LINEAR REGRESSION TECHNIQUES FOR IDENTIFYING INFLUENTIAL DATA

and Applications in Commercial Data Analysis

by

Michael Jacobs

Presented to the
UNIVERSITY OF CAPE TOWN
in fulfillment of the
requirements for the degree of
DOCTOR OF PHILOSOPHY

February 1983

University of Cape Town

The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.

University of Cape Town

I wish to express my sincere thanks to the following:

To my supervisor, Professor Cas Troskie, for many first class ideas, and for guidance over the years;

To Mrs. Cousins, for typing this thesis during her holidays;

To Professor Arthur Money and Professor John Simpson, for their encouragement and support.

I certify that the thesis is my own work and that all references are accurately reported.

MICHAEL JACOBS.

SYNOPSIS

Recent literature contains many publications on techniques for identifying extreme data points (outliers) and influential observations or groups in sample data sets. This thesis begins by reviewing the statistics and distributional properties of the standard techniques, viz. the standardized residual as a test for outliers, and Cook's distance as a measure of influence. An outlier test which is distributionally neater than the standardized residual is proposed.

In practical applications, ordinary least squares regression is often inappropriate, and the use of biased estimators may be preferable. In this thesis, the existing theory is extended to several alternative regression techniques. Ridge regression and generalized inverse regression are suitable techniques when the cross product matrix is ill-conditioned. Restricted least squares regression, with exact or stochastic prior information, is used in many econometric applications. Models with selected variables are used to eliminate design faults or to reduce computational effort. New statistics are developed for all these techniques, the distributional results are proved, and computational formulae are developed.

Computational problems may arise in the actual use of the various techniques, and these are investigated. Computer programs written in BASIC and suitable for microcomputer use are presented, making the techniques accessible to vitually any commercial environment.

The performance of the various techniques is examined, using a controlled simulation study and a number of practical data sets drawn from several areas of South African commerce. This is, as far as can be ascertained, the first extensive practical South African study on the effects of influential data.

It is shown that the presence of outliers or influential data can bias the results of any study significantly. It is recommended that no data analysis should be attempted without a preliminary scan of outliers and influential observations. The techniques presented can be used advantageously even in data sets where the ultimate analysis does not involve linear regression. It is shown that influential data are not merely of nuisance value in the analysis, but may contain valuable information in their own right.

CONTENTS

Chapter 1

INTRODUCTION

The business analyst is constantly faced with problems that require the collection of some form of sample data. Decisions will be based on the results of the analysis of this data.

Unfortunately, all stochastic sample data sets possess inherent variability, and all observations in the sample will not have the same influence upon the results of the analysis. Thus it may well be that conclusions are ultimately heavily biased by a small subset of extreme observations, rather than being representative of the whole population. The analyst should be aware of the presence of extreme cases in his data sets, and the effects of the inclusion thereof.

A number of recent books and other publications have dealt with the problem of detecting influential data in multivariate linear regression models. For examples, see Barnett and Lewis (1978), Hawkins (1980) or Belsley, Kuh and Welsch (1980). Linear regression is a statistical technique widely used in time series analysis, forecasting, econometrics and many other applications. In this study, we shall demonstrate that linear regression techniques for detecting influential data may be applied advantageously to a variety of commercial data sets, as a preliminary screening procedure, even if the

ultimate analysis of the data does not involve regression analysis.

Outliers and influential observations:

Any set of sample data drawn for statistical analysis will possess some inherent variability. Although a certain spread in data observations is expected, it will often be found that some data points are very far removed from the rest of the data set. Extreme values are generally referred to as outliers.

In models where the technique of multiple linear regression is applied, observations are vectors of the form $(y_i, x_{i1}, x_{i2},\ldots,x_{iq})$. It is generally not immediately obvious which vector observations are outliers. One high or low value amongst the $q+1$ elements would not necessarily make the whole vector substantially different from the rest of the data set. More sophisticated tests for the detection of outliers will have to be used.

If outlying observations are detected in a data sample, how should they be treated? Outliers may be observation errors, but, on the other hand, they may be valid results in the tails of the population distribution.

The analyst should thus examine carefully all outlying data points, in case genuine execution errors have been committed. Examples of execution errors are inaccurate measurements, transcription errors, changes in physical conditions and

sampling from biased populations. Because one really bad data error can wreck an entire study, the best action to take upon the proven detection of a genuine error would be to discard the observation/s completely. The sampling act has failed and no useful information has been obtained, unless we wish to study the errors for their own sakes.

In most practical situations, however, it is impossible to tell whether an outlying observation is a real error or merely an extreme sample value. It may still be advantageous to exclude outliers, as the analysis can be heavily biased by even one extreme value, especially if the sample size is small.

The analyst must obviously be aware of the effect of the exclusion of an outlying data point. It may well be that the removal of a single observation from a large data set will not change the results of the analysis significantly. Conversely, the results may be changed drastically.

If the removal of a single data point causes significant changes in the results of the analysis, then that data point will be referred to as an influential observation. An observation may be influential without being an outlier. A meaningful quantitative measure of influence must be defined.

Scope of the study:

The literature currently available presents various techniques for detecting influential data using ordinary least squares

regression analysis. However, in many practical situations the data possess collinearity problems, and there will be weaknesses in the ordinary least squares regression techniques. In such situations, techniques such as ridge regression and generalized inverse regression would be more appropriate. We shall therefore develop a theoretical basis and propose statistics for detecting outliers and influential observations in a ridge or generalized inverse regression framework.

In economic applications, models incorporating prior information are often used, as are models involving selected variables. Hence we shall also develop a similar theory for restricted least squares regression and for models with selected variables.

A primary objective of this study is to examine how the various techniques may be applied in practical commercial situations. Several illustrative data sets will be presented, and it will be shown that highly influential observations are not merely of nuisance value in the analysis, but are of considerable interest in their own right.

The theoretical development and distributional properties of the various statistics are essential, but the practicalities of using them are ultimately more important to the data analyst. Therefore, we shall present a discussion of some of the computational considerations of using the techniques in a computer analysis, particularly microcomptuer analysis.

Original programs have been written in BASIC, and structured for easy modification to virtually any computer, or even microcomputer. The programs are self-documented and reproduced in full.

Chapter 2

OUTLIERS IN LINEAR REGRESSION

Suppose we fit by least squares the model

$$y = X\beta + \varepsilon \tag{2.1}$$

where y is an $n\times 1$ vector of responses

X is an $n\times q$ matrix of known constants

β is an $q\times 1$ vector of regression parameters

ε is an $n\times 1$ vector of random errors.

Assume also that $E(\varepsilon) = 0$ and $var(\varepsilon) = \sigma^2 I$.

The least squares estimate of the unknown vector β is given by

$$b = (X'X)^{-1}X'y \ . \tag{2.2}$$

In order to locate outlying observations, we shall need to examine the vector of observed residuals, r, where

$$\begin{aligned} r &= y - \hat{y} \\ &= y - Xb \\ &= y - X[(X'X)^{-1}X'y] \end{aligned}$$

$$\therefore \quad r = (I-V)y \tag{2.3}$$

where

$$V = X(X'X)^{-1}X'$$

is commonly known as the projection or "hat" matrix, and will feature prominently in later analysis.

Note that $r = y - Vy$

$$= X\beta+\varepsilon - V(X\beta+\varepsilon)$$

$$= X\beta+\varepsilon - X(X'X)^{-1}X'X\beta - V\varepsilon$$

$$= (I-V)\varepsilon \, . \qquad (2.4)$$

Hence $E(r) = 0$ as $E(\varepsilon) = 0$

$$\text{var}(r) = (I-V)\sigma^2 \qquad (2.5)$$

as $\text{var}(\varepsilon) = \sigma^2 I$ and $(I-V)$ is idempotent.

Thus, for each element of the vector r

$$\text{var}(r_i) = (1-v_{ii})\sigma^2$$

$$= \sigma^2 - \text{var}(\hat{y}_i) \qquad i = 1,\ldots,n$$

where

r_i is the i^{th} element of the residual vector r

v_{ii} is the i^{th} diagonal element of the projection matrix V.

Specifically,

$$v_{ii} = x_i(X'X)^{-1}x_i' \qquad (2.6)$$

where x_i is the i^{th} row of X.

Observations associated with residuals of large absolute magnitude would naturally be suspected as being potential outliers. However, the variances of the residuals are not equal. The most commonly used statistic for testing whether a single vector observation $(y_i, x_{i1},\ldots,x_{iq})$ is an outlier for some i is

$$z_i = \frac{r_i}{s\sqrt{1-v_{ii}}} \qquad i = 1,\ldots,n \tag{2.7}$$

where $s^2 = \Sigma r_i^2/(n-q) = r'r/(n-q)$. (2.8)

This statistic is known as the standardized residual. It is the ratio of the residual r_i to the sample standard deviation of r_i. We can define z_i to be zero when $(1-v_{ii})s^2 = 0$, as it can easily be shown that $r_i = 0$ in this case.

Let $Z = \max|z_i|$. A large value of Z would indicate that the observation associated with Z is a potential outlier. How large must Z be to indicate the presence of a significant outlier? Critical points for the statistic Z are required. However, the distribution of Z is unknown. Prescott (1975) developed a first-order Bonferroni upper bound for Z, and suggested that this bound would be adequately close to the true critical value. Lund (1975) compiled tables for this upper bound. Doornbos (1981) showed that the outlier test based on this Bonferroni upper bound has significance between α and $\alpha - \frac{1}{2}\alpha^2$. Cook and Prescott (1981) provide a relatively simple test for the accuracy of the bound.

In the model

$$y = X\beta + \theta u_i + \varepsilon \tag{2.9}$$

where u_i is an $n\times 1$ vector of zeros with a single 1 in the i^{th} position, the normal theory likelihood ratio test for $\theta = 0$ is given by

$$F_i = (n-q-1)z_i^2/(n-q-z_i^2). \qquad (2.10)$$

where, under normality, F_i follows an F distribution with 1 and n-q-1 degrees of freedom. As F_i is a monotonic increasing function of z_i, it provides another test statistic for detection of outliers.

Note that in (2.8), the estimate of the residual variance is a function of all observed residuals, r_i, $i = 1,\ldots,n$. If the i^{th} vector observation is suspected of being an outlier, then it would be wise to exclude its effects from the calculation of z_i. Belsley, Kuh and Welsch (1980) suggest an estimate of σ^2 free of the i^{th} observation

$$s_{-i}^2 = \frac{1}{n-q-1} \sum_{k \neq i} (y_k - x_k b_{-i})^2 \qquad (2.11)$$

where x_k is the k^{th} row of X

b_{-i} is the least squares estimate of β calculated with the i^{th} observation deleted.

Beckman and Trussell (1974) show that (2.11) can be written in the form

$$(n-q-1)s_{-i}^2 = (n-q)s^2 - r_i^2/(1-v_{ii})$$

and therefore

$$s_{-i}^2 = \frac{1}{n-q-1} [r'r - r_i^2/(1-v_{ii})]. \qquad (2.12)$$

Using this estimate of σ^2, another statistic for detecting outliers would be

$$t_i = \frac{r_i}{s_{-i}\sqrt{1-v_{ii}}} \qquad i = 1,\ldots,n. \qquad (2.13)$$

Ellenberg (1973) has shown that s_{-i} is independent of r_i, and also that $s^2_{-i} \sim \chi^2_{n-q-1}$. Hence under normality, the statistic

$$t_i^2 = \frac{r_i^2}{s^2_{-i}(1-v_{ii})}$$

follows the F distribution with 1 and $n-q-1$ degrees of freedom. Alternatively, t_i the "studentized residual" follows the familiar Student's t distribution. Ellenberg (1976) has proved the equivalence of the test statistics z_i and t_i.

Whether we choose to use t_i or t_i^2, we require the upper α/n points of the t (or F) distribution, assuming that the location of outliers is *a priori* unknown. Tables of such points have been prepared for various convenient values of n,q and α, and are presented in appendix A.

Chapter 3

INFLUENTIAL OBSERVATIONS IN LINEAR REGRESSION

If an observation is a (potential) outlier, what will be the effects of removing it from the data set? More generally, any observation can be considered to be influential if important features of the analysis are altered substantially when it is deleted from the data.

Let b_{-i} denote the least squares estimate of β calculated with the i^{th} observation deleted.

$$b_{-i} = (X'_{-i}X_{-i})^{-1}X'_{-i}y_{-i} \qquad (3.1)$$

where X_{-i} is the $(n-1)\times q$ matrix derived from X by deleting the i^{th} row x_i, and y_{-i} is similarly derived from y.

Cook (1977) proposed the following measure of the influence of the i^{th} observation:

$$D_i = \frac{(b_{-i}-b)'X'X(b_{-i}-b)}{qs^2} \qquad (3.2)$$

The statistic D_i can also be written as

$$D_i = \frac{1}{qs^2}(\hat{y}_{-i}-\hat{y})'(\hat{y}_{-i}-\hat{y}) \qquad (3.3)$$

showing that D_i is proportional to the squared Euclidean distance that the estimate of the y vector moves when the

i^{th} observation is deleted. Thus, D_i is referred to as "Cook's distance".

Now, the $100(1-\alpha)\%$ confidence ellipsoid for the unknown parameter β is given by the set of all vectors b^* that satisfy

$$\frac{(b^*-b)'X'X(b^*-b)}{qs^2} \leq F_{1-\alpha}(q,n-q) \qquad (3.4)$$

where $F_{1-\alpha}(q,n-q)$ is the upper α point of the F distribution with q and $n-q$ degrees of freedom.

It is readily seen that D_i provides a measure of the distance between b_{-i} and b in terms of significance levels. For example, if D_i is near to $F_{.4}(q,n-q)$ then the removal of the i^{th} observation has the effect of moving the estimate for β to the edge of the 40% confidence ellipsoid for based on the full data set, which may be cause for some concern. Ideally, we would like all the b_{-i} to remain within say a 5% confidence region.

It is obviously inefficient to perform $n+1$ regressions in order to compute the value of b_{-i}, $i = 1,\ldots,n$. Cook (1977) showed that D_i can be written as

$$D_i = \frac{1}{q}\left(\frac{r_i}{s\sqrt{1-v_{ii}}}\right)^2 \frac{v_{ii}}{1-v_{ii}}$$

$$= \frac{1}{q}\, z_i^2\, \frac{v_{ii}}{1-v_{ii}} \qquad (3.5)$$

where z_i is the i^{th} standardized residual and v_{ii} is the

i^{th} diagonal element of the hat matrix, as before.

The statistic D_i will have a large value if z_i is large or if v_{ii} is large. We have seen in chapter 2 that a large value of z_i indicates a potential outlier. Following the conventions of chapter 2, we would recommend that the estimate s^2_{-i} be used in place of s^2, and hence the measure of influence is to be defined as

$$\Delta_i = \frac{(b_{-i}-b)'X'X(b_{-i}-b)}{qs^2_{-i}} \tag{3.6}$$

$$= \frac{1}{q}\, t_i^2\, \frac{v_{ii}}{1-v_{ii}} \,. \tag{3.7}$$

The statistic Δ_i was proposed by Belsley, Kuh and Welsch (1980). It has precisely the same geometric interpretation as D_i, with degrees of freedom q and $n-q-1$ in (3.4). The distribution of t_i^2 is known to be $F(1,n-q-1)$.

Hoaglin and Welsch (1978) have examined the hat matrix V in some detail. They refer to a data point with large v_{ii} as a "high leverage point". Cook (1977) interpreted the ratio $v_{ii}/(1-v_{ii})$ as being $var(\hat{y}_i)/var(r_i)$, a measure of the sensitivity of estimation at x_i. This result follows directly from equation (2.5). This ratio can also be written in the form (from Beckman and Trussell (1974))

$$v_{ii}/(1-v_{ii}) = x_i(X'_{-i}X_{-i})^{-1}x_i' \,. \tag{3.8}$$

Thus, $v_{ii}/(1-v_{ii})$ is the distance from x_i to the centre of the remaining $n-1$ points in the sample. This explains why a high leverage point is potentially influential.

Cook (1979) offers a geometric interpretation of the v_{ii}. Assuming that the intercept is in the model, let $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_{q-1}$ denote the eigenvalues of the corrected cross product matrix, and let $S_1,\ldots,S_{q-1}$ denote the corresponding eigenvectors. By the spectral decomposition of the corrected cross product matrix

$$v_{ii} = \frac{1}{n} + \sum_{j=1}^{q-1} \left(\frac{S_j'(x_i-\bar{x})}{\sqrt{\lambda_j}} \right)^2 \qquad (3.9)$$

where $\bar{x}$ is the vector of sample averages.

Let θ_{ji} denote the angle between S_j and $(x_i-\bar{x})$. Then

$$\cos \theta_{ji} = \frac{S_j'(x_i-\bar{x})}{[(x_i-\bar{x})'(x_i-\bar{x})]^{\frac{1}{2}}} \qquad (3.10)$$

and $$v_{ii} = \frac{1}{n} + (x_i-\bar{x})'(x_i-\bar{x}) \sum_{j=1}^{q-1} \frac{\cos^2\theta_{ji}}{\lambda_j} . \qquad (3.11)$$

Hence v_{ii} may be large for two reasons. If x_i is far removed from the bulk of the data points, $(x_i-\bar{x})$ will be large. Secondly, v_{ii} will be large if x_i is in a direction of an eigenvector corresponding to a small eigenvalue. Note that if $(x_i-\bar{x})$ is small, then v_{ii} will be small regardless of its direction.

