2$	$[\hat{\dot{\epsilon}}'\dot{W}^{-1}\hat{\dot{\epsilon}}]/(n-r+\ell)$
$V(\hat{\beta}_G) - V\{\hat{\beta}_G\}$	$\sigma^2S_kH'D_kHS_k$

(Note that σ^2S_k, and $\sigma^{-2}D_k$ are S, and D with the variance terms, reduced by (6.3.17). The subscript k denotes that the associated matrices are known)

An estimate of $\nu = h - H\beta$ is the GLSE of ν, namely $\hat{\nu}_G$, where

$$\begin{aligned}\hat{\nu}_G &= h - H\hat{\beta}_G \\ E(\hat{\nu}_G) = E(h - H\hat{\beta}_G) &= H\beta + \delta - H\beta - HSX'W_\epsilon^{-1}\tau \\ &= \delta - HSX'W_\epsilon^{-1}\tau \\ &= \delta \quad (\text{when } X'W_\epsilon^{-1}\tau = 0)\end{aligned} \tag{6.3.30}$$

$$\begin{aligned}
\mathrm{Var}(\hat{\nu}_G) &= \mathrm{Var}(h - H\hat{\beta}_G) \\
&= \mathrm{Var}(H\beta + \nu) + H\mathrm{Var}(\hat{\beta}_G)H' \\
&= \mathrm{Var}(\nu) + \sigma^2 H\{X'W_\epsilon^{-1}X\}^{-1})H'] \\
&= \sigma^2 W_\nu + \sigma^2 H\{X'W_\epsilon^{-1}X\}^{-1})H'] \\
&= \sigma^2[W_\nu + H\{X'W_\epsilon^{-1}X\}^{-1})H'] \\
&= \sigma^2 D_k^{-1} \qquad (6.3.31)
\end{aligned}$$

Thus $\hat{\nu}_G \sim (\delta, \sigma^2 D_k^{-1})$, and if we assume normality then $\hat{\nu}_G \sim N(\delta, \sigma^2 D_k^{-1})$ and we note that

$$(h - H\hat{\beta}_G)'[\sigma^{-2}D_k](h - H\hat{\beta}_G) \text{ is } \chi^2[r(D_k), \tfrac{1}{2}\delta' D_k \delta/\sigma^2] \qquad (6.3.32)$$

as $[\sigma^{-2}D_k][\sigma^2 D_k^{-1}] = I$, is an idempotent matrix. (Theorem 2, p57 of Searle (1971)). Also $r(D_k) = \ell$, and using the results in 6.3.2.3.1, the MSE-I criteria is precisely

$$\varphi_k = (h - H\beta)'[\sigma^{-2}D_k](h - H\beta) \leq 1$$

or if we denote the non-centrality parameter by λ,

$$\lambda = \varphi_k/2 = (h - H\beta)'[\sigma^{-2}D_k](h - H\beta)/2 \leq \tfrac{1}{2} \qquad (6.3.33)$$

If we assume normality for $\hat{\nu}_G$, we have shown that MSE-I criterion requires $\lambda \leq \frac{1}{2}$ for the non-centrality parameter of the χ^2 statistic (with ℓ degrees of freedom) that tests the consistency of the stochastically observed auxiliary information with the data. We have established that $\lambda \leq \frac{1}{2}$ for the AGLSE to be better than the GLSE in the sense of MSE-I criterion. This result is given by Toro-Vizcarrondo and Wallace (1968).

Under the MSE-II criterion we establish in (6.3.25)

$$\varphi \leq a_r \mathrm{tr}[SH'DHS]$$

$$\mathrm{tr}[SH'DHS] = \sigma^2 \mathrm{tr}[[X'W^{-1}X]^{-1}H'[W_\nu + H[X'W^{-1}X]^{-1}H']^{-1}H[X'W^{-1}X]^{-1}] \tag{6.3.34}$$

The MSE-III criterion established in 6.3.2.3.3, under the covariance structure of (6.3.17), and taking the weight matrix, W as $\sigma^2[X'W_\epsilon^{-1}X]$ the MSE-III criterion reduces to

$$\mathrm{tr}\{\sigma^2[X'W_\epsilon^{-1}X]\mathrm{MSE}(\hat{\beta}_G)\} - \mathrm{tr}\{\sigma^2[X'W_\epsilon^{-1}X]\mathrm{MSE}(\hat{\beta}_G)\}$$

$$= \sigma^2 \mathrm{tr}\{[X'W_\epsilon^{-1}X]S_k H'D_k[D_k^{-1} - \delta\delta']D_k HS_k\}$$

$$= \sigma^2 \mathrm{tr}\{H'D_k[D_k^{-1} - \delta\delta']D_k HS_k\}$$

$$= \sigma^2 \mathrm{tr}\{HS_k H'D_k\} - \sigma^2 \mathrm{tr}\{\delta'D_k HS_k H'D_k \delta\} \tag{6.3.35}$$

where $D_k = [W_\nu + H[X'W_\epsilon^{-1}X]^{-1}H']^{-1}$

$S_k = [X'W_\epsilon^{-1}X]^{-1}$

6.3.3.2 $\Sigma_\epsilon = \sigma^2 W_\epsilon$, W_ϵ , Σ_ν known

Assume that Σ takes the form

$$\Sigma = \begin{bmatrix} \sigma^2 W_\epsilon & 0 \\ 0 & \Sigma_\nu \end{bmatrix} \tag{6.3.36}$$

where σ^2 is unknown and W_ϵ and Σ_ν is pd and known. As Σ_ν is the covariance matrix associated with the prior information, it is feasible that it can be known *a priori*. Then the AGLSE of (6.3.8) becomes

$$\hat{\beta}_G = [\sigma^{-2}X'W_\epsilon^{-1}X + H'\Sigma_\nu^{-1}H]^{-1}[\sigma^{-2}X'W_\epsilon^{-1}Y + H'\Sigma_\nu^{-1}h] \tag{6.3.37}$$

an estimator which is a function of known matrices, and unknown σ^2. We refer to the effect of the unknown term σ^2 by adopting the notation $\hat{\beta}_G(\sigma^2)$ for the estimator in (6.3.37). The properties of $\hat{\beta}_G(\sigma^2)$ are summarised in the following table:

Table 6.2 Properties of $\hat{\beta}_G$, for Σ as in (6.3.36)

$\hat{\beta}_G$	$= [\sigma^{-2}X'W_\epsilon^{-1}X + H'\Sigma_\nu^{-1}H]^{-1}[\sigma^{-2}X'W_\epsilon^{-1}Y + H'\Sigma_\nu^{-1}h]$
	$= \hat{\beta}_G - SH'D[H\hat{\beta}_G - h]$
$\hat{\beta}_G$	$= \{X'W_\epsilon^{-1}X\}^{-1}X'W_\epsilon^{-1}Y = \sigma^{-2}SX'W_\epsilon^{-1}Y$
$\mathrm{Var}(\hat{\beta}_G)$	$[\sigma^{-2}X'W_\epsilon^{-1}X + H'\Sigma_\nu^{-1}H]^{-1}$
S	$\sigma^2[X'W_\epsilon^{-1}X]^{-1}$ and
D	$[\Sigma_\nu + \sigma^2H[X'W_\epsilon^{-1}X]^{-1}H']^{-1}$
$E(\hat{\beta}_G)$	$\beta - SH'D\delta$
bias	$- SH'D\delta$
$\mathrm{MSE}(\hat{\beta}_G)$	$[\sigma^{-2}X'W_\epsilon^{-1}X + H'\Sigma_\nu^{-1}H]^{-1} + SH'D\delta\delta'DHS$
$\mathrm{TMSE}(\hat{\beta}_G)$	$\mathrm{tr}([\sigma^{-2}X'W_\epsilon^{-1}X + H'\Sigma_\nu^{-1}H]^{-1}) + \delta'DHSSH'D\delta$
$\dot{\epsilon}$	$\hat{\dot{\epsilon}} = [I - \dot{M}]\dot{\epsilon}$
$\dot{M}$	$\dot{X}[\sigma^{-2}X'W_\epsilon^{-1}X + H'\Sigma_\nu^{-1}H]^{-1}[\sigma^{-2}X'W_\epsilon^{-1}\ H'\Sigma_\nu^{-1}]$
$V(\hat{\beta}_G) - V\{\hat{\beta}_G\}$	$\sigma^2SH'DHS$

If we consider the random error variables $\hat{\nu}_G$ it can be shown that

$$\begin{aligned}\mathrm{Var}(\hat{\nu}_G) &= \mathrm{Var}(h - H\hat{\beta}_G)\\ &= \mathrm{Var}(H\beta + \nu) + H\mathrm{Var}(\hat{\beta}_G)H'\\ &= \mathrm{Var}(\nu) + \sigma^2H\{X'W_\epsilon^{-1}X\}^{-1})H']\\ &= \Sigma_\nu + HSH' = D^{-1}\end{aligned} \qquad (6.3.38)$$

Thus $\hat{\nu}_G \sim (\delta, D^{-1})$. Furthermore note that $DD^{-1} = I$, which is an idempotent matrix, thus if we assume $\hat{\nu}_G \sim N(\delta, D^{-1})$, then

$$(h - H\hat{\beta}_G)'D(h - H\hat{\beta}_G) \text{ is } \chi^2[r(D), \tfrac{1}{2}\delta'D\delta]$$

Also $r(D) = \ell$, and the MSE-I, II and III criteria are given by (6.3.22), (6.3.25) and (6.3.27).

The only unknown quantity in (6.3.36) is the scalar parameter σ^2, which can be (i) replaced by using the sample information, (ie we are seeking an unbiased estimator of σ^2), or (ii) replaced by a constant f, stochastic or fixed, generating the so called f-class estimators (Theil (1963)).

6.3.3.2.1 Estimating σ^2 using the sample information

The residuals obtained after fitting an estimate of β, are usually used to estimate σ^2. We will distinguish between two ways of estimating σ^2. Note that in the previous section, any estimator of σ^2 was denoted by $\hat{\sigma}^2$. In this section, we will change the notation slightly for clarity:

1. $\hat{\sigma}^2 = (Y - X\hat{\beta}_G)'W_\epsilon^{-1}(Y - X\hat{\beta}_G)/(n-r)$

2. $\hat{\sigma}^2_G = (Y - X\hat{\beta}_G(s^2))'W_\epsilon^{-1}(Y - X\hat{\beta}_G(s^2))/(n - tr(C))$

Method 1: An unbiased estimator of σ^2 is obtained by using the residuals resulting from the OLS (GLS) fit. Note that when $\epsilon \sim (\tau, \sigma^2 W_\epsilon)$, with W_ϵ known, we can orthogonalise the model so that GLSE in the original model is equivalent to OLS in the transformed framework. In (1.4.10) we have shown that $\hat{\sigma}^2$ is a unbiased estimator of σ^2. If this estimator of σ^2 is used to

make $\hat{\beta}_G(\sigma^2)$ operational, we denote the operationalised form by $\hat{\beta}(\hat{\sigma}^2)$.

$$\begin{aligned}\hat{\beta}_G(\hat{\sigma}^2) &= [\hat{\sigma}^{-2}X'W_\epsilon^{-1}X + H'\Sigma_\nu^{-1}H]^{-1}[\hat{\sigma}^{-2}X'W_\epsilon^{-1}Y + H'\Sigma_\nu^{-1}h] \\ &= [X'W_\epsilon^{-1}X + \hat{\sigma}^2H'\Sigma_\nu^{-1}H]^{-1}[X'W_\epsilon^{-1}Y + \hat{\sigma}^2H'\Sigma_\nu^{-1}h] \qquad (6.3.39)\end{aligned}$$

$$\begin{aligned}\hat{\beta}_G(\hat{\sigma}^2) - \beta &= [X'W_\epsilon^{-1}X + \hat{\sigma}^2H'\Sigma_\nu^{-1}H]^{-1}[X'W_\epsilon^{-1}Y + \hat{\sigma}^2H'\Sigma_\nu^{-1}h] \\ &\quad - [X'W_\epsilon^{-1}X + \hat{\sigma}^2H'\Sigma_\nu^{-1}H]^{-1}[X'W_\epsilon^{-1}X\beta + \hat{\sigma}^2H'\Sigma_\nu^{-1}H\beta] \\ &= [X'W_\epsilon^{-1}X + \hat{\sigma}^2H'\Sigma_\nu^{-1}H]^{-1}X'W_\epsilon^{-1}\{Y - X\beta\} \\ &\quad + \hat{\sigma}^2[X'W_\epsilon^{-1}X + \hat{\sigma}^2H'\Sigma_\nu^{-1}H]^{-1}H'\Sigma_\nu^{-1}\{h - H\beta\} \\ &= [X'W_\epsilon^{-1}X + \hat{\sigma}^2H'\Sigma_\nu^{-1}H]^{-1}X'W_\epsilon^{-1}\epsilon \\ &\quad + \hat{\sigma}^2[X'W_\epsilon^{-1}X + \hat{\sigma}^2H'\Sigma_\nu^{-1}H]^{-1}H'\Sigma_\nu^{-1}\nu\end{aligned}$$

Assuming that $\epsilon \sim N(\tau, \sigma^2W_\epsilon)$, then by Theorem 3 of Searle (1971) $\hat{\sigma}^2 = \epsilon'\{(I - M')W_\epsilon^{-1}\}\epsilon$ and $X'W_\epsilon^{-1}\epsilon$ are distributed independently because

$$\begin{aligned}[X'W_\epsilon^{-1}][\sigma^2W_\epsilon]\{(I - M')W_\epsilon^{-1}\} &= \sigma^2[X'(I - M')W_\epsilon^{-1}] \\ &= \sigma^2[X' - X'W_\epsilon^{-1}X\{X'W_\epsilon^{-1}X\}^{-1}X']W_\epsilon^{-1} \\ &= 0.\end{aligned}$$

Also $\hat{\sigma}^2$ and ν are independent because of assumption (6.3.4).

Thus if $E[X'W_\epsilon^{-1}X + \hat{\sigma}^2H'\Sigma_\nu^{-1}H]^{-1}$ exists

$$\begin{aligned}E[\hat{\beta}_G(\hat{\sigma}^2) - \beta] &= E[X'W_\epsilon^{-1}X + \hat{\sigma}^2H'\Sigma_\nu^{-1}H]^{-1}E[X'W_\epsilon^{-1}\epsilon] \\ &\quad + E[\hat{\sigma}^2[X'W_\epsilon^{-1}X + \hat{\sigma}^2H'\Sigma_\nu^{-1}H]^{-1}E[H'\Sigma_\nu^{-1}\nu] \\ &= E[X'W_\epsilon^{-1}X + \hat{\sigma}^2H'\Sigma_\nu^{-1}H]^{-1}[X'W_\epsilon^{-1}\tau] \\ &\quad + E[\hat{\sigma}^2[X'W_\epsilon^{-1}X + \hat{\sigma}^2H'\Sigma_\nu^{-1}H]^{-1}[H'\Sigma_\nu^{-1}\delta] \\ &= 0 \text{ (when } X'W_\epsilon^{-1}\tau = 0 \text{ and } H'\Sigma_\nu^{-1}\delta = 0)\end{aligned}$$

Thus $\hat{\beta}_G(\hat{\sigma}^2)$ is unbiased whenever $X'W_\epsilon^{-1}\tau = 0$ and $H'\Sigma_\nu^{-1}\delta = 0$.

Another way to show the unbiasedness of $\hat{\beta}_G(\hat{\sigma}^2)$ is the procedure of Kakwani (1967). Similarly the procedure used by Theil (1963) shows that

$$\hat{\beta}_G(\hat{\sigma}^2) = \hat{\beta}_G(\sigma^2) + O(n^{-1}). \qquad (6.3.40)$$

If $\hat{\sigma}^2$ is a random variable and if the difference

$$\hat{\sigma}^{-2} - \sigma^{-2} \text{ is } O(n^{-\frac{1}{2}}) \text{ in probability,}$$

Theil then shows that $\hat{\beta}_G(\hat{\sigma}^2)$ is asymptotically unbiased (if $\hat{\beta}_G(\sigma^2)$ is unbiased) and that the bias is $O(n^{-1})$. Furthermore $\hat{\beta}_G(\sigma^2)$ and $\hat{\beta}_G(\hat{\sigma}^2)$ have asymptotically the same covariance matrix.

Nagar and Kakwani (1964) show that the bias (where the authors assume τ and $\delta = 0$) to be zero if the disturbance terms (ϵ,ν) are symmetrically distributed about zero, even if they are not normally distributed.

Vinod and Ullah (1981) give the conditional (on $\hat{\sigma}^2$) covariance matrix of $\hat{\beta}_G(\hat{\sigma}^2)$ and comment that the mathematical expressions for the unconditional covariance matrix, as given by Swamy and Mehta (1976) and Charette (1978), are too complicated to draw useful conclusions.

Given this asymptotic convergence of $\hat{\beta}_G(\hat{\sigma}^2)$ to $\hat{\beta}_G(\sigma^2)$, and a sufficient sample size, we expect the properties discussed in 6.3.3 (ie Var dif, MSE-I, MSE-II and MSE-III between this estimator and OLS) to hold in general.

Method 2: We use the residuals obtained from the restricted estimator $\hat{\beta}_G(\hat{\sigma}^2)$. Thus

$$\hat{\sigma}_G^2 = (Y - X\hat{\beta}_G(\hat{\sigma}^2))'W_\epsilon^{-1}(Y - X\hat{\beta}_G(\hat{\sigma}^2))/(n - tr(C))$$

where (6.3.41)

$$C = \hat{\sigma}^{-2}S^{-1}[\hat{\sigma}^{-2}S^{-1} + H'\Sigma_\nu^{-1}H]^{-1}$$

The proof of result (6.3.41) can be found in the Appendix of Theil (1963). Theil (1963) shows that the expectation of (6.3.41) is

$$E[\hat{\sigma}_G^2] = \sigma^2 + O(n^{-\frac{1}{2}}) \qquad (6.3.42)$$

Using (6.3.41) we obtain the operational estimator $\hat{\beta}_G(\hat{\sigma}_G^2)$. Optimizing this estimator would involve an iteration process, i.e first estimate $\hat{\sigma}^2$ (or any starting value) by least squares, then use this estimator to compute $\hat{\beta}_G(\hat{\sigma}^2)$, then compute $\hat{\sigma}_G^2$, then fit $\hat{\beta}_G(\hat{\sigma}_G^2)$ and continue this process until the required convergence criterion is satisfied.

Because $\hat{\sigma}_G^2$ is based on prior information, and because $\hat{\beta}_G(\hat{\sigma}_G^2)$ is obtained via a convergence criterion, one might think that $\hat{\sigma}_G^2$ is asymptotically more efficient estimator than $\hat{\sigma}^2$ for σ^2. This view however is not valid, as described by Theil (1963).

6.3.3.2.2 Replace σ^2 by f

If the unknown scalar quantity σ^{-2} in (6.3.37) is replaced by a constant f, where f can be stochastic or non-stochastic, we obtain the family of f-class

estimators (Theil (1963)). Thus the estimator $\hat{\beta}_G(f)$ is defined as

$$\hat{\beta}_G(f) = [fX'W_\epsilon^{-1}X + H'\Sigma_\nu^{-1}H]^{-1}[fX'W_\epsilon^{-1}Y + H'\Sigma_\nu^{-1}h] \qquad (6.3.43)$$

It is clear from (6.3.43) that if f increases indefinitely, the estimator $\hat{\beta}_G(f)$ converges to the GLSE. On the other hand if f approaches zero ($\sigma^2 \to \infty$) the prior information become very important, and the limit of $\hat{\beta}_G(f)$ is $[H'\Sigma_\nu^{-1}H]^{-1}H'\Sigma_\nu^{-1}h$.

A particular member of the f-class was considered in section 6.3.3.2.1, ie

$$f = 1/\hat{\sigma}^2 \qquad (6.3.44)$$

In that case f is based on the sample and this particular estimator and its properties were discussed in the previous section and in Theil (1963). It is in fact the only stochastic estimator that we will consider in the family of f-class estimators.

The case when f is a fixed constant (ie non-stochastic, $f \geq 0$) is discussed in detail by Toutenberg (1982). If we let $f = c$, $c \geq 0$, then $\hat{\beta}_G(c)$ is

$$\begin{aligned}\hat{\beta}_G(c) &= [cX'W_\epsilon^{-1}X + H'\Sigma_\nu^{-1}H]^{-1}[cX'W_\epsilon^{-1}Y + H'\Sigma_\nu^{-1}h] \\ &= \beta + [cX'W_\epsilon^{-1}X + H'\Sigma_\nu^{-1}H]^{-1}[cX'W_\epsilon^{-1}\epsilon + H'\Sigma_\nu^{-1}\nu] \qquad (6.3.45)\end{aligned}$$

Thus we define a family $F_c = \{\hat{\beta}_G(c)\}$, of estimators that are operational. The members of the family is indexed by the value of the chosen scalar c.

The properties of $\hat{\beta}_G(c)$ are summarised as follows:

Table 6.3 Properties of $\hat{\beta}_G(c)$, Σ in the form of (6.3.36)

$\hat{\beta}_G(c)$	$= [cX'W_\epsilon^{-1}X + H'\Sigma_\nu^{-1}H]^{-1}[cX'W_\epsilon^{-1}Y + H'\Sigma_\nu^{-1}h]$
	$= \beta + [cX'W_\epsilon^{-1}X + H'\Sigma_\nu^{-1}H]^{-1}[cX'W_\epsilon^{-1}\epsilon + H'\Sigma_\nu^{-1}\nu]$
$\hat{\beta}_G$	$= \{X'W_\epsilon^{-1}X\}^{-1}X'W_\epsilon^{-1}Y$
$Var[\hat{\beta}_G(c)]$	$= [cX'W_\epsilon^{-1}X+H'\Sigma_\nu^{-1}H]^{-1}[c^2\sigma^2X'W_\epsilon^{-1}X + H'\Sigma_\nu^{-1}H][cX'W_\epsilon^{-1}X+H'\Sigma_\nu^{-1}H]^{-1}$
	$= M_c^{-1}[c^2\sigma^2X'W_\epsilon^{-1}X + H'\Sigma_\nu^{-1}H]M_c^{-1}$
	$= c\sigma^2(c\sigma^2 - 1)M_c^{-1}S^{-1}M_c^{-1} + M_c^{-1}$
M_c	$[cX'W_\epsilon^{-1}X + H'\Sigma_\nu^{-1}H] = [c\sigma^2S^{-1} + H'\Sigma_\nu^{-1}H]$
S	$\sigma^2[X'W_\epsilon^{-1}X]^{-1}$ and
D	$[\Sigma_\nu + \sigma^2H[X'W_\epsilon^{-1}X]^{-1}H']^{-1}$
$E(\hat{\beta}_G(c)))$	$\beta + M_c^{-1}[cX'W_\epsilon^{-1}\tau + H'\Sigma_\nu^{-1}\delta]$
Bias	$M_c^{-1}[cX'W_\epsilon^{-1}\tau + H'\Sigma_\nu^{-1}\delta]$
$MSE[\hat{\beta}_G(c)]$	$M_c^{-1}[c^2\sigma^2X'W_\epsilon^{-1}X + H'\Sigma_\nu^{-1}H]M_c^{-1} +$ $M_c^{-1}[c^2X'W_\epsilon^{-1}\tau\tau'W_\epsilon^{-1}X + H'\Sigma_\nu^{-1}\delta\delta'\Sigma_\nu^{-1}H]M_c^{-1}$
$TMSE[\hat{\beta}_G(c)]$	$tr\{M_c^{-1}[c^2\sigma^2X'W_\epsilon^{-1}X + H'\Sigma_\nu^{-1}H]M_c^{-1} +$ $M_c^{-1}[c^2X'W_\epsilon^{-1}\tau\tau'W_\epsilon^{-1}X + H'\Sigma_\nu^{-1}\delta\delta'\Sigma_\nu^{-1}H]M_c^{-1}\}$
$\hat{\epsilon}(c)$	$= [I - \dot{M}]\hat{\epsilon}$
$\dot{M}$	$\dot{X}M_c^{-1}[cX'W_\epsilon^{-1}\ H'\Sigma_\nu^{-1}]$

Note that $Var[\hat{\beta}_G(c)] \geq Var[\hat{\beta}_G(\sigma^2)]$ as $\hat{\beta}_G(\sigma^2)$ is the BLUE.