If the matrix $X'X$ is ill-conditioned, i.e. contains one or more small eigenvalues, then this will impact upon the leverage, and hence influence, of certain observations. In chapters 5 and 6 we shall examine the use of other regression techniques in such situations, and investigate their effects

in the detection of influential observations.

In summary, an observation is considered influential if its deletion causes substantial changes in the analysis. A data point may be influential if it is an outlier or a high leverage point. Cook's distance D_i, or alternatively the statistic Δ_i, provide measures of the influence of the i^{th} data point. Both statistics can be expressed in terms of confidence ellipsoids for β centred on the full sample estimate b.

It is possible that two or more data points may be jointly influential, even if the individual observations are not. For example, consider figures 1 and 2 below. In figure 1, if any individual point A, B or C is deleted, the fitted model will change very little. However, if all three points are deleted, estimates of parameters may change significantly. Conversely, in figure 2, if either point D or E is deleted, the fitted line will change, while if both are deleted, it will move only slightly.

Figure 1

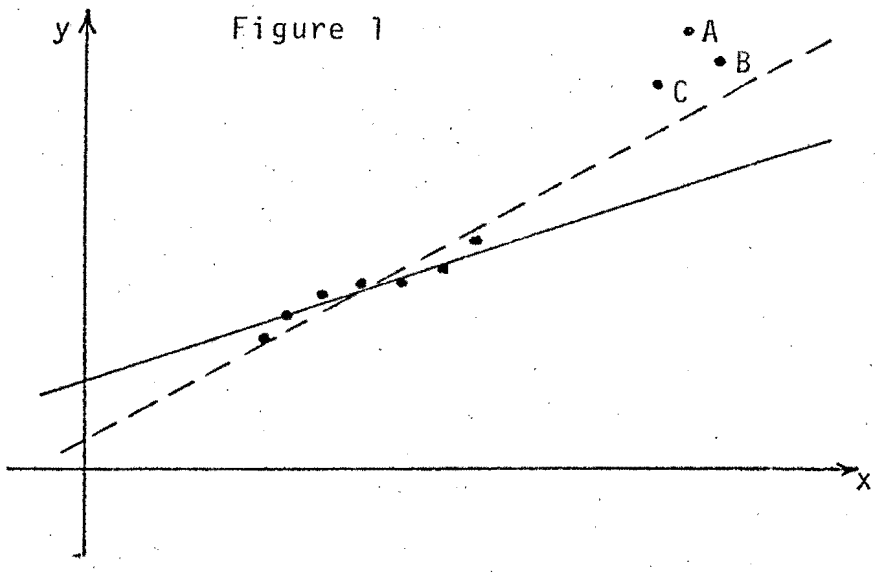
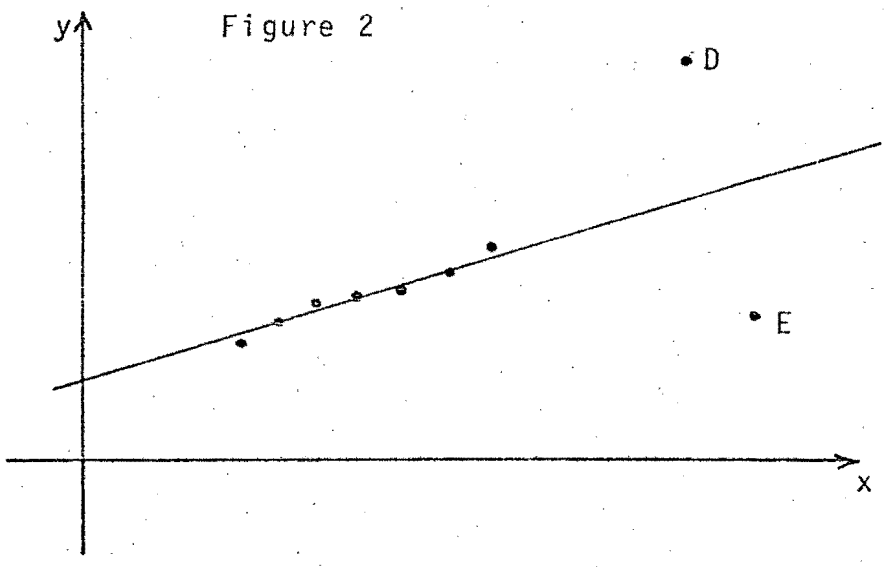


Figure 2



Let I be an m-vector specifying the indices of observations to be deleted. The subscript $-I$ will denote a submatrix with the m observation indexed by I deleted, while the subscript I will denote that only the observations indexed by I remain. For example, b_{-I} is the estimate for β based on the truncated sample of $n-m$ data points; V_I is

the $m \times m$ submatrix of V formed by the rows and columns indexed by I.

The distance measure Δ_i can be generalized to

$$\Delta_I = \frac{(b_{-I}-b)'X'X(b_{-I}-b)}{qs^2_{-I}} \tag{3.12}$$

where $s^2_{-I} = [r'r - r_I'(I-V_I)^{-1}r_I]/(n-q-m)$. (3.13)

The geometric interpretation of Δ_I is the same as that of Δ_i with degrees of freedom q and $n-q-m$ in (3.4). An influential subset of the sample is one with large Δ_I.

Cook and Weisberg (1980) show how the statistic Δ_I can be broken down to aid its interpretation. Bingham (1977) derived a number of identities, leading to the result

$$b_{-I}-b = -(X'X)^{-1}X_I'(I-V_I)^{-1}r_I \ . \tag{3.14}$$

The complete proof of this result is given in Jacobs (1982); it is the special case $k=0$ in theorem 5.4 proved later.

It follows that

$$\Delta_I = \frac{r_I'(I-V_I)^{-1}V_I(I-V_I)^{-1}r_I}{qs^2_{-I}} \ . \tag{3.15}$$

There exists an $m \times m$ diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_m)$ with $0 \le \lambda_1 \le \ldots \le \lambda_m \le 1$ and an orthogonal matrix Γ, such that

$$V_I = \Gamma'\Lambda\Gamma \ . \tag{3.16}$$

If $\lambda_m = 1$, then $(I-V_I)$ is singular; hence if the data

points indexed by I are removed, the resulting data are rank deficient and a unique b_{-I} does not exist. Therefore, if $\lambda_m = 1$, define $\Delta_I = \infty$.

If $\lambda_m < 1$ then (3.15) can be written as

$$\begin{aligned}\Delta_I &= r_I'(\Gamma'\Gamma-\Gamma'\Lambda\Gamma)^{-1}\Gamma'\Lambda\Gamma(\Gamma'\Gamma-\Gamma'\Lambda\Gamma)^{-1}r_I/qs_{-I}^2 \\ &= (\Gamma r_I)'(I-\Lambda)^{-1}\Lambda(I-\Lambda)^{-1}(\Gamma r_I)/qs_{-I}^2 \\ &= g'(I-\Lambda)^{-1}\Lambda(I-\Lambda)^{-1}g/qs_{-I}^2\end{aligned} \tag{3.17}$$

where $g = (g_j) = \Gamma r_I$.

Hence $$\Delta_I = \frac{1}{qs_{-I}^2}\sum_{j=1}^{m} g_j^2 \frac{\lambda_j}{(1-\lambda_j)^2} \ . \tag{3.18}$$

Each g_j is a linear combination of the elements of r_I. Furthermore,

$$\text{var}(g) = \text{var}(\Gamma r_I) = \sigma^2\Gamma\Gamma'(I-\Lambda)\Gamma\Gamma' = \sigma^2(I-\Lambda). \tag{3.19}$$

Therefore, each g_j is uncorrelated with $\text{var}(g_j) = \sigma^2(1-\lambda_j)$.

Let $$h_j = \frac{g_j}{s_{-I}\sqrt{1-\lambda_j}} \ . \tag{3.20}$$

Then (3.18) may be rewritten

$$\begin{aligned}\Delta_I &= \frac{1}{q}\sum_{j=1}^{m}\left(\frac{g_j}{s_{-I}\sqrt{1-\lambda_j}}\right)^2\frac{\lambda_j}{1-\lambda_j} \\ &= \frac{1}{q}\sum_{j=1}^{m} h_j^2\frac{\lambda_j}{1-\lambda_j} \ .\end{aligned} \tag{3.21}$$

(3.21) should be compared to (3.7). The role of the t_i statistic is taken by the h_j, while $v_{ii}/(1-v_{ii})$ is

replaced by $\lambda_j/(1-\lambda_j)$. In (3.21), a sum over m orthogonal directions is required, while in (3.7) $m = 1$.

A generalization of the studentized residuals is given by

$$\Sigma h_j^2 = g'(I-\Lambda)^{-1}g/s_{-I}^2$$
$$= r_I'(I-V_I)^{-1}r_I/s_{-I}^2 \quad . \tag{3.22}$$

Let $$t_I^2 = \frac{r_I'(I-V_I)^{-1}r_I/m}{s_{-I}^2} \quad . \tag{3.23}$$

We shall now show that t_I^2 follows an F distribution with m and $n-q-m$ degrees of freedom.

Theorem 3.1: (Khatri (1962)).

If $x \sim N(0,I)$ then $x'Qx+m'x+d$ will have a noncentral χ^2 distribution if and only if

(a) $Q^2 = Q$

(b) $m' = m'Q$

(c) $d = \frac{1}{4}m'm$.

The degrees of freedom will be $f = tr(Q)$ and the noncentrality $\nu = d$.

Theorem 3.2:

$$\frac{r_I'(I-V_I)^{-1}r_I}{\sigma^2} \sim \text{central} \quad \chi_m^2 \quad . \tag{3.24}$$

Proof:

From (2.4), $r = (I-V)\varepsilon$.

Without loss of generality it may be assumed that

$$r_I = (I_m \; 0)r$$

$$= (I_m \; 0)(I-V)\varepsilon \; .$$

Let $(I-V) = N$ be partitioned as $N = \begin{bmatrix} N_m & N_a \\ N_a' & N_{n-m} \end{bmatrix}$.

Then $(I-V_I) = N_m$.

$$r_I'(I-V_I)^{-1}r_I = \varepsilon'(I-V)\begin{pmatrix} I_m \\ 0 \end{pmatrix}N_m^{-1}(I_m \; 0)(I-V)\varepsilon$$

$$= \varepsilon' NN^*N\varepsilon \qquad (3.25)$$

where $N^* = \begin{bmatrix} N_m^{-1} & 0 \\ 0 & 0 \end{bmatrix}$.

$NN^*NNN^*N = NN^*NN^*N$ since $N = I-V$ is idempotent

$= NN^*N$ since $N^*NN^* = N^*$. (3.26)

Hence, condition (a) of theorem 3.1 is satisfied and conditions (b) and (c) are trivial.

$$tr(NN^*N) = tr(NN^*)$$

$$= tr\begin{pmatrix} N_m & N_a \\ N_a' & N_{n-m} \end{pmatrix}\begin{pmatrix} N_m^{-1} & 0 \\ 0 & 0 \end{pmatrix}$$

$$= tr\begin{pmatrix} I_m & 0 \\ N_a'N_m^{-1} & 0 \end{pmatrix} = m \; . \qquad (3.27)$$

Q.E.D.

Theorem 3.3:

$r_I'(I-V_I)^{-1}r_I$ and S_{-I}^2 are independently distributed, (3.28)

where $S_{-I}^2 = r'r - r_I'(I-V_I)^{-1}r_I$.

Proof:

It is sufficient to show that the product of the matrices determining the quadratic forms is equal to zero (Graybill (1961)).

$$r_I'(I-V_I)^{-1}r_I = \varepsilon' NN^*N\varepsilon \qquad \text{from (3.25)}$$

$$S^2_{-I} = \varepsilon'(N-NN^*N)\varepsilon \qquad \text{from (2.4) and (3.25)} .$$

$$\begin{aligned} NN^*N(N-NN^*N) &= NN^*N - NN^*NNN^*N \\ &= NN^*N - NN^*N \qquad \text{from (3.26)} \\ &= 0 . \end{aligned}$$

Q.E.D.

Since S^2_{-I}/σ^2 has a central χ^2 distribution with $n-q-m$ degrees of freedom (Ellenberg (1973)), it follows from (3.24) and (3.28) that t_I^2 follows an F distribution with m and $n-q-m$ degrees of freedom. The statistic t_I^2 is also the likelihood ratio test for testing $H_0 : \theta_I = 0$ in the model

$$y = X\beta + \theta_I + \varepsilon$$

where θ_I is an $n\times 1$ vector of zeros with unknown parameters in the positions indexed by I. (cf. Gentleman and Wilk (1975), Barnett and Lewis (1978)).

If we are examining groups of size m, there will be $N = \binom{n}{m}$ possibilities to be considered. If the location of outlying groups is unknown beforehand, we should use the upper α/N point of the $F(m,n-q-m)$ distribution to obtain critical values for t_I^2. As N may be very large, this practice may

give critical values so large that groups will seldom, if ever, be found to be statistically significant as outliers. We shall thus recommend a procedure to overcome this problem.

If a group of observations has small influence, then it does not really matter whether or not it is outlying, as its impact on the analysis will be negligible either way. Hence, the first step in the analysis would be to locate influential groups. Say there are N^* groups with influence greater than a predetermined value (the 5% confidence ellipsoid for example). Then a suitable critical value for the t_I^2 statistic would be the upper α/N^* point of the appropriate F distribution.

Finally, a note on the distributional properties of Δ_I. Now $E(g_j) = 0$ and $var(g_j) = \sigma^2(1-\lambda_j)$ from (3.19). Hence, under normality $g_j^2/\sigma^2(1-\lambda_j)$ has a χ^2 distribution with 1 degree of freedom, and the distribution is also independent of s_{-I}^2. Hence, from (3.21)

$$\Delta_I = \frac{1}{q} \sum_{j=1}^{m} \frac{g_j^2}{s_{-I}^2(1-\lambda_j)} \cdot \frac{\lambda_j}{1-\lambda_j} \qquad (3.29)$$

follows a weighted sum of $F(1,n-q-m)$ distributions.

Chapter 4

BIASED ESTIMATORS

Suppose one column of the X matrix is a linear combination of other independent variables in the model. In this case, the matrix X would be rank deficient, $X'X$ would be singular, and no unique estimate for β would exist. The existence of one or more linear relationships in addition to the hypothesised linear relationship $y = X\beta+\varepsilon$ is known as the problem of collinearity.

In practice, we are unlikely to find an exact linear relationship between the columns of the X matrix. It happens often, however, that some of the variables are highly correlated, and $X'X$ may be ill-conditioned; i.e. $|X'X|$ will be very small, or alternatively, one or more eigenvalues of $X'X$ will be very close to zero.

The expected squared Euclidean distance between β and the least squares estimate b can be expressed as

$$E(b-\beta)'(b-\beta) = \sigma^2 tr(X'X)^{-1}$$
$$= \sigma^2 \Sigma(1/\lambda_i) \qquad (4.1)$$

where the λ_i are the eigenvalues of $X'X$. Hence, if one or more eigenvalues are small, the estimate b may be very far from the true value of β.

In addition

$$\text{var}(b) = \sigma^2 (X'X)^{-1} \qquad (4.2)$$

and therefore the ordinary least squares estimator b may have very large variance if the matrix $X'X$ is ill-conditioned, although it is the best linear unbiased estimator for β.

In the presence of near collinearity, it may pay us to use a biased estimator for β if the variance of the estimator is less than that of b. In the following two chapters, we shall examine two such techniques in detail : ridge regression and generalized inverse regression.

There remains the problem of defining when the matrix $X'X$ is considered to be ill-conditioned. The determinant $|X'X|$ is not a reliable measure of the conditioning of a matrix, as it can be made arbitrarily large or small by multiplying through by a constant. Several authors, e.g. Forsythe and Moler (1967), Marshall and Olkin (1971), recommend the condition number as a measure of conditioning. The condition number is defined as the ratio of the largest eigenvalue to the smallest

$$\kappa(X'X) = \lambda_q/\lambda_1 \ . \qquad (4.3)$$

Belsley, Kuh and Welsch (1980) determined empirically that medium to strong relations between the independent variables are associated with condition numbers between 30 and 100.

The ill-conditioning of the $X'X$ matrix is not the only circumstance in which the use of a biased estimator can be justified.

In many applications in econometrics, the analyst may have prior information on particular regression coefficients or linear combinations thereof, and this information should be incorporated into the model together with the sample data. Alternatively, the prior information may form constraints upon the regression coefficients.

Restricted estimators may be biased, but will have smaller variance than the ordinary least squares estimator. The use of (biased) restricted estimators in econometrics will be examined in chapters 7 and 8. In chapter 7 we shall discuss the case where the prior information is exact, i.e. in the form

$$H\beta = h \tag{4.4}$$

while in chapter 8 we shall discuss the case where the prior information is stochastic, i.e. in the form

$$H\beta + u = h \tag{4.5}$$

with $u \sim N(\delta, \sigma^2 R)$.

In a model with q independent variables, a variable selection procedure can be used to reduce the model to $p < q$ variables. This may be done to remove experimental design faults (such as near collinearity) or simply to reduce the computational effort. The estimator based on p variables may be biased due to the removal of predictor variables, but it is preferred to the full model estimator under certain conditions. This situation will be examined in chapter 9.

Chapter 5

RIDGE REGRESSION

In this chapter we shall examine a biased estimation technique to be followed when the matrix $X'X$ appears to be ill-conditioned. This technique, proposed by Hoerl and Kennard (1970), is known as "ridge regression".

It is assumed that the matrix $X'X$ is in the form of the correlation matrix between the independent variables, and similarly $X'y$ is the vector of correlations between X and y. This can be achieved with the following transformation of $X = (x_{ij})$:

$$x^*_{ij} = (x_{ij} - \bar{x}_j)/s_{jj}^{\frac{1}{2}} \tag{5.1}$$

where $\bar{x}_j = \frac{1}{n} \sum_i x_{ij}$

$$s_{jj} = \sum_i (x_{ij} - \bar{x}_j)^2$$

and a similar transformation of the y vector. In the theoretical development, we shall assume that both X and y have been transformed as above.

The ridge estimate for β is given by

$$\tilde{b} = (X'X+kI)^{-1}X'y \tag{5.2}$$
$$= WX'y$$

where $W = (X'X+kI)^{-1}$ (5.3)

$k \geq 0$ is an arbitrary constant.

For $k > 0$, $\tilde{b}$ is a biased estimator for β. The expected squared distance between $\tilde{b}$ and β can be expressed as the sum of bias and variance terms. As k increases, the bias term increases and the variance term decreases. There exists an optimal k which minimizes the expected squared distance between $\tilde{b}$ and β, but it is a function of the unknown parameter β.

Several authors, e.g. Theobald (1974), Gunst and Mason (1976), have suggested using the generalized mean square error criterion as a standard for making comparisons among estimators. Theobald (1974) showed that these exists some $K > 0$ such that wherever $0 < k < K$ the estimate $\tilde{b}$ is preferred to b under this criterion. Unfortunately K is a function of the unknown β. Price (1982) showed that b is never preferred to $\tilde{b}$ under this criterion, for any value of k.

Let $$\begin{aligned}\tilde{r} &= y - X\tilde{b} \\ &= (I-XWX')y \\ &= (I-XWX')(X\beta+\varepsilon) \\ &= X(I-WX'X)\beta + (I-XWX')\varepsilon \\ &= kXW\beta + (I-XWX')\varepsilon \quad \text{as} \quad WX'X = I - kW . \end{aligned} \tag{5.4}$$

Therefore, $E(\tilde{r}) \neq 0$ as $\beta \neq 0$ in general,

and $$\begin{aligned}\text{var}(\tilde{r}) &= (I-XWX')(I-XWX')'\sigma^2 \\ &= (I-2XWX'+XWX'XWX')\sigma^2 \\ &= (I-2XWX'+X(I-kW)WX')\sigma^2 \\ &= (I-XWX'-kXWWX')\sigma^2 \\ &= (I-Q)\sigma^2 \end{aligned} \tag{5.5}$$

where $Q = XWX' + kXWWX'$

$= XW(I+kW)X'$. (5.6)

Troskie, Coutsourides, Jacobs and Dunne (1982) have proved the following three results:

Theorem 5.1:

$\text{var}(\tilde{r}) - \text{var}(r)$ is positive definite for $k > 0$ (5.7)

where r is the vector of ordinary least squares residuals.

Theorem 5.2:

The total variance of $\tilde{r}$ is greater than the total variance of r for all $k > 0$; specifically

$$\text{tr}(\text{var}(\tilde{r})) - \text{tr}(\text{var}(r)) = \sigma^2 \sum_{i=1}^{q} \left(\frac{k}{\lambda_{i+k}}\right)^2 > 0 \qquad (5.8)$$

where the λ_i are the eigenvalues of $X'X$.

Theorem 5.3:

There exists $K > 0$ such that the generalized mean square error of $\tilde{r}$ is less than that of r whenever $0 < k < K$.

Hence, the ridge residuals $\tilde{r}$ are biased with more variability than the ordinary least squares residuals. However, under certain conditions, $\tilde{r}$ is a better estimator of the unknown ε than the estimator r.

The following statistic is suggested for testing whether the i^{th} data point is an outlier:

$$\tilde{t}_i = \frac{\tilde{r}_i}{s_{-i}\sqrt{1-q_{ii}}} \qquad i = 1,\ldots,n \qquad (5.9)$$

where q_{ii} is the i^{th} diagonal element of Q, the ridge equivalent of the hat matrix. Note that a least squares estimate for σ^2 is used, following theorems 5.1 and 5.2. The estimate s^2_{-i} is preferred to s^2 for reasons stated in chapter 2. The distribution of s^2_{-i} is known to be χ^2_{n-q-1}, and it is distributed independently of $\tilde{r}_i$ (Troskie et al (1982)).

The statistic $\tilde{t}_i$ follows a non-central F distribution with 1 and n-q-1 degrees of freedom. This result is proved in the general case theorems 5.5 and 5.6 later in this chapter. From (5.4), the non-centrality is a function of k and the unknown β. If k is small, the non-centrality will be small, and the tables of appendix A should provide a reasonable approximation of critical values for $\tilde{t}_i$.

Assuming that $k > 0$ is fixed, and that ridge regression is used to estimate β, we define a distance measure for the influence of the i^{th} observation:

$$\tilde{\Delta}_i = \frac{(\tilde{b}_{-i}-\tilde{b})'X'X(\tilde{b}_{-i}-\tilde{b})}{qs^2_{-i}} \qquad i = 1,\ldots,n \qquad (5.10)$$

where the subscripts have the same meaning as in chapter 3. Using (3.4) again, we can express the distance between $\tilde{b}_{-i}$ and $\tilde{b}$ in terms of confidence ellipsoids centred on b.

The effect on b of deleting the i^{th} observation is

$$\tilde{b}_{-i}-\tilde{b} = -(1-\tilde{v}_{ii})^{-1}(X'X+kI)^{-1}x_i'\tilde{r}_i \qquad (5.11)$$

where $\tilde{v}_{ii} = x_i(X'X+kI)^{-1}x_i'$. This result is a special case

for $m = 1$ in the general theorem 5.4 proved later.

Therefore, $\tilde{\Delta}_i$ cannot be simplified to give an expression analogous to (3.7). If, however, we define

$$\Delta_i^* = \frac{(\tilde{b}_{-i}-\tilde{b})'(X'X+kI)(\tilde{b}_{-i}-\tilde{b})}{qs^2_{-i}} \tag{5.12}$$

then we can use (5.11) to simplify this to

$$\begin{aligned}\Delta_i^* &= \frac{\tilde{r}_i x_i (X'X+kI)^{-1}(X'X+kI)(X'X+kI)^{-1}x_i'\tilde{r}_i}{qs^2_{-i}(1-\tilde{v}_{ii})^2} \\ &= \frac{\tilde{v}_{ii}\tilde{r}_i^2}{qs^2_{-i}(1-\tilde{v}_{ii})^2} \\ &= \frac{1}{q}\left(\frac{\tilde{r}_i}{s_{-i}\sqrt{1-q_{ii}}}\right)^2\left(\frac{\tilde{v}_{ii}}{1-\tilde{v}_{ii}}\right)\left(\frac{1-q_{ii}}{1-\tilde{v}_{ii}}\right) \\ &= \frac{1}{q}\,\tilde{t}_i^2\left(\frac{\tilde{v}_{ii}}{1-\tilde{v}_{ii}}\right)\left(\frac{1-q_{ii}}{1-\tilde{v}_{ii}}\right) .\end{aligned} \tag{5.13}$$

The expression (5.13) is equivalent to (3.7), with a bias $(1-q_{ii})/(1-\tilde{v}_{ii})$ introduced by the ridge effect. Although computationally convenient, it is doubtful whether Δ_i^* is an appropriate measure, as (3.4) no longer applies. Hence the interpretation of the distance in terms of confidence ellipsoids is not exact.

However, it is possible to express $\tilde{\Delta}_i$ as

$$\tilde{\Delta}_i = \frac{\tilde{r}_i x_i (X'X+kI)^{-1} X'X (X'X+kI)^{-1} x_i' \tilde{r}_i}{q s_{-i}^2 (1-\tilde{v}_{ii})^2} \quad \text{from (5.10) and (5.11)}$$

$$= \frac{1}{q s_{-i}^2} \left(\frac{\tilde{r}_i}{1-\tilde{v}_{ii}} \right)^2 x_i (X'X+kI)^{-1} X'X (X'X+kI)^{-1} x_i'$$

$$= \frac{1}{q s_{-i}^2} \left(\frac{\tilde{r}_i}{1-\tilde{v}_{ii}} \right)^2 x_i W X'X W x_i'$$

$$= \frac{1}{q s_{-i}^2} \left(\frac{\tilde{r}_i}{1-\tilde{v}_{ii}} \right)^2 x_i (I-kW) W x_i' \qquad (5.14)$$

since $WX'X = I-kW$.

Turning now to the case where $m > 1$ observations are deleted, and these are indexed by the set I, (5.10) can be generalized as before:

$$\tilde{\Delta}_I = \frac{(\tilde{b}_{-I}-\tilde{b})' X'X (\tilde{b}_{-I}-\tilde{b})}{q s_{-I}^2} \qquad (5.15)$$

where the notation follows earlier conventions.

Theorem 5.4:

$$\tilde{b}_{-I}-\tilde{b} = -(X'X+kI)^{-1} X_I' (I-\tilde{V}_I)^{-1} \tilde{r}_I \qquad (5.16)$$

where $W = (X'X+kI)^{-1}$ as before

$$\tilde{V}_I = X_I W X_I' = X_I (X'X+kI)^{-1} X_I' \ . \qquad (5.17)$$

Proof:

Let $W_{-I} = (X_{-I}'X_{-I}+kI)^{-1}$ (5.18)

$U = X_I W_{-I} X_I'$. (5.19)

Now $$X'X+kI = X'_{-I}X_{-I} + X'_I X_I + kI \tag{5.20}$$

$$I-U(I+U)^{-1} = (I+U)(I+U)^{-1}-U(I+U)^{-1}$$

$$= (I+U)^{-1}. \tag{5.21}$$

Let $$W^* = W_{-I} - W_{-I}X'_I(I+U)^{-1}X_I W_{-I} . \tag{5.22}$$

Multiply (5.20) by (5.22):

$$\begin{aligned}(X'X+kI)W^* &= (X'_{-I}X_{-I}+kI)W_{-I}-(X'_{-I}X_{-I}+kI)W_{-I}X'_I(I+U)^{-1}X_I W_{-I} \\ &\quad + (X'_I X_I)W_{-I}-(X'_I X_I)W_{-I}X'_I(I+U)^{-1}X_I W_{-I} \\ &= I-X'_I(I+U)^{-1}X_I W_{-I}+(X'_I X_I)W_{-I}-X'_I U(I+U)^{-1}X_I W_{-I} \\ &= I-X'_I[(I+U)^{-1}+I-U(I+U)^{-1}]X_I W_{-I} \\ &= I-X'_I[(I+U)^{-1}-(I+U)^{-1}]X_I W_{-I} \quad \text{from (5.21)} \\ &= I .\end{aligned}$$

Hence $$W^* = W = W_{-I}-W_{-I}X'_I(I+U)^{-1}X_I W_{-I} . \tag{5.23}$$

Then $$\begin{aligned}WX'_I &= W_{-I}X'_I-W_{-I}X'_I(I+U)^{-1}X_I W_{-I}X'_I \quad \text{from (5.23)} \\ &= W_{-I}X'_I[I-(I+U)^{-1}U] \\ &= W_{-I}X'_I(I+U)^{-1} \qquad \text{using (5.21).}\end{aligned} \tag{5.24}$$

$$\begin{aligned}(I-\tilde{V}_I) &= I - X_I W X'_I \quad \text{by definition (5.17)} \\ &= I - X_I W_{-I}X'_I(I+U)^{-1} \quad \text{from (5.24)} \\ &= I - U(I+U)^{-1} \\ &= (I+U)^{-1} \quad \text{using (5.21).}\end{aligned} \tag{5.25}$$

$$\tilde{r}_I = y_I - X_I\tilde{b}$$

$$= y_I - X_I WX'y$$

$$= y_I - X_I W(X'_{-I}y_{-I}+X'_I y_I)$$

$$= (I-X_I WX'_I)y_I - X_I WX'_{-I}y_{-I}$$

$$= (I-\tilde{V}_I)y_I - (I+U)^{-1}X_I W_{-I}X'_{-I}y_{-I} \quad \text{from (5.17) and (5.24)}$$

$$= (I-\tilde{V}_I)(y_I-X_I\tilde{b}_{-I}) \quad \text{from (5.25).} \qquad (5.26)$$

Now $\tilde{b} = WX'y = W(X'_{-I}y_{-I}+X'_I y_I)$

$$= (W_{-I}-W_{-I}X'_I(I+U)^{-1}X_I W_{-I})(X'_{-I}y_{-I}+X'_I y_I) \quad \text{from (5.23)}$$

$$= W_{-I}X'_{-I}y_{-I}-W_{-I}X'_I(I+U)^{-1}X_I W_{-I}X'_{-I}y_{-I}$$

$$+ W_{-I}X'_I y_I - W_{-I}X'_I(I+U)^{-1}X_I W_{-I}X'_I y_I$$

$$= \tilde{b}_{-I}-W_{-I}X'_I[(I+U)^{-1}X_I\tilde{b}_{-I}-y_I+(I+U)^{-1}Uy_I]$$

$$= \tilde{b}_{-I}-W_{-I}X'_I[(I+U)^{-1}X_I\tilde{b}_{-I}-(I+U)^{-1}y_I] \quad \text{using (5.21)}$$

$$= \tilde{b}_{-I}+W_{-I}X'_I(I+U)^{-1}(y_I-X_I\tilde{b}_{-I})$$

$$= \tilde{b}_{-I}+WX'_I(y_I-X_I\tilde{b}_{-I}) \quad \text{from (5.24)}$$

$$= \tilde{b}_{-I}+WX'_I(I-\tilde{V}_I)^{-1}\tilde{r}'_I \quad \text{from (5.26) .}$$

Q.E.D.

Substituting (5.16) into (5.15) we obtain

$$\tilde{\Delta}_I = \frac{1}{qs^2_{-I}}\,\tilde{r}'_I(I-\tilde{V}_I)^{-1}X_I(X'X+kI)^{-1}X'X(X'X+kI)^{-1}X'_I(I-\tilde{V}_I)^{-1}\tilde{r}_I \qquad (5.27)$$

$$= \frac{1}{qs^2_{-I}}\,\tilde{r}'_I(I-\tilde{V}_I)^{-1}X_I WX'XWX'_I(I-\tilde{V}_I)^{-1}\tilde{r}_I$$

$$= \frac{1}{qs^2_{-I}}\,\tilde{r}'_I(I-\tilde{V}_I)^{-1}X_I(I-kW)WX'_I(I-\tilde{V}_I)^{-1}\tilde{r}_I$$

$$= \frac{1}{qs^2_{-I}}[\tilde{r}'_I(I-\tilde{V}_I)^{-1}\tilde{V}_I(I-\tilde{V}_I)^{-1}\tilde{r}_I - k\tilde{r}'_I(I-\tilde{V}_I)^{-1}X_I WWX'_I(I-\tilde{V}_I)^{-1}\tilde{r}_I]. \tag{5.28}$$

Recall that in the ordinary least squares case (3.15), we were able to write Δ_I in terms of r_I and the eigenvalues and eigenvectors of $V_I = X_I(X'X)^{-1}X'_I$. However, in (5.28) we are unable to obtain a simplified expression, on account of the $X_I WWX'_I$ term. Even the generalized form of (5.12) cannot be simplified into an expression analogous to (3.21).

In generalizing the studentized residuals, a statistic for testing whether a group of m observations is an outlying group is

$$\tilde{t}^2_I = \tilde{r}'_I(I-Q_I)^{-1}\tilde{r}_I/ms^2_{-I} \tag{5.29}$$

where $Q_I = X_I W(I+kW)X'_I$ from (5.6).

Theorem 5.5:

$$\frac{\tilde{r}'_I(I-Q_I)^{-1}\tilde{r}_I}{\sigma^2} \sim \chi^2_f(\nu) \tag{5.30}$$

with degrees of freedom $f = m$

and noncentrality $\nu = k^2\beta'WX'_I(I-Q_I)^{-1}X_I W\beta$.

Proof:

$$\tilde{r} = kXW\beta + (I-XWX')\varepsilon \quad \text{from (5.4)}$$

$$= \mu + (I-\tilde{V})\varepsilon \tag{5.31}$$

where $\mu = kXW\beta$ (5.32)

$\tilde{V} = X(X'X+kI)^{-1}X'$.