To obtain an optimal choice of c, the quantity that we want to minimize with respect to c is $MSE[\hat{\beta}_G(c)]$ or $TMSE[\hat{\beta}_G(c)]$. For simplicity we will assume

$X'W_\epsilon^{-1}\tau = 0$ and $H'\Sigma_\nu^{-1}\delta = 0$, thus we minimize $M_c^{-1}[c^2\sigma^2X'W_\epsilon^{-1}X+H'\Sigma_\nu^{-1}H]M_c^{-1}$ using

$$\begin{aligned}\frac{\partial}{\partial c}(\mathrm{TMSE}[\hat{\beta}_G(c)]) &= \frac{\partial}{\partial c}(\mathrm{tr}\{c\sigma^2(c\sigma^2 - 1)M_c^{-1}S^{-1}M_c^{-1}+M_c^{-1}+ \\ &\qquad M_c^{-1}[c^2X'W_\epsilon^{-1}\tau\tau'W_\epsilon^{-1}X + H'\Sigma_\nu^{-1}\delta\delta'\Sigma_\nu^{-1}H]M_c^{-1}\}) \\ &= \frac{\partial}{\partial c}(\mathrm{tr}\{c\sigma^2(c\sigma^2 - 1)M_c^{-1}S^{-1}M_c^{-1}+M_c^{-1}) \qquad (6.3.46)\end{aligned}$$

Now from theorem A.53, B.1.10 and B.3.2 of Toutenberg (1982), we have that

$$\begin{aligned}\frac{\partial}{\partial c}(\mathrm{tr}\{M_c^{-1}\}) &= \mathrm{tr}(\frac{\partial}{\partial c}M_c^{-1}) \\ &= \mathrm{tr}(- M_c^{-1}[\frac{\partial}{\partial c}M_c]M_c^{-1}) \\ &= - \mathrm{tr}(M_c^{-1}[\frac{\partial}{\partial c}[c\sigma^2S^{-1} + H'\Sigma_\nu^{-1}H]M_c^{-1}) \\ &= - \sigma^2\mathrm{tr}(M_c^{-1}S^{-1}M_c^{-1}) \qquad (6.3.47)\end{aligned}$$

and

$$\begin{aligned}&\frac{\partial}{\partial c}(\mathrm{tr}\{c\sigma^2(c\sigma^2 - 1)M_c^{-1}S^{-1}M_c^{-1}\}) \\ &\quad = \mathrm{tr}(\frac{\partial}{\partial c}\{(c^2\sigma^4 - c\sigma^2)\}M_c^{-1}S^{-1}M_c^{-1} + (c^2\sigma^4 - c\sigma^2)\frac{\partial}{\partial c}[M_c^{-1}S^{-1}M_c^{-1}]) \\ &\quad = \mathrm{tr}(\{2c\sigma^4-\sigma^2\}M_c^{-1}S^{-1}M_c^{-1} + (c^2\sigma^4-c\sigma^2)([\frac{\partial}{\partial c}M_c^{-1}]S^{-1}M_c^{-1}+ M_c^{-1}S^{-1}\frac{\partial}{\partial c}M_c^{-1})) \\ &\quad = \mathrm{tr}(\{2c\sigma^4 - \sigma^2\}M_c^{-1}S^{-1}M_c^{-1}- 2\sigma^2(c^2\sigma^4-c\sigma^2)M_c^{-1}S^{-1}M_c^{-1}S^{-1}M_c^{-1}) \\ &\quad = \sigma^2\mathrm{tr}(\{2c\sigma^2 - 1\}M_c^{-1}S^{-1}M_c^{-1}- 2(c^2\sigma^4-c\sigma^2)M_c^{-1}S^{-1}M_c^{-1}S^{-1}M_c^{-1}) \qquad (6.3.48)\end{aligned}$$

Thus

$$\begin{aligned}\frac{\partial}{\partial c}(\mathrm{TMSE}[\hat{\beta}_G(c)]) &= 2\sigma^2\mathrm{tr}(\{c\sigma^2 - 1\}M_c^{-1}S^{-1}M_c^{-1} - (c^2\sigma^4 - c\sigma^2)M_c^{-1}S^{-1}M_c^{-1}S^{-1}M_c^{-1}) \\ &= 2\sigma^2\mathrm{tr}(\{c\sigma^2 - 1\}M_c^{-1}S^{-1}M_c^{-1} - c\sigma^2(c\sigma^2 - 1)M_c^{-1}S^{-1}M_c^{-1}S^{-1}M_c^{-1}) \\ &= 2\sigma^2\mathrm{tr}(\{c\sigma^2 - 1\}M_c^{-1}S^{-1}[S - c\sigma^2M_c^{-1}]S^{-1}M_c^{-1}) \qquad (6.3.49)\end{aligned}$$

In table 6.3 $\mathrm{Var}[\hat{\beta}_G(c)]$ is given as

$$\mathrm{Var}[\hat{\beta}_G(c)] = c\sigma^2(c\sigma^2 - 1)M_c^{-1}S^{-1}M_c^{-1} + M_c^{-1}$$
$$\geq 0 \quad \text{(ie psd)}$$

as M_c is psd, and $c \geq 0$, we can conclude that $c\sigma^2 - 1 \geq 0$. Furthermore

$$[S - c\sigma^2 M_c^{-1}] = \sigma^2[X'W_\epsilon^{-1}X]^{-1} - c\sigma^2[cX'W_\epsilon^{-1}X + H'\Sigma_\nu^{-1}H]^{-1}$$
$$= \sigma^2\{[X'W_\epsilon^{-1}X]^{-1} - [X'W_\epsilon^{-1}X + c^{-1}H'\Sigma_\nu^{-1}H]^{-1}\} \qquad (6.3.50)$$
$$\geq 0 \quad \text{(psd (Theorem A.12, Toutenberg (1982)))}.$$

Thus $M_c^{-1}S^{-1}[S - c\sigma^2 M_c^{-1}]S^{-1}M_c^{-1}$ is psd and symmetric and

$$\mathrm{tr}\{M_c^{-1}S^{-1}[S - c\sigma^2 M_c^{-1}]S^{-1}M_c^{-1}\} = a \text{ (say)} \geq 0$$

From $\{c\sigma^2 - 1\}a \geq 0$ (6.3.51)

we note (6.3.49) is at a minimum when $c = \sigma^{-2}$, is monotonically increasing in c when $c > \sigma^{-2}$ and monotonically decreasing when $c < \sigma^{-2}$. Toutenberg (1982) then assumes that prior information on σ^2 is available in the form of bounds

$$\sigma_L^2 < \sigma^2 < \sigma_U^2$$

where σ_L^2 and σ_U^2 is known *a priori*, and shows that an estimator with lower TMSE that all other estimators is obtained when

$$c^*(\lambda) = \sigma_U^{-2} + \lambda(\sigma_L^{-2} - \sigma_U^{-2}), \quad \text{and } (0 < \lambda < 1) \qquad (6.3.52)$$

6.3.3.3 Heteroscedasticity

Assume that Σ, takes on the following form

$$\Sigma = \text{diag}[\sigma_1^2, \sigma_2^2, \ldots, \sigma_{n+\ell}^2] \qquad (6.3.53)$$

Some of the σ_i^2 may be equal, and the structures discussed in 6.3.3.1 and 6.3.3.2 can be considered as special cases of (6.3.53). Dispersion matrices like (6.3.53) are also described by the term heteroscedasticity. The topic of heteroscedasticity is wide and considerable literature on this subject is available *inter alia* in Putter (1967), and several texts on Econometrics (eg Judge *et al.* (1980), Schmidt (1967) and Theil (1971)).

In (6.3.8) the AGLSE was defined and its properties discussed in section 6.3.1. It was pointed out that (6.3.8) was non-operational and that an operational estimator is obtained when a suitable structure and estimate of Σ is available. In estimating Σ, we have to estimate $(n+\ell)$ variance terms. Such a process is impossible, and it would exhaust all the degrees of freedom. In applied work it is common to restrict the σ_i^2 still further, see Judge *et al.* (1980, Chapter 4).

Before considering any restrictions on the σ_i^2's or even before accepting (6.3.53) it is recommended to explore for evidence heteroscedasticity in the model. In section 6.3.3.3.1 we will consider tests for heteroscedasticity and in section 6.3.3.3.2 ways of estimating Σ.

6.3.3.3.1 Test for Heteroscedasticity

Various methods are available to detect heteroscedasticity. Some of these methods will be briefly mentioned here, with possible references for implementing them.

1. The most common method of detecting heteroscedasticity is by plotting the fitted residuals against the independent variables or the fitted value $\hat{Y}$. When non-constancy of the error variance occurs the residual

plot is usually of the trapezoidal type (eg the variance increases as the independent variable increases). Plots like these are discussed in Neter and Wasserman (1974).

2. **Glejser's test:** Glejser (1969) suggests regressing the absolute value of the fitted residuals (we are still thinking of the augmented model, thus $n+\ell$ residuals) on a number of alternative functions of one of the regressor variables, ie

$$|\hat{\epsilon}| = Z\alpha + \xi$$

where $|\hat{\epsilon}|$ is a $(n+\ell)\times 1$ vector consisting of the absolute values of the residuals. Thus

$$|\hat{\epsilon}| = [|\hat{\epsilon}_1|, |\hat{\epsilon}_2|, \ldots, |\hat{\epsilon}_{n+\ell}|]$$

α is a$\times$1 vector of unknowns and
Z is a $(n+\ell)\times$a matrix of non-stochastic variables that may be identical to or functions of the X matrix. Note that the first column of Z is a column of ones.

The null hypothesis

$$H_0: \alpha_2 = \alpha_3 = \ldots = \alpha_a = 0$$

in the Glejser model is tested by usual methods. If the test rejects the null hypothesis, heteroscedasticity is presumed not to be a serious problem.

Judge *et al.* (1980) viewed Glejser's: Z as a matrix of alternative functions of one of the independent regressors whereas Vinod and Ullah (1981) view the Z matrix as a known function of one or more regressors. Glejser's intention was to form functions of one of the regressors, usually polynomial.

One problem of implementing this test is the question of how one defines Z. Trying a number of functions of one or more independent variables involves preliminary test considerations and can lead to results that are influenced by the user. It is unclear if arbitrary

choices of Z, will lead to the same conclusions under the null hypothesis.

3. **Bartlett's test:** If the $n+\ell$ residuals can be grouped into t subgroups, the Bartlett (1937) test is designed to test if heteroscedasticity exists between the t groups ie the likelihood ratio test is testing the null hypothesis that $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_t^2$ (where the subscript i here indicates the i-th group and not the individual variance of the i-th diagonal element of Σ). For the implementation of this test see Bartlett (1937) or Judge *et al.* (1980).

   The Bartlett's test is an old well-established test and its usefulness depends on whether the residuals can be grouped into sets with constant variances. Anderson and McLean (1974) warned that Bartlett's test is sensitive under non-normality, and when the tails of the distribution are too long it tends to achieve significance too often.

4. **Burr-Foster Q-test of Homogeneity:** This test was described by Burr and Foster (1972). It is published by Anderson and McLean (1974) and is easy to perform. However it does not have the sensitivity to normality departures encountered with Bartlett's test. Furthermore a zero sample variance (in one of the groups) does not disrupt this test as in the case of Bartlett's test.

   As in Bartlett's test the residuals are divided into t groups and the test statistic, q is computed:

$$q = (s_1^4 + s_2^4 + \ldots + s_t^4) / (s_1^2 + s_2^2 + \ldots + s_t^2)^2$$

   where s_i^2 is the estimated variance of the i-th group. Large values of q lead to rejection of the hypothesis of equal population variances, critical values of q are given in Anderson and McLean (1974). For implementation of this test, when the number of residuals in the t groups are unequal, see Anderson and McLean (1974).

5. **Breusch-Pagan test:** Breusch and Pagan (1979) regress squared scaled residuals on the specified function of Z_j' (the j-th row of Z)

$$\hat{\epsilon}_j^2/\hat{\sigma}^2 = Z_j' a.$$

The authors show that one half the regression sum of squares is distributed asymptotically as χ^2 with a degrees of freedom. The $\hat{\epsilon}_j$ are the OLS residuals from the augmented model and $\hat{\sigma}^2$ the usual sample variance of these residuals. The null hypothesis is that all a coefficients are zero.

6. **Goldfeld-Quandt Test:** Goldfeld and Quandt (1965, 1972) ranked the residuals in order of increased variance. Then the s central residuals are omitted and two separate regressions are run on the first and the last residuals. Calculate $R = S_2/S_1$ where S_1 and S_2 is the residual sums of squares from the first and second regressions, respectively. The statistic R has the F distribution with $[\{(n+\ell) - s - r\}/2\}$ and $[\{(n+\ell) - s - r\}/2]$ degrees of freedom.

Theil (1971) used a test similar to that of Goldfeld and Quandt, partitioning the Y vector and the X matrix into two equal sets (eg $Y' = [Y_A' \; Y_B']$) then computing OLS residuals for both sets (A and B) and the ratio of the squared sum of residuals for sets A and B. The ratio is distributed as an F with $(n + \ell - r)/2$ and $(n + \ell - r)/2$ degrees of freedom. Note that Theil only omits one central observation if the number of observations is odd, and does not use any ordering of the residuals.

The Goldfeld-Quandt test depends on one's ability to rank the observations according to increasing variance. This writer would prefer to test for heteroscedasticity before having the burden of estimating the $(n+\ell)$ variances. We hope that by testing for heteroscedasticity, the unknown $(n+\ell)$ diagonal elements of Σ can be plausibly reduced, to $\sigma^2 I$.

7. **White test:** White (1980) consider the following artificial regression

$$\hat{\epsilon}_i^2 = a_0 + \sum_{j=1}^{r} \sum_{k=j}^{r} a_{jk} X_{ij} X_{ik} , \quad \text{for } i = 1,2,\ldots,n \qquad (6.3.54)$$

$\hat{\epsilon}_i$ is the i-th residual obtain from OLS and X_{ij} is (i,j)th element of the X matrix (in position i-th row and the j-th column). The a's are $\frac{1}{2}r(r+1) + 1$ parameters that are to be estimated by OLS. White then tests the joint null hypothesis

$$a_{11}=a_{12}=\ldots=a_{1r}=a_{22}=\ldots=a_{(r-1)r}=a_{rr}=0$$

Then under 6 specific assumptions in White (1980) and if, in addition, ϵ_i is independent of the i-th row of X, $E[\epsilon_i^2] = \sigma^2$ and $E[\epsilon_i^4] = \mu_4$ (ie ϵ_i are homokurtic for all i), then

$$nR^2 \sim \chi^2 \text{ with } r(r+1)/2 \text{ degrees of freedom}$$

where R^2 is the (constant-adjusted) squared multiple correlation coefficient from the regression (6.3.54). For a general discussion of this test and comparison with others available in the literature see White (1980). If the null hypothesis is accepted, one can proceed by using the usual OLS variance of the β's for hypothesis testing of the β. If the null hypothesis is rejected one can calculate White's heteroscedasticity-consistent-covariance matrix, which is given in the estimation section below.

There are numerous other tests available to detect heteroscedasticity. Depending on the circumstances the user must be careful in specifying the model, and if, heteroscedasticity is expected, or obvious from residual plots, it is recommended that a general test is used, like the test of White, using no preset structure. There is no easy-and-best test and the issues are open for further research.

6.3.3.3.2 Estimation of Σ

After detecting heteroscedasticity, we need to specify a model, and then to estimate Σ. When an *a priori* structure of Σ is available, the *a priori* specification will be assumed true, and tested. In this section we will concentrate on estimating Σ in general form, ie we want an estimate of the n diagonal elements of Σ without assuming any restriction on Σ.

1. White (1980) adopts the LRM of 1.1 with the following four assumptions:

(a) (X_i, ϵ_i) is a sequence of independent not (necessarily) identically distributed (i.n.i.d) random vectors, such that $E(X_i' \epsilon_i) = 0$.

There exist positive finite constants ξ and Ψ such that, for all i,

(b) $E(|\epsilon_i^2|^{1+\xi}) < \Psi$ and $E(|X_{ij}X_{ik}|^{1+\xi}) < \Psi$ $(j,k = 1,\ldots,r)$; $\sum_{i=1}^{n} E(X_i'X_i)/n$ is non-singular for (all) n sufficiently large.

(c) $E(|\epsilon_i^2 X_{ij} X_{ik}|^{1+\xi}) < \Psi$ $(j,k = 1,\ldots,r)$; V_n (defined in (6.3.55)) is non-singular for n sufficiently large.

(d) $E(|X_{ij}^2 X_{ik} X_{ih}|^{1+\xi}) < \Psi$ $(j,k,h = 1,\ldots,r)$.

White only considers the OLS estimator $\hat{\beta} = (X'X)^{-1}X'Y$, with average covariance matrix

$$V_n = (X'\Sigma X)/n = \sum_{i=1}^{n} E(\epsilon_i^2 X_i' X_i)/n \qquad (6.3.55)$$

Note that V_n is a function of S in our notation. The estimators for the diagonal elements of Σ, do not depend on a formal model of the structure of the heteroscedasticity. What is actually required is to estimate an average of expectations ie V_n. Under the conditions stated by White a consistent

estimator for $\sum_{i=1}^{n} E(\epsilon_i^2 X_i' X_i)/n$ is $\sum_{i=1}^{n} \epsilon_i^2 X_i' X_i/n$. The error term ϵ_i^2 is not observable, but can be estimated by the OLS residuals:

$$\hat{V}_n = n^{-1} \sum_{i=1}^{n} \hat{\epsilon}_i^2 X_i' X_i \qquad (6.3.56)$$

thus the i-th diagonal element of Σ, σ_i^2 is replaced with $\hat{\epsilon}_i^2$.

White (Theorem 1, 1980) assumes (a) through to (d) and then proves that

(i) $|\hat{V}_n - V_n| \xrightarrow{a.s} 0$

(ii) $|(X'X/n)^{-1}\hat{V}_n(X'X/n)^{-1} - M_n^{-1}V_nM_n^{-1}| \xrightarrow{a.s} 0$

where $M_n^{-1} = n^{-1} \sum_{i=1}^{n} E(X_i' X_i)$ (where X must be an observable matrix and is not necessarily fixed)

(iii) Given the general hypotheses

$$H_0: C\beta = m \quad \text{vs} \quad H_1: C\beta \neq m$$

where C is any q×r full row rank matrix and m is a q×1 vector of constants, then

$$n(C\hat{\beta} - m)'[C(X'X/n)^{-1}\hat{V}_n(X'X/n)^{-1}C']^{-1}(C\hat{\beta} - m) \dot{\sim} \chi_q^2$$

To summarise: $\hat{V}_n$ consistently estimates V_n, the heteroscedasticity-consistent covariance matrix and that to test linear hypotheses in the usual way give asymptotically correct results.

As pointed out earlier, White only considers the OLS estimator, without taking into account the covariance structure, and technically the GLS

estimator is the appropriate estimator under the heteroscedasticity model. It is unclear if this estimator of White for diagonal Σ, namely the square of the OLS residuals, would be a good estimator for Σ. Furthermore in some of the assumptions made by White, he defines or rather uses the term non-singular and determinant $> \xi > 0$ as being synonymous. The determinant is excessively sensitive to scaling, and a small determinant may imply little or nothing about the invertibility of a matrix (Stewart (1987), or Thiart (1990)).

2. Rao (1970) consider the LRM with $\epsilon \sim (0,\Sigma)$ and for the estimation of the heteroscedastic variances introduces a method known as MINQUE (MInimum, Norm Quadratic Unbiased Estimation). In order to define the MINQUE, we need to define the following matrices:

(a) the projection matrix $P = [I - X(X'X)^{-}X']$, with (i,j)th element p_{ij}, where $(X'X)^{-}$ is the generalised inverse, and when X is full rank, $(X'X)^{-}$ is the same as $(X'X)^{-1}$.

(b) the vector v is the vector of squared residuals $\hat{\epsilon}_i^2$ where $\hat{\epsilon} = PY$, σ is the vector containing the diagonal elements of Σ, and

(c) the matrix F is the Hadamard product P*P, ie F consists of all the element of P squared, $F = (p_{ij}^2)$.

The MINQUE is defined (Rao (Lemma 5, 1970)

> Let $\sigma_1^2,\ldots,\sigma_n^2$ be all different, then the MINQUE of $\sigma_1^2,\ldots,\sigma_n^2$ are solutions of the equation $F\sigma = v$ provided F is non-singular.

Thus the MINQUE is a linear combination of the squares of residuals. When all the σ_i^2 are not different, one can use the above mentioned method, or the collapsed procedure given as lemma 6 of Rao (1970).

Conditions for the non-singularity of F are given by Hartley *et al.* (1969) and by Rao (lemma 7 and 8, 1970).

3. Fuller and Rao (1978) assume that $\Sigma = \text{Diag}[\sigma_1^2 I_{n_1}, \ldots, \sigma_k^2 I_{n_k}]$; the observations fall into k groups with constant error variance within a group. In estimation of the GLE, the OLS estimator is firstly computed and the OLS residuals are used to estimate the covariance matrix. This estimate of Σ is then used to calculate an operational GLE. Fuller and Rao refer to the operational GLE as the two step weighted least squares (WLS) estimator of β:

$$\hat{\Sigma} = \text{Diag}[\hat{\sigma}_1^2 I_{n_1}, \ldots, \hat{\sigma}_k^2 I_{n_k}], \text{ where}$$

$$\hat{\sigma}_i^2 = n_i^{-1} \sum_{j=1}^{n_j} \hat{\epsilon}_{ij}^2 \tag{6.3.57}$$

for $\hat{\epsilon}' = [\hat{\epsilon}_{11}, \ldots, \hat{\epsilon}_{1n_1}; \ldots, \hat{\epsilon}_{k1}, \ldots, \hat{\epsilon}_{kn_k}]$

and the operational GLS estimator is

$$\hat{\beta}_G = [X'\hat{\Sigma}^{-1}X]^{-1}X'\hat{\Sigma}^{-1}Y \tag{6.3.58}$$

Fuller and Rao (1978) then introduce the weighted GLS estimator, where it seems that the weights are corrections for the number of observations falling into the k groups. Thus

$$\hat{\beta}_{G_w} = [X'\hat{\Sigma}^{-1}WX]^{-1}X'\hat{\Sigma}^{-1}WY \tag{6.3.59}$$

where $W = \text{Diag}[w_1 I_{n_1}; \ldots; w_k I_{n_k}]$ and $w_i = g(n_i)$. When there are an equal number of observations in the k-groups then $w_i = 1$, and $\hat{\beta}_{G_w}$ reduces to $\hat{\beta}_G$. The main result of Fuller and Rao (1978) is the asymptotic distribution of $\hat{\beta}_{G_w}$:

Let ϵ_i's be independently and normally distributed, with mean zero and variance σ_i^2, and assuming

(a) the sequences $\{\hat{\sigma}_i^2\}$ and $\{n_i\}$ satisfy $0 < \sigma_L^2 \leq \sigma_i^2 \leq \sigma_U^2 < \infty$ and $3 \leq n_i < n^* < \infty$ for all i

(b) the rows of X, $(X_{ij1},\ldots,X_{ijr})$, form a fixed sequence with $\sum_{t=1}^{r} X_{ijt}^2 < \partial < \infty$ for all (i,j)

(c) the limits (as $k \to \infty$) of the matrices $X'X/n$, $X'\Sigma X/n$, $X'WX/n$, $X'WG\Sigma^{-1}X/n$, $X'W\Sigma^{-1}LWX/n$ and $X'\Sigma_w^{-1}X/n$ exist and are positive definite where

$$G = \text{Diag}[(n_1-2)^{-1}I_{n_1} ;\ldots; (n_k-2)^{-1}I_{n_k}]$$

$$L = \text{Diag}[n_1(n_1-2)^{-1}I_{n_1} ;\ldots; n_k(n_k-2)^{-1}I_{n_k}]$$

$$\Sigma_w^{-1} = \text{Diag}[n_1w_1(n_1-2)^{-1}\sigma_1^{-2}I_{n_1} ;\ldots; n_kw_k(n_k-2)^{-1}\sigma_k^{-2}I_{n_k}];$$

then

$$\sqrt{n}(\hat{\beta}_{G_w} - \beta) \dot{\sim} N(0, H)$$

where $$H = \lim_{k\to\infty} n(X'\Sigma_w^{-1}X)^{-1}D(X'\Sigma_w^{-1}X)^{-1}$$

$$D = X'W\Sigma^{-1}LWX + 2(M + M') + 4(X'WG\Sigma^{-1}X)(X'X)^{-1}(X'\Sigma X)(X'X)^{-1}(X'WG\Sigma^{-1}X)$$

and $M = X'WG\Sigma^{-1}X(X'X)^{-1}X'WX)$.

The asymptotic covariance matrix of $\hat{\beta}_{G_w}$ simplifies when $n_i = c \geq 3$:

$$V[\hat{\beta}_{G_w}] = V[\hat{\beta}_G]$$
$$= (1 + 2c^{-1} - 8c^{-2})(X'\Sigma^{-1}X)^{-1} + 4c^{-2}(X'X)^{-1}(X'\Sigma^{-1}X)(X'X)^{-1}$$

4. **Van der Genugten** (1993) assumed a LRM of which the errors are independent and symmetrically distributed. The variance matrix of the errors is not specified, and no assumptions are made about the variances.

The author then investigates the iterated weighted least square (IWLS) estimator, $\hat{\beta}_G(q+1)$, the IWLS estimator of β at step (q+1), as

$$\hat{\beta}_G(q+1) = \{\sum_t^r X_t X_t' f(\hat{\sigma}_t^2(q)\}^{-1} \sum_t^r X_t Y_t f(\hat{\sigma}_t^2(q)) \quad \text{for } q = 0,1,\ldots \tag{6.3.60}$$

Thus, in terms of our notation, Van der Genugten uses the iterated operational AGLSE at the q+1 iteration.

The starting value of this procedure is OLS (the 0-th iteration). The estimator of the j-th variance term obtain in the q-iteration is

$$\hat{\sigma}_j^2(q) = \sum_{i \epsilon A} w_j \hat{\epsilon}_{j+i}^2(q) \tag{6.3.61}$$

where $\hat{\epsilon}_{j+i}^2(q)$ is the (j+i)-th residual obtained in the q-th iteration and A is a set of integers as far as possible symmetrically placed around 0. A is defined as follow: Let $m \geq 1$ be a constant, then

$$A = \{-[(m-1)/2], \ldots, [m/2]-1, [m/2]\}$$

Van den Genugten (1993) defines A as above. We interpreted A as the set of integers ranging from $-[(m-1)/2]$ to $[m/2]$, where $[\bullet]$ denotes the integer part of the argument inside the brackets. Thus for example if m = 4, then

$$A = \{-1,0,1,2\}$$

and w_j is defined as one or other weighting factor, which has to be chosen in advance independently of the data (a natural choice for $w_j = 1/m$, $j \epsilon A$).

The function $f(\bullet)$ is a function that ensure that we do not divide by 0 if $\hat{\sigma}_j^2(q)$ is zero. That is

$$f(z) = \frac{1}{h+z}$$

with $h > 0$. When $h \to 0$ $f(z) \to 1/z$.

Then under appropriate conditions

$$\sqrt{n}(\hat{\beta}_G(q) - \beta) \longrightarrow N_r(0,\Sigma(q))$$

The number of iterations q is determined when $tr[\hat{\Sigma}(q)]$ is minimal.

The idea of using residuals to improve the efficiency in the case of unknown heteroscedasticity was suggested by Rao (1970). In applying IWLS the main idea is to use a method which is not optimal for a particular form of heteroscedasticity but is good for a broad class of alternatives.

A consistent and simplest estimator $\hat{\Sigma}(q)$ of $\Sigma(q)$ can be based on OLS residuals $\hat{\epsilon}(0)$. This basis leads to estimators

$$\hat{C}_\gamma = n^{-1}\sum_{t=1}^{n} X_t X_t' \hat{\epsilon}(0)_t^{2\gamma}, \quad \text{for } \gamma = -1,0,1$$

$$\hat{V}_{a\iota} = n^{-1}\sum_{t=1}^{n} X_t X_t' \hat{\epsilon}(0)_t^{2a} (f(\hat{\sigma}(0)_t^2)^{\iota}) \text{ for } (a,\iota) = (0,1),(1,1),(1,2)$$

$$\hat{W}_{11} = w_o n^{-1}\sum_{t=1}^{n} X_t X_t' \hat{\epsilon}(0)_t^{2a} \frac{\partial f}{\partial \hat{\sigma}}(\hat{\sigma}(0)_t^2))$$

$$\hat{\Sigma}(q) = \hat{A}(q)\hat{V}_{12}\hat{A}(q)' + \hat{A}(q)\hat{V}_{11}\hat{B}(q)' + \hat{B}(q)\hat{V}_{11}\hat{A}(q)' + \hat{B}(q)\hat{C}_1\hat{B}(q)'$$

and

$$\hat{A}(0) = 0$$

$$\hat{A}(q) = \sum_{j=0}^{q-1} (2\hat{V}_{01}^{-1}\hat{W}_{11})^j \hat{V}_{01}^{-1}$$

$$\hat{B}(q) = (2\hat{V}_{01}^{-1}\hat{W}_{11})^q \hat{C}_0^{-1}$$

The assumptions and proof of these results can be found in Van der Genugten (1993). The most important consideration, when one applies IWLS is that h and w are chosen in advance independently of the data. It is not clear how large n must be to justify the asymptotic approximations given by Van der Genugten. Van der Genugten's results are based on a diagonal matrix Σ. When Σ is not diagonal (as in the case of autocorrelations, see section 6.3.3.4), the effect of IWLS has not yet been investigated, and should be a field of further research.

6.3.3.4 Autocorrelation

In 6.3.3.3 it was assumed that Σ was a diagonal matrix. However when the error terms are correlated, Σ is not diagonal and we referred to this phenomenon as autocorrelation. If the autocorrelation admits a specific family of structures it can be described by one of the Autoregressive Moving Average (ARMA) models. The most common use of ARMA or MA (moving average) models, is in time series data, and is discussed in many texts eg Box and Jenkins (1976).

The first order autoregressive (AR(1)) may be described as

$$\epsilon_t = \rho\epsilon_{t-1} + \nu_t \quad t = \ldots,-2,-1,0,1,2,\ldots \tag{6.3.62}$$

(ie ϵ_t regressed on lagged values of the series) where ν_t is a stochastic process such that

$$E(\nu_t) = 0,\ E(\nu_t^2) = \sigma_t^2,\ E(\nu_t\nu_r) = 0\ (t \neq r)$$

and $|\rho| < 1$ is the autocorrelation coefficient which has to be estimated.

First order moving average process (MA(1)) is defined by

$$\epsilon_t = \nu_t + \theta_1 \nu_{t-1} \qquad (6.3.63)$$

the right hand side of (6.3.63) contains a moving (weighted) sum of the error terms only. When AR and MA processes are mixed, we refer to this phenomenon as an ARMA(p,q) process and is defined as

$$\epsilon_t = \rho_1 \epsilon_{t-1} + \ldots + \rho_p \epsilon_{t-p} + \theta_1 \epsilon_t + \ldots + \theta_q \epsilon_{t-q}$$

The most common test for the presence of autocorrelated errors is the test of Durbin and Watson (1950, 1951)). A list of further references is given by Vinod and Ullah (1981).

In the usual linear hypothesis (see §6.3.5.3) on the regression coefficients, ie $C\beta = m$, the F statistic can be severely affected by the presence of ARMA errors. With ARMA type errors the user is obliged to fit the GLSE because $\Sigma \neq \sigma^2 I$, the correct approach model would be to estimate Σ, and then base the GLSE on this estimated variance structure, which as shown else where in this chapter is quite difficult, and if possible the estimation of Σ should be a avoided. Vinod and Ullah (1981) shown that if one is mainly interested in significance tests for regression coefficients, it may be possible to rely on the simpler OLS estimator. The authors then devise bounds on the t and F values based on OLS, that ensure that the conclusions made under the null hypothesis would not be reversed if we had estimated Σ, and then computed the GLSE and used significance levels based on the GLSE computations. For a derivation of these bounds see Vinod and Ullah (1981) and for tables of these bounds see Vinod and Ullah (1981) and Kiviet (1980).

We do not elaborate on autocorrelation any further, and the interested reader may examine the references mentioned above. It is, however important that the user, in specifying a plausible model, be aware that heteroscedasticity and ARMA type structures can all be part of one single structure.

6.3.4 Special cases of prior stochastic information (H, h, Σ)

A variety of biased estimators can be generated as special cases of (6.3.5) and (6.3.7). In (6.3.5) we presented the augmented model

$$\begin{bmatrix} Y \\ h \end{bmatrix} = \begin{bmatrix} X \\ H \end{bmatrix}\beta + \begin{bmatrix} \epsilon \\ \nu \end{bmatrix}$$

or

$$\dot{Y} = \dot{X}\beta + \dot{\epsilon} \tag{6.3.5}$$

where $\dot{Y} = \begin{bmatrix} Y \\ h \end{bmatrix}:(n+\ell)\times 1$; $\dot{X} = \begin{bmatrix} X \\ H \end{bmatrix}:(n+\ell)\times r$; and $\dot{\epsilon} = \begin{bmatrix} \epsilon \\ \nu \end{bmatrix}:(n+\ell)\times 1$ are the augmented matrices. The expectation and covariance of $\dot{\epsilon}$ are then

$$E(\dot{\epsilon}) = \begin{bmatrix} \tau \\ \delta \end{bmatrix} \tag{6.3.6}$$

$$\mathrm{Cov}(\dot{\epsilon}\dot{\epsilon}') = \Sigma = \begin{bmatrix} \Sigma_\epsilon & 0 \\ 0 & \Sigma_\nu \end{bmatrix} \tag{6.3.7}$$

and the AGLSE is

$$\hat{\beta}_G = \{\dot{X}'\Sigma^{-1}\dot{X}\}^{-1}\dot{X}'\Sigma^{-1}\dot{Y} \tag{6.3.8}$$

The following are some special cases:

6.3.4.1 Ridge regression

In (6.3.5) let $h = 0$, $H = I$, $E(\dot{\epsilon}) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and

$$\mathrm{Cov}(\dot{\epsilon}\dot{\epsilon}') = \sigma^2\begin{bmatrix} I_n & 0 \\ 0 & I_\ell/k \end{bmatrix}$$

that is $\epsilon \sim (0,\sigma^2 I)$ and $\nu \sim (0,(\sigma^2/k)I)$. The estimator defined in (6.3.8) will then reduce to

$$\hat{\beta}_G = [X'X + kI]^{-1}X'Y$$

the ridge estimator, defined in Chapter 2. Properties of the ridge estimator, and ways to estimate k were discussed in Chapter 2, or see Thiart (1990).

6.3.4.2 Generalised Ridge regression

In (6.3.5) let h = 0, H = I, and $E(\dot{\epsilon}) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and

$$Cov(\dot{\epsilon}\dot{\epsilon}') = \sigma^2 \begin{bmatrix} I_n & 0 \\ 0 & K^{-1} \end{bmatrix},$$

where K is an n×n diagonal matrix, and the i-th diagonal element of K is k_i, then the AGLSE, defined in (6.3.8) will reduce to

$$\hat{\beta}_G = [X'X + K]^{-1}X'Y$$

the generalised ridge estimator, defined in Chapter 2. Properties of the GRE, and ways to estimate k_i were given in Chapter 2, or see Thiart (1990).

6.3.4.3 Principal components

In (6.3.5) let h = 0, $H = V_2'$, where V, the matrix of eigenvectors of X'X is partitioned as $V = [V_1 \; V_2]$:r×r, and V_2:r×ℓ, comprises are the eigenvectors associated with the ℓ-smallest eigenvalues of X'X. Let $E(\dot{\epsilon}) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and

$Cov(\dot{\epsilon}\dot{\epsilon}') = \sigma^2 \begin{bmatrix} I_n & \\ 0 & \lambda^{-1}I_\ell \end{bmatrix}$, where $\sqrt{\lambda}$ is large, then the AGLSE, defined in (6.3.8 and using the form of 6.3.9) will reduce to

$$\begin{aligned}
\hat{\beta}_G &= [S - SH'DHS][\sigma^{-2}X'Y + H'I_\ell h] \\
&= \sigma^{-2}[S - SH'DHS]X'Y \quad (h = 0) \\
&= \sigma^{-2}SX'Y - \sigma^{-2}SH'DHSX'Y \\
&= [X'X]^{-1}X'Y - \sigma^2[X'X]^{-1}H'DH[X'X]^{-1}X'Y \quad (S = \sigma^2[X'I_nX]^{-1}) \\
&= V\Delta^{-1}U'Y - \sigma^2V\Delta^{-2}V'H'DHV\Delta^{-2}V'V\Delta U'Y \quad \text{(using SVD of X)} \\
&= V\Delta^{-1}[I - \sigma^2\Delta^{-1}V'H'DHV\Delta^{-1}]U'Y \\
&= V\Delta^{-1}[I - \sigma^2\Delta^{-1}V'V'DHV\Delta^{-1}]U'Y \qquad (6.3.64)
\end{aligned}$$

and

$$[I - \sigma^2\Delta^{-1}V'H'DHV\Delta^{-1}] = [I - \sigma^2\Delta^{-1}V'V_sDV_s'V\Delta^{-1}]$$
$$= I - \begin{bmatrix} 0 & 0 \\ 0 & \sigma^2\Delta_2^{-1}D\Delta_2^{-1} \end{bmatrix}$$
$$= \text{diag}[1,1,\ldots 1,(\lambda_{r-\ell+1})/(\lambda+\lambda_{r-\ell+1}),\ldots,\lambda_r/(\lambda+\lambda_r)] \tag{6.3.65}$$

as $D = [\Sigma_\nu + HSH']^{-1}$

$$= \sigma^{-2}[I_\ell/\lambda + V_2'V\Delta^{-2}V'V_2]^{-1}$$
$$= \sigma^{-2}[I_\ell/\lambda + [0\ I]\begin{bmatrix}\Delta_1^{-2} & 0 \\ 0 & \Delta_2^{-2}\end{bmatrix}\begin{bmatrix}0 \\ I\end{bmatrix}]^{-1}$$
$$= \sigma^{-2}[I_\ell/\lambda + \Delta_2^{-2}]^{-1}$$

and $\sigma^2\ \Delta_2^{-1}D\Delta_2^{-1} = \Delta_2^{-1}[I_\ell/\lambda + \Delta_2^{-2}]^{-1}\Delta_2^{-1}$

$$= \{\Delta_2[I_\ell/\lambda + \Delta_2^{-2}]\Delta_2\}^{-1}$$
$$= \{\lambda^{-1}\Delta_2^2 + I\}^{-1}$$
$$= \text{diag}[\lambda/(\lambda+\lambda_{r-\ell+1}),\ldots,\lambda/(\lambda+\lambda_r)]$$

It is clear that as $\lambda \to \infty$, that (6.3.65) reduces to

$$\text{diag}[1,1,\ldots 1,0,,\ldots,0]$$

Askin and Montgomery (1980) suggested $\sqrt{\lambda} = 40$. Thus the role of λ is to downweight the small eigenvalues of X'X. Thus (6.3.64) becomes asymptotically

$$\hat{\beta}_G = V\Delta^{-1}(\text{diag}[1,1,\ldots 1,0,,\ldots,0)U'Y$$
$$= \sum_{i=1}^{r-\ell} v_i u_i'Y/\sqrt{\lambda_i}$$

which is the PCE (see Thiart (1990)). Properties of the PCE, and ways to eliminate PC's are given in Thiart (1990).

6.3.4.4 Shrunken estimators

In (6.3.5) let h = 0, H = V′, where V contains the eigenvectors of X. Let $E(\dot{\epsilon}) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $Cov(\dot{\epsilon}\dot{\epsilon}') = \sigma^2 \begin{bmatrix} I_n & \\ 0 & W^{-1} \end{bmatrix}$, where the i-th element of diagonal matrix W, $w_i = d/\{\lambda_i(1-d)\}$ then the AGLSE, defined in (6.3.8) reduces to

$$\begin{aligned}
\hat{\beta}_G &= [S - SH'DHS][\sigma^{-2}X'Y + H'I_\ell h] \\
&= \sigma^{-2}[S - SH'DHS]X'Y \quad (h = 0) \\
&= \sigma^{-2}SX'Y - \sigma^{-2}SH'DHSX'Y \\
&= \hat{\beta} - \sigma^2 V\Delta^{-2}V'VDV'V\Delta^{-2}V'V\Delta U'Y \quad \text{(using SVD of X)} \\
&= \hat{\beta} - V\Delta^{-2}[W + \Delta^{-2}]^{-1}\Delta^{-2}\Delta U'Y \qquad (6.3.66)
\end{aligned}$$

Now

$\Delta^{-2}[W + \Delta^{-2}]^{-1}$ is a diagonal matrix with i-th diagonal element

$$\frac{1}{\lambda_i}\left\{\frac{w_i\lambda_i+1}{\lambda_i}\right\}^{-1} = \left\{\frac{1}{w_i\lambda_i+1}\right\} = \left\{\frac{1-d}{d+(1-d)}\right\} = 1-d$$

thus $\Delta^{-2}[W + \Delta^{-2}]^{-1} = (1-d)I$, and substituting this matrix back in (6.3.66) we have

$$\begin{aligned}
\hat{\beta}_G &= \hat{\beta} - (1-d)V\Delta^{-2}V'V\Delta U'Y \\
&= \hat{\beta} - (1-d)\hat{\beta} \\
&= d\hat{\beta}
\end{aligned}$$

the shrinkage estimator of β. Properties of the SHE, and an optimal choice of d was given in Chapter 2 and discussed in Thiart (1990).

6.3.4.5 Fractional PC estimators

In (6.3.5) let h = 0, H = V′, let

$$E(\dot{\epsilon}) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \text{ and } Cov(\dot{\epsilon}\dot{\epsilon}') = \sigma^2 \begin{bmatrix} I_n & \\ 0 & W \end{bmatrix},$$

then the AGLSE, defined in (6.3.8) reduces to

$$\begin{aligned}\hat{\beta}_G &= \sigma^{-2}[S - SH'DHS]X'Y \\ &= [I - SH'DH]\hat{\beta} \\ &= V[I - \Delta^{-2}D]V'\hat{\beta} \\ &= VFV'\hat{\beta}\end{aligned}$$

which is the fractional PC estimator defined in (5.2.5) of Thiart (1990) and Chapter 2 with

$$F = [I - \Delta^{-2}D]$$

F is a diagonal matrix with i-th diagonal element given by

$$\begin{aligned}f_i &= 1 - \lambda_i^{-1}\{(w_i\lambda_i + 1)/\lambda_i\}^{-1} \\ &= 1 - 1/(w_i\lambda_i + 1) \\ &= (w_i\lambda_i)/(w_i\lambda_i + 1) \\ &= \lambda_i/(\lambda_i + 1/w_i)\end{aligned}$$

thus for FPE $1/w_i = s^2/[\bullet]_i^2$ is found iteratively in that $[\bullet]_i$ is the ridge or the generalised ridge estimator of β at the t-th iteration. For further details see Chapter 2 or Thiart (1990).

6.3.5 Exact prior information

In (6.3.2) the auxiliary (prior) information was given in the form of a stochastic model, that is

$$h = H\beta + \nu$$

where H is an $\ell \times r$ known design matrix $r(H) = \ell \leq r$, h is a $\ell \times 1$ stochastic vector which is known, ν is the unobserved $\ell \times 1$ vector of uncorrelated error variables. When we have exact prior information the stochastic model reduces to the fixed form

$$h = H\beta \qquad (6.3.67)$$

where h is now an $\ell \times 1$ vector of constants. In many applications h = 0. Just as in the case of the stochastic model we seek the Restricted Least square (RLS) estimator, $\hat{\beta}_{RLS}$, that would minimize

$$(Y - X\beta)'\Sigma_\epsilon^{-1}(Y - X\beta)$$

subject to

$$h = H\beta$$

If we view this restriction in terms of the augmented model we get

$$\begin{bmatrix} Y \\ h \end{bmatrix} = \begin{bmatrix} X \\ H \end{bmatrix}\beta + \begin{bmatrix} \epsilon \\ 0 \end{bmatrix}$$

or

$$\dot{Y} = \dot{X}\beta + \dot{\epsilon}$$

where $\dot{Y} = \begin{bmatrix} Y \\ h \end{bmatrix}:(n+\ell)\times 1$; $\dot{X} = \begin{bmatrix} X \\ H \end{bmatrix}:(n+\ell)\times r$; and $\dot{\epsilon} = \begin{bmatrix} \epsilon \\ 0 \end{bmatrix}:(n+\ell)\times 1$ are the augmented matrices.