Without loss of generality it may be assumed that

$$\tilde{r}_I = (I_m\ 0)\tilde{r}$$

$$= (I_m\ 0)(\mu+(I-\tilde{V})\varepsilon) .$$

Now $I-Q = (I-\tilde{V})(I-\tilde{V})'$ from (5.5).

Let $I-Q = N$ be partitioned as $N = \begin{bmatrix} N_m & N_a \\ N_a' & N_{n-m} \end{bmatrix}$. (5.33)

Then $I-Q_I = N_m$.

$$\tilde{r}_I' N_m^{-1} \tilde{r}_I = (\mu'+\varepsilon'(I-\tilde{V}))\begin{pmatrix} I_m \\ 0 \end{pmatrix} N_m^{-1} (I_m\ 0)(\mu+(I-\tilde{V})\varepsilon)$$

$$= \varepsilon'(I-\tilde{V})N^*(I-\tilde{V})\varepsilon+2\mu'N^*(I-\tilde{V})\varepsilon+\mu'N^*\mu \qquad (5.34)$$

where $N^* = \begin{bmatrix} N_m^{-1} & 0 \\ 0 & 0 \end{bmatrix}$.

(a) $(I-\tilde{V})N^*(I-\tilde{V})(I-\tilde{V})N^*(I-\tilde{V}) = (I-\tilde{V})N^*NN^*(I-\tilde{V})$ from (5.5)

$$= (I-\tilde{V})N^*(I-\tilde{V})$$

since $N^*NN^* = N^*$ by simple algebra.

(b) $2\mu'N^*(I-\tilde{V})(I-\tilde{V})N^*(I-\tilde{V}) = 2\mu'N^*NN^*(I-\tilde{V})$

$$= 2\mu'N^*(I-\tilde{V}) .$$

(c) $\frac{1}{4}[2\mu'N^*(I-\tilde{V})][2\mu'N^*(I-\tilde{V})]' = \mu'N^*(I-\tilde{V})(I-\tilde{V})N^*\mu$

$$= \mu'N^*\mu .$$

Hence the conditions of theorem 3.1 are satisfied.

$$\mu'N^*\mu = k^2\beta'WX'\begin{pmatrix} I_m \\ 0 \end{pmatrix} N_m^{-1} (I_m\ 0)XW\beta$$

$$= k^2\beta'WX_I'(I-Q_I)^{-1}X_IW\beta .$$

$$\mathrm{tr}((I-\tilde{V})N^*(I-\tilde{V})) = \mathrm{tr}(NN^*) = m . \quad \text{(c.f. (3.27))}.$$

Q.E.D.

Theorem 5.6:

Let $S^2_{-I} = r'r - r'_I(I-V_I)^{-1}r_I$ as before.

Then $\tilde{r}'_I(I-Q_I)^{-1}\tilde{r}_I$ and S^2_{-I} are independently distributed. (5.35)

Proof:

Let $I-V = M$ be partitioned as $M = \begin{bmatrix} M_m & M_a \\ M'_a & M_{n-m} \end{bmatrix}$. (5.36)

Then $I-V_I = M_m$.

From (2.4), $r = (I-V)\varepsilon$

and r_I can be written $r_I = (I_m\ 0)(I-V)\varepsilon$.

Hence $S^2_{-I} = \varepsilon'M\varepsilon - \varepsilon'M\begin{pmatrix} I_m \\ 0 \end{pmatrix}M_m^{-1}(I_m\ 0)M\varepsilon$ as M is idempotent

$$= \varepsilon'(M-MM^*M)\varepsilon \qquad (5.37)$$

where $M^* = \begin{bmatrix} M_m^{-1} & 0 \\ 0 & 0 \end{bmatrix}$.

To prove the result, it is sufficient to show that the product of the matrices determining the quadratic forms (5.34) and (5.37) is equal to zero (Graybill (1961)).

Now $(I-\tilde{V})M = (I-XWX')(I-X(X'X)^{-1}X')$ by definition

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

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

$$= M . \qquad (5.38)$$

$$\therefore (I-\tilde{V})N^*(I-\tilde{V})[M-MM^*M] = (I-\tilde{V})N^*(I-\tilde{V})M-(I-\tilde{V})N^*(I-\tilde{V})MM^*M$$

$$= (I-\tilde{V})N^*M-(I-\tilde{V})N^*MM^*M \quad \text{from (5.38)}$$

$$= (I-\tilde{V})N^*M-(I-\tilde{V})N^*M \quad \text{since } N^*MM^* = N^*$$

$$= 0 .$$

Q.E.D.

From (5.30) and (5.35), the statistic $\tilde{t}_I^2$ follows a non-central F distribution with m and n-q-m degrees of freedom. The noncentrality is a function of k and the unknown β. As $k \to 0$, (5.29) approaches the ordinary least squares statistic (3.23), which has a central F distribution. Thus, for small k, the central F distribution may be used as an approximation.

Determining a value for k:

The problem remains of specifying a value for the ridge constant k. The "optimal" value for k cannot be determined, and it may well be that different results are obtained for various values of k.

Now, the ridge estimate for the regression parameters is a function of k. To obtain a unique estimate for β requires a specification of k. However, in the context of this study, we are not primarily concerned with parameter estimation, but the identification of influential data.

Hence, it is recommended that the analysis be performed for a number of different values of k. A table, or preferably a graph, should be prepared showing the values of the statistics

$\tilde{t}_i^2$ and $\tilde{\Delta}_i$ (or $\tilde{t}_I^2$ and $\tilde{\Delta}_I$ as the case may be) for varying k. If either statistic is large for any value of k, then there is evidence that the observation (or set of observations) may be outlying or influential. This conclusion is reinforced if the statistics stabilise over a range of k values. A graph of $\tilde{t}_i^2$ or $\tilde{\Delta}_i$ (or $\tilde{t}_I^2$ or $\tilde{\Delta}_I$) versus k will be referred to as a "ridge trace" in the examples.

Chapter 6

GENERALIZED INVERSE REGRESSION

When the matrix $X'X$ is ill-conditioned, an alternative to ridge regression is a technique known as generalized inverse regression or principal component regression, (Marquardt (1970)).

Again assume that $X'X$ is in the form of a correlation matrix, applying the transformation (5.1) to the raw data if necessary. Let $S = (S_q,\ldots,S_2,S_1)$ be the matrix of orthonormal eigenvectors of $X'X$, and let

$$S'X'XS = \Lambda = \mathrm{diag}(\lambda_q,\ldots,\lambda_2,\lambda_1) \tag{6.1}$$

where $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_q$ are the eigenvalues of $X'X$. Suppose it is decided that $c < q$ of the eigenvalues are significantly greater than zero. Partition S and Λ as

$$S = (S_c \quad S_{q-c})$$

$$\Lambda = \begin{pmatrix} \Lambda_c & 0 \\ 0 & \Lambda_{q-c} \end{pmatrix} . \tag{6.2}$$

The "generalized inverse" of $X'X$ is given by

$$(X'X)^- = S_c\Lambda_c^{-1}S_c'$$

$$= \sum_{j>q-c} \frac{1}{\lambda_j} S_jS_j' . \tag{6.3}$$

The generalized inverse (or g-inverse) possesses some useful

properties (see for example Graybill (1976)), particularly

$$(X'X)^{-}X'X(X'X)^{-} = (X'X)^{-} . \tag{6.4}$$

The g-inverse estimator of β is

$$b^{-} = (X'X)^{-}X'y . \tag{6.5}$$

Johnson, Reimer and Rothrock (1972) have shown that the g-inverse estimator can be expressed as

$$b^{-} = b - (X'X)^{-1}S_c(S_c'(X'X)^{-1}S_c)^{-1}S_c'b \tag{6.6}$$

and hence

$$E(b^{-}) = [I-(X'X)^{-1}S_c(S_c'(X'X)^{-1}S_c)^{-1}S_c']\beta \neq \beta \tag{6.7}$$

$$\begin{aligned} var(b^{-}) = \sigma^2[&(X'X)^{-1} \\ &-(X'X)^{-1}S_c(S_c'(X'X)^{-1}S_c')^{-1}S_c(X'X)^{-1}]. \end{aligned} \tag{6.8}$$

The g-inverse estimator b^{-} is therefore biased, but has smaller variance than the ordinary least squares estimator b.

Toro-Vizcarrondo and Wallace (1968) showed that b^{-} is preferred to b under the generalized mean square error criterion if and only if

$$(S_c'\beta)'(S_c'(X'X)^{-1}S_c)^{-1}(S_c'\beta)/\sigma^2 \leq 1 .$$

This condition depends upon the unknown parameters β and σ^2, and cannot be verified exactly in practice.

Let $r^{-} = y - Xb^{-}$

$$\begin{aligned} &= (I - XS_c\Lambda_c^{-1}S_c'X')y \\ &= (I - XS_c\Lambda_c^{-1}S_c'X')(X\beta+\varepsilon) \\ &= X(I-S_c\Lambda_c^{-1}S_c'X'X)\beta + (I-XS_c\Lambda_c^{-1}S_c'X')\varepsilon . \end{aligned} \tag{6.9}$$

Hence $E(r^-) \neq 0$

and $var(r^-) = (I-XS_c\Lambda_c^{-1}S_c'X')\sigma^2$

$$= (I-P)\sigma^2 \qquad (6.10)$$

where $P = XS_c\Lambda_c^{-1}S_c'X' = X(X'X)^-X'$ (6.11)

is the g-inverse equivalent of the hat matrix, and $(I-P)$ is idempotent.

Troskie et al (1982) have proved the following three results:

Theorem 6.1:

$$var(r^-) - var(r) \text{ is nonnegative definite} \qquad (6.12)$$

where r is the vector of ordinary least squares residuals.

Theorem 6.2:

The total variance of r^- is greater than the total variance of r;

$$tr(var(r^-)) - tr(var(r)) = \sigma^2(n-c) - \sigma^2(n-q)$$

$$= \sigma^2(q-c) > 0 . \qquad (6.13)$$

Theorem 6.3:

The generalized mean square error of r^- is less than that of r provided that

$$\sum_{i \leq q-c} \frac{\lambda_i(S_i'\beta)^2}{\sigma^2} < 1 . \qquad (6.14)$$

Although the g-inverse residuals are biased with more variability than the ordinary least squares residuals, r^- is a better estimator of ε than the estimator r under the

condition (6.14). This condition is dependent upon the unknown β and σ^2.

The following statistic is proposed for testing whether the i^{th} data point is an outlier:

$$t_i^- = \frac{r_i^-}{s_{-i}\sqrt{1-p_{ii}}} \qquad i = 1,\ldots,n \qquad (6.15)$$

where p_{ii} is the i^{th} diagonal element of $P = X(X'X)^-X'$. The statistic $(t_i^-)^2$ follows a non-central F distribution with 1 and $n-q-1$ degrees of freedom. From (6.9) the non-centrality is a function of c and the unknown β. The distributional results are proved in the general case in theorems 6.5 and 6.6 later in this chapter.

If $c < q$ is fixed, a measure of the influence of the i^{th} observation is:

$$\Delta_i^- = \frac{(b_{-i}^- - b^-)'X'X(b_{-i}^- - b^-)}{qs_{-i}^2} \qquad i = 1,\ldots,n \qquad (6.16)$$

where the subscripts have their usual meanings. The special case ($m = 1$) of theorem 6.4, proved later, is

$$b_{-i}^- - b^- = -(1-p_{ii})^{-1}(X'X)^-x_i'r_i^- \quad . \qquad (6.17)$$

Hence,

$$\Delta_i^- = \frac{(r_i^-)'x_i(X'X)^-X'X(X'X)^-x_i'(r_i^-)}{qs_{-i}^2(1-p_{ii})^2}$$

$$= \frac{p_{ii}(r_i^-)^2}{qs_{-i}^2(1-p_{ii})^2}$$

$$= \frac{1}{q}\left(\frac{r_i^-}{s_{-i}\sqrt{1-p_{ii}}}\right)^2\left(\frac{p_{ii}}{1-p_{ii}}\right)$$

$$= \frac{1}{q}\left(t_i^-\right)^2\frac{p_{ii}}{1-p_{ii}} \qquad (6.18)$$

which is the g-inverse equivalent of (3.7). The studentized residual t_i^- is the outlier statistic, and $p_{ii}/(1-p_{ii})$ is the measure of leverage.

In the case where $m > 1$ observations are deleted, and these are indexed by the set I, (6.16) can be generalized as before:

$$\Delta_I^- = \frac{(b_{-I}^- - b^-)'X'X(b_{-I}^- - b^-)}{qs_{-I}^2} \qquad (6.19)$$

where the notation follows earlier conventions.

Theorem 6.4:

$$b_{-I}^- - b^- = -(X'X)^-X_I'(I-P_I)^{-1}r_I^- \qquad (6.20)$$

where $P_I = X_I(X'X)^-X_I'$. (6.21)

Proof: Let $W_{-I} = (X_{-I}'X_{-I})^-$ (6.22)

$$U = X_I W_{-I} X_I'. \qquad (6.23)$$

Then $W_{-I}X_{-I}'X_{-I}W_{-I} = W_{-I}$ (6.24)

$$X'X = X_{-I}'X_{-I} + X_I'X_I \qquad (6.25)$$

$$I - U(I+U)^{-1} = (I+U)(I+U)^{-1} - U(I+U)^{-1}$$

$$= (I+U)^{-1}. \qquad (6.26)$$

Let $W^* = W_{-I} - W_{-I}X_I'(I+U)^{-1}X_I W_{-I}$. (6.27)

Then $W^*X'XW^* = W_{-I}(I-X_I'(I+U)^{-1}X_I W_{-I})(X_{-I}'X_{-I}+X_I'X_I)(I-W_{-I}X_I'(I+U)^{-1}X_I)W_{-I}$

from (6.25) and (6.27)

$$= W_{-I}[X_{-I}'X_{-I} - X_{-I}'X_{-I}W_{-I}X_I'(I+U)^{-1}X_I + X_I'X_I$$
$$- X_I'X_I W_{-I}X_I'(I+U)^{-1}X_I - X_I'(I+U)^{-1}X_I W_{-I}X_{-I}'X_{-I}$$
$$+ X_I'(I+U)^{-1}X_I W_{-I}X_{-I}'X_{-I}W_{-I}X_I'(I+U)^{-1}X_I$$
$$- X_I'(I+U)^{-1}X_I W_{-I}X_I'X_I$$
$$+ X_I'(I+U)^{-1}X_I W_{-I}X_I'X_I W_{-I}X_I'(I+U)^{-1}X_I]W_{-I}$$
$$= W_{-I}[I+X_I'\{-(I+U)^{-1} + I - U(I+U)^{-1} - (I+U)^{-1}$$
$$+ (I+U)^{-1}U(I+U)^{-1} - (I+U)^{-1}U$$
$$+ (I+U)^{-1}UU(I+U)^{-1}\}X_I W_{-I}]$$

using (6.23) and (6.24)

$$= W_{-I}[I+X_I'\{-(I+U)^{-1} + (I+U)^{-1}$$
$$- (I+U)^{-1}[I-U(I+U)^{-1} + U(I-U(I+U)^{-1})]\}X_I W_{-I}]$$
$$= W_{-I}[I-X_I'(I+U)^{-1}\{I-U(I+U)^{-1} + U(I+U)^{-1}\}X_I W_{-I}]$$
$$= W_{-I} - W_{-I}X_I'(I+U)^{-1}X_I W_{-I} = W^* .$$

Hence $W^* = W_{-I} - W_{-I}X_I'(I+U)^{-1}X_I W_{-I} = (X'X)^-$ (6.28)

from property (6.4).

Then $(X'X)^-X_I' = W_{-I}X_I' - W_{-I}X_I'(I+U)^{-1}X_I W_{-I}X_I'$ from (6.28)

$$= W_{-I}X_I'[I-(I+U)^{-1}U]$$ from (6.23)

$$= W_{-I}X_I'(I+U)^{-1}$$ using (6.26). (6.29)

$$(I-P_I) = I - X_I(X'X)^{-}X_I' \qquad \text{by definition (6.21)}$$

$$= I - X_I W_{-I} X_I'(I+U)^{-1} \qquad \text{from (6.29)}$$

$$= I - U(I+U)^{-1} \qquad \text{from (6.23)}$$

$$= (I+U)^{-1} \qquad \text{using (6.26).} \qquad (6.30)$$

$$r_I^{-} = y_I - X_I b^{-}$$

$$= y_I - X_I(X'X)^{-}X'y \qquad \text{from (6.5)}$$

$$= y_I - X_I(X'X)^{-}(X_{-I}'y_{-I}+X_I'y_I)$$

$$= (I-X_I(X'X)^{-}X_I')y_I - X_I(X'X)^{-}X_{-I}'y_{-I}$$

$$= (I-P_I)y_I - (I+U)^{-1}X_I W_{-I} X_{-I}'y_{-I} \qquad \text{from (6.21) and (6.29)}$$

$$= (I-P_I)(y_I - X_I b_{-I}^{-}) \qquad \text{from (6.30).} \qquad (6.31)$$

Now $b^{-} = (X'X)^{-}X'y$

$$= (X'X)^{-}(X_{-I}'y_{-I}+X_I'y_I)$$

$$= (W_{-I}-W_{-I}X_I'(I+U)^{-1}X_I W_{-I})(X_{-I}'y_{-I}+X_I'y_I) \qquad \text{from (6.28)}$$

$$= W_{-I}X_{-I}'y_{-I} - W_{-I}X_I'(I+U)^{-1}X_I W_{-I}X_{-I}'y_{-I}$$
$$+ W_{-I}X_I'y_I - W_{-I}X_I'(I+U)^{-1}X_I W_{-I}X_I'y_I$$

$$= b_{-I}^{-} - W_{-I}X_I'[(I+U)^{-1}X_I b_{-I}^{-} - (I-(I+U)^{-1}U)y_I] \qquad \text{from (6.23)}$$

$$= b_{-I}^{-} - W_{-I}X_I'[(I+U)^{-1}X_I b_{-I}^{-} - (I+U)^{-1}y_I] \qquad \text{using (6.26)}$$

$$= b_{-I}^{-} + W_{-I}X_I'(I+U)^{-1}(y_I - X_I b_{-I}^{-})$$

$$= b_{-I}^{-} + (X'X)^{-}X_I'(y_I - X_I b_{-I}^{-}) \qquad \text{from (6.29)}$$

$$= b_{-I}^{-} + (X'X)^{-}X_I'(I-P_I)^{-1}r_I^{-} \qquad \text{from (6.31).}$$

Q.E.D.

From (6.19) and (6.20)

$$\Delta_I^- = \frac{1}{qs_{-I}^2}(r_I^-)'(I-P_I)^{-1}X_I(X'X)^-X_I'(I-P_I)^{-1}(r_I^-)$$

$$= \frac{1}{qs_{-I}^2}(r_I^-)'(I-P_I)^{-1}P_I(I-P_I)^{-1}(r_I^-) \quad \text{from (6.21).} \tag{6.32}$$

This is the g-inverse equivalent of (3.15).

Let $P_I = \Gamma'\Lambda\Gamma$ (6.33)

where Γ is orthogonal and $\Lambda = \text{diag}(\lambda_1,\ldots,\lambda_m)$ with

$0 \leq \lambda_1 \leq \ldots \leq \lambda_m < 1$, and $g = (g_j) = \Gamma r_I^-$. (6.34)

Then $$\Delta_I^- = \frac{1}{qs_{-I}^2}\sum_{j=1}^{m} g_j^2 \frac{\lambda_j}{(1-\lambda_j)^2} \tag{6.35}$$

$$= \frac{1}{q}\sum_{j=1}^{m}\left(\frac{g_j}{s_{-I}\sqrt{1-\lambda_j}}\right)^2 \frac{\lambda_j}{1-\lambda_j}$$

$$= \frac{1}{q}\sum_{j=1}^{m} h_j^2 \frac{\lambda_j}{1-\lambda_j}, \tag{6.36}$$

where $h_j = \dfrac{g_j}{s_{-I}\sqrt{1-\lambda_j}}$. (6.37)

The generalization of the studentized residuals is given by

$$\Sigma h_j^2 = (r_I^-)'(I-P_I)^{-1}(r_I^-)/s_{-I}^2 .$$

Let $(t_I^-)^2 = \dfrac{(r_I^-)'(I-P_I)^{-1}(r_I^-)}{ms_{-I}^2}$. (6.38)

Theorem 6.5:

$$\frac{(r_I^-)'(I-P_I)^{-1}(r_I^-)}{\sigma^2} \sim \chi^2_f(\nu) \qquad (6.39)$$

with degrees of freedom $f = m$

and noncentrality $\nu = \beta'(I-(X'X)^-X'X)X_I'(I-P_I)^{-1}X_I(I-(X'X)^-X'X)\beta$.

Proof:

$$r^- = X(I-(X'X)^-X'X)\beta + (I-X(X'X)^-X')\varepsilon \text{ from (6.9)}$$

$$= \mu + (I-P)\varepsilon \qquad (6.40)$$

where $\mu = X(I-(X'X)^-X'X)\beta$ (6.41)

and $P = X(X'X)^-X'$ as before.

Without loss of generality it may be assumed that

$$r_I^- = (I_m \; 0)r^-$$

$$= (I_m \; 0)(\mu+(I-P)\varepsilon).$$

Let $(I-P) = N$ be partitioned as $N = \begin{bmatrix} N_m & N_a \\ N_a' & N_{n-m} \end{bmatrix}$. (6.42)

Note that $N_m = (I-P_I)$.

Now $(r_I^-)'N_m^{-1}(r_I^-) = (\mu'+\varepsilon'N)\begin{pmatrix} I_m \\ 0 \end{pmatrix}N_m^{-1}(I_m \; 0)(\mu+N\varepsilon)$

$$= \varepsilon'NN^*N\varepsilon + 2\mu'N^*N\varepsilon + \mu'N^*\mu \qquad (6.43)$$

where $N^* = \begin{bmatrix} N_m^{-1} & 0 \\ 0 & 0 \end{bmatrix}$.

(a) $NN^*NNN^*N = NN^*NN^*N$ since $N = (I-P)$ is idempotent

$= NN^*N$ since $N^*NN^* = N^*$.

(b) $2\mu'N^*NNN^*N = 2\mu'N^*NN^*N$

$= 2\mu'N^*N.$

(c) $\frac{1}{4}(2\mu'N^*N)(2\mu'N^*N)' = \mu'N^*NNN^*\mu$

$= \mu'N^*\mu$.

Hence the conditions of theorem 3.1 are satisfied.

$$\mu'N^*\mu = \beta'(I-(X'X)^-X'X)X'\begin{pmatrix} I_m \\ 0 \end{pmatrix}N_m^{-1}(I_m \ 0)X(I-(X'X)^-X'X)\beta$$

$$= \beta'(I-(X'X)^-X'X)X_I'(I-P_I)^{-1}X_I(I-(X'X)^-X'X)\beta \ .$$

$tr(NN^*N) = tr(NN^*) = m$ (c.f. (3.27)).

Q.E.D.

Theorem 6.6:

$(r_I^-)'(I-P_I)^{-1}(r_I^-)$ and S_{-I}^2 are independently distributed. (6.44)

Proof:

$S_{-I}^2 = \varepsilon'(M-MM^*M)\varepsilon$ from (5.37)

where $M = I-V = I-X'(X'X)^{-1}X'$

and $M^* = \begin{bmatrix} M_m^{-1} & 0 \\ 0 & 0 \end{bmatrix}$

$NM = [I-X(X'X)^-X'][I-X(X'X)^{-1}X']$ by definition

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

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

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

$= M$. (6.45)

$NN^*N(M-MM^*M) = NN^*NM - NN^*NMM^*M$

$= NN^*M - NN^*MM^*M$ from (6.45)

$= NN^*M - NN^*M = 0$ using $N^*MM^* = N^*$.

Q.E.D.

From (6.39) and (6.44), the statistic $(t_I^-)^2$ follows a noncentral F distribution with m and n-q-m degrees of freedom. The noncentrality is a function of c and the unknown β. Hence, it is impossible to determine accurate critical values for this statistic. Approximations may be obtained by estimating the noncentrality and then using tables for the noncentral F distribution.

The noncentrality parameter will be zero if $\lambda_1 = \ldots = \lambda_{q-c} = 0$. Thus the smaller the roots, the closer the noncentrality will be to zero. In the case of all $\lambda_1 = \ldots = \lambda_{q-c} = 0$, b^- is an unbiased estimator of β, (Judge, Griffiths, Hill and Lee (1980)).

Chapter 7

RESTRICTED LEAST SQUARES REGRESSION

In many applications in econometrics, the analyst may have prior information on particular regression coefficients or linear combinations thereof, and may wish to incorporate this information into the model. Alternatively, the prior information may form constraints upon the regression coefficients.

Again assume the linear model (2.1), and in addition let

$$H\beta = h \tag{7.1}$$

where h is an $\ell \times 1$ vector of known constants

H is an $\ell \times q$ design matrix that expresses the structure of the prior information on β.

Goldberger (1964) has shown that the least squares estimator for β, combining the model (2.1) and the restrictions (7.1) is

$$b^* = b + (X'X)^{-1}H'(H(X'X)^{-1}H')^{-1}(h-Hb) \tag{7.2}$$

where b is the unrestricted least squares estimator. For notational convenience we define

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

$$Z = (HWH')^{-1} \tag{7.3}$$

$$\delta = h - H\beta \tag{7.4}$$

then $$b^* = b + WH'Z(h-Hb) \; . \tag{7.5}$$

From (7.5) we have

$$E(b^*) = \beta + WH'Z\delta$$

$$\text{var}(b^*) = \sigma^2(W-WH'ZHW).$$

Therefore, b^* is a biased estimator of β unless $\delta = 0$, i.e. the restrictions on β are exactly correct. But

$$\text{var}(b) - \text{var}(b^*) = \sigma^2 WH'ZHW$$

which is nonnegative definite, and so b^* has smaller variance than the ordinary least squares estimator b, whether or not b^* is biased.

Turning now to the residuals, let

$$\begin{aligned} r^* &= y - Xb^* \\ &= y - X(b+WH'Z(h-Hb)) \qquad \text{from (7.5)} \\ &= y - XWX'y - XWH'Zh + XWH'ZHWX'y \\ &= [I - X(W-WH'ZHW)X']y - XWH'Zh \\ &= [I - X(W-WH'ZHW)X'](X\beta+\varepsilon) - XWH'Zh \\ &= XWH'ZH\beta + [I-X(W-WH'ZHW)X']\varepsilon - XWH'Zh \; . \end{aligned} \tag{7.6}$$

Hence $E(r^*) = XWH'Z(H\beta-h) \neq 0$ unless $\delta = 0$

$$\begin{aligned} \text{var}(r^*) &= \sigma^2[I-X(W-WH'ZHW)X'] \\ &= \sigma^2(I-T) \end{aligned} \tag{7.7}$$

where $T = X(W-WH'ZHW)X'$ is the equivalent of the hat matrix and $(I-T)$ is idempotent.

Theorem 7.1:

$$\text{var}(r^*) - \text{var}(r) \text{ is nonnegative definite.} \tag{7.8}$$

Proof:

$$\text{var}(r^*)-\text{var}(r) = \sigma^2(I-XWX'+XWH'ZHWX') - \sigma^2(I-XWX')$$

from (7.7) and (2.5)

$$= \sigma^2 XWH'ZHWX' \qquad (7.9)$$

which is a nonnegative definite quadratic form.

Q.E.D.

Theorem 7.2:

The total variance of r^* is greater than the total variance of r;

$$\text{tr}(\text{var}(r^*)) - \text{tr}(\text{var}(r)) = \sigma^2 \ell > 0 \ . \qquad (7.10)$$

Proof:

$$\begin{aligned} \text{tr}(\text{var}(r^*))-\text{tr}(\text{var}(r)) &= \sigma^2 \text{tr}(XWH'ZHWX') \quad \text{from (7.9)} \\ &= \sigma^2 \text{tr}(WX'XWH'ZH) \\ &= \sigma^2 \text{tr}(HWH'Z) \\ &= \sigma^2 \ell \quad \text{from (7.3).} \end{aligned}$$

Q.E.D.

Theorem 7.3: (Judge, Griffiths, Hill and Lee (1980)).

The generalized mean square error of r^* is less than that of r provided that

$$\delta'[H(X'X)^{-1}H']\delta/\sigma^2 \leq 1 \ . \qquad (7.11)$$

If the condition (7.11) is true, then the restricted least squares residual r^* is a better estimator of the unknown ε than the ordinary least squares residual r, under the generalized mean square error criterion. Note that if the restrictions are exact, i.e. $\delta = 0$, then (7.11) is always

true. However, r^* is biased unless $\delta = 0$, and from (7.8) and (7.10) r^* has larger variance than r. This would suggest that it is preferable to use the least squares residuals r in the estimation of σ^2.

The following statistic may be used for testing whether the i^{th} data point is an outlier:

$$t_i^* = \frac{r_i^*}{s_{-i}\sqrt{1-t_{ii}}} \qquad i = 1,\ldots,n \tag{7.12}$$

where t_{ii} is the i^{th} diagonal element of T. The statistic $(t_i^*)^2$ follows a noncentral F distribution with 1 and $n-q-1$ degrees of freedom. From (7.6) the noncentrality is a function of the known parameters H and h, and the unknown β. If $\delta = 0$, then $(t_i^*)^2$ will have a central F distribution. These results are proved in the general case in theorems 7.4 and 7.8.

Consider any subset r_I^* of m elements of the residuals r^* indexed by the set I. Without loss of generality, we may write

$$r_I^* = (I_m \; 0)r^* \; .$$

Now

$$\begin{aligned} r^* &= XWH'Z(H\beta-h) + (I-T)\varepsilon \qquad \text{from (7.6)} \\ &= \mu + (I-T)\varepsilon \end{aligned}$$

where $\mu = XWH'Z\delta$. (7.13)

Therefore $r_I^* = (I_m \; 0)(\mu+(I-T)\varepsilon)$. (7.14)

Let $(I-T) = N$ be partitioned as $N = \begin{bmatrix} N_m & N_a \\ N_a' & N_{n-m} \end{bmatrix}$,

then $I-T_I = N_m$, where $T_I = X_I(W-WH'ZHW)X_I'$. (7.15)

Note that

$$\begin{aligned}\mu'(I-T) &= (H\beta-h)'ZHWX'(I-XWX'+XWH'ZHWX') \\ &= (H\beta-h)'(ZHWX'-ZHWX'XWX'+ZHWX'XWH'ZHWX') \\ &= (H\beta-h)'ZHWX' \\ &= \mu' \ . \end{aligned} \tag{7.16}$$

Theorem 7.4:

$$\frac{(r_I^*)'(I-T_I)^{-1}(r_I^*)}{\sigma^2} \sim \chi_m^2(\nu) \tag{7.17}$$

with noncentrality $\nu = \delta'ZHWX_I'(I-T_I)^{-1}X_IWH'Z\delta$.

Proof:

$$(r_I^*)'N_m^{-1}(r_I^*) = (\mu'+\varepsilon'N)\begin{pmatrix} I_m \\ 0 \end{pmatrix}N_m^{-1}(I_m \ 0)(\mu+N\varepsilon) \quad \text{from (7.14)}$$

$$= \varepsilon'NN^*N\varepsilon + 2\mu'N^*N\varepsilon + \mu'N^*\mu \tag{7.18}$$

where $N^* = \begin{bmatrix} N_m^{-1} & 0 \\ 0 & 0 \end{bmatrix}$.

(a) $NN^*NNN^*N = NN^*NN^*N$ since $N = (I-T)$ is idempotent

$= NN^*N$ since $N^*NN^* = N^*$.

(b) $2\mu'N^*NNN^*N = 2\mu'N^*NN^*N$

$= 2\mu'N^*N$.

(c) $\frac{1}{4}[2\mu'N^*N][2\mu'N^*N]' = \mu'N^*NNN^*\mu$

$= \mu'N^*\mu$.

Hence the conditions of theorem 3.1 are satisfied.

$$\mu'N^*\mu = \delta'ZHWX'\begin{pmatrix} I_m \\ 0 \end{pmatrix} N_m^{-1} (I_m \; 0) XWH'Z\delta$$

$$= \delta'ZHWX_I'(I-T_I)^{-1}X_IWH'Z\delta \; . \qquad (7.19)$$

$$tr(NN^*N) = tr(NN^*) = m \; . \qquad (7.20)$$

Q.E.D.

Theorem 7.5:

$$\frac{(r^*)'(r^*)}{\sigma^2} \sim \chi_f^2(\nu) \qquad (7.21)$$

with degrees of freedom $f = n-q+\ell$

and noncentrality $\nu = \delta'Z\delta$.

Proof:

$$(r^*)'(r^*) = [\mu+(I-T)\varepsilon]'[\mu+(I-T)\varepsilon] \quad \text{from (7.13)}$$

$$= \varepsilon'(I-T)\varepsilon + 2\mu'(I-T)\varepsilon + \mu'\mu \; .$$

(a) $(I-T)$ is idempotent.

(b) $2\mu'(I-T)(I-T) = 2\mu'(I-T)$.

(c) $\frac{1}{4}[2\mu'(I-T)][2\mu'(I-T)]' = \mu'(I-T)\mu$

$= \mu'\mu$ from (7.16).

Hence the conditions of theorem 3.1 are satisfied.

$$\mu'\mu = \delta'ZHWX'XWH'Z\delta$$

$$= \delta'Z\delta \; . \qquad (7.22)$$

$$tr(I-T) = tr(I_n - XWX' + XWH'ZHWX')$$

$$= n - tr(WX'X) + tr(WH'ZHWX'X)$$

$$= n-q+tr(ZHWH')$$

$$= n-q+\ell \; . \qquad (7.23)$$

Q.E.D.

Theorem 7.6:

$$\frac{(S^*_{-I})^2}{\sigma^2} \sim \chi^2_f(\nu) \tag{7.24}$$

with degrees of freedom $f = n-q-m+\ell$

and noncentrality $\nu = \delta'Z[I-HWX_I'(I-T_I)^{-1}X_IWH']Z]\delta$,

where $(S^*_{-I})^2 = (r^*)'(r^*) - (r^*_I)'(I-T_I)^{-1}(r^*_I)$.