The expectation and covariance of $\dot{\epsilon}$ are then

$$E(\dot{\epsilon}) = \begin{bmatrix} \tau \\ 0 \end{bmatrix} \quad Cov(\dot{\epsilon}\dot{\epsilon}') = \Sigma = \begin{bmatrix} \Sigma_\epsilon & 0 \\ 0 & 0 \end{bmatrix} \tag{6.3.68}$$

Thus using (6.3.9) we can write the RLS estimator as

$$\hat{\beta}_{RLS} = \hat{\beta}_G - SH'D[H\hat{\beta}_G - h] \tag{6.3.69}$$

where $S = [X'\Sigma_\epsilon^{-1}X]^{-1}$ and $D = [HSH']^{-1}$.

Note as before we have not claimed any structure for Σ_ϵ. Thus only after certain assumptions on Σ_ϵ and replacing the unknown quantities in Σ_ϵ by an

constant or estimate based on the sample, can we obtain an operational form of (6.3.69).

Note that $(\dot{Y} - \dot{X}\beta)'\Sigma^{-1}(\dot{Y} - \dot{X}\beta) \equiv (Y - X\beta)'\Sigma_{\epsilon}^{-1}(Y - X\beta)$.

6.3.5.1 Properties of $\hat{\beta}_{RLS}$

1. Relationship to GLSE

The relationship is given in (6.3.69)

2. Expectation

Since h is now a vector of constants, with expectation, E(h) = h, the expectation of $\hat{\beta}_{RLS}$ is

$$\begin{aligned} E(\hat{\beta}_{RLS}) &= E[(I - SH'DH)\hat{\beta}_G] + SH'Dh \\ &= (I - SH'DH)(\beta + SX'\Sigma_{\epsilon}^{-1}\tau) + SH'Dh \quad \text{from (1.4.2)} \\ &= \beta - SH'D(H\beta - h) + (I - SH'DH)SX'\Sigma_{\epsilon}^{-1}\tau \end{aligned} \tag{6.3.70}$$

Thus $\hat{\beta}_{RLS}$ is in general biased for β, and we denote the bias of $\hat{\beta}_{RLS}$ by

$$\theta = - SH'D(H\beta - h) + (I - SH'DH)SX'\Sigma_{\epsilon}^{-1}\tau \tag{6.3.71}$$

The RLS estimator is unbiased if the restrictions are true ($H\beta = h$) and $X'\Sigma_{\epsilon}^{-1}\tau = 0$.

3. Variance

Since $var(\hat{\beta}_G) = S$ and $HSH' = D^{-1}$ the variance of $\hat{\beta}_{RLS}$ is

$$\begin{aligned} var\{\hat{\beta}_{RLS}\} &= var\{(I - SH'DH)\hat{\beta}_G] + SH'Dh\} \\ &= [I - SH'DH]var(\hat{\beta}_G)[I - SH'DH]' + 0 \\ &= var(\hat{\beta}_G) - SH'DHvar(\hat{\beta}_G) - var(\hat{\beta}_G)H'DHS + SH'DHvar(\hat{\beta}_G)H'DHS \\ &= S - SH'DHS - SH'DHS + SH'DHSH'DHS \\ &= S - SH'DHS \\ &= var(\hat{\beta}_G) - SH'DHS \end{aligned} \tag{6.3.72}$$

4. Mean squared error of $\hat{\beta}_{RLS}$

$$MSE(\hat{\beta}_{RLS}) = var(\hat{\beta}_G) - SH'DHS + \theta\theta' \tag{6.3.73}$$

where θ as defined in (6.3.71).

5. Total mean squared error of $\hat{\beta}_{RLS}$

$$\begin{aligned} TMSE(\hat{\beta}_{RLS}) &= tr(var(\hat{\beta}_G) - SH'DHS + \theta\theta') \\ &= tr(var(\hat{\beta}_G)) - tr(SH'DHS - \theta\theta') \end{aligned} \tag{6.3.74}$$

6. Residuals of the RLSE

$$\begin{aligned} \hat{\epsilon}_{RLS} &= Y - X\hat{\beta}_{RLS} \\ &= Y - X\hat{\beta}_G - XSH'D[H\hat{\beta}_G - h] \\ &= X\beta + \epsilon - XSX'\Sigma_\epsilon^{-1}(X\beta + \epsilon) - XSH'D[HSX'\Sigma_\epsilon^{-1}(X\beta + \epsilon) - H\beta] \\ &= \epsilon - XSX'\Sigma_\epsilon^{-1}\epsilon + XSH'DHSX'\Sigma_\epsilon^{-1}\epsilon \\ &= [I - XSX'\Sigma_\epsilon^{-1} + XSH'DHSX'\Sigma_\epsilon^{-1}]\epsilon \\ &= [I - X(S - SH'DHS)X'\Sigma_\epsilon^{-1}]\epsilon \end{aligned} \tag{6.3.75}$$

6.3.5.2 Comparison with the GLSE

The matrix variance difference between the RLSE and the GLSE is

$$\begin{aligned} V(\hat{\beta}_G) - V\{\hat{\beta}_{RLS}\} &= S - S + SH'DHS \\ &= SH'DHS \end{aligned} \tag{6.3.76}$$

which is psd and reflects the increased precision associated with additional information.

MSE-I: $$MSE[\hat{\beta}_G] - MSE(\hat{\beta}_{RLS}) = S + SX'\Sigma_\epsilon^{-1}\tau\tau'\Sigma_\epsilon^{-1}XS - S + SH'DHS - \theta\theta'$$
$$= SX'\Sigma_\epsilon^{-1}\tau\tau'\Sigma_\epsilon^{-1}XS + SH'DHS - \theta\theta'$$

where θ was defined in (6.3.71) and

$$\theta\theta' = [SX'\Sigma_\epsilon^{-1}\tau - SH'DHSX'\Sigma_\epsilon^{-1}\tau - SH'D(H\beta - h)][SX'\Sigma_\epsilon^{-1}\tau - SH'DHSX'\Sigma_\epsilon^{-1}\tau - SH'D(H\beta - h)]'$$

$$\begin{aligned} &= SX'\Sigma_\epsilon^{-1}\tau\tau'\Sigma_\epsilon^{-1}XS - SX'\Sigma_\epsilon^{-1}\tau\tau'\Sigma_\epsilon^{-1}XSH'DHS - SX'\Sigma_\epsilon^{-1}\tau(H\beta - h)'DHS \\ &\quad - SH'DHSX'\Sigma_\epsilon^{-1}\tau\tau'\Sigma_\epsilon^{-1}XS + SH'DHSX'\Sigma_\epsilon^{-1}\tau\tau'\Sigma_\epsilon^{-1}XSH'DHS \\ &\quad + SH'DHSX'\Sigma_\epsilon^{-1}\tau(H\beta - h)'DHS - SH'D(H\beta - h)\tau'\Sigma_\epsilon^{-1}XS \\ &\quad + SH'D(H\beta - h)\tau'\Sigma_\epsilon^{-1}XSH'DHS + SH'D(H\beta - h)(H\beta - h)'DHS \end{aligned} \tag{6.3.77}$$

Thus

$$\begin{aligned} MSE[\hat{\beta}_G] - MSE(\hat{\beta}_{RLS}) &= SH'DHS + SX'\Sigma_\epsilon^{-1}\tau\tau'\Sigma_\epsilon^{-1}XSH'DHS + SX'\Sigma_\epsilon^{-1}\tau(H\beta - h)'DHS \\ &\quad + SH'DHSX'\Sigma_\epsilon^{-1}\tau\tau'\Sigma_\epsilon^{-1}XS - SH'DHSX'\Sigma_\epsilon^{-1}\tau\tau'\Sigma_\epsilon^{-1}XSH'DHS \\ &\quad - SH'DHSX'\Sigma_\epsilon^{-1}\tau(H\beta - h)'DHS + SH'D(H\beta - h)\tau'\Sigma_\epsilon^{-1}XS \\ &\quad - SH'D(H\beta - h)\tau'\Sigma_\epsilon^{-1}XSH'DHS + SH'D(H\beta - h)(H\beta - h)'DHS \end{aligned}$$

If $X'\Sigma_\epsilon^{-1}\tau = 0$, this difference reduces to

$$\begin{aligned} MSE[\hat{\beta}_G] - MSE(\hat{\beta}_{RLS}) &= SH'DHS - SH'D(H\beta - h)(H\beta - h)'DHS \\ &= SH'D\{D^{-1} - (H\beta - h)(H\beta - h)'\}DHS \end{aligned} \tag{6.3.78}$$

The MSE-difference is psd if

$$\lambda = (H\beta - h)'D(H\beta - h)/2 \leq \tfrac{1}{2} \tag{6.3.79}$$

using the same arguments as in the previous section, as shown by Toro-Vizcarrondo and Wallace (1968).

A necessary and sufficient condition for the RLSE to be preferred to $\hat{\beta}$ is that

$$(H\beta - h)'D(H\beta - h)/2 \leq \tfrac{1}{2}$$

Also see the comments on (6.3.79) in the section on general linear hypothesis testing.

Vinod and Ullah (1981) consider the reversed difference (6.3.78)

$$\begin{aligned} MSE(\hat{\beta}_{RLS}) - MSE[\hat{\beta}_G] &= SH'D(H\beta - h)(H\beta - h)'DHS - SH'DHS \\ &= SH'D\{(H\beta - h)(H\beta - h)' - D^{-1}\}DHS \end{aligned} \qquad (6.3.80)$$

which will be psd if and only if for any η:r×1 non-zero constant vector

$$\frac{\eta'SH'D(H\beta - h)(H\beta - h)'DHS\eta}{\eta'SH'DD^{-1}DHS\eta} \geq 1 \qquad (6.3.81)$$

(for a proof see Rao (1973, p60)). When there is only one restriction (ie when $\ell = 1$), the quantity $\eta'SH'D$ (1×r,r×r,r×1,1×1) is a scalar (note that D is also a scalar) and thus the left hand side of (6.3.81) becomes

$$\frac{(H\beta - h)^2}{D^{-1}} \geq 1$$

and $D^{-1} = [HSH']$, a scalar. It is possible to estimate β so that the left hand side will satisfy the stated condition when $\ell = 1$.

If there is more than one restriction ($\ell \geq 2$), Vinod and Ullah (1981) show by using the result of Rao (1973), that the infimum on the left hand side must be greater or equal to one. However Rao has shown that the infimum of $\frac{\eta'SH'D(H\beta - h)(H\beta - h)'DHS\eta}{\eta'SH'DD^{-1}DHS\eta}$ over $[\eta'SH'D]$ is zero. There are no parameter values for which OLS is preferred to the RLSE when $\ell \geq 2$ on condition that the restrictions are true.

By considering (6.3.80) simultaneously with (6.3.78) the advantages of RLSE are obvious (Vinod and Ullah (1981, p67).

MSE-II: $\text{TMSE}(\hat{\beta}_G) - \text{TMSE}(\hat{\beta}_{RLSG}) = \tau'\Sigma_\epsilon^{-1}XSSX'\Sigma_\epsilon^{-1}\tau + \text{tr}(SH'DHS - \theta\theta')$

and by using (6.3.78) the

$$\begin{aligned}\text{tr}[\theta\theta'] = \text{tr}[\tau'\Sigma_\epsilon^{-1}XSSX'\Sigma_\epsilon^{-1}\tau] &- \text{tr}[\tau'\Sigma_\epsilon^{-1}XSH'DHSSX'\Sigma_\epsilon^{-1}\tau] - \text{tr}[(H\beta-h)'DHSSX'\Sigma_\epsilon^{-1}\tau] \\ &- \text{tr}[\tau'\Sigma_\epsilon^{-1}XSSH'DHSX'\Sigma_\epsilon^{-1}\tau] + \text{tr}[\tau'\Sigma_\epsilon^{-1}XSH'DHSSH'DHSX'\Sigma_\epsilon^{-1}\tau] \\ &+ \text{tr}[(H\beta-h)'DHSSH'DHSX'\Sigma_\epsilon^{-1}\tau] - \text{tr}[\tau'\Sigma_\epsilon^{-1}XSSH'D(H\beta-h)] \\ &+ \text{tr}[\tau'\Sigma_\epsilon^{-1}XSH'DHSSH'D(H\beta-h)] + \text{tr}[(H\beta-h)'DHSSH'D(H\beta-h)]\end{aligned}$$

If we assume $X'\Sigma_\epsilon^{-1}\tau = 0$, the difference in TMSE is

$$\text{TMSE}(\hat{\beta}_G) - \text{TMSE}(\hat{\beta}_{RLSG}) = \text{tr}(SH'DHS) - (H\beta-h)'DHSSH'D(H\beta-h) \qquad (6.3.82)$$

Now as before $\lambda = (H\beta-h)'D(H\beta-h)/2 \leq \frac{1}{2}$ and $S^{-1} = [X'\Sigma_\epsilon^{-1}X]$, denote the eigenvalues of S^{-1} by $s_1 \geq s_2 \geq \ldots \geq s_r$. Then by using Theorem A.13 of Toutenberg (1982), we obtain

$$s_r \leq \frac{(H\beta-h)'DHSS^{-1}SH'D(H\beta-h)}{(H\beta-h)'DHSSH'D(H\beta-h)} \leq s_1$$

$$s_r \leq \frac{(H\beta-h)'D(H\beta-h)}{(H\beta-h)'DHSSH'D(H\beta-h)} \leq s_1$$

$$s_1^{-1} \leq \frac{(H\beta-h)'DHSSH'D(H\beta-h)}{(H\beta-h)'D(H\beta-h)} \leq s_r^{-1}$$

$$(H\beta-h)'DHSSH'D(H\beta-h) \leq s_r^{-1}(H\beta-h)'D(H\beta-h) = 2s_r^{-1}\lambda$$

Thus the upper bound of $(H\beta-h)'DHSSH'D(H\beta-h)$ is $2s_r^{-1}\lambda$ and for (6.3.82) to be greater or equal to 0, we want

$$\text{tr}(SH'DHS) \geq (H\beta-h)'DHSSH'D(H\beta-h)$$

If we replace the right hand side by its upper bound, then

$$tr(SH'DHS) \geq 2s_r^{-1}\lambda$$

or

$$\lambda \leq \tfrac{1}{2}s_r tr(SH'DHS)$$

The right hand side includes the matrix Σ_ϵ^{-1}, and at this stage we have assumed no restrictions other than symmetry and positive definiteness for Σ_ϵ^{-1}.

MSE-III: We note simply that the form of the conditions should be the same as in the general case of the stochastic model.

6.3.5.3 The general linear hypothesis

For the GLRM $Y = X\beta + \epsilon$ we are usually interested in linear hypothesis on the β's. That is we might formulate hypotheses like H_0: $\beta = 0$, H_0: $\beta_i = \beta_j$ and so on. All these hypotheses can be combined in the general, simultaneous hypothesis

$$H_0: \quad C\beta = m \tag{6.3.83}$$

where C is any q×r full row matrix (sometimes referred to as the contrast matrix) and m is a q×1 vector of constants.

The restriction, $H\beta = h$, can be viewed firstly as the general hypothesis (6.3.83) with $C = H$, $m = h$ and $q = \ell$. Secondly we will consider having the restriction and the null hypothesis (6.3.83) simultaneously.

The only limitation on C in (6.3.83) is that it must have full row rank, that is $r(C) = q$ (which means that the linear functions of β in the hypothesis must be linearly independent). This limitation is merely formal, and guarantees that the hypothetical equations are consistent. Firstly we will briefly introduce the F-statistic to test the hypothesis H_0: $C\beta = m$, analogously to Searle (1971). Searle only considers the case where $\epsilon \sim N(0,\sigma^2 I)$, but we will assume the more general form $\epsilon \sim N(\tau,\Sigma_\epsilon)$. Where this

general form leads to terms that do not have the desired F distribution, we will adopt the form of Searle ie $\epsilon \sim N(0,\sigma^2 I)$. Secondly we will consider having the null hypothesis (6.3.83) and the restrictions holding simultaneously.

The general hypothesis

To generalise the result of Searle using our notation:

$$Y \sim N(X\beta + \tau;\ \Sigma_\epsilon)$$

and $$\hat{\beta}_G \sim N(\beta + SX'\Sigma_\epsilon^{-1}\tau;\ S) \quad \text{and } S = \{X'\Sigma_\epsilon^{-1}X\}^{-1}$$

Therefore

$$C\hat{\beta}_G - m \sim N(\{C\beta - m\} + CSX'\Sigma_\epsilon^{-1}\tau\ ;\ CSC')$$

or

$$C\hat{\beta}_G - m \sim N(\{C\beta - m\} + CSX'\Sigma_\epsilon^{-1}\tau\ ;\ D^{-1})$$

where $D = [CSC']^{-1}$.

The quadratic Q (sum of squares under the hypothesis) is

$$Q = (C\hat{\beta}_G - m)'D(C\hat{\beta}_G - m) \qquad (6.3.84)$$

From Theorem 2 of Searle (1971), Q will follow a $\chi^2(r[D],\ \lambda)$ distribution if DD^{-1} is idempotent. Clearly $DD^{-1} = I$ is an idempotent matrix and the $r[D] = q$, the non-centrality parameter is then

$$\lambda = (\{C\beta - m\} + CSX'\Sigma_\epsilon^{-1}\tau)'D(\{C\beta - m\} + CSX'\Sigma_\epsilon^{-1}\tau)/2 \qquad (6.3.85)$$

The residuals obtained when we fit $\hat{\beta}_G$ are

$$(Y - X\hat{\beta}_G) = [I - M]\epsilon \qquad \text{using (1.4.6)}$$

where $M = X\{X'\Sigma_\epsilon^{-1}X\}^{-1}X'\Sigma_\epsilon^{-1}$. Hence

$$E(Y - X\hat{\beta}_G) = [I - M]\tau$$

$$\text{var}(Y - X\hat{\beta}_G) = [I - M]\Sigma_\epsilon[I - M]'$$

and when $\epsilon \sim N(\tau,\Sigma_\epsilon)$, then

$$(Y - X\hat{\beta}_G) \sim N([I - M]\tau;\ [I - M]\Sigma_\epsilon[I - M]') \tag{6.3.86}$$

The error sum of squares (SSE) under $\hat{\beta}_G$ is

$$\begin{aligned} SSE &= (Y - X\hat{\beta}_G)'\Sigma_\epsilon^{-1}(Y - X\hat{\beta}_G) \\ &= \epsilon'\{[I - M]'\Sigma_\epsilon^{-1}[I - M]\}\epsilon \qquad \text{using (1.4.6)} \end{aligned}$$

SSE is a quadratic (and is in fact $(n-r)\hat{\sigma}^2$, defined in §6.3.3.2.1), which follows a $\chi^2(r([I-M]\Sigma_\epsilon^{-1}),\lambda)$ distribution if $\Sigma_\epsilon^{-1}[I - M]\Sigma_\epsilon[I - M]'$ is idempotent. This results follows from

$$\begin{aligned} &\Sigma_\epsilon^{-1}[I - M]\Sigma_\epsilon[I - M]'\Sigma_\epsilon^{-1}[I - M]\Sigma_\epsilon[I - M]' \\ &\qquad = \Sigma_\epsilon^{-1}[I - M]\Sigma_\epsilon\Sigma_\epsilon^{-1}[I - M][I - M]\Sigma_\epsilon[I - M]' \\ &\qquad = \Sigma_\epsilon^{-1}[I - M]^3\ \Sigma_\epsilon[I - M]' \\ &\qquad = \Sigma_\epsilon^{-1}[I - M]\Sigma_\epsilon[I - M]' \end{aligned} \tag{6.3.87}$$

Satisfying idempotency. Without any further proof we state the following result from Searle (1971):

If
$$\begin{aligned} w &= \frac{Q/(q\sigma^2)}{SSE/(\sigma^2(n-r))} \\ &= \frac{Q}{q\hat{\sigma}^2} \end{aligned} \tag{6.3.88}$$

then $w \sim F(q,(n-r);\lambda)$

where $\lambda = \{C\beta - m\}'D\{C\beta - m\}/2\sigma^2$ (for $\tau = 0$, and $D = [C(X'X)^{-1}C']^{-1}$)

This hypothesis test can be applied to any linear hypothesis. The only limitations are the consistency of the hypothetical equations, and Gaussian

distribution of error terms. Thus if we have the LRM $Y = X\beta + \epsilon$ and the restrictions $H\beta = h$, we find that the general linear hypothesis with $C = H$, $m = h$ and $q = \ell$, is in effect a hypothesis on the restrictions, and the estimator of β under the null hypothesis is the RLSE.

We extend this case to consider situations in which we have the general hypothesis and some restrictions simultaneously. That is we have the LRM

$$Y = X\beta + \epsilon, \text{ subject to } H\beta = h,$$

and we now consider the general linear hypothesis

$$H_0: \ C\beta = m$$

Assume that the rows of C are linearly independent of the rows of H, with $q+\ell \leq r$. Then analogously to Walker and O'Brien (1992) the sum of squares for the linear hypothesis now becomes

$$\begin{aligned} Q_{RLS} &= (C\hat{\beta}_{RLS} - m)'[C\{S - SH'DHS\}C']^{-1}(C\hat{\beta}_{RLS} - m) \\ &= (C\{\hat{\beta}_G - SH'D[H\hat{\beta}_G - h]\} - m)'[C\{S - SH'DHS\}C']^{-1}(C\{\hat{\beta}_G - SH'D[H\hat{\beta}_G - h]\} - m) \end{aligned} \tag{6.3.89}$$

where $S = [X'\Sigma_\epsilon^{-1}X]^{-1}$ and $D = [HSH']^{-1}$.

If $E[\hat{\beta}_G] = \beta$ (with $X'\Sigma_\epsilon^{-1}\tau = 0$; from (1.4.2)) then

$$\begin{aligned} E[C\{\hat{\beta}_G - SH'D[H\hat{\beta}_G - h]\} - m] &= E[C\hat{\beta}_G - CSH'D[H\hat{\beta}_G - h] - m] \\ &= C\{\beta - SH'D[H\beta - h]\} - m \end{aligned}$$

and

$$\begin{aligned} \text{var}[C\{\hat{\beta}_G - SH'D[H\hat{\beta}_G - h]\} - m] &= \text{var}[C\hat{\beta}_G - CSH'DH\hat{\beta}_G] \\ &= CSC' - CSH'DHSC' \\ &= C\{S - SH'DHS\}C' \end{aligned}$$

Now $[C\{S - SH'DHS\}C']^{-1}\{C\{S - SH'DHS\}C'\} = I_q$, an idempotent matrix, and if we assume that the error terms follow a normal distribution, then according to Searle (1971), Theorem 2, $Q_{RLS} \sim \chi^2(q,\lambda_{RLS}]$ where

$$q = r[C\{S - SH'DHS\}C']$$

and

$$\lambda_{RLS} = [C\{\beta - SH'D[H\beta-h]\} - m]'[C\{S - SH'DHS\}C']^{-1}[C\{\beta - SH'D[H\beta-h]\} -m]/2.$$

Now $\lambda_{RLS} = 0$ if and only if $C\{\beta - SH'D[H\beta-h]\} = m$, because λ_{RLS} is a positive quadratic form. Thus the null hypothesis, when used in RLSE, is in effect a test of

$$H_0(RLS): \quad C\{\beta - SH'D[H\beta - h]\} = m \qquad (6.3.90)$$

or $H_0: \quad C\beta = m \quad$ given $H\beta = h$

As before, to enable us to make use of the F tables, we assume $\epsilon \sim N(0, \sigma^2 I)$

In both null hypotheses the denominator is the same unbiased estimator of σ^2 (ie SSE/(n-r)) to ensure a non-central F test. The consequences of using an estimator of σ^2 based on the RLS residuals will be a doubly non-central F distribution (Mittelhammer (1984)).

Walker and O'Brien (1992), use SSE/(n-r) as the denominator, in the F-statistic, and compared H_0 and $H_0(RLS)$ making the following comments:

1. If the rows of C are orthogonal to those of H, ie if $CSH' = 0$ then $Q_{RLS} = Q$ and $\lambda_{RLS} = \lambda$, thus the two tests, (6.3.84) and (6.3.89) are identical. This result is true regardless of whether the restriction, $H\beta = h$, holds in the population.

2. When $H\beta = h$ is true and $CSH' \neq 0$ then H_0 and $H_0(RLS)$ are equivalent but the non-centrality parameter for $H_0(RLS)$ is greater or equal to the non-centrality parameter of H_0, as shown below. If $H\beta = h$ then:

$$\lambda = \{C\beta - m\}'[C(X'X)^{-1}C']^{-1}\{C\beta - m\}/2\sigma^2$$

$$\lambda_{RLS} = [C\beta - m]'[CSC' - CSH'DHSC']^{-1}[C\beta - m]/2\sigma^2 \quad \text{and}$$

$$[CSC'-CSH'DHSC']^{-1} = [CSC']^{-1} + [CSC']^{-1}CSH'(D^{-1}-HSC'[CSC']^{-1}CSH')^{-1}HSC'[CSC']^{-1}$$

so that

$$\lambda_{RLS} = \lambda + [C\beta - m]'[CSC']^{-1}CSH'(D^{-1} - HSC'[CSC']^{-1}CSH')^{-1}HSC'[CSC']^{-1}[C\beta - m]$$
$$= \lambda + \text{positive quantity}$$
$$\geq \lambda$$

3. When $H\beta \neq h$ and $CSH' \neq 0$ then H_o and $H_o(RLS)$ are not equivalent. Imposing the (false) restrictions transforms H_o to $H_o(RLS)$.

6.3.5.4 Model misspecification

In the LRM of (1.1) $Y = X\beta + \epsilon$, it sometimes happens that important variables are not included (underfitting) or it might happen that we have unnecessary variables included in the model (overfitting). To compensate for overfitting, we usually use the technique of subset selection, as introduced in Chapter one. In this section, we do not discuss model misspecification as such, but we will show only that model misspecification can be viewed as RLS. In the rest of this section we will assume that $\epsilon \sim (0, \sigma^2 I)$. Readers interested in model misspecification can consult Miller (1990).

6.3.5.4.1 Underfitting

Underfitting occurs when important regressors are omitted from the linear model. To examine this phenomenon from the viewpoint of RLS, we assume that we can partition the X matrix as

$$X = [X_I \; X_E] \tag{6.3.91}$$

where $X_I : n \times r_I$ contains those regressor variables included in our model, and $X_E : n \times r_E$ holds those regressors omitted form the model, but technically are observable. Similarly $\beta' = [\beta_I' \; \beta_E']$. Thus the correct model would be $Y = X\beta + \epsilon$, but the model that fitted is

$$Y = X\beta + \epsilon \quad \text{subject to} \quad \beta_E = 0. \tag{6.3.92}$$

In terms of RLS notation $H\beta = h$, with $H = [0 \;\; I]$ and $h = 0$. Note that 0 is a $r_E \times r_I$ null matrix and I is a $r_E \times r_E$ identity matrix. Thus if we partition all the matrices conformably with X we obtain the following partitioned matrices

$$S = \begin{bmatrix} X_I'X_I & X_I'X_E \\ X_E'X_I & X_E'X_E \end{bmatrix}^{-1} \qquad SH' = \begin{bmatrix} -(X_I'X_I)^{-1}X_I'X_ET^{-1} \\ T^{-1} \end{bmatrix}$$

where $T = X_E'X_E - X_E'X_I(X_I'X_I)^{-1}X_I'X_E$, $D = [HSH']^{-1} = T$ then

$$SH'DH = \begin{bmatrix} -(X_I'X_I)^{-1}X_I'X_ET^{-1} \\ T^{-1} \end{bmatrix} T[0 \quad I] = \begin{bmatrix} (X_I'X_I)^{-1}X_I'X_ET0 & -(X_I'X_I)^{-1}X_I'X_ET^{-1}TI \\ T^{-1}T0 & T^{-1}TI \end{bmatrix}$$

$$= \begin{bmatrix} 0 & -(X_I'X_I)^{-1}X_I'X_E \\ 0 & I \end{bmatrix} \begin{matrix} r_I \\ r_E \end{matrix}$$
$$\qquad\qquad\qquad\qquad r_I \qquad\qquad r_E$$

Thus

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_I \\ \hat{\beta}_E \end{bmatrix}$$

$$= \begin{bmatrix} X_I'X_I & X_I'X_E \\ X_E'X_I & X_E'X_E \end{bmatrix}^{-1} \begin{bmatrix} X_I' \\ X_E' \end{bmatrix} Y$$

$$= \begin{bmatrix} (X_I'X_I)^{-1} + (X_I'X_I)^{-1}X_I'X_ET^{-1}X_E'X_I(X_I'X_I)^{-1} & -(X_I'X_I)^{-1}X_I'X_ET^{-1} \\ -T^{-1}X_E'X_I(X_I'X_I)^{-1} & T^{-1} \end{bmatrix} \begin{bmatrix} X_I'Y \\ X_E'Y \end{bmatrix}$$

$$= \begin{bmatrix} \{(X_I'X_I)^{-1} + (X_I'X_I)^{-1}X_I'X_ET^{-1}X_E'X_I(X_I'X_I)^{-1}\}X_I'Y - (X_I'X_I)^{-1}X_I'X_ET^{-1}X_E'Y \\ -T^{-1}X_E'X_I(X_I'X_I)^{-1}X_I'Y \quad + \quad T^{-1}X_E'Y \end{bmatrix} \qquad (6.3.93)$$

thus $\hat{\beta}_{RLS} = \hat{\beta} - SH'DH\hat{\beta}$

$$\hat{\beta}_{RLS} = \begin{bmatrix} \hat{\beta}_{RLSI} \\ \hat{\beta}_{RLSE} \end{bmatrix}$$

$$= \begin{bmatrix} \hat{\beta}_I \\ \hat{\beta}_E \end{bmatrix} - \begin{bmatrix} 0 & -(X_I'X_I)^{-1}X_I'X_E \\ 0 & I \end{bmatrix} \begin{bmatrix} \hat{\beta}_I \\ \hat{\beta}_E \end{bmatrix}$$

$$\begin{aligned} \hat{\beta}_{RLSI} &= \hat{\beta}_I + (X_I'X_I)^{-1}X_I'X_E\hat{\beta}_E \\ &= (X_I'X_I)^{-1}X_I'Y + (X_I'X_I)^{-1}X_I'X_E T^{-1}X_E'X_I(X_I'X_I)^{-1}X_I'Y - (X_I'X_I)^{-1}X_I'X_E T^{-1}X_E'Y \\ &\qquad - (X_I'X_I)^{-1}X_I'X_E T^{-1}X_E'X_I(X_I'X_I)^{-1}X_I'Y + (X_I'X_I)^{-1}X_I'X_E T^{-1}X_E'Y \\ &= (X_I'X_I)^{-1}X_I'Y \end{aligned} \qquad (6.3.94)$$

and $\hat{\beta}_{RLSE} = 0$

Then the expectation of $\hat{\beta}_{RLS}$ is

$$E(\hat{\beta}_{RLS}) = \begin{bmatrix} \beta_I - (X_I'X_I)^{-1}X_I'X_E\beta_E \\ 0 \end{bmatrix}$$

thus the bias of $\hat{\beta}_{RLS}$ is

$$\theta = \begin{bmatrix} -(X_I'X_I)^{-1}X_I'X_E\beta_E \\ \beta_E \end{bmatrix} \qquad (6.3.95)$$

It is clear that the bias of $\hat{\beta}_{RLS}$ is a function of the parameters excluded from the model. If the excluded regressors are (statistically) orthogonal to the included regressors ($X_I'X_E = 0$) then $\hat{\beta}_{RLSI}$ is an unbiased estimator of β_I.

The variance of $\hat{\beta}_{RLS}$ is from (6.3.72) ($\Sigma_\epsilon = \sigma^2 I$)

$$\text{var}\{\hat{\beta}_{RLS}\} = \text{var}(\hat{\beta}) - SH'DHS$$

$$SH'DHS = \sigma^2 \begin{bmatrix} 0 & (X_I'X_I)^{-1}X_I'X_E \\ 0 & I \end{bmatrix} \begin{bmatrix} X_I'X_I & X_I'X_E \\ X_E'X_I & X_E'X_E \end{bmatrix}^{-1}$$

$$= \sigma^2 \begin{bmatrix} -(X_I'X_I)^{-1}X_I'X_E T^{-1} X_E'X_I'(X_I'X_I)^{-1} & (X_I'X_I)^{-1}X_I'X_E T^{-1} \\ -T^{-1}X_E'X_I'(X_I'X_I)^{-1} & T^{-1} \end{bmatrix}$$

$$\text{var}(\hat{\beta}) = \sigma^2 \begin{bmatrix} (X_I'X_I)^{-1} + (X_I'X_I)^{-1}X_I'X_E T^{-1} X_E'X_I (X_I'X_I)^{-1} & -(X_I'X_I)^{-1}X_I'X_E T^{-1} \\ -T^{-1}X_E'X_I (X_I'X_I)^{-1} & T^{-1} \end{bmatrix}$$

thus

$$\text{var}\{\hat{\beta}_{RLS}\} = \sigma^2 \begin{bmatrix} (X_I'X_I)^{-1} & 0 \\ 0 & 0 \end{bmatrix} \qquad (6.3.96)$$

The conditions under which $\hat{\beta}_{RLS}$ is better than $\hat{\beta}$ in terms of MSE criteria have been shown previously.

Under hypothesis testing of $\hat{\beta}_{RLSI}$

$$H_0: \; C_I \beta_I = m_I$$

or equivalently

$$H_0: \; C\beta = m_I$$

with $C = [C_I \; 0]$ then the actual null hypothesis tested is

$$H_0(RLS): \; C\{\beta - SH'D[H\beta - h]\} = m_I$$

with $C\{\beta - SH'D[H\beta - h]\} = C_I \beta_I + C_I (X_I'X_I)^{-1}X_I'X_E \beta_E$.

Thus it is clear $X_E\beta_E$ influences the estimation and testing of β_I. Only when the two sets of regressors are (statistically) orthogonal is the hypothesis about β_I unaffected. Thus by omitting relevant regressors, we obtain an estimate that is biased, and although it is better than the estimator when we fitted the full model, we see that any hypotesis test on the parameters in the model is likely to be influenced by the parameters not included in the model.

There is no hypothesis that can be set up to test underfitting. The only way the researcher and the statistician (analyst) can ensure that no relevant regressors are excluded from the model is by consultation between the two parties, and by being overcautious. We might rather include all the available regressors, and then by subset selection extract the appropriate model.

6.3.5.4.2 Overfitting

Overfitting occurs when non-relevant regressors are included in the linear model. To view this phenomenon from the viewpoint of RLS, we assume that we can partition the X matrix as

$$X = [X_R \; X_U] \tag{6.3.97}$$

where X_R:n×r_R contains those relevant regressors in our model, and X_U:n×r_U contains non-relevant regressors also included in the model. Similarly let $\beta' = [\beta_R' \; \beta_U']$. Thus the wrong model would be $Y = X\beta + \epsilon$, and the true model is

$$Y = X\beta + \epsilon \quad \text{subject to } \beta_U' = 0 \tag{6.3.98}$$

in terms of RLS notation $H\beta = h$, with $H = [0 \;\; I]$ and $h = 0$.

The OLSE of the 'wrong' model will be unbiased, but the variance of the OLSE will be inflated by the presence of X_U. If we consider the hypotesis

$$H_0: \; C_R\beta_R = 0$$

we find that the statistic derived from the overfitted model will test this hypothesis, but the non-centrality parameter under the overfitted ('false') model will be less than the non-centrality parameter of the restricted ('true') model (as established in section 6.3.5.3). Hence the overfitted model will have lower power than the restricted model.

In practice the probability of finally choosing an overfitted model can be reduced if we use subset selection. It is clear from the previous paragraphs that subset selection can be viewed as restricted regression. In the spirit of subset selection, we do not have one 'true' model, but a set of models each of which may be just as good as the others. These sets of possible models can be viewed as restricted least squares where the ordering and the partitioning of the X matrix as well as the dimensions of $H = [0\ I]$ are changed for each model in the subset. It is unclear to the writer, if we use RLS, whether one will have a defensible criterion to choose between the set of best RLS models for such overfitting situations.

6.3.6 Compatibility of sample and prior information

The AGLSE of β, $\hat{\beta}_G$ has smaller variance than the GLSE $\hat{\beta}_G$ whatever the restriction $h = H\beta + \nu$ implies (Toutenburg(1982)). Thus to ensure that the imposed prior model is relevant, and before the mixed estimator $(\hat{\beta}_G)$ is accepted, one has to check whether the prior and sample information are possibly in conflict with each other. The null-hypothesis that tests if the sample and prior information is compatible is due to Theil (1963):

H_0: prior and sample information are in agreement

Under this null hypothesis there are two independent estimates of $H\beta$, ie the known vector h (prior information) and the GLS estimator of $H\hat{\beta}_G$. If sample and prior information are compatible, then the difference between h and $H\hat{\beta}_G$

is near zero, and this difference will have

$$E[h - H\hat{\beta}_G] = H\beta + \delta - H\beta - HSX'\Sigma_\epsilon^{-1}\tau$$
$$= \delta - HSX'\Sigma_\epsilon^{-1}\tau$$
$$= 0 \text{ (when } \delta = 0 \text{ and } X'\Sigma_\epsilon^{-1}\tau = 0)$$

$$Cov[h - H\hat{\beta}_G] = \Sigma_\nu + H[X'\Sigma_\epsilon^{-1}X]H'$$

Under normality

$$(h - H\hat{\beta}_G) \sim N(\delta - HSX'\Sigma_\epsilon^{-1}\tau \;\; ; \; \Sigma_\nu + H[X'\Sigma_\epsilon^{-1}X]H')$$

thus a non-centrality component ∂ may be estimated as

$$\hat{\partial} = (h - H\hat{\beta}_G)'[\Sigma_\nu + H[X'\Sigma_\epsilon^{-1}X]H']^{-1}(h - H\hat{\beta}_G) \qquad (6.3.99)$$

where $\hat{\partial}$ is the compatibility statistic, and

$$\hat{\partial} \sim \chi^2(r[(\Sigma_\nu + H[X'\Sigma_\epsilon^{-1}X]H')^{-1}], \; \lambda)$$

where

$$\lambda = (\delta - HSX'\Sigma_\epsilon^{-1}\tau)'[\Sigma_\nu + H[X'\Sigma_\epsilon^{-1}X]H']^{-1}(\delta - HSX'\Sigma_\epsilon^{-1}\tau)/2 \qquad (6.3.100)$$

$r[(\Sigma_\nu + H[X'\Sigma_\epsilon^{-1}X]H')^{-1}] = \ell$. When $\delta = 0$ and $X'\Sigma_\epsilon^{-1}\tau = 0$, then $\lambda = 0$ and $\partial \sim \chi^2(\ell)$.

If we use the GLS estimator of σ^2, $\hat{\sigma}^2$, then under the null hypothesis ($\hat{\partial} = 0$) the test statistic is

$$F = \hat{\partial}/(\ell\hat{\sigma}^2)$$

which has a non-central F distribution with ℓ and n-r degrees of freedom and non-centrality parameter λ defined in (6.3.100).

6.4 Outliers in the AGLS model

In Chapter 2 it was pointed out that the collinearity structure of the data can be strongly affected by a few observations (Belsley *et al.* (1980), Mason and Gunst (1985), Draper and John (1981)). The term influential is used to describe an observation whose inclusion in a data set substantially changes regression coefficient estimates, predicted responses, or the results of inferential procedures (Mason and Gunst (1985)). Not all outliers are necessarily collinearity-influential points and *vice versa*. Various methods of detecting outliers are available in the literature. Common methods include graphical representation of the residuals versus the individual fitted Y and observed X variables, and normal probability plots (see for instance Cook and Weisberg (1982) and Daniel and Wood (1980)). Various methods are based on the diagonal elements h_{ii} of the Hat matrix, which are referred to as leverage values. Other methods include outlier sum of squares, the Andrews-Pregibon statistic, Cook's statistic, DFFITS, variance inflation factors and condition indices. All these methods are discussed in Thiart (1990). In this section outliers will be viewed under the general scenario of restricted least squares.

6.4.1 The mean shift model

Consider the augmented matrices of (6.3.5)

$$\begin{bmatrix} Y \\ h \end{bmatrix} = \begin{bmatrix} X \\ H \end{bmatrix}\beta + \begin{bmatrix} \epsilon \\ \nu \end{bmatrix}$$

or

$$\dot{Y} = \dot{X}\beta + \dot{\epsilon} \qquad (6.3.5)$$

where $\dot{Y} = \begin{bmatrix} Y \\ h \end{bmatrix}:(n+\ell)\times 1$; $\dot{X} = \begin{bmatrix} X \\ H \end{bmatrix}:(n+\ell)\times r$; and $\dot{\epsilon} = \begin{bmatrix} \epsilon \\ \nu \end{bmatrix}:(n+\ell)\times 1$.

The expectation and covariance of $\dot{\epsilon}$ are then

$$E(\dot{\epsilon}) = \begin{bmatrix} \tau \\ \delta \end{bmatrix} \qquad (6.3.6)$$

$$Cov(\dot{\epsilon}\dot{\epsilon}') = \Sigma = \begin{bmatrix} \Sigma_\epsilon & 0 \\ 0 & \Sigma_\nu \end{bmatrix} \qquad (6.3.7)$$

$$\hat{\beta}_G = \{\dot{X}'\Sigma^{-1}\dot{X}\}^{-1}\dot{X}'\Sigma^{-1}\dot{Y} \qquad (6.3.8)$$

Suppose that after inspection we identify s outliers (for the moment the s outliers may be part of the sample data or the prior information). One useful framework to study such outliers is the mean shift model (Cook and Weisberg (1982))

$$\dot{Y} = \dot{X}\beta + Z\alpha + \dot{\epsilon} \qquad (6.4.1)$$

where $Z:(n+\ell)\times s$ is a matrix of zeros and one's. Each column of Z contains a one in only one position, determined by the outlier. The vector $\alpha:s\times 1$ is a vector of unknown parameters. Partition Z as

$$Z = \begin{bmatrix} Z_s \\ Z_p \end{bmatrix}$$

where $Z_s:n\times s$ indicates the outliers within the sample information and $Z_p:\ell\times s$ models the outliers within the prior information. Thus (6.4.1) becomes

$$\begin{bmatrix} Y \\ h \end{bmatrix} = \begin{bmatrix} X & Z_s \\ H & Z_p \end{bmatrix}\begin{bmatrix} \beta \\ \alpha \end{bmatrix} + \begin{bmatrix} \epsilon^* \\ \nu^* \end{bmatrix}$$

or

$$\dot{Y} = X^*\beta^* + \dot{\epsilon}^* \qquad (6.4.2)$$

where $\dot{Y} = \begin{bmatrix} Y \\ h \end{bmatrix}:(n+\ell)\times 1$; $X^* = \begin{bmatrix} X & Z_s \\ H & Z_p \end{bmatrix}:(n+\ell)\times(r+s)$; $\beta^* = \begin{bmatrix} \beta \\ \alpha \end{bmatrix}:(r+s)\times 1$ and $\dot{\epsilon}^* = \begin{bmatrix} \epsilon^* \\ \nu^* \end{bmatrix}:(n+\ell)\times 1$. For simplicity the covariance matrice of ϵ^*, Σ, will be in the form of a unknown scalar (σ^2) multiple of a known matrix, such that Σ is fixed up to a constant factor σ^2. Let Σ be partitioned conformably to the sample and the prior information, ie

$\Sigma = \begin{bmatrix} \Sigma_s & 0 \\ 0 & \Sigma_p \end{bmatrix}$ which is in effect equivalent to the $\begin{bmatrix} \Sigma_\epsilon & 0 \\ 0 & \Sigma_\nu \end{bmatrix}$ development but we now emphasize the sample and prior components. For generality here we only assume that the sample and the prior information are uncorrelated thus we do not assume that either Σ_ϵ or Σ_ν is diagonal.