Proof:

$$\begin{aligned}(S^*_{-I})^2 &= (r^*)'(I-N^*)(r^*)\\ &= (\mu+N\varepsilon)'(I-N^*)(\mu+N\varepsilon) \quad \text{from (7.13)}\\ &= \varepsilon'(N-NN^*N)\varepsilon + 2\mu'(I-N^*)N\varepsilon + \mu'(I-N^*)\mu. \end{aligned} \tag{7.25}$$

(a) $(N-NN^*N)(N-NN^*N) = N-2NN^*N+NN^*NNN^*N$

$= N-NN^*N$ using $N^2 = N$ and $N^*NN^* = N^*$ (7.26)

(b) $2\mu'(I-N^*)N(N-NN^*N) = 2\mu'(I-N^*)(N-NN^*N)$

$= 2\mu'(N-N^*N-NN^*N+N^*NN^*N)$

$= 2\mu'(NN-NN^*N)$ since $N = NN$

$= 2\mu'N(I-N^*)N$

$= 2\mu'(I-N^*)N$ using (7.16).

(c) $\frac{1}{4}[2\mu'(I-N^*)N][2\mu'(I-N^*)N]' = \mu'(I-N^*)NN(I-N^*)\mu$

$= \mu'(N-N^*N)(N-NN^*)\mu$

$= \mu'N(N-N^*N)(N-NN^*)N\mu$

using (7.16)

$= \mu'(N-NN^*N)(N-NN^*N)\mu$

$= \mu'(N-NN^*N)\mu$ from (7.26)

$= \mu'(NN-NN^*N)\mu$

$= \mu'(I-N^*)\mu$ using (7.16).

Hence the conditions of theorem 3.1 are satisfied.

$$\mu'(I-N^*)\mu = \mu'\mu - \mu'N^*\mu$$
$$= \delta'Z[I-HWX_I'(I-T_I)^{-1}X_IWH'Z]\delta \quad \text{from (7.22) and (7.19)}$$

$$tr(N-NN^*N) = tr(N) - tr(NN^*N)$$
$$= n-q+\ell-m \quad \text{from (7.23) and (7.20).}$$

Q.E.D.

Theorem 7.7:

$(r_I^*)'(I-T_I)^{-1}(r_I^*)$ and $(S_{-I}^*)^2$ are independently distributed. (7.27)

Proof:

$(r_I^*)'(I-T_I)^{-1}(r_I^*) = \varepsilon'NN^*N\varepsilon + 2\mu'N^*N\varepsilon + \mu'N^*\mu$ from (7.18)

$(S_{-I}^*)^2 = \varepsilon'(N-NN^*N)\varepsilon + 2\mu'(I-N^*)N\varepsilon + \mu'(I-N^*)\mu$ from (7.25)

But $NN^*N(N-NN^*N) = NN^*N - NN^*N = 0$ using $N^2 = N$

and $N^*NN^* = N^*$.

Q.E.D.

Theorem 7.8:

$(r_I^*)'(I-T_I)^{-1}(r_I^*)$ and S_{-I}^2 are independently distributed. (7.28)

Proof:

$$S_{-I}^2 = \varepsilon'(M-MM^*M)\varepsilon \quad \text{from (5.37)}$$

where $M = I-V = I-X(X'X)^{-1}X'$

and $M^* = \begin{bmatrix} M_m^{-1} & 0 \\ 0 & 0 \end{bmatrix}$.

$$NN^*N(M-MM^*M) = NN^*NM - NN^*NMM^*M$$
$$= NN^*M - NN^*MM^*M$$
since $NM = M$ by direct multiplication
$$= NN^*M - NN^*M = 0$$
since $N^*MM^* = N^*$.

Q.E.D.

We are now able to generalize the test for outliers. The extension of (7.12) to a group of m observations is

$$(t_I^*)^2 = \frac{(r_I^*)'(I-T_I)^{-1}(r_I^*)}{ms_{-I}^2} \quad . \tag{7.29}$$

From (7.17) and (7.28), $(t_I^*)^2$ follows a noncentral F distribution with m and $n-q-m$ degrees of freedom. If the restrictions on β are exact, then the distribution will be central F.

From (7.8) and (7.10) it would appear that the least squares residuals should always be used to estimate σ^2. The estimator $s_{-I}^2 = S_{-I}^2/(n-q-m) \sim$ central $\sigma^2\chi^2$ with $n-q-m$ degrees of freedom. Note, however, from (7.24) that the estimator $(s_{-I}^*)^2 = (S_{-I}^*)^2/(n-q-m+\ell) \sim$ central $\sigma^2\chi^2$ with $n-q-m+\ell$ degrees of freedom if the restrictions on β are exact. Hence, if $\delta = 0$, we may gain ℓ degrees of freedom (where ℓ is the number of linear restrictions on β) by using the estimator $(s_{-I}^*)^2$. This may be valuable in cases where the number of degrees of freedom is small. The statistic

$$(t_I^{**})^2 = \frac{(r_I^*)'(I-T_I)^{-1}(r_I^*)}{m(s_{-I}^*)^2} \tag{7.30}$$

may be used for outlier testing if $H\beta = h$. From (7.24) and (7.27), $(t_I^{**})^2$ follows a central χ^2 distribution with m and $n-q-m+\ell$ degrees of freedom provided $\delta = 0$. When $m = 1$, (7.12) may be amended similarly.

If $\delta \neq 0$, then $(t_I^{**})^2$ is doubly non-central, and the non-centrality is unknown. In this case, it is recommended that the statistic $(t_I^{*})^2$ be used.

Furthermore, some analysts may prefer to use the estimator $(s^*)^2 = (r^*)'(r^*)/(n-q+\ell)$. From (7.21), this estimator yields a further gain in degrees of freedom, and is computationally easy to use. The distribution of the statistic $(r_I^*)'(I-T_I)^{-1}(r_I^*)/m(s^*)^2$ is a monotonic function of $F(m,n-q+\ell)$ c.f. (2.10). We prefer a statistic which follows the easily recognizable F distribution directly, and hence do not recommend the use of the estimator $(s^*)^2$.

Let us now consider a measure of the influence of a group of m observations indexed by the set I. The influence of a single observation is simply the special case $m = 1$. An appropriate measure would be

$$\Delta_I^* = \frac{(b_{-I}^* - b^*)'X'X(b_{-I}^* - b^*)}{qs_{-I}^2} \tag{7.31}$$

where the subscripts have their usual meanings. Δ_I^* is a simple extension of (3.12).

From (7.5), the nature of the estimator b^* does not allow a simple expression for $(b_{-I}^* - b^*)$. From (7.5)

$$b^* = (I - WH'ZH)b + WH'Zh \quad . \tag{7.32}$$

Let $W_{-I} = (X'_{-I}X_{-I})^{-1}$ (7.33)

$$Z_{-I} = (HW_{-I}H')^{-1} \tag{7.34}$$

then $b^*_{-I} = (I - W_{-I}H'Z_{-I}H)b_{-I} + W_{-I}H'Z_{-I}h$. (7.35)

Now $H(b^*_{-I} - b^*) = Hb_{-I} - HW_{-I}H'Z_{-I}Hb_{-I} + HW_{-I}H'Z_{-I}h$

$- Hb + HWH'ZHb - HWH'Zh$ from (7.32) and (7.35)

$= Hb_{-I} - Hb_{-I} + h - Hb + Hb - h$

using (7.3) and (7.34)

$= 0$. (7.36)

The equation (7.36) is a system of ℓ simultaneous equations in q unknowns. An analytic solution for $(b^*_{-I} - b^*)$ will exist only if the rank of H is $\geq q$; that is, the number of constraints must be at least as great as the number of variables in the regression. In practice, the number of constraints ℓ will tend to be small, so we shall not be able to find $(b^*_{-I} - b^*)$ in general.

We would nevertheless like a computational expression for b^*_{-I} which would enable us to compute its value without the need for performing a complete regression on the truncated data set.

Bingham (1977) showed that

$$W_{-I} = W + WX'_I(I - V_I)^{-1}X_I W \tag{7.37}$$

where $V_I = X_I W X'_I$ as before.

Using (7.37) we have

$$Z_{-I} = (HWH' + HWX_I'(I-V_I)^{-1}X_IWH')^{-1} . \qquad (7.38)$$

Now $b^*_{-I} = (I-W_{-I}H'Z_{-I}H)b_{-I} + W_{-I}H'Z_{-I}h$ from (7.35)

$= (I-W_{-I}H'Z_{-I}H)(b-WX_I'(I-V_I)^{-1}r_I)+W_{-I}H'Z_{-I}h$ from (3.14)

$= [I-(W+WX_I'(I-V_I)^{-1}X_IW)H'Z_{-I}H](b-WX_I'(I-V_I)^{-1}r_I)$
$+ (W+WX_I'(I-V_I)^{-1}X_IW)H'Z_{-I}h$ from (7.37)

$= [I-(I+WX_I'(I-V_I)^{-1}X_I)WH'Z_{-I}H]b$
$-[I-(I+WX_I'(I-V_I)^{-1}X_I)WH'Z_{-I}H]WX_I'(I-V_I)^{-1}r_I$
$+(I+WX_I'(I-V_I)^{-1}X_I)WH'Z_{-I}h$

$= (I-CH)b - (I-CH)WX_I'(I-V_I)^{-1}r_I + Ch$

$= b + C(h-Hb) - (I-CH)WX_I'(I-V_I)^{-1}r_I \qquad (7.39)$

where $C = (I+WX_I'(I-V_I)^{-1}X_I)WH'Z_{-I}$. $\qquad (7.40)$

Hence the procedure for computing b^*_{-I} would be

1. compute the value of $Z_{-I}(\ell\times\ell)$ from (7.38);
2. compute the value of $C\ (q\times\ell)$ from (7.40);
3. compute b^*_{-I} from (7.39).

The procedure seems clumsy, with many matrix multiplications and inversion of the matrices $(I-V_I)$ and $(HWH'+HWX_I'(I-V_I)^{-1}X_IWH')$. However, these matrices are $m\times m$ and $\ell\times\ell$ respectively and hence small, and in fact many sequences of multiplication of small matrices are repeated and need to be computed once only.

Chapter 8

STOCHASTIC PRIOR INFORMATION

In many practical situations, prior information on the regression parameters will not be exact, or exact restrictions are inappropriate. In such cases, stochastic prior information may be incorporated into the model. Nagar and Kakwani (1964) have described how such stochastic information may be used.

Suppose, for example, we know with 95% probability that one element of β (β_1 say) lies between 0 and 1. Then the range of β_1 is approximately $\frac{1}{2}\pm 2\times\frac{1}{4}$. It is clear from (8.1) below how such a constraint can be incorporated into the model.

In many studies we find a situation where identical or similar analysis has been performed in another place or at a previous time. Hence prior information on the model is available, and it may be advantageous to augment our study with this information. By its very nature, any previous statistical analysis will be stochastic, that is information about parameters will not be exact; hence the form of restricted least squares regression presented in the previous chapter will not be applicable.

Suppose we assume that

$$H\beta + u = h \tag{8.1}$$

where h is an $\ell \times 1$ vector of known constants

H is an $\ell \times q$ design matrix

u is an $\ell \times 1$ normally distributed random vector with mean δ and covariance $\sigma^2 R$ with R known. Also assume that R is nonsingular.

In addition, we have the model from (2.1)

$$y = X\beta + \varepsilon$$

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

Combining (2.1) and (8.1), we may rewrite the model as

$$\begin{bmatrix} y \\ h \end{bmatrix} = \begin{bmatrix} X \\ H \end{bmatrix} \beta + \begin{bmatrix} \varepsilon \\ u \end{bmatrix} \tag{8.2}$$

or $$y^* = X^*\beta + \varepsilon^* \, . \tag{8.3}$$

$$E(\varepsilon^*) = E\begin{bmatrix} \varepsilon \\ u \end{bmatrix} = \begin{bmatrix} 0 \\ \delta \end{bmatrix} .$$

$$E(\varepsilon^*\varepsilon^{*\prime}) = \sigma^2 \begin{bmatrix} I & 0 \\ 0 & R \end{bmatrix} = \sigma^2 \Omega \, . \tag{8.4}$$

Note that if $R = I$ and $\delta = 0$, then (8.3) reduces to ordinary least squares regression. Thus we shall assume that $R \neq I$ and $\delta \neq 0$ for the following analysis to be meaningful.

The Aitken estimator for β is

$$b^* = (X'X + H'R^{-1}H)^{-1}(X'y + H'R^{-1}h) \, . \tag{8.5}$$

We shall use the generalized linear regression form, viz.

$$b^* = (X^{*\prime}\Omega^{-1}X^*)^{-1}X^{*\prime}\Omega^{-1}y^* \, . \tag{8.6}$$

For notational convenience, define

$$A = (X^{*\prime}\Omega^{-1}X^{*})^{-1} = (X'X+H'R^{-1}H)^{-1} \tag{8.7}$$

then $$b^{*} = AX^{*\prime}\Omega^{-1}y^{*} \ . \tag{8.8}$$

Now $$E(b^{*}) = \beta + AH'R^{-1}\delta$$

$$\text{var}(b^{*}) = \sigma^{2}A \tag{8.9}$$

therefore b^{*} is biased, unless $\delta = 0$. The effect of adding information to the model (2.1) will generally be that b^{*} has smaller variance than the ordinary least squares estimator b.

Let $$r^{*} = y^{*} - X^{*}b^{*}$$

$$= (I-X^{*}AX^{*\prime}\Omega^{-1})y^{*} \qquad \text{from (8.8)}$$

$$= (I-X^{*}AX^{*\prime}\Omega^{-1})(X^{*}\beta^{*}+\varepsilon^{*})$$

$$= (I-X^{*}AX^{*\prime}\Omega^{-1})\varepsilon^{*}$$

$$= (I-B)\varepsilon^{*} \tag{8.10}$$

where $B = X^{*}AX^{*\prime}\Omega^{-1}$, and $(I-B)$ is idempotent . (8.11)

Note that

$$(I-B)\Omega(I-B)' = (I-X^{*}AX^{*\prime}\Omega^{-1})\Omega(I-\Omega^{-1}X^{*}AX^{*\prime})$$

$$= \Omega-2X^{*}AX^{*\prime} + X^{*}AX^{*\prime}\Omega^{-1}X^{*}AX^{*\prime}$$

$$= \Omega-X^{*}AX^{*\prime}$$

$$= (I-B)\Omega \ . \tag{8.12}$$

From (8.10), $E(r^{*}) = (I-X^{*}AX^{*\prime}R^{-1})\delta \neq 0$ unless $\delta = 0$.

$$\text{var}(r^{*}) = \sigma^{2}(I-B)\Omega(I-B)' \tag{8.13}$$

$$= \sigma^{2}(I-B)\Omega \qquad \text{from (8.12).} \tag{8.14}$$

Now $\text{var}(r^{*})$ is a $(n+\ell)\times(n+\ell)$ matrix while $\text{var}(r)$ is $n\times n$, and thus the two cannot be compared directly. The

difference in the total variance between the ordinary and generalized least squares residuals is dependent upon the structure of the prior information, i.e. the structure of H and R.

Theorem 8.1: (Judge and Bock (1978))

The generalized mean square error of r^* is less than that of r provided that

$$\frac{\delta'[H(X'X)^{-1}H'+R]^{-1}\delta}{\sigma^2} \leq 1 . \qquad (8.15)$$

If the condition (8.15) is true, then the generalized least squares residual r^* is a better estimate of the unknown ε than the ordinary least squares residual r, under the generalized mean square error criterion. Note that if the prior information (or restrictions) are correct on average, i.e. $\delta = 0$, then (8.15) is always true. However, r^* is biased unless $\delta = 0$, and its variability relative to r is unknown.

We shall consider the outlier test in the general case for groups of m observations indexed by the set I. The case of a single outlier is then the special case $m = 1$. Note that outliers (and influential observations) will be sought only amongst the sample data, that is the first n elements of r^*.