The GLS estimator of (6.4.2) is found analogously to (6.3.8) as

$$\hat{\beta}^*_G = [X^{*\prime}\Sigma^{-1}X^*]^{-1}X^{*\prime}\Sigma^{-1}Y \qquad (6.4.3)$$

It is clear that (6.4.3) is invariant over the unknown scalar σ^2. Note that

$$\begin{aligned}[X^{*\prime}\Sigma^{-1}X^*] &= \begin{bmatrix} X' & H' \\ Z_s' & Z_p' \end{bmatrix}\begin{bmatrix} \Sigma_s^{-1} & 0 \\ 0 & \Sigma_p^{-1} \end{bmatrix}\begin{bmatrix} X & Z_s \\ H & Z_p \end{bmatrix} \\ &= \begin{bmatrix} X'\Sigma_s^{-1}X+H'\Sigma_p^{-1}H & X'\Sigma_s^{-1}Z_s+H'\Sigma_p^{-1}Z_p \\ Z_s'\Sigma_s^{-1}X+Z_p'\Sigma_p^{-1}H & Z_s'\Sigma_s^{-1}Z_s+Z_p'\Sigma_p^{-1}Z_p \end{bmatrix} \\ &= \begin{bmatrix} E & F \\ F' & T \end{bmatrix} \quad \text{(say)}\end{aligned}$$

thus

$$[X^{*\prime}\Sigma^{-1}X^*]^{-1} = \begin{bmatrix} E^{-1}+E^{-1}FG^{-1}F'E^{-1} & -E^{-1}FG^{-1} \\ -G^{-1}F'E^{-1} & G^{-1} \end{bmatrix} \qquad (6.4.4)$$

where $G = T - F'E^{-1}F$ and

$$\begin{aligned}\hat{\beta}^* &= \begin{bmatrix} E^{-1}+E^{-1}FG^{-1}F'E^{-1} & -E^{-1}FG^{-1} \\ -G^{-1}F'E^{-1} & G^{-1} \end{bmatrix}\begin{bmatrix} X'\Sigma_s^{-1} & H'\Sigma_p^{-1} \\ Z_s'\Sigma_s^{-1} & Z_p'\Sigma_p^{-1} \end{bmatrix}\begin{bmatrix} Y \\ h \end{bmatrix} \\ &= \begin{bmatrix} (E^{-1}+E^{-1}FG^{-1}F'E^{-1})(X'\Sigma_s^{-1}Y+H'\Sigma_p^{-1}h) & -E^{-1}FG^{-1}(Z_s'\Sigma_s^{-1}Y+Z_p'\Sigma_p^{-1}h) \\ -G^{-1}F'E^{-1}(X'\Sigma_s^{-1}Y+H'\Sigma_p^{-1}h) & G^{-1}(Z_s'\Sigma_s^{-1}Y+Z_p'\Sigma_p^{-1}h) \end{bmatrix} \\ &= \begin{bmatrix} \tilde{\beta} \\ \hat{a} \end{bmatrix} \qquad (6.4.5)\end{aligned}$$

At this stage we do not simplify the equations for $\hat{\beta}^*$, as we are more interested in what will happen for special cases of Z.

6.4.1.1 Applications to prior observations

In (6.4.2) let $Z = \begin{bmatrix} 0 \\ I_\ell \end{bmatrix}$, where $s = \ell$, thus all the prior information is envisaged as outlying. Under this mean shift model it can be shown that:

(a) there are two values available for h, the given value (subject to error) and $H\hat{\beta}$. Under the null hypothesis that

H_0: $\alpha = 0$, we want the difference $h - H\hat{\beta}$ (an estimate of α) to be as close to zero as possible. If we consider $h - H\hat{\beta}$ we have

$$E(h - H\hat{\beta}) = H\beta + \delta - H\beta - HSH'D\delta - S(I - H'DHS)X'\Sigma_\epsilon^{-1}\tau$$
$$= (I - HSH'D)\delta - S(I - H'DHS)X'\Sigma_\epsilon^{-1}\tau \qquad (6.4.6)$$

by using (6.3.11). By using (6.3.9) the variance is

$$\text{var}(h - H\hat{\beta}) = \text{var}(h - H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}(X'\Sigma_\epsilon^{-1}Y + H'\Sigma_\nu^{-1}h))$$
$$= \text{var}([I - H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}H'\Sigma_\nu^{-1}]h - H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}X'\Sigma_\epsilon^{-1}Y)$$
$$= [I - H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}H'\Sigma_\nu^{-1}]\Sigma_\nu[I - \Sigma_\nu^{-1}H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}H'] +$$
$$H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}X'\Sigma_\epsilon^{-1}\Sigma_\epsilon\Sigma_\epsilon^{-1}X(\dot{X}'\Sigma^{-1}\dot{X})^{-1}H' \text{ (by independence)}$$
$$= \Sigma_\nu - 2H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}H' + H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}H'\Sigma_\nu^{-1}H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}H'$$
$$+ H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}X'\Sigma_\epsilon^{-1}X(\dot{X}'\Sigma^{-1}\dot{X})^{-1}H'$$
$$= \Sigma_\nu - H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}H' -$$
$$H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}[(\dot{X}'\Sigma^{-1}\dot{X}) - H'\Sigma_\nu^{-1}H - X'\Sigma_\epsilon^{-1}X](\dot{X}'\Sigma^{-1}\dot{X})^{-1}H'$$
$$= \Sigma_\nu - H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}H' - 0 \qquad (6.4.7)$$

Now if we assume normality then

$$h - H\hat{\beta} \sim N(E(h - H\hat{\beta}),\ (\Sigma_\nu - H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}H'))$$

where $E(h - H\hat{\beta})$ is given in (6.4.6). Note that if $\delta = 0$ and $X'\Sigma_\epsilon^{-1}\tau = 0$, then $E(h - H\hat{\beta}) = 0$. By using Theorem 2, p 57 of Searle (1971) we have that

$$Q = (h - H\hat{\beta})'[\Sigma_\nu - H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}H']^{-1}(h - H\hat{\beta})$$
$$\sim \chi^2\{k,\lambda\} \text{ where}$$
$$k = r[(\Sigma_\nu - H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}H')^{-1}] \text{ and}$$
$$\lambda = \tfrac{1}{2}E(h - H\hat{\beta})'[\Sigma_\nu - H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}H']^{-1}E(h - H\hat{\beta})\} \qquad (6.4.8)$$

which will reduce to the central χ^2 distribution of $E(h-H\hat{\beta}) = 0$.

Chalton (1990) acting on a suggestion of Schall (1987) show that $\mathbb{Q}$ defined in (6.4.8) is in fact the compatibility statistic, $\hat{\partial}$, defined in (6.3.100). The result is not surprising as when the prior information is compatible (or in agreement) with the sample information, then in fact there should be no outliers in the prior information. For completeness, and using our notation we show that $\mathbb{Q} = \hat{\partial}$. First consider

$$\begin{aligned}
H\hat{\beta} &= H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}H'\Sigma_\nu^{-1}h + H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}X'\Sigma_\epsilon^{-1}Y \\
&= H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}H'\Sigma_\nu^{-1}h + H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}(X'\Sigma_\epsilon^{-1}X)\hat{\beta}_G \\
&= H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}H'\Sigma_\nu^{-1}h + H[I + (X'\Sigma_\epsilon^{-1}X)^{-1}H'\Sigma_\nu^{-1}H]^{-1}(H'\Sigma_\nu^{-1}H)^{-1}(H'\Sigma_\nu^{-1}H)\hat{\beta}_G \\
&= H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}H'\Sigma_\nu^{-1}h + H[H'\Sigma_\nu^{-1}H + H'\Sigma_\nu^{-1}H(X'\Sigma_\epsilon^{-1}X)^{-1}H'\Sigma_\nu^{-1}H]^{-1}H'\Sigma_\nu^{-1}H\hat{\beta}_G \\
&= H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}H'\Sigma_\nu^{-1}h + H[(H'\Sigma_\nu^{-1}H)^{-1} - (X'\Sigma_\epsilon^{-1}X + H'\Sigma_\nu^{-1}H)^{-1}]H'\Sigma_\nu^{-1}H\hat{\beta}_G \\
&= H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}H'\Sigma_\nu^{-1}h + H\hat{\beta}_G - H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}H'\Sigma_\nu^{-1}H\hat{\beta}_G \\
&= H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}H'\Sigma_\nu^{-1}h + \{I - H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}H'\Sigma_\nu^{-1}\}H\hat{\beta}_G \qquad (6.4.9)
\end{aligned}$$

thus

$$\begin{aligned}
h - H\hat{\beta} &= [I - H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}H'\Sigma_\nu^{-1}]h - [I - H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}H'\Sigma_\nu^{-1}]H\hat{\beta}_G \\
&= [I - H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}H'\Sigma_\nu^{-1}][h - H\hat{\beta}_G] \\
&= [\Sigma_\nu - H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}H']\Sigma_\nu^{-1}[h - H\hat{\beta}_G] \qquad (6.4.10)
\end{aligned}$$

so that

$$\begin{aligned}
\mathbb{Q} &= (h - H\hat{\beta})'[\Sigma_\nu - H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}H']^{-1}(h - H\hat{\beta}) \\
&= [h - H\hat{\beta}_G]'\Sigma_\nu^{-1}[\Sigma_\nu - H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}H']\Sigma_\nu^{-1}[h - H\hat{\beta}_G] \\
&= [h - H\hat{\beta}_G]'[\Sigma_\nu^{-1} - \Sigma_\nu^{-1}H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}H'\Sigma_\nu^{-1}][h - H\hat{\beta}_G] \\
&= [h - H\hat{\beta}_G]'[\Sigma_\nu + H(X'\Sigma_\epsilon^{-1}X)^{-1}H']^{-1}[h - H\hat{\beta}_G] \\
&= \hat{\partial}
\end{aligned}$$

(b) If we consider model (6.4.3) with $Z = \begin{bmatrix} 0 \\ I_\ell \end{bmatrix}$ then

$$[X^{*\prime}\Sigma^{-1}X^*] = \begin{bmatrix} X' & H' \\ 0' & I_\ell \end{bmatrix}\begin{bmatrix} \Sigma_\epsilon^{-1} & 0 \\ 0 & \Sigma_\nu^{-1} \end{bmatrix}\begin{bmatrix} X & 0 \\ H & I_\ell \end{bmatrix}$$

$$= \begin{bmatrix} X'\Sigma_\epsilon^{-1}X+H'\Sigma_\nu^{-1}H & H'\Sigma_\nu^{-1} \\ \Sigma_\nu^{-1}H & \Sigma_\nu^{-1} \end{bmatrix}$$

Let $E = \dot{X}'\Sigma^{-1}\dot{X}$ and $G = \Sigma_\nu^{-1} - \Sigma_\nu^{-1}H(\dot{X}'\Sigma^{-1}\dot{X})^{-1}H'\Sigma_\nu^{-1} = [\Sigma_\nu + H(X'\Sigma_\epsilon^{-1}X)^{-1}H']^{-1}$ thus

$$\hat{\beta}^* = \begin{bmatrix} E^{-1}+E^{-1}H'\Sigma_\nu^{-1}G^{-1}\Sigma_\nu^{-1}HE^{-1} & -E^{-1}H'\Sigma_\nu^{-1}G^{-1} \\ -G^{-1}\Sigma_\nu^{-1}HE^{-1} & G^{-1} \end{bmatrix}\begin{bmatrix} X'\Sigma_\epsilon^{-1} & H'\Sigma_\nu^{-1} \\ 0 & \Sigma_\nu^{-1} \end{bmatrix}\begin{bmatrix} Y \\ h \end{bmatrix}$$

$$= \begin{bmatrix} (E^{-1}+E^{-1}H'\Sigma_\nu^{-1}G^{-1}\Sigma_\nu^{-1}HE^{-1})(X'\Sigma_\epsilon^{-1}Y+H'\Sigma_\nu^{-1}h) - E^{-1}H'\Sigma_\nu^{-1}G^{-1}\Sigma_\nu^{-1}h \\ -G^{-1}\Sigma_\nu^{-1}HE^{-1}(X'\Sigma_\epsilon^{-1}Y+H'\Sigma_\nu^{-1}h) + G^{-1}\Sigma_\nu^{-1}h \end{bmatrix}$$

$$= \begin{bmatrix} \tilde{\beta} \\ \hat{a} \end{bmatrix}$$

If we now consider the general null hypothesis (defined in §6.3.5.3)

$$H_0: C\beta^* = m$$

choosing both $C = [0\ I_\ell]$ and $m = 0$ is in effect the null hypothesis $a = 0$. If we assume normality then

$$C\hat{\beta}^* - m \sim N(E(C\hat{\beta}^* - m),\ C[X^{*\prime}\Sigma^{-1}X^*]^{-1}C'))$$

and $C[X^{*\prime}\Sigma^{-1}X^*]^{-1}C' = G^{-1} = [\Sigma_\nu + H(X'\Sigma_\epsilon^{-1}X)^{-1}H']$

Now $(C\hat{\beta}^* - m)'G(C\hat{\beta}^* - m) = \hat{a}'G\hat{a} \sim \chi^2(\ell,\ \lambda)$

where $r[\Sigma_\nu+H(X'\Sigma_\epsilon^{-1}X)^{-1}H'] = \ell$ and $\lambda = \frac{1}{2}E(C\hat{\beta}^* - m)'GE(C\hat{\beta}^* - m)$.

We consider $\hat{a}'G\hat{a}$. Firstly note that

$$\begin{aligned}
\hat{a} &= -G^{-1}\Sigma_\nu^{-1}HE^{-1}(X'\Sigma_\epsilon^{-1}Y + H'\Sigma_\nu^{-1}h) + G^{-1}\Sigma_\nu^{-1}h \\
&= G^{-1}\Sigma_\nu^{-1}[(I - HE^{-1}H'\Sigma_\nu^{-1})h - HE^{-1}X'\Sigma_\epsilon^{-1}X\hat{\beta}_G] \\
&= G^{-1}\Sigma_\nu^{-1}[(I-HE^{-1}H'\Sigma_\nu^{-1})h - H(H'\Sigma_\nu^{-1}H+H'\Sigma_\nu^{-1}H(X'\Sigma_\epsilon^{-1}X)^{-1}H'\Sigma_\nu^{-1}H)^{-1}H'\Sigma_\nu^{-1}H\hat{\beta}_G] \\
&= G^{-1}\Sigma_\nu^{-1}[(I-HE^{-1}H'\Sigma_\nu^{-1})h - H[(H'\Sigma_\nu^{-1}H)^{-1} - (X'\Sigma_\epsilon^{-1}X+H'\Sigma_\nu^{-1}H)^{-1}]H'\Sigma_\nu^{-1}H\hat{\beta}_G] \\
&= G^{-1}\Sigma_\nu^{-1}[(I-HE^{-1}H'\Sigma_\nu^{-1})h - H\hat{\beta}_G + HE^{-1}H'\Sigma_\nu^{-1}H\hat{\beta}_G] \\
&= G^{-1}\Sigma_\nu^{-1}[(I-HE^{-1}H'\Sigma_\nu^{-1})h - (I - HE^{-1}H'\Sigma_\nu^{-1})H\hat{\beta}_G] \\
&= G^{-1}\Sigma_\nu^{-1}(I-HE^{-1}H'\Sigma_\nu^{-1})(h - H\hat{\beta}_G) \qquad (6.4.11)
\end{aligned}$$

thus

$$\begin{aligned}
\hat{a}'G\hat{a} &= (h - H\hat{\beta}_G)'(I-\Sigma_\nu^{-1}HE^{-1}H')\Sigma_\nu^{-1}G^{-1}\Sigma_\nu^{-1}(I-HE^{-1}H'\Sigma_\nu^{-1})(h - H\hat{\beta}_G) \\
&= (h - H\hat{\beta}_G)'(\Sigma_\nu^{-1} - \Sigma_\nu^{-1}HE^{-1}H'\Sigma_\nu^{-1})G^{-1}(\Sigma_\nu^{-1} - \Sigma_\nu^{-1}HE^{-1}H'\Sigma_\nu^{-1})(h - H\hat{\beta}_G) \\
&= (h - H\hat{\beta}_G)'GG^{-1}G(h - H\hat{\beta}_G) \\
&= (h - H\hat{\beta}_G)'[\Sigma_\nu + H(X'\Sigma_\epsilon^{-1}X)^{-1}H']^{-1}(h - H\hat{\beta}_G) \\
&= \hat{\partial}
\end{aligned}$$

Thus we have shown that testing the compatibility of the prior information with the sample information, is in effect the same in both approaches. In (b) we test the direct effect of $a = 0$, where as in (a) we consider whether $h - H\hat{\beta}$, where the structure of $\hat{\beta}$ contains no reference to Z, at all, is approximately 0. In effect we test if the estimate of $H\beta$, is near the given h.

In the structure of $Z = [0\ I_\ell]'$, we have implied that there are outliers in all the rows of the prior information. If outliers were present, and we reject H_0, we have shown that prior and sample information are not compatible. In examining the prior information, we do not consider other structures of Z. For instance, we will not consider, $Z = [0\ Z_p]$, where Z_p is as defined before. However we do not lose any generality because by rearranging the rows it would be feasible to have $Z_p = [I_s\ 0]'$, where $s < \ell$.

6.4.1.2 Applications to outliers in the sample

If we only envisage outliers in the sample information then $Z = [Z_s \; 0]'$. Say there are s outliers then by rearrangement of the rows of Y, X and the rows and columns of Σ_ϵ, we obtain the following partitioning of (6.4.2)

$$\begin{bmatrix} Y_o \\ Y_c \\ h \end{bmatrix} = \begin{bmatrix} X_0 & I_s \\ X_c & 0 \\ H & 0 \\ r & s \end{bmatrix} \begin{bmatrix} \beta \\ a \end{bmatrix} + \begin{bmatrix} \epsilon_0^* \\ \epsilon_c^* \\ \nu^* \end{bmatrix} \tag{6.4.12}$$

where the subscript ^{0} indicates those observations suspected to be outliers and the subscript c indicate the "clean" data. Furthermore Y_0 and ϵ_0^* are s×1 vectors, Y_c and ϵ_c are (n-s)×1 vectors, X_0:s×r and X_c:(n-s)×r matrix are corresponding matrices. The vector $\epsilon^* = [\epsilon_0^{*\prime} \; \epsilon_c^{*\prime} \; \nu^*]'$ is in general different from vector $\dot{\epsilon}$. The covariance matrix of ϵ^* can be partitioned as

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & 0 \\ \Sigma_{21} & \Sigma_{22} & 0 \\ 0 & 0 & \Sigma_\nu \\ s & (n-s) & \ell \end{bmatrix} \begin{matrix} s \\ (n-s) \\ \ell \end{matrix} \tag{6.4.13}$$

and

$$\Sigma^{-1} = \begin{bmatrix} S_{11} & S_{12} & 0 \\ S_{21} & S_{22} & 0 \\ 0 & 0 & \Sigma_\nu^{-1} \end{bmatrix} = \begin{bmatrix} S_1 \\ S_2 \\ S_3 \end{bmatrix} \tag{6.4.14}$$

with $S_{11} = [\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}]^{-1}$; $S_{12} = \Sigma_{11}^{-1}\Sigma_{12}[\Sigma_{22} - \Sigma_{21}\Sigma_{22}^{-1}\Sigma_{12}]^{-1}$; $S_{21} = S_{12}'$ and $S_{22} = [\Sigma_{22} - \Sigma_{21}\Sigma_{22}^{-1}\Sigma_{12}]^{-1}$.

Using the partitioning of (6.4.12) we have that

$$[X^{*\prime}\Sigma^{-1}X^*] = \begin{bmatrix} \dot{X}'\Sigma^{-1}\dot{X} & \dot{X}'S_1' \\ S_1\dot{X} & S_{11} \end{bmatrix}$$

Let $E = \dot{X}'\Sigma^{-1}\dot{X}$ and $N_{11} = S_{11} - S_1\dot{X}E^{-1}\dot{X}'S_1'$

where the matrix N is defined as

$$N = \Sigma^{-1} - \Sigma^{-1}\dot{X}(\dot{X}'\Sigma^{-1}\dot{X})^{-1}\dot{X}'\Sigma^{-1}$$

$$= \begin{bmatrix} S_{11}-S_1\dot{X}E^{-1}\dot{X}'S_1' & S_{12}-S_1\dot{X}E^{-1}\dot{X}'S_2' & -S_1\dot{X}E^{-1}\dot{X}'S_3' \\ S_{21}-S_2\dot{X}E^{-1}\dot{X}'S_1' & S_{22}-S_2\dot{X}E^{-1}\dot{X}'S_2' & -S_2\dot{X}E^{-1}\dot{X}'S_3' \\ -S_3\dot{X}E^{-1}\dot{X}'S_1' & -S_3\dot{X}E^{-1}\dot{X}'S_2' & \Sigma_\nu^{-1}-S_3\dot{X}E^{-1}\dot{X}'S_3' \end{bmatrix}$$

$$= \begin{bmatrix} N_{11} & N_{12} & N_{13} \\ N_{21} & N_{22} & N_{23} \\ N_{31} & N_{31} & N_{33} \end{bmatrix} = \begin{bmatrix} N_1 \\ N_2 \\ N_3 \end{bmatrix} \qquad (6.4.15)$$

so that

$$\hat{\beta}^* = \begin{bmatrix} E^{-1}+E^{-1}\dot{X}'S_1'N_{11}^{-1}S_1\dot{X}E^{-1} & -E^{-1}\dot{X}'S_1'N_{11}^{-1} \\ -N_{11}^{-1}S_1\dot{X}E^{-1} & N_{11}^{-1} \end{bmatrix} \begin{bmatrix} X_0'S_1 + X_c'S_2 + H'\Sigma_\nu^{-1} \\ S_1 \end{bmatrix} \dot{Y}$$

$$= \begin{bmatrix} (E^{-1}+E^{-1}\dot{X}'S_1'N_{11}^{-1}S_1\dot{X}E^{-1})(X_0'S_1+X_c'S_2+H'\Sigma_\nu^{-1})\dot{Y} - E^{-1}\dot{X}'S_1N_{11}^{-1}S_1\dot{Y} \\ -N_{11}^{-1}S_1\dot{X}E^{-1}(X_0'S_1+X_c'S_2+H'\Sigma_\nu^{-1})\dot{Y} + N_{11}^{-1}S_1\dot{Y} \end{bmatrix}$$

$$= \begin{bmatrix} \tilde{\beta} \\ \hat{a} \end{bmatrix} \qquad (6.4.16)$$

If we now consider the general null hypothesis (defined in §6.3.5.3)

$$H_0: C\beta^* = m$$

choosing both $C = [0\ I_s]$ and $m = 0$ is in effect the null hypothesis $a = 0$. If we assume normality then

$$C\hat{\beta}^* - m \sim N(E(C\hat{\beta}^* - m),\ C[X^{*\prime}\Sigma^{-1}X^*]^{-1}C')$$

and $C[X^{*\prime}\Sigma^{-1}X^*]^{-1}C' = N_{11}^{-1}$

Now $(C\hat{\beta}^* - m)'N_{11}(C\hat{\beta}^* - m) = \hat{a}'N_{11}\hat{a} \sim \chi^2(s,\ \lambda)$ where $r[N_{11}^{-1}] = s$ and $\lambda = E(C\hat{\beta}^* - m)'N_{11}E(C\hat{\beta}^* - m)$.