Consider any subset r_I^* of m elements from the first n elements of r^*. Without loss of generality

$$r_I^* = (I_m \; 0)r^*$$

and $r^* = (I-B)\varepsilon^*$ from (8.10)

$$r_I^* = (I_m \; 0)(I-B)\varepsilon^* \; . \tag{8.16}$$

Let $M = I-B$

$$N = (I-B)\Omega(I-B)' = M\Omega M' \; . \tag{8.17}$$

Partition N as $N = \begin{bmatrix} N_m & N_a \\ N_a' & N_{n+\ell-m} \end{bmatrix}$

Then N_m is the $m \times m$ submatrix of N formed by the rows and columns indexed by I. From (8.13)

$$\text{var}(r_I^*) = \sigma^2 N_m \; .$$

Note that from (8.12)

$$N = \Omega - X^* A X^{*\prime}$$

$$= \begin{bmatrix} I & 0 \\ 0 & R \end{bmatrix} - \begin{bmatrix} X \\ H \end{bmatrix} A \begin{bmatrix} X' & H' \end{bmatrix} \quad \text{from (8.2) and (8.4)}$$

$$= \begin{bmatrix} I-XAX' & -XAH' \\ -HAX' & R-HAH' \end{bmatrix} \; .$$

As the index set I will contain elements $\leq n$ only, we may write

$$N_m = I - X_I A X_I' \; . \tag{8.18}$$

Theorem 8.2: (Searle (1971))

If $x \sim N(\mu, \Omega)$ then $x'Ax$ will have a noncentral χ^2 distribution with $f = \text{tr}(A\Omega)$ degrees of freedom and noncentrality $\nu = \frac{1}{2}\mu' A \mu$ if and only if

$$\Omega A \Omega A \Omega = \Omega A \Omega \; .$$

In addition, if Ω is nonsingular, then the above condition is

$$A \Omega A = A \tag{8.19}$$

Theorem 8.3:

$$\frac{(r_I^*)' N_m^{-1} (r_I^*)}{\sigma^2} \sim \chi_m^2(\nu) \qquad (8.20)$$

with noncentrality $\nu = \frac{1}{2}(0\ \delta')(I-B)' \begin{pmatrix} N_m^{-1} & 0 \\ 0 & 0 \end{pmatrix} (I-B) \begin{pmatrix} 0 \\ \delta \end{pmatrix}$.

Proof:

$$(r_I^*)' N_m^{-1} (r_I^*) = \varepsilon^{*\prime} M' \begin{pmatrix} I_m \\ 0 \end{pmatrix} N_m^{-1} (I_m\ 0) M\varepsilon^* \qquad \text{from (8.16)}$$

$$= \varepsilon^{*\prime} M' N^* M \varepsilon^* \qquad (8.21)$$

where $N^* = \begin{bmatrix} N_m^{-1} & 0 \\ 0 & 0 \end{bmatrix}$.

Now $M'N^*M\Omega M'N^*M = M'N^*NN^*M$ since $M\Omega M' = N$

$$= M'N^*M \qquad \text{since } N^*NN^* = N^* \qquad (8.22)$$

Hence condition (8.19) of theorem 8.2 is established.

The degrees of freedom are

$$\operatorname{tr}(M'N^*M\Omega) = \operatorname{tr}(N^*N) = m \ . \qquad (8.23)$$

Q.E.D.

Theorem 8.4:

$$\frac{(r^*)' \Omega^{-1} (r^*)}{\sigma^2} \sim \chi_f^2(\nu) \qquad (8.24)$$

with degrees of freedom $f = n-q+\ell$

and noncentrality $\nu = \frac{1}{2}(0\ \delta')(I-B)' \Omega^{-1} (I-B) \begin{pmatrix} 0 \\ \delta \end{pmatrix}$.

Proof:

$$(r^*)'\Omega^{-1}(r^*) = \varepsilon^*'M'\Omega^{-1}M\varepsilon^* \quad \text{from (8.10)} \tag{8.25}$$

Now $M'\Omega^{-1}M\Omega M'\Omega^{-1}M = M'\Omega^{-1}M\Omega\Omega^{-1}M$ from (8.12)

$$= M'\Omega^{-1}M \quad \text{since } M \text{ is idempotent.} \tag{8.26}$$

Hence condition (8.19) of theorem 8.2 is satisfied.

Also $\operatorname{tr}(M'\Omega^{-1}M\Omega) = \operatorname{tr}(\Omega^{-1}M\Omega)$ from (8.12)

$$= \operatorname{tr}(M)$$

$$= \operatorname{tr}(I - X^*AX^{*\prime}\Omega^{-1})$$

$$= n+\ell - \operatorname{tr}(AX^{*\prime}\Omega^{-1}X^*)$$

$$= n+\ell-q \quad \text{using (8.7)} \tag{8.27}$$

Q.E.D.

Theorem 8.5:

Let $(S^*_{-I})^2 = (r^*)'\Omega^{-1}(r^*) - (r^*_I)'N_m^{-1}(r^*_I)$.

$$\frac{(S^*_{-I})^2}{\sigma^2} \sim \chi^2_f(\nu) \tag{8.28}$$

with degrees of freedom $f = n-q-m+\ell$

and noncentrality $\nu = \frac{1}{2}(0\ \delta')(I-B')\begin{pmatrix} I-N_m^{-1} & 0 \\ 0 & R^{-1} \end{pmatrix}(I-B)\begin{pmatrix} 0 \\ \delta \end{pmatrix}$.

Proof:

$$(S^*_{-I})^2 = \varepsilon^*'M'\Omega^{-1}M\varepsilon^* - \varepsilon^*'M'N^*M\varepsilon^* \quad \text{from (8.25) and (8.21)}$$

$$= \varepsilon^*'M'(\Omega^{-1}-N^*)M\varepsilon^* \ . \tag{8.29}$$

Now $M'(\Omega^{-1}-N^*)M\Omega M'(\Omega^{-1}-N^*)M$

$$= M'\Omega^{-1}M\Omega M'\Omega^{-1}M - M'\Omega^{-1}M\Omega M'N^*M - M'N^*M\Omega M'\Omega^{-1}M$$

$$+ M'N^*M\Omega M'N^*M \ .$$

But $M'\Omega^{-1}M\Omega M'\Omega^{-1}M = M'\Omega^{-1}M$ from (8.26)

$M'N^*M\Omega M'N^*M = M'N^*M$ from (8.22)

$M'N^*M\Omega M'\Omega^{-1}M = M'N^*M\Omega\Omega^{-1}M$ using (8.12)

$= M'N^*M$ since M is idempotent. (8.30)

Similarly $M'\Omega^{-1}M\Omega M'N^*M = M'N^*M$.

Then $M'(\Omega^{-1}-N^*)M\Omega M'(\Omega^{-1}-N^*)M = M'(\Omega^{-1}-N^*)M$

and condition (8.19) of theorem 8.2 is established.

Also $tr(M'(\Omega^{-1}-N^*)M\Omega) = n+\ell-q-m$ from (8.27) and (8.23).

Q.E.D.

Theorem 8.6:

$(r_I^*)'N_m^{-1}(r_I^*)$ and $(S_{-I}^*)^2$ are independently distributed. (8.31)

Proof:

$(r_I^*)'N_m^{-1}(r_I^*) = \varepsilon^{*\prime}M'N^*M\varepsilon^*$ from (8.21)

$(S_{-I}^*)^2 = \varepsilon^{*\prime}M'(\Omega^{-1}-N^*)M\varepsilon^*$ from (8.29)

Now $M'N^*M\Omega M'(\Omega^{-1}-N^*)M = M'N^*M-M'N^*M$ from (8.30) and (8.22)

$= 0$.

This establishes the independence, since $M'N^*M$ and $M'(\Omega^{-1}-N^*)M$ are symmetric, (Searle (1971)).

Q.E.D.

A statistic for testing whether a group of m observations is an outlying group is

$$(t_I^*)^2 = \frac{(r_I^*)'N_m^{-1}(r_I^*)}{m(s_{-I}^*)^2} \qquad (8.32)$$

where $(s_{-I}^*)^2 = [(r^*)'\Omega^{-1}(r^*)-(r_I^*)'N_m^{-1}(r_I^*)]/(n-q-m+\ell)$. (8.33)

From (8.20), (8.28) and (8.31), the statistic $(t_I^*)^2$ follows a doubly noncentral F distribution with m and $n-q-m+\ell$ degrees of freedom.

In order to have critical points for the statistic (8.32), we need to know the noncentrality parameters. From (8.20) and (8.28), the noncentrality is a function of the unknown δ. However, if the restrictions on β are unbiased, i.e. $\delta = 0$, then the noncentrality will be zero, and the statistic $(t_I^*)^2$ will follow a central F distribution with m and $n-q-m+\ell$ degrees of freedom. Therefore, if it is practically possible, we should endeavour to ensure that prior information or restrictions on the parameters are unbiased.

If $\delta = 0$, the statistic $(t_I^*)^2$ is also the likelihood ratio test for testing $H_0 : \theta_I = 0$ in the model

$$\begin{pmatrix} y \\ h \end{pmatrix} = \begin{pmatrix} X \\ H \end{pmatrix}\beta + \theta_I + \varepsilon^*$$

where θ_I is an $(n+\ell)\times 1$ vector of zeros with unknown parameters in the positions indexed by I.

In testing whether a single observation is an outlier, we should use the statistic

$$t_i^* = \frac{r_i^*}{s_{-i}^* \sqrt{n_{ii}}} \quad , \qquad i = 1,\ldots,n \tag{8.34}$$

where n_{ii} is the i^{th} diagonal element of $N = (I-B)\Omega(I-B)'$,

$$s_{-i}^{*2} = [(r^*)'\Omega^{-1}(r^*) - (r_i^*)^2/n_{ii}]/(n-q-1+\ell). \tag{8.35}$$

The statistic $(t_i^*)^2$ will have a doubly noncentral F distribution with 1 and $n-q-1+\ell$ degrees of freedom. Again, if $\delta = 0$, the noncentrality will be zero. Note that $n_{ii} = 1-x_i A x_i'$ (from (8.18)), where x_i is the i^{th} row of X.

Turning now to the measure of influence, an appropriate statistic would be

$$\Delta_I^* = \frac{(b_{-I}^*-b^*)'X'X(b_{-I}^*-b^*)}{q(s_{-I}^*)^2} \tag{8.36}$$

where the subscripts have their usual meanings, and s_{-I}^* is defined in (8.33). The geometric interpretation of Δ_I^* in terms of confidence ellipsoids is the same as that of Δ_I in (3.12), with degrees of freedom q and $n-q-m+\ell$ in (3.4).

Theorem 8.7:

$$b_{-I}^*-b^* = -AX_I'N_m^{-1}r_I^* \tag{8.37}$$

where $N_m^{-1} = (I-X_I AX_I')^{-1}$ as before.

Proof:

Let $$A_{-I} = (X_{-I}'X_{-I} + H'R^{-1}H)^{-1} \tag{8.38}$$

$$U = X_I A_{-I} X_I' . \tag{8.39}$$

Now $$X'X+H'R^{-1}H = X_{-I}'X_{-I} + X_I'X_I + H'R^{-1}H \tag{8.40}$$

$$I-U(I+U)^{-1} = (I+U)(I+U)^{-1} - U(I+U)^{-1}$$

$$= (I+U)^{-1} . \tag{8.41}$$

Let $$A^* = A_{-I} - A_{-I}X_I'(I+U)^{-1}X_I A_{-I} \tag{8.42}$$

Multiply (8.40) by (8.42):

$$(X'X+H'R^{-1}H)A^* = (X'_{-I}X_{-I}+H'R^{-1}H)A_{-I}$$
$$- (X'_{-I}X_{-I}+H'R^{-1}H)A_{-I}X'_I(I+U)^{-1}X_IA_{-I}$$
$$+ (X'_IX_I)A_{-I} - (X'_IX_I)A_{-I}X'_I(I+U)^{-1}X_IA_{-I}$$
$$= I-X'_I(I+U)^{-1}X_IA_{-I}+(X'_IX_I)A_{-I}-X'_IU(I+U)^{-1}X_IA_{-I}$$
$$= I-X'_I[(I+U)^{-1}+I-U(I+U)^{-1}]X_IA_{-I}$$
$$= I-X'_I[(I+U)^{-1}-(I+U)^{-1}]X_IA_{-I} \quad \text{from (8.41)}$$
$$= I .$$

Hence $A^* = A = A_{-I} - A_{-I}X'_I(I+U)^{-1}X_IA_{-I}$. (8.43)

Then $AX'_I = A_{-I}X'_I - A_{-I}X'_I(I+U)^{-1}X_IA_{-I}X'_I$ from (8.43)

$$= A_{-I}X'_I[I-(I+U)^{-1}U]$$
$$= A_{-I}X'_I(I+U)^{-1} \quad \text{using (8.41).} \qquad (8.44)$$

$N_m = I-X_IAX'_I$ from (8.18)

$$= I-X_IA_{-I}X'_I(I+U)^{-1} \quad \text{from (8.44)}$$
$$= I-U(I+U)^{-1}$$
$$= (I+U)^{-1} \quad \text{using (8.41).} \qquad (8.45)$$

$r^*_I = y_I - X_Ib^*$

$$= y_I - X_IA(X'y+H'R^{-1}h) \quad \text{from (8.5)}$$
$$= y_I - X_IA(X'_{-I}y_{-I}+X'_Iy_I+H'R^{-1}h)$$
$$= (I-X_IAX'_I)y_I-X_IA(X'_{-I}y_{-I}+H'R^{-1}h)$$
$$= N_my_I-(I+U)^{-1}X_IA_{-I}(X'_{-I}y_{-I}+H'R^{-1}h) \quad \text{from (8.18) and (8.44)}$$
$$= N_m(y_I-X_Ib^*_{-I}) \quad \text{from (8.45) .} \qquad (8.46)$$

Now $b^* = A(X'y+H'R^{-1}h)$

$= A(X_{-I}'y_{-I}+X_I'y_I+H'R^{-1}h)$

$= (A_{-I}-A_{-I}X_I'(I+U)^{-1}X_IA_{-I})(X_{-I}'y_{-I}+X_I'y_I+H'R^{-1}h)$

$= A_{-I}(X_{-I}'y_{-I}+H'R^{-1}h)-A_{-I}X_I'(I+U)^{-1}X_IA_{-I}(X_{-I}'y_{-I}+H'R^{-1}h)$

$+ A_{-I}X_I'y_I-A_{-I}X_I'(I+U)^{-1}X_IA_{-I}X_I'y_I$

$= b_{-I}^*-A_{-I}X_I'[(I+U)^{-1}X_Ib_{-I}^*-y_I+(I+U)^{-1}Uy_I]$

$= b_{-I}^*-A_{-I}X_I'[(I+U)^{-1}X_Ib_{-I}^*-(I+U)^{-1}y_I]$ using (8.41)

$= b_{-I}^*+A_{-I}X_I'(I+U)^{-1}(y_I-X_Ib_{-I}^*)$

$= b_{-I}^*+AX_I'(y_I-X_Ib_{-I}^*)$ from (8.44)

$= b_{-I}^*+AX_I'N_m^{-1}r_I^*$ from (8.46)

Q.E.D.

From (8.36) and (8.37)

$$\Delta_I^* = \frac{1}{q(s_{-I}^*)^2}(r_I^*)'N_m^{-1}X_IAX'XAX_I'N_m^{-1}(r_I^*) . \qquad (8.47)$$

This computational form is very similar to the influence measure in ridge regression (c.f. (5.27)), and likewise it does not simplify to an expression analogous to (3.21)

To measure the influence of a single observation in the sample data, the above theory applies with $m = 1$. The influence measure will be

$$\Delta_i^* = \frac{(b_{-i}^*-b^*)'X'X(b_{-i}^*-b^*)}{q(s_{-i}^*)^2} \qquad (8.48)$$

$$= \frac{(r_i^*)'x_iAX'XAx_i'(r_i^*)}{qn_{ii}^2(s_{-i}^*)^2} \qquad (8.49)$$

where x_i is the i^{th} row of X

$n_{ii} = 1 - x_i A x_i'$ from (8.18)

s^*_{-i} is as in (8.35).

Chapter 9

MODELS WITH SELECTED VARIABLES

Suppose we have a near collinearity problem; say the i^{th} column of the X matrix is almost exactly a linear combination of other columns of X. If this variable were to be eliminated from the model, we would lose little information, as almost all information in the i^{th} variable is contained in other variables of the model. By eliminating the i^{th} variable, we would improve the conditioning of X'X, and obtain more reliable estimates of all parameters.

In many situations in econometrics, an economic variable of interest may be dependent upon an almost infinite number of other variables. Prediction of the variable may be improved by adding more and more variables to the prediction model. However, for practical limitations, we must choose a subset of variables which will give "adequate" prediction.

There are numerous well-known variable selection techniques (see, for example Graybill (1976) or Judge et al (1980)). Discussion of these techniques is beyond the scope of this study. Models with selected variables are used frequently in practice, to remove experimental design faults or to reduce computational effort, as described above.

In a model with q independent variables, suppose that

$p < q$ variables are retained and the rest omitted. Partition X and β as

$$X = (X_1 \; X_2) \qquad \beta = (\beta_1 \; \beta_2)$$

where X_1 is the $n \times p$ matrix of observations in the retained variables etc. Then the true model is

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon \tag{9.1}$$

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

As X_2 is discarded, the least squares estimate for β_1 is

$$b_1 = (X_1'X_1)^{-1}X_1'y \; . \tag{9.2}$$

Now

$$E(b_1) = (X_1'X_1)^{-1}X_1'(X_1\beta_1+X_2\beta_2)$$

$$= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 \tag{9.3}$$

$$\text{var}(b_1) = \sigma^2(X_1'X_1)^{-1} \; . \tag{9.4}$$

Judge et al (1980) show that b_1 is unbiased, and $\text{var}(b) - \text{var}(b_1)$ is nonnegative definite if $\beta_2 = 0$. If $\beta_2 \neq 0$ then b_1 is preferred to b under the generalized mean square error criterion if

$$\frac{\beta_2'(X_2'X_2-X_2'X_1(X_1'X_1)^{-1}X_1'X_2)^{-1}\beta_2}{\sigma^2} \lesssim q-p \; . \tag{9.5}$$

Let $r_1 = y - X_1b_1$

$$= (I-X_1(X_1'X_1)^{-1}X_1')y$$

$$= (I-X_1(X_1'X_1)^{-1}X_1')(X_1\beta_1+X_2\beta_2+\varepsilon)$$

$$= (I-X_1(X_1'X_1)^{-1}X_1')(X_2\beta_2+\varepsilon)$$

$$= (I-V_1)(X_2\beta_2+\varepsilon) \tag{9.6}$$

where $V_1 = X_1(X_1'X_1)^{-1}X_1'$.