Consider $\hat{a}'N_{11}\hat{a}$. Firstly note that

$$\begin{aligned}
\hat{a} &= -N_{11}^{-1}S_1\dot{X}E^{-1}(X_0'S_1+X_c'S_2 + H'\Sigma_\nu^{-1})\dot{Y} + N_{11}^{-1}S_1\dot{Y} \\
&= N_{11}^{-1}[(S_{11}Y_0 + S_{12}Y_c - S_1\dot{X}E^{-1}[(X_0'S_{11}+X_c'S_{21})Y_0 + (X_0'S_{12}+X_c'S_{22})Y_c + H'\Sigma_\nu^{-1}h] \\
&= N_{11}^{-1}[(S_{11}-S_1\dot{X}E^{-1}\dot{X}'S_1')Y_0 + (S_{12} -S_1\dot{X}E^{-1}\dot{X}'S_2')Y_c - S_1\dot{X}E^{-1}\dot{X}'S_3h] \\
&= N_{11}^{-1}[N_{11}Y_0 + N_{12}Y_c - N_{13}h] \\
&= N_{11}^{-1}N_1\dot{Y} \\
&= N_{11}^{-1}N_1(\hat{\dot{\epsilon}} + \dot{X}\hat{\beta}_G) \quad (\hat{\dot{\epsilon}} = \dot{Y} - \dot{X}\hat{\beta}_G)) \\
&= N_{11}^{-1}N_1\hat{\dot{\epsilon}} + N_{11}^{-1}N_1\dot{X}\hat{\beta}_G \\
&= N_{11}^{-1}N_1\hat{\dot{\epsilon}} \qquad (6.4.17)
\end{aligned}$$

as $$\begin{aligned}
N_1\dot{X} &= [(S_{11}-S_1\dot{X}E^{-1}\dot{X}'S_1')X_0 + (S_{12} -S_1\dot{X}E^{-1}\dot{X}'S_2')X_c - S_1\dot{X}E^{-1}\dot{X}'S_3h] \\
&= [(S_{11}X_0 + S_{12}X_c - S_1\dot{X}E^{-1}\dot{X}'\Sigma^{-1}\dot{X}] \\
&= [S_1\dot{X} - S_1\dot{X}] \\
&= 0 \qquad (6.4.18)
\end{aligned}$$

thus

$$\begin{aligned}
\hat{a}'N_{11}\hat{a} &= \hat{\dot{\epsilon}}N_1'N_{11}^{-1}N_{11}N_{11}^{-1}N_1\hat{\dot{\epsilon}} \\
&= \hat{\dot{\epsilon}}'N_1'N_{11}^{-1}N_1\hat{\dot{\epsilon}} \qquad (6.4.19)
\end{aligned}$$

Gentleman and Wilk (1975), called the quantity defined in (6.4.19), the outlier sum of squares.

The term $\tilde{\beta}$ can be expressed as

$$\begin{aligned}
\tilde{\beta} &= (E^{-1}+E^{-1}\dot{X}'S_1'N_{11}^{-1}S_1\dot{X}E^{-1})(X_0'S_1+X_c'S_2 + H'\Sigma_\nu^{-1})\dot{Y} - E^{-1}\dot{X}'S_1N_{11}^{-1}S_1\dot{Y} \\
&= E^{-1}(X_0'S_1+X_c'S_2 + H'\Sigma_\nu^{-1})\dot{Y} - E^{-1}\dot{X}'S_1[N_{11}^{-1}S_1\dot{Y} - N_{11}^{-1}S_1\dot{X}E^{-1}(X_0'S_1+X_c'S_2+H'\Sigma_\nu^{-1})\dot{Y}] \\
&= E^{-1}\dot{X}'\Sigma^{-1}\dot{Y} - E^{-1}\dot{X}'S_1'\hat{a} \\
&= \hat{\beta}_G - E^{-1}\dot{X}'S_1'\hat{a} \qquad (6.4.20)
\end{aligned}$$

In terms of the ANOVA structure we can partition the sum of squares:

ANOVA Table

Source	SS	df	MS
Regression	$SSR = \hat{\beta}^{*\prime}X^{*\prime}\Sigma^{-1}\dot{Y} - n\bar{\dot{Y}}^2$	$(r+s)-1$	$MSR = SSR/(r+s-1)$
Error	$SSE = \dot{Y}'\Sigma^{-1}\dot{Y} - \hat{\beta}^{*\prime}X^{*\prime}\Sigma^{-1}\dot{Y}$	$(n+\ell)-(r+s)$	$MSE = SSE/(n+\ell-r-s)$
Total	$SSTO = \dot{Y}'\Sigma^{-1}\dot{Y} - n\bar{\dot{Y}}^2$	$(n+\ell)-1$	

Note that by using the relationship $\hat{\dot{\epsilon}} = \dot{Y} - \dot{X}\hat{\dot{\beta}}_G$, the SSE can be written as

$$
\begin{aligned}
SSE &= \dot{Y}'\Sigma^{-1}\dot{Y} - \hat{\beta}^{*\prime}X^{*\prime}\Sigma^{-1}\dot{Y} \\
&= \hat{\dot{\epsilon}}'\Sigma^{-1}\hat{\dot{\epsilon}} + (\dot{Y} - \dot{X}\hat{\dot{\beta}}_G)'\Sigma^{-1}\dot{X}\hat{\dot{\beta}}_G + \hat{\dot{\beta}}_G'EE^{-1}\dot{X}'\Sigma^{-1}\dot{Y} - [\tilde{\beta}'\ \hat{a}']X^{*\prime}\Sigma^{-1}\dot{Y} \\
&= \hat{\dot{\epsilon}}'\Sigma^{-1}\hat{\dot{\epsilon}} + \dot{Y}'\Sigma^{-1}\dot{X}\hat{\dot{\beta}}_G - \hat{\dot{\beta}}'_G\dot{X}'\Sigma^{-1}\dot{X}\hat{\dot{\beta}}_G + \hat{\dot{\beta}}_G'\dot{X}'\Sigma^{-1}\dot{X}\hat{\dot{\beta}}_G - \tilde{\beta}'\dot{X}'\Sigma^{-1}\dot{Y} - \hat{a}'[I\ 0\ 0]\Sigma^{-1}\dot{Y} \\
&= \hat{\dot{\epsilon}}'\Sigma^{-1}\hat{\dot{\epsilon}} + \dot{Y}'\Sigma^{-1}\dot{X}E^{-1}\dot{X}'\Sigma^{-1}\dot{Y} - (\hat{\dot{\beta}}_G - E^{-1}\dot{X}'S_1'\hat{a})'\dot{X}'\Sigma^{-1}\dot{Y} - \hat{a}'S_1\dot{Y} \\
&= \hat{\dot{\epsilon}}'\Sigma^{-1}\hat{\dot{\epsilon}} + \hat{\dot{\beta}}'_G\dot{X}'\Sigma^{-1}\dot{Y} - \hat{\dot{\beta}}_G'\dot{X}'\Sigma^{-1}\dot{Y} + \hat{a}'S_1\dot{X}E^{-1}\dot{X}'\Sigma^{-1}\dot{Y} - \hat{a}'S_1\dot{Y} \\
&= \hat{\dot{\epsilon}}'\Sigma^{-1}\hat{\dot{\epsilon}} - \hat{a}'[S_1 - S_1\dot{X}E^{-1}\dot{X}'\Sigma^{-1}]\dot{Y} \\
&= \hat{\dot{\epsilon}}'\Sigma^{-1}\hat{\dot{\epsilon}} - \hat{a}'N_{11}N_{11}^{-1}N_1\dot{Y} \\
&= \hat{\dot{\epsilon}}'\Sigma^{-1}\hat{\dot{\epsilon}} - \hat{a}'N_{11}\hat{a} \quad \text{(from 6.4.17)} \\
&= \hat{\dot{\epsilon}}'\Sigma^{-1}\hat{\dot{\epsilon}} - \hat{\dot{\epsilon}}'N_1'N_{11}^{-1}N_1\hat{\dot{\epsilon}} \qquad (6.4.21)
\end{aligned}
$$

The F statistic, associated with the hypothesis

$$H_0: \ a = 0 \text{ is}$$

$$F = \frac{\hat{\dot{\epsilon}}'N_1'N_{11}^{-1}N_1\hat{\dot{\epsilon}}/s}{SSE/(n+\ell-r-s)}$$

Note that

$$\begin{aligned} SSR(\beta^*) - SSR(\hat{\beta}_G) &= \hat{\beta}^{*\prime}X^{*\prime}\Sigma^{-1}\dot{Y} - \hat{\beta}_G'\dot{X}'\Sigma^{-1}\dot{Y} \\ &= \hat{\beta}_G'\dot{X}'\Sigma^{-1}\dot{Y} + \hat{a}'S_1\dot{X}E^{-1}\dot{X}'\Sigma^{-1}\dot{Y} - \hat{a}'S_1\dot{Y} - \hat{\beta}_G'\dot{X}'\Sigma^{-1}\dot{Y} \\ &= \hat{a}'S_1\dot{X}E^{-1}\dot{X}'\Sigma^{-1}\dot{Y} - \hat{a}'S_1\dot{Y} \\ &= \hat{\epsilon}'N_1'N_{11}^{-1}N_1\hat{\epsilon} \end{aligned}$$

If we assume that $\lambda = 0$, ie $E(C\hat{\beta}^* - m)' = 0$ then the F-statistic will follow a central F-distribution with s and $(n+\ell-r-s)$ degrees of freedom. In general the F-statistic will follow a doubly non-central F-distribution.

6.4.2 Standard regression packages and outliers

Standard regression packages can be use to identify outliers, by augmenting the regression model. Two scenarios will be discussed in this section.

First of all consider the structure of the covariance matrix. If $\Sigma = \sigma^2 I$, proceed with fitting the model $\dot{Y} = \dot{X}\beta_G + \dot{\epsilon}$, if $\Sigma = \sigma^2 V$, V known, then as V is psd it is possible to write V as $V^{\frac{1}{2}}V^{\frac{1}{2}}$, then by multiplying the model $\dot{Y} = \dot{X}\beta_G + \dot{\epsilon}$ with $V^{-\frac{1}{2}}$, the model reduces to the LRM, where the OLS fit is in order. When there is no prior information the model will reduce to the LRM. Identify any suspected outliers, augment the $\dot{X}$ model (mean shift model) and then by using any stepwise procedures (forward, backward and all subsets) identify the best set of outliers.

In finding the best set of outliers two procedures can be followed. First: keep the X variables fixed and only perform the stepwise procedures on the outliers (the Z matrix). Then after finding the set of most likely outliers, perform a stepwise procedure, on the X matrix, keeping the likely outliers fixed (fitting the mean shift model is equivalent to having deleted the set of likely outliers)

Alternatively: do a stepwise procedure on the complete or unified mean shift model, ie simultaneously find the best subset of the X and Z variables.

The opinion of the author is that the second procedure should give better subsets than the first, as any confounding between the X matrix and the outliers is substantially eliminated. The difference between these two approaches is a field for further research.

6.5 Summary

In this chapter a general model was described. It was shown that various misspecifications can be addressed by employing the theory of restricted least squares. Several models resulting from various assumptions under the general model were investigated, and fields for further research indicated.

REFERENCES

ABDELMALEK N.N.(1971): Linear L_1 approximation for a discrete point set and L_1 solutions of overdetermined linear equations. *Journal of the Association for Computing Machinery*, 18, 41-47.

AITKEN A.C.(1935): On least squares and linear combinations of observations. *Proceedings of the Royal Society of Edinburgh*, 55, 42-48.

ANDERSON V.L. AND McLEAN R.A.(1974): *Design of Experiments.* Marcel Dekker, Inc, New York.

ANDREWS D.F. AND PREGIBON D.(1978): Finding the Outliers that Matter. *Journal of the Royal Statistical Society, Series* B, 40, No. 1, 85-93.

APPA G. AND SMITH C.(1973): On L_1 and Chebychev estimation. *Mathematical Programming,* 5, 73-87.

ARMSTRONG R.D. AND KUNG M.T.(1980): A dual method for discrete Chebychev curve fitting. *Mathematical Programming,* 19, 186-199.

ARMSTRONG R.D., FROME E.L. AND KUNG D.S.(1979): A revised simplex algorithm for the absolute deviation curve fitting problem. *Communications in Statistics, Part B - Simulation and Computation,* 8, 175-190.

ASKIN R.G. AND MONTGOMERY D.C.(1980): Augmented robust estimators. *Technometrics,* 22, 333-341.

ATKINSON A.C.(1987): *Plots, Transformations and Regression. An introduction to graphical methods of diagnostic regression analysis.* Oxford University Press, Oxford.

BARR G.D.I.(1981): A contribution to adaptive robust estimation. PhD Thesis. University of Cape Town, Cape Town.

BARRODALE I. AND PHILLIPS C.(1974): An improved algorithm for discrete Chebychev linear approximation. Proceedings of the Fourth Manitoba Conference on Numerical Mathematics, 4, 177-190.

BARRODALE I. AND PHILLIPS C.(1975): Solution of an overdetermined system of linear equations in the Chebychev norm. *ACM Transactions on Mathematical Software* 1, 264-270.

BARRODALE I. AND ROBERTS F.D.K.(1970): Applications of Mathematical Programming to L_p approximation In: J.B. Rosen, O.L. Mangasarian and K. Ritter eds. *Nonlinear programming,* Academic Press, New York, 447-464.

BARRODALE I. AND ROBERTS F.D.K.(1973): An improved algorithm for discrete L_1 linear approximation. *SIAM Journal of Numerical Analysis* 10, 839-848.

BARRODALE I. AND ROBERTS F.D.K.(1974): Algorithm 478: Solution of an overdetermined system of equations in the L_1-norm. *Communications of the Association for Computing Machinery*, 17, 319-320.

BARRODALE I. AND ROBERTS F.D.K.(1978): An efficient algorithm for discrete L_1 linear approximation with linear constraints. *Siam Journal of Numerical Analysis* 15, 603-611.

BARTELS R.H. CONN A.R. AND SINCLAIR J.W.(1978): Minimization techniques for piecewise differentiable functions: The L_1 solution to an overdetermined linear system. *SIAM Journal of Numerical Analysis*, 15, 224-241.

BARTLETT M.S.(1937): Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London,* A160, 268-282.

BASSETT G. AND KOENKER R.(1978): Asymptotic theory of least absolute error regression. *Journal of the American Statistical Association*, 73, 618-622.

BELSLEY D.A.(1984): Demeaning conditioning diagnostics through centering, and Reply. *The American Statistician*, 38, 73-77 and 90-93.

BELSLEY D.A.(1986): Centering the constant, first-differencing, and assessing collinearity, in: D.A. Belsley and E. Kuh (Eds.), *Model Reliability* (MIT Press. Cambridge, MA, 1986).

BELSLEY D.A., KUH E. AND WELSCH R.E.(1980): *Regression diagnostics: identifying influential data and sources of collinearity.* John Wiley & Sons, New York.

BLOOMFIELD P. AND STEIGER W.L.(1983): Least absolute deviations: Theory, applications and algorithms. Birkhäuser, Boston Massachusetts.

BOSCOVICH R.J.(1757): De litteraria expeditione per pontificiam ditionem et synopsis amplioris operis, ac habentur plura ejus ex exemplaria etiam sensorum impressa, Bononiensi Scientiarum et Artum Instituto atque Academia Commentarii, 4, 353-396.

BOX G.E.P. AND JENKINS G.M.(1976): Time series analysis forecasting and control. San Francisco: Holden Day)

BOX G.E.P. AND MULLER M.E.(1958): A note on the generation of random normal deviates. *Annals of Mathematical Statistics*, 29, 610-611.

BRENT R.P.(1973): *Algorithms for Minimization without Derivatives.* Englewood Cliffs, N.J.: Prentice-Hall.

BREUSCH T.S. AND PAGAN A.R.(1979): A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47, 1287-1294.

BRODLIE K.W.(1977): Unconstrained minimization. In Proceedings of the conference on The State of the Art in Numerical Analysis held in April, 1976, (D. Jacobs, ed.).

BROYDEN C.G.(1967): Quasi-Newton methods and their application to function minimization. *Mathematics of Computation*, 21, 368-381.

BROYDEN C.G.(1970): The convergence of a class of double-rank minimization algorithms. *Journal of the Institute of Mathematics and Applications*, 6, 222-231.

BURR I.W. AND FOSTER L.A.(1972): A test for equality of variances, Department of Statistics Mimeo Series No. 282, Purdue University, Lafayette, Indiana.

CHALTON D.O.(1990): Contributions to influence, outliers and Bayesian analysis in the multiple linear regression model. Ph.D. Thesis, University of Cape Town.

CHALTON D.O. AND TROSKIE C.G.(1992): Q plots: A graphical aid for regression analysis. *Communications in Statistics, Theory and Methods*, 21, 626-636.

CHARNES A., COOPER W.W. AND FERGUSON R.(1955): Optimal estimation of executive compensation by linear programming. *Management Science*, 1, 138-151.

CHARETTE M.(1978): The exact finite sample properties of the mixed estimator, unpublished Ph.D. thesis, 1978, University of Western Ontario.

CHAY S.C., FARDO R.D. AND MAZUMDAR M.(1975): On using the Box-Muller transformation with multiplicative congruential Pseudo-random number generators. *Applied Statistics,* 24, 132-135.

CHEBYCHEV P.L.(1854): Theorie des mecanismes connus sous le nom de parallelogrammes. Reprinted in: Ouvres de P.L. Chebychev ed. A. Markoff and N. Sonin Vol I, (1899), 111-143. Imprimerie de l'Academie Imperiale des Sciences St. Petersbourg.

COOK R.D.(1977): Detection of influential observations in linear regression. *Technometrics*, 19, No. 1, February 1977, 15-18.

COOK R.D. AND WEISBERG S.(1982): *Residuals and influence in regression.* Chapman and Hall, New York.

COX D.R. AND HINKLEY D.V.(1974): *Theoretical Statistics.* Chapman and Hall, London.

CRAMER H.(1946): *Mathematical methods of statistics.* Princeton University Rress.

DANIEL C. AND WOOD F.S.(1980): *Fitting equations to data,* 2nd ed. John Wiley & Sons, New York.

DAHLQUIST G. AND BJORCK, A.(1974): *Numerical methods.* Translated by N. Anderson, Prentice Hall, New Jersey.

DAVIDON W.C.(1959): Variable metric method for minimization. A.E.C. Research and development Report, ANL-5990 (Rev).

DIELMAN T.E.(1984): Least absolute value estimation in regression models: An annotated bibliography. *Communications in Statistics, Part A - Theory and Methods*, 13, 513-541.

DIELMAN T.E. AND PFAFFENBERGER R.C.(1983): LAV estimation with correlated independent variables. *American Statistical Association, Business and Economic Statistics, Proceedings,* 709-713.

DIGBY P.G.N.(1992): Personal communication.

DIXON L.C.W.(1972): Quasi-Newton algorithms generate identical points. *Mathematical Programming,* 2.

DOORNBOS R.(1981): Testing for a single outlier in a linear model *Biometrics,* 37, 705-711.

DRAPER N.R. AND JOHN J.A.(1981): Influential observations and outliers in regression. *Technometrics*, 23, No. 1, February 1981, 21-26.

DRAPER N.R. AND SMITH H.(1981): *Applied Regression Analysis.* John Wiley & Sons, New York. (2nd ed).

DURBIN J. AND WATSON G.S.(1950): Testing for serial correlation in least squares regression I. *Biometrika*, 37, 409-428.

DURBIN J. AND WATSON G.S.(1951): Testing for serial correlation in least squares regression II. *Biometrika*, 38, 159-178.

DURBIN J. AND WATSON G.S.(1971): Testing for serial correlation in least squares regression III. *Biometrika*, 58, 1-42.

EKBLOM H.(1973): Calculation of linear best L_p-approximations. *BIT*, 13, 292-300.

EKBLOM H.(1974): L_p-methods for robust regression. *BIT*, 14, 22-32.

FAREBROTHER R.W.(1985): Unbiased L_1 and L_∞ estimation. *Communications in Statistics, Part A - Theory and Methods*, 14, 1941-1962.

FARRAR D.E. AND GLAUBER R.R.(1967): Multicollinearity in Regression Analysis: The Problem Revisited. *Review of Economics and Statistics*, 49, 92-107.

FISHER J.(1981): An algorithm for discrete linear L_p approximations. *Numerische Mathematik* 38, 129-139.

FLETCHER R.(1970): A new approach to variable metric algorithms. *Computer Journal* 13, 317-322.

FLETCHER R. AND POWELL M.J.D.(1963): A rapidly convergent method for minimization. *Computer Journal*, 6, 163-168.

FORSYTHE A.B.(1972): Robust estimation of straight line regression coefficients by minimizing p-th power deviations. *Technometrics* 14, 159-166.

FRISCH R.(1934): *Statistical Confluence Analysis by Means of Complete Regression Systems.* Oslo: Universitetets Okonomiske Institutt, Oslo, Norway.

FULLER W.A. AND RAO J.N.K.(1978): Estimation for a linear regression model with unknown diagonal. *Annals of Statistics,* 6, 1149-1158.

GAUSS C.F.(1806): II Comet von Jahr 1805. Monatliche Correspondenz zur Befördering der Erd-und Himmelskunde, 14, 181-186.

GENTLE J.E., KENNEDY W.J. AND SPOSITO V.A (1977): On least absolute values estimation. *Communications in Statistics, Part A - Theory and Methods*, 6, 313-328.

GENTLEMAN J.F. AND WILK M.B.(1975): Detecting Outliers II. Supplementing the direct analysis of residuals. *Biometrics,* 31, 387-410.

GLEJSER H.(1969): A new test for heteroscedasticity. *Journal of the American Statistical Association*, 64, 316-323.

GOLDBERGER A.S.(1964): *Econometric Theory*, Wiley, London.

GOLDER E.R. AND SETTLE (1976): The Box-Muller method for generating pseudo-random normal deviates. *Applied Statistics.* 5, 12-20.

GOLDFARB D.(1970): A Family of variable metric methods derived by variational means. *Mathematics of Computation*, 24, 23-26.

GOLDFELD S.M. AND QUANDT R.E.(1965): Some tests for homoscedasticity. *Journal of the American Statistical Association*, 60, 539-547.

GOLDFELD S.M. AND QUANDT R.E.(1972): Nonlinear methods in Econometrics. Amsterdam: North-Holland.

GOLUB G.H., KLEMA V. AND STEWART G.W.(1976): Rank degeneracy and least squares problems. *Technical Report* TR-751, Dept. Computer Science, Univ. Maryland.

GOLUB G.H. AND VAN LOAN C.F.(1980): An analysis of the total least squares problem. *SIAM Journal on Numerical Analysis*, 17, 883-893.

GOLUB G.H. AND VAN LOAN C.F.(1983): *Matrix Computations.* Johns Hopkins University Press. Baltimore, MD, 1983.

GONIN R. AND MONEY A.H.(1985): Nonlinear L_p-norm estimation: Part I - On the choice of the exponent, p, where the errors are additive. *Communications in Statistics, Part A - Theory and Methods*, Volume 14, Part 4, 827-840.

GONIN R. AND MONEY A.H.(1989): *Nonlinear L_p-Norm Estimation.* Marcel Dekker, Inc., New York and Basel.

GRAY J.B. AND LING R.F.(1984): K-clustering as a detection tool for influential subsets in regression (with discussion). *Technometrics* 26, 305-330.

GRAYBILL F.A.(1976): *Theory and Applications of the Linear Model.* Duxbury press, Belmont, California.

GUNST R.F.(1983): Regression analysis with multicollinear predictor variables: definition, detection, and effects. *Communications in Statistics, Part A-Theory Methods,* 12, no. 19, 2217-2260.

GUNST R.F. AND MASON R.L.(1980): *Regression analysis and its application. A data-oriented approach.* Statistics: Textbooks and Monographs, 34. Marcel Dekker, New York.

HAND M.L.(1978): Aspects of linear regression estimation under the criterion of minimizing the maximum absolute residual. Ph.D thesis, Iowa State University.

HARVEY A.C.(1978): On the unbiasedness of robust regression estimators. *Communications in Statistics, Part A - Theory and Methods*, 7, 779-783.

HARTER H L (1977): The nonuniqueness of absolute values regression. *Communications in Statistics, Part B - Simulation and Computation*, 6, 829-838.

HARTLEY H.O., RAO J.N.K. AND KIEFER G.(1969): Variance estimation with one unit per stratum. *Journal of the American Statistical Association*, 64, 841-851.

HAWKINS D.M.(1980): *Identification of outliers*. London: Chapman and Hall.

HAWKINS D.M., BRADU D. AND KASS G.V.(1984): Location of several outliers in multiple regression data using elemental sets. *Technometrics* 26, 197-208.

HENRIKSSON S.(1972): On a Generalization of L_p-Approximation and Estimation, Thesis, Dept. of Computer Sciences, Lund University, Sweden.

HINKLEY D.V.(1977): Jackknifing in unbalanced situations. *Technometrics*, 19, No. 3, 285-292.

HOAGLIN D.C. AND WELSCH R.E.(1978): The Hat matrix in regression and ANOVA. *The American Statistician*, 32, 17-22.

HOERL A.E., KENNARD R.W. AND BALDWIN K.F.(1975): Ridge regression: some simulations. *Communications in Statistics*, 4, 105-123.

HOERL R.W., SCHUENEMEYER J.H. AND HOERL A.E.(1986): A simulation of biased estimation and subset selection regression techniques. *Technometrics*, 28, 369-380.

HOGG R.V.(1974): Adaptive robust procedures: A partial review and some suggestions for future applications and theory. *Journal of the American Statistical Association*, 69, 909-927.

HORN R.A. AND JOHNSON C.R.(1987): *Matrix Analysis.* Cambridge University Press, Cambridge.

HUBER P.J.(1973): *Robust Statistics.* John Wiley and Sons.

HUBER P.J.(1981): *Robust Statisics*, 2nd ed. John Wiley and Sons.

IBM CORPORATION(1968): IBM System/360. Scientific subroutine package.

JACOBS D.(1977): The state of the art in numerical analysis. Academic press, London.

JOHNSON N.L. AND KOTZ S.(1970): *Distributions in statistics.* John Wiley and Sons, New York.

JOHNSTON J.(1963): *Econometric Methods.* McGraw-Hill, New York.

JUDGE G.G. AND BOCK M.E.(1978): *The statistical implications of pre-test and Stein-rule estimators in econometrics.* North-Holland, Amsterdam.

JUDGE G.G. AND TAKAYAMA T.(1966): Inequality restrictions in regression analysis. *Journal of the American Statistical Association,* 61, 166-181.

JUDGE G.G., GRIFFITHS W.E., HILL R.C. AND LEE T.C.(1980): The theory and practice of econometrics. John Wiley and Sons, Inc., New York.

KADIYALA K.(1984): A class of almost unbiased and efficient estimators of regression coefficients. *Economics Letters,* 16, 293-296.

KAKWANI N.C.(1967): The unbiasedness of Zellner's seemingly unrelated regression equations estimators. *Journal of the American Statistical Association,* 62, 141-142.

KENDALL M.G. AND STUART A.(1966): *The advanced theory of statistics.* Volume II. London, Griffin.

KENNARD R.W.(1971): A note on the C_p statistic. *Technometrics,* 13, 899-900.

KIOUNTOUZIS E.A.(1971): Optimal L_p approximation. Techniques and data analysis. *Extrait du Bull. De La Soc. Mathematique de Grece, Novelle Serie*, Tome 12, Fasc. 1, 191-206.

KIVIET J.F.(1980): Effects of ARMA errors on tests for regression coefficients: Comments on Vinod's article; Improved and additional results. *Journal of the American Statistical Association*, 75, 353-358.

KNUTH D.E.(1969): The art of computer programming. Vol. 2: Seminumerical Algorithms. Addison-Wesley Publishing Company.

KRIPKE R. AND RIVLIN T.J.(1965): Approximations in the metric of $L_1(X,\mu)$. *Transactions of the American Mathematical Society,* 119,101-122.

LAPLACE P.S.(1786): Exposition du systeme du monde, Paris.

LAWRENCE K.D.(1979): A comparison of minimum absolute deviations and least squares estimation methods with data structures that violate traditional regression. Rutgers University the State U. of New Jersey.

LAWLESS J.F. AND WANG P.(1976): A simulation study of ridge and other regression estimators. *Communications in Statistics,* 5, 307-323.

LAWSON C.L. AND HANSON R.J.(1974): *Solving Least-Squares problems.* Prentice-Hall, Inc., Englewood Cliffs, N.J.

LEE W.W. AND BIRCH J.B.(1988): Fractional principal components regression: A general approach to biased estimators. *Communications in Statistics, Part B - Simulation and Computation,* 17, 713-727.

LEGENDRE A.M.(1805): Nouvelles methodes pour la determination des orbites des cometes Courcier Paris. *Appendice sur la methode des moindres quarres*, 72-80.

LIEW C.K.(1976): Inequality constrained least squares estimation. *Journal of the American Statistical Association*, 71, 746-751.

MACLAREN M.D. AND MARSAGLIA G. (1965): Uniform random number generators. *Journal of the Association for Computing Machinery*, 12, 83-89.

MALLOWS C.L.(1964): Choosing Variables in a Linear Regression: A Graphical Aid, presented at the Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, Kansas, May 7-9, 1964.

MANTEL N.(1969): Restricted least squares regression and convex quadratic programming. *Technometrics*, 11, 763-773.

MARQUARDT D.W.(1970): Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation. *Technometrics*, 12, 591-612.

MARQUARDT D.W.(1980): You should standardize the predictor variables in your regression models (discussion of a paper by G. Smith and F. Campbell). *Journal of the American Statistical Association*, 75, 87-91.

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

MASON R.L. AND GUNST R.F.(1985): Outlier-Induced Collinearities. *Technometrics* 27, 401-407.

MASON R.L. GUNST R.F. AND WEBSTER J.T.(1975): Regression analysis and problems of multicollinearity. *Communications in Statistics*, 4, 277-292.

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

McDONALD G.C. AND GALARNEAU D.I.(1975): A Monte-Carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association*, 70, 407-416.

McKEAN J.W. AND SCHRADER R.M.(1987): Least absolute errors analysis of variance. In: Y.Dodge, ed. *Statistical data analysis - Based on the L_1-norm and related methods*, North Holland, Amsterdam 297-305.

MERLE G. AND SPäTH H.(1973): Computational experiences with discrete L_p-approximation. *Computing* 12, 315-321.

MILLER A.J.(1990): *Subset Selection in Regression* Chapman and Hall.

MITTELHAMMER R.C.(1984): Restricted least squares, pre-test, OLS and Stein rule estimators, risk comparisons under model misspecification. *Journal of Econometrics*, 25, 151-164.

MONEY A.H., AFFLECK-GRAVES J.F., HART M.L. AND BARR G.D.I.(1982): The linear regression model: L_p norm estimation and the choice of p. *Communications in Statistics, Part B - Simulation and Computation*, 11, 89-109.

MOOD A.M., GRAYBILL F.A. AND BOES D.C.(1974): Introduction to the theory of Statistics. McGraw-Hill Kogakusha, Ltd.

MULLET G.M.(1976): Why regression coefficients have the wrong sign. *Journal of Quality Technology*, 8, 121-126

NAGAR A.L. AND KAKWANI N.C.(1964): The bias and moment matrix of a mixed regression estimator. *Econometric*, 32, 174-182.

NEAVE H.R.(1972): A random number package. *Computer Applications in the Natural and Social Sciences*, 14.

NEAVE H.R.(1973): On using the Box-Muller transformation with multiplicative congruential pseudo-random number generators. *Applied Statistics*, 22, 92-97.

NEWHOUSE J.P. AND OMAN S.D.(1971): An evaluation of Ridge estimators. *Technical report* No. R-716-PR, The Rand Corporation, Santa Monica, Calif.

NETER J. AND WASSERMAN W.(1974): *Applied Linear Statistical Models.* Richard D. Irwin, Inc. Ontario.

NOMURA M.(1988): On the almost unbiased ridge regression estimator. *Communications in Statistics, Series B - Simulation and Computation*, 17, 729-743.

NYQUIST H.(1980): Recent studies on L_p-norm estimation. Doctoral thesis, University of Umea, Sweden.

NYQUIST H.(1983): The optimal L_p norm estimator in linear regression models. *Communications in Statistics, Part A - Theory and Methods*, 12, 2511-2524.

OHTANI K.(1986): On small sample properties of the almost unbiased generalized ridge estimator. *Communication in Statistics, Part A - Theory and Methods*, 15, 1571-1578.

PAYNE R.W.(chair)(1988): *GENSTAT 5 Reference manual*, Oxford University Press, Oxford.

PORTER M.A. AND WINSTANLEY D.J.(1979): Remarks on AS 110: L_p norm fit of a straight line. *Applied Statistics*, 28 112-113.

POWELL M.J.D.(1971): Recent advances in unconstrained optimization. *Mathematical Programming*, 1, 26-57.

PUTTER J.(1967): Orthonormal bases of error spaces and their use for investigating the normality and variances of residuals. *Journal of the American Statistical Association*, 62, 1022-1036.

PRESS S.J. (1972): *Applied multivariate analysis.* Holt, Rinehart and Winston, Inc., New York.

PRESS W.H., FLANNERY B.P., TEUKOLSKY S.A. AND VETTERLING W.T. (1985): *Numerical Recipes: The Art of Scientific Computing.* Cambridge University Press.

QUENOULLE M.H.(1956): Notes on bias in estimation. *Biometrika*, 43, 353-360.

RANDALL J.H. AND RAYNER A.A.(1987): The accuracy of least squares calculations with the Cholesky algorithm. *Technical report*, University of Natal.

RAO C.R.(1970): Estimation of heteroscedastic variances in linear models. *Journal of the American Statistical Association*, 65, 161-172.

RAO C.R.(1973): Linear statistical inference and its applications. John Wiley & Sons.

RAWLINGS J.(1988): *Applied regression analysis: a research tool.* Wadsworth & Brooks/Cole: Pacific Grove, California.

RICE J.A.(1988): *Mathematical statistics and data analysis.* Wadsworth & Brooks/Cole Advance Books & Software, California.

RICE J.R.(1964): *The approximation of functions, vol 1: Linear Theory.* Addison-Wesley, Reading, Massachusetts.

RIVLIN T.J. (1969): *An introduction to the approximation of functions.* Blaisdell publishing company, New york.

RONNER A.E. (1977): P-norm estimators in a linear regression model. Doctoral thesis, Rijksuniversiteit te Groningen.

ROSENBERG B. AND CARLSON D.(1971): The sampling distribution of least absolute residuals regression estimates. Working paper no IP-164, Institute of Business and Economic research, University of California, Berkeley.

ROSENBERG B. AND CARLSON D. (1977): A simple approximation of the sampling distribution of least absolute residuals regression estimates. *Communication in Statistics, Part B - Simulation and computations*, 6,421-437.

SCHALL R. AND DUNNE T.T.(1987b): Variance inflation and collinearity in regression. *Technical Report* 5/87, Institute for Biostatistics of the South African Medical Research Council, Tygerberg, Republic of South Africa.

SCHMIDT P.(1967): *Econometrics*. Marcel Dekker, Inc., New York.

SCHLOSSMACHER E.J.(1973): An iterative technique for absolute deviation curve fitting. *Journal of the American Statistical Association*, 68, 857-859.

SEARLE S.R.(1971): *Linear Models*. John Wiley & Sons, New York.

SHANNO D.F.(1970): Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, 24, 647-656.

SIELKEN R.L. AND HARTLEY H.O.(1973): Two linear programming algorithms for unbiased estimation of linear models. *Journal of the American Statistical Association*, 68, 639-641.

SILVEY S.D.(1969): Multicollinearity and imprecise estimation. *Journal of the Royal Statistical Society, Series* B, 31, 539-552.

SIMON S.D. AND LESAGE J.P.(1988): The impact of collinearity involving the intercept term on the numerical accuracy of regression. *Computer Science in Economics and Management* 1, 137-152.

SINGH B., CHAUBEY Y.P. AND DWIVEDI T.D.(1986): An almost unbiased ridge estimator. *Sankhya, Series B*, 48, 342-346.

SNYMAN J.A.(1993): LFOPCON. Department of Mechanical Engineering, University of Pretoria.

SPOSITO V.A.(1975): Linear and nonlinear programming. The Iowa State University Press, Ames, Iowa.

SPOSITO V.A.(1976): Minimizing the maximum absolute deviation. *Sigmap,* 20, 51-53.

SPOSITO V.A. (1982): On unbiased L_p regression estimators. *Journal of the American Statistical Association*, 77, 652-654.

SPOSITO V.A., HAND M.L. AND SKARPNESS B.(1983): On the efficiency of using the sample kurtosis in selecting optimal L_p estimators. *Communications in Statistics, Part B - Simulation and Computation*, 12, 265-272.

SPOSITO V.A., KENNEDY W.J. AND GENTLE J.E.(1977): L_p norm fit of a straight line. *Applied Statistics,* 26, 114-118.

SPOSITO V.A. AND TVEITE M.D.(1986): On the estimation of the variance of the median unsed in L_1 linear inference procedures. *Communications in Statistics, Part A - Theory and Methods*, Volume 15, part 4, 1367-1375.

STEWART G.W.(1973): *Introduction to matrix computations.* Academic Press, New York.

STEWART G.W.(1987): Collinearity and least squares regression. With discussion by D.A. Belsley, A.S. Hadi, D.W. Marquardt, P.F. Velleman, R.A. Thisted, and with a reply by the authors. *Statistical Science.* 2, no. 1, 68-100.

SWAMY P.A.V.B AND MEHTA J.S.(1976): Minimum average risk estimators for coefficients in linear models. *Communications in Statistics,* Ser. A, 5, 803-818.

THEIL H.(1963): On the use of incomplete prior information in regression analysis. *Journal of the American Statistical Association*, 58, 401-414.

THEIL H.(1971): Principles of Economics. New York, Wiley.

THIART C.(1990): Collinearity and consequences for estimation: A study and simulation. MSc Thesis, University of Cape Town.

THIART C., DUNNE T.T., TROSKIE C.G. AND CHALTON D.O.(1993): A simulation study of biased estimators against the ordinary least square estimator. *Communications in Statistics, Simulation and Computation*, 22, No 2, 569-589.

TORO-VIZCARRONDO C. AND WALLACE T.D.(1968): A test of the mean square error criterion for restrictions in linear regression. *Journal of the American Statistical Association*, 63, 558-572.

TOUTENBERG H.(1982): *Prior information in linear models.* John Wiley & Sons.

TROSKIE C.G.(1971): Regression and correlation. *Proceedings of the Third Symposium on Mathematical Statistics,* NRIMS, Wisk. 89, 21-50.

TROSKIE C.G.(1990): Personal communication.

TUKEY J.W.(1958): Bias and confidence in not quite large samples. *The Annals of Mathematical Statistics,* 29, 614.

VAN DER GENUGTEN B.B.(1993): Efficiency of iterated WLS in the linear model with completely unknown heteroskedasticity. *Statistica Neerlandica*, Volume 47, nor 2, 111-125.

VINOD H.D. AND ULLAH A.(1981): *Recent advances in regression methods.* Marcell Dekker Inc., New York.

WAGNER H.M.(1959): Linear programming techniques for regression analysis. *Journal of the American Statistical Association*, 54, 206-212.

WALKER E. AND O'BRIEN R.G.(1992): Using restricted least squares to delineate the effects of misspecification in linear models. *The Statistician,* 41, 467-476.

WALLACE T.D.(1972): Weaker Criteria and tests for linear restrictions in regression. *Ecomometrica,* 40, 689-698.

WATSON G.A.(1980): *Approximation theory and numerical methods.* John Wiley, New York.

WHITE H.(1980): Heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrica*, 48, 817-838.

WICHMANN B.A. AND HILL I.D.(1982): Algorithm AS 183: An efficient and portable pseudo-random number generator. *Applied Statistics*, 31, 188-189.

WICHERN D.W. AND CHURCHILL G.A.(1978): A comparison of Ridge estimators. *Technometrics*, 20, 301-311.

WU C.F.J.(1986): Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, Vol. 14, No. 4, 1261-1295.

YANCEY T.A., JUDGE G.G. AND BOCK M.E.(1973): Wallace's weak mean square error criterion for testing linear restrictions in regression: A tighter bound. *Econometrica,* 41, 1203-1206.

Appendix I

USEFUL FORMULAE AND DERIVATIONS

I.1 If $x \sim N(\mu,V)$ then for A symmetric and conformable

$$E(x'Ax) = tr(AV) + \mu'A\mu$$
$$V(x'Ax) = 2tr(AV)^2 + 4\mu'AVA\mu$$

(Searle (1971, pp55-57))

I.2 Let A be an nxn matrix with eigenvalues $\lambda_1,\lambda_2,\ldots\lambda_n$, then

$$tr(A) = \Sigma\ \lambda_i .$$

If A is symmetric and $\lambda_i > 0 \quad \forall_i$

then $tr\ (A^{-1}) = \Sigma\ \lambda_i^{-1}$

(Graybill (1969, pp223-225))

APPENDIX II

II.1 Let $\tau_i^2 = \lambda_i^2/\sigma^2$ and $\nu = n-r$, then the first moment of $[\hat{\delta}_0]_i$ is

$$E([\hat{\delta}_0]_i) = [\delta_i \exp(-\tau_i^2/2)/\{\sqrt{\pi}\ \Gamma(\nu/2)\}] \times \sum_{g=0}^{\infty} [2^{g+1}(\tau_i^2)^g \Gamma(g+(\nu+3)/2)/(2g+1)!] \times \int_0^1 [(t+2(1-t)/\nu)/(t+(1-t)/\nu)^2] \times t^{g+3/2}(1-t)^{\nu/2-1}dt$$

II.2 For τ_i^2 and ν as above the second moment of $[\hat{\delta}_0]_i$ is

$$E([\hat{\delta}_0]_i^2) = [\delta_i^2 \exp(-\tau_i^2/2)/\{\sqrt{\pi}\ \Gamma(\nu/2)\}] \times \sum_{g=0}^{\infty} [2^{g+1}(\tau_i^2)^{g-1} \Gamma(g+(\nu+3)/2)/(2g)!] \times \int_0^1 [(t+2(1-t)/\nu)^2/(t+(1-t)/\nu)^4] \times t^{g+5/2}(1-t)^{\nu/2-1}dt$$

The variance of the random variable is hence easily obtained.

II.3 If we minimize the TMSE of the shrinkage estimator (given in table 2.4) with respect to d, we obtain:

$$\frac{\delta\ \mathrm{TMSE}(\hat{\beta}_{SH})}{\delta\ d} = \sigma^2 2d \sum_{i=1}^{r} 1/\lambda_i - 2(1-d)\beta'\beta = 0$$

thus

$$\sigma^2 d \sum_{i=1}^{r} 1/\lambda_i + d\beta'\beta = \beta'\beta$$

$$d = \frac{\beta'\beta}{\beta'\beta + \mathrm{tr}(\mathrm{Var}(\hat{\beta}))}$$

II.4 In Table 2.2 the fractional principal component (FPC) estimator is defined as $F\hat{\delta}$. In the iterative fractional principal component estimator (FPCG) denoted by $\hat{\delta}_{FPCG}$, with the limiting fraction matrix F_{PCG}, is

$$\hat{\delta}_{FPCG} = F_{PCG}\hat{\delta}$$

where $F_{PCG} = Diag(f^*_{1,PCG},\ldots,f^*_{r,PCG})$, and $f^*_{j,PCG} = \lim_t [f_{j,PC}(t+1)]$,

for $j = 1,\ldots,r$. The FPCG estimator is thus derived from the combined concepts of the PC estimator and of the iterative generalized ridge estimator. For the FPCG estimator the iterative scheme of the optimal fraction is:

$$f_{j,GR}(t+1) = \frac{\lambda_j}{\lambda_j + s^2/[\hat{\delta}_K(t)]^2_j}, \qquad t=0,1,2,\ldots$$

where t denotes the iteration number, s^2 is the OLS estimate of σ^2 and $[\hat{\delta}_K(t)]_j$ is the generalized ridge estimate of δ_j at the t-th iteration with $[\hat{\delta}_K(0)]_j = \hat{\delta}_j$. The iteration continues until there is stability achieved in the length of the generalized ridge estimator ($[\hat{\delta}_K(t)]_j$).

In the presence of collinearity the starting values of OLS may be severely perturbed and therefore it may be more beneficial to use a biased estimator as initial value. If one considers a PCE as the initial value, then the iterative scheme $f_{j,GR}(t+1)$ can be used to compute the fractions, where the starting value for $[\hat{\delta}_K(0)]_j$ would be $[\hat{\delta}_{PC}]_j$ and s^2 is replaced with $s^2(t) = (Y-Za(t))'(Y-Za(t))/(n-r)$. Here $a(t)$ is the estimate of δ at the t-th iteration, $a(0) = \hat{\delta}_{PC}$. Fractions obtained in this way will be denoted by $f_{j,PC}(t+1)$.

II.5 The second biased estimator due to Lee and Birch (1988) is based on the iterative ridge estimator concept. In this scheme the fraction becomes

$$f_{j,R}(t+1) = \frac{\lambda_j}{\lambda_j + s^2/[\hat{\delta}_R(t)'\hat{\delta}_R(t)/r]}, \qquad t = 0,1,2,\ldots$$

where $\hat{\delta}_R(t)$ is the ridge estimate of δ at the t-th iteration with $\hat{\delta}_R(0) = \hat{\delta}$. Just as did the authors of the FPCG estimator, Lee and Birch (1988) replaced the OLSE by the PCE ($\hat{\delta}_R(0) = \hat{\delta}_{PC}$) and the s^2 by $s^2(t)$, and denoted the fractions by $f_{j,PCR}(t+1)$. The resulting estimator, denoted by $\hat{\delta}_{FPCR}$, is defined as:

$$\hat{\delta}_{FPCR} = F_{PCR}\hat{\delta}$$

where $F_{PCR} = \text{Diag}(f^*_{1,PCR},\ldots,f^*_{r,PCR})$, and $f^*_{j,PCR} = \lim_t [f_{j,PCR}(t+1)]$, for $j = 1,\ldots,p$.