Then $E(r_1) = 0$ only if $\beta_2 = 0$

$$\text{var}(r_1) = \sigma^2(I-V_1) \quad \text{as} \quad (I-V_1) \quad \text{is idempotent.} \qquad (9.7)$$

The total variance of r_1 is greater than that of the full model residuals r, since

$$\begin{aligned} \text{tr}(\text{var}(r_1)) - \text{tr}(\text{var}(r)) &= \sigma^2(n-p)-\sigma^2(n-q) \\ &= \sigma^2(q-p) > 0. \end{aligned} \qquad (9.8)$$

Having selected p independent variables and discarded all others in the original model, the analysis will proceed as if the $q-p$ discarded variables had not existed. Any of the foregoing regression techniques may be applied to the model with p independent variables. If $(X_1'X_1)$ is still ill-conditioned, then ridge regression or g-inverse regression may be chosen. Assume, however, that ordinary least squares regression will suffice. Then, following chapters 2 and 3, statistics for testing for outlying observations and groups are

$$(t_1)_i^2 = \frac{(r_1)_i^2}{(s_1)_{-i}^2[1-(v_1)_{ii}]} \qquad i = 1,\ldots,n \qquad (9.9)$$

$$(t_1)_I^2 = \frac{(r_1)_I'[I-(V_1)_I]^{-1}(r_1)_I}{m(s_1)_{-I}^2} \qquad (9.10)$$

where the notation follows earlier conventions, and the subscript 1 denotes the use of the p-variate model. The statistics $(t_1)_i^2$ and $(t_1)_I^2$ follow central $F(1,n-p-1)$ and $F(m,n-p-m)$ distributions respectively.

Measures of the influence of single observations and groups are

$$(\Delta_1)_i = \frac{1}{p}(t_1)_i^2(v_1)_{ii}/[1-(v_1)_{ii}] \quad \text{from (3.7)} \qquad (9.11)$$

$$(\Delta_1)_I = \frac{(r_1)_I'[I-(V_1)_I]^{-1}(V_1)_I[I-(V_1)_I]^{-1}(r_1)_I}{p(s_1)^2_{-I}} \quad \text{from (3.15)} \qquad (9.12)$$

As $\beta_2 \neq 0$ in general, we shall actually be working with biased statistics; we shall choose to ignore this fact in practice, as the discarded variables are in effect non-existent.

In eliminating certain variables, what we have in fact done is taken the full model and applied the prior restriction $\beta_2 = 0$. This is a special case of (7.1) in chapter 7. As $\beta_2 \neq 0$ in general, the restriction is not exactly correct, and therefore the theoretical considerations of chapter 7 apply with $\delta \neq 0$.

Of particular interest is the recommendation of chapter 7 that the estimator s^2_{-i} (or s^2_{-I}), i.e. the unrestricted least squares estimator for σ^2, be used when $\delta \neq 0$. In the present context, this means that we should use the unrestricted, full model to obtain an estimate for σ^2. This recommendation is supported by (9.8). The expected value of the restricted estimate of σ^2 is: (Judge et al (1980))

$$E(s_1^2) = \frac{\sigma^2+\beta_2'X_2'(I-V_1)X_2\beta_2}{n-p} \qquad (9.13)$$

and hence all estimates of σ^2 based on the p-variate model will be biased when $\beta_2 \neq 0$.

For practical purposes, it is always assumed that critical points for the statistics (9.9), (9.10), (9.11) and (9.12) can be obtained from tables of the central F distribution. However, if $\beta_2 \neq 0$, then the distributions are in fact doubly noncentral. By using an estimate of σ^2 based on the unrestricted model, we can make the denominator of the F statistic central, thus removing one source of inaccuracy.

Although the unrestricted estimator of σ^2 is distributionally preferable, it has in fact fewer degrees of freedom, and involves additional computational effort. In most practical situations, it is unlikely that we would wish to perform calculations using the full model, particularly if variable selection was used because of computational limitations. However, in situations where computational effort is not a problem, or where the full model is analysed anyway as part of the computer package, we may obtain more reliable estimates by using the unrestricted or full model residuals in estimating σ^2.

Chapter 10

MEASURES OF LEVERAGE

In chapter 3 we discussed the concept of leverage. A data point of high leverage was defined to be one with a large value of the ratio $v_{ii}/(1-v_{ii})$, where v_{ii} is the i^{th} diagonal element of the "hat" matrix $V = X(X'X)^{-1}X'$.

It was shown in chapter 3 that if $X'X$ is ill-conditioned, then this measure of leverage may be unreliable, due to the inversion of the $X'X$ matrix. When $X'X$ is ill-conditioned, it may be advantageous to define a g-inverse or ridge type measure of leverage.

The hat matrix V is important in the ordinary least squares regression theory, since

$$r = (I-V)\varepsilon \tag{10.1}$$

$$\text{var}(r) = (I-V)\sigma^2 \tag{10.2}$$

$$\text{var}(Xb) = V\sigma^2. \tag{10.3}$$

For notational convenience we shall denote the measure of the leverage of the i^{th} data point as

$$\ell_i = v_{ii}/(1-v_{ii}) \tag{10.4}$$

where v_{ii} is the i^{th} diagonal element of $V = X(X'X)^{-1}X'$.

Cook (1977) interpreted ℓ_i as being

$$\ell_i = \text{var}(\hat{y}_i)/\text{var}(r_i) \tag{10.5}$$

a measure of the sensitivity of estimation at x_i, the i^{th} row of X. The result (10.5) follows from (10.2) and (10.3).

From the interpretation of (3.7) and (3.21), the leverage of a group of m observations indexed by the set I will be defined as

$$\ell_I = \sum_{j=1}^{m} \frac{\lambda_j}{1-\lambda_j} \tag{10.6}$$

where (λ_j) are the eigenvalues of V_I.

The statistics ℓ_i and ℓ_I are functions of X, which is assumed to be a matrix of known constants. They do not follow any known statistical distributions, and hence we cannot determine whether a "large" value of ℓ_i (or ℓ_I) is statistically significant. Data points or groups with relatively high leverage indicate to the analyst the possibility of experimental design problems. They also indicate whether the influence of a particular point or group is the result of an outlier in the dependent variable or a design problem in X.

The statistics ℓ_i and ℓ_I are easily extended to g-inverse regression, since

$$r^- = f(\beta) + (I-P)\varepsilon \tag{10.7}$$

$$\text{var}(r^-) = (I-P)\sigma^2 \tag{10.8}$$

$$\text{var}(Xb^-) = P\sigma^2 \tag{10.9}$$

where $P = X(X'X)^-X'$. Hence the matrix P is the g-inverse equivalent of the hat matrix V, and measures of leverage

$\bar{\ell}_i$ and $\bar{\ell}_I$ can be defined in terms of the diagonal elements and eigenvalues of P, as in (10.4) and (10.6).

In ridge regression we have

$$\tilde{r} = f(\beta) + (I-\tilde{V})\varepsilon \qquad (10.10)$$

$$\text{var}(\tilde{r}) = (I-\tilde{V})^2\sigma^2 = (I-\tilde{V}-\tilde{U})\sigma^2 \qquad (10.11)$$

$$\text{var}(X\tilde{b}) = (\tilde{V}-\tilde{U})\sigma^2 \qquad (10.12)$$

where $\tilde{V} = X(X'X+kI)^{-1}X'$

$\tilde{U} \equiv X(X'X+kI)^{-2}X'$.

Hence there is no obvious ridge equivalent of the hat matrix. To be consistent with (10.5), we shall therefore define

$$\tilde{\ell}_i = (\tilde{v}_{ii}-\tilde{u}_{ii})/(1-(\tilde{v}_{ii}+\tilde{u}_{ii})) \qquad (10.13)$$

where $\tilde{v}_{ii}$ and $\tilde{u}_{ii}$ are the diagonal elements of $\tilde{V}$ and $\tilde{U}$ respectively. Obviously, there are other possible measures of leverage, but these do not follow Cook's interpretation (10.5). A possible measure of the leverage of a group of observations is -

$$\tilde{\ell}_I = \Sigma\lambda_j/(1-\mu_j) \qquad (10.14)$$

where (λ_j) are the eigenvalues of $\tilde{V}_I-\tilde{U}_I$

(μ_j) are the eigenvalues of $\tilde{V}_I+\tilde{U}_I$.

In restricted least squares regression with exact prior information, the matrix

$T = X[(X'X)^{-1}-(X'X)^{-1}H'(H(X'X)^{-1}H')^{-1}H(X'X)^{-1}]X'$ is the equivalent of the hat matrix, since

$$r^* = f(\beta,h) + (I-T)\varepsilon \qquad (10.15)$$

$$\text{var}(r^*) = (I-T)\sigma^2 \tag{10.16}$$

$$\text{var}(Xb^*) = T\sigma^2 \ . \tag{10.17}$$

Hence, there is no difficulty in defining measures of leverage analogous to (10.4) and (10.6).

However, if prior information is stochastic then

$$r^* = (I-U^*\Omega^{-1})\varepsilon \tag{10.18}$$

$$\text{var}(r^*) = (I-U^*\Omega^{-1})\Omega\sigma^2$$

$$= (\Omega-U^*)\sigma^2 \tag{10.19}$$

$$\text{var}(X^*b^*) = U^*\sigma^2 \tag{10.20}$$

where $U \equiv X^*(X^{*\prime}\Omega^{-1}X^*)^{-1}X^{*\prime}$ and $\Omega = \begin{bmatrix} I & 0 \\ 0 & R \end{bmatrix}$ from (8.4).

Although there is no direct equivalent of the hat matrix, we can define measures of leverage in terms of the matrices Ω and U^*:

$$\ell_i^* = u_{ii}^*/(1-u_{ii}^*) \tag{10.21}$$

where u_{ii}^* is the i^{th} diagonal element of U^*. We are concerned only with the first n observations of X^*, and hence all diagonal elements of Ω corresponding to these observations will be 1. Similarly,

$$\ell_I^* = \Sigma\lambda_j/(1-\lambda_j) \tag{10.22}$$

where (λ_j) are the eigenvalues of U_I^*.

Chapter 11

COMPUTATIONAL CONSIDERATIONS

It is implicit that a computer, possibly a microcomputer, will be used to screen data for outliers and influential observations. Two kinds of computational difficulties can arise for large data sets. Firstly, the size of the computer's memory may restrict the storage of matrices used in the calculations. Secondly, the number of calculations involved may consume an unjustifiable amount of computer time.

Storage problems:

If the sample size is large, the $n \times n$ hat matrix $V = X(X'X)^{-1}X'$ may create storage problems if a large computer memory is not available. In the case where single influential observations are considered, only the diagonal elements of V are used in the calculations; hence, it is unnecessary to store the full matrix. Where sets of influential observations are considered, it is necessary to form submatrices of V at various stages of the calculations.

It is possible to find a $q \times q$ upper triangular matrix U, with rank q, such that $X'X = U'U$. Since $v_{ij} = x_i(X'X)^{-1}x_j'$, it follows that

$$v_{ij} = (x_i U^{-1})(x_j U^{-1})' \qquad (11.1)$$

where x_i is the i^{th} row of X as before. Hence the n row vectors $x_i U^{-1}$, $i = 1,\ldots,n$, can be stored (occupying $n \times q$ storage locations), and the v_{ij} can be computed, when needed, as the inner product of the i^{th} and j^{th} row vectors. As m, the number of observations in an influential set, will tend to be small, and the submatrices V_I are $m \times m$, the number of computations will not be excessive.

Should the storage of row vectors $x_i U^{-1}$ still occupy too much memory space, then the v_{ij} can be calculated from $v_{ij} = x_i (X'X)^{-1} x_j'$ each time. Here only X and $(X'X)^{-1}$ need be stored. Savings on memory space are traded off against extra computational effort.

Upper bounds for Δ_I:

The sheer number of calculations in the analysis may render it uneconomically time-consuming, even for a large computer. For example, in a data set of size $n = 50$, there are 20875 subsets of size $m \leq 3$; if $n = 100$, there are 166750 subsets of this size to be considered, any of which may be influential. If the ridge regression technique is used, the matrix calculations are laborious and several values of the ridge constant k need to be examined, so the volume of computational effort is multiplied. Similarly, other regression techniques involve laborious computation. Simplifying techniques are needed to reduce the computational effort. Cook and Weisberg (1980) derive upper bounds for the statistic Δ_I as follows:

For some index set I, let $0 \leq \lambda_1 \leq \ldots \leq \lambda_m \leq 1$ be the eigenvalues of V_I, as in (3.16). From (3.18)

$$\Delta_I = \frac{1}{qs_{-I}^2} \sum_{j=1}^{m} g_j^2 \frac{\lambda_j}{(1-\lambda_j)^2} \quad \text{where} \quad g = (g_j) = \Gamma r_I,$$

$$\leq \frac{1}{qs_{-I}^2} \frac{\lambda_m}{(1-\lambda_m)^2} \sum_{j=1}^{m} g_j^2 \tag{11.2}$$

since $\lambda_m/(1-\lambda_m)^2 \geq \lambda_j/(1-\lambda_j)^2$ for all $j = 1,\ldots,m$.

Furthermore, $\Sigma g_j^2 = r_I'\Gamma'\Gamma r_I = \sum_{i\in I} r_i^2$, and hence

$$\Delta_I \leq \frac{1}{qs_{-I}^2} \frac{\lambda_m}{(1-\lambda_m)^2} \sum_{i\in I} r_i^2 = B_1 . \tag{11.3}$$

Assuming that $tr(V_I) < 1$, we have that $\lambda_m \leq tr(V_I)$ and

$$\Delta_I \leq \frac{1}{qs_{-I}^2} \frac{\Sigma v_{ii}}{(1-\Sigma v_{ii})^2} \Sigma r_i^2 = B_2 \tag{11.4}$$

where summation is performed over $i\in I$. The upper bound B_2 is valid only if $tr(V_I) < 1$; for any subset with $tr(V_I) \geq 1$, Δ_I must be calculated exactly.

For fixed m, let $r_*^2 = \max_I\left(\sum_{i\in I} r_i^2\right)$, where I varies over all subsets of size m. Then

$$\Delta_I \leq \frac{r_*^2}{qs_{-I}^2} \frac{\Sigma v_{ii}}{(1-\Sigma v_{ii})^2} = B_3 . \tag{11.5}$$

Further bounds for Δ_I can be found, but these are of little use in practice. The relationships between the bounds are $\Delta_I \leq B_1 \leq B_2 \leq B_3$. B_3 requires little computational effort, as does B_2. On the other hand, B_1 requires calculation of

the largest eigenvalue, and Δ_I requires inversion of the matrix $(I-V_I)$, both procedures involving considerably more computation.

If the bound B_3 is smaller than a certain cutoff point, then Δ_I will be small, and the data set indexed by I will be uninfluential. Similarly for the bounds B_2 and B_3. The problem is to determine suitable cutoff points for these bounds.

The bounds were compared to exact values of Δ_I for all data sets presented in chapter 13, which involved over 40 000 calculations of Δ_I. The distributions of B_1/Δ_I, B_2/Δ_I and B_3/Δ_I were examined to ascertain how closely the bounds approximate the value of Δ_I. The 1^{st}, 2^{nd}, 5^{th} and 10^{th} percentiles are listed in table 11.1 below.

Table 11.1: Percentiles of the ratios B_1/Δ_I, B_2/Δ_I and B_3/Δ_I:

	1%	2%	5%	10%
m = 1:				
B_1/Δ_I	1.000	1.000	1.000	1.000
B_2/Δ_I	1.000	1.000	1.000	1.000
B_3/Δ_I	1.000	1.138	1.572	2.379
m = 2:				
B_1/Δ_I	1.000	1.001	1.002	1.006
B_2/Δ_I	1.097	1.142	1.246	1.383
B_3/Δ_I	2.772	3.228	4.310	5.677
m = 3:				
B_1/Δ_I	1.009	1.017	1.041	1.086
B_2/Δ_I	1.671	2.007	2.727	3.425
B_3/Δ_I	5.047	6.447	9.846	14.774

Say, for example, that we wish to identify groups of observations which, when deleted, would change the estimate of β beyond its 10% confidence ellipsoid. Thus, from (3.12), the critical value for Δ_I would be the 10% point of the F distribution with q and n-q-m degrees of freedom. Call this value F*. If m = 3, then 95% of groups with $B_1 < (1.041)F^*$ will have $\Delta_I < F^*$ and will be uninfluential. Thus suitable cutoff points for B_1, B_2 and B_3 may be established. If the probability that any group is influential is denoted p, then the probability that it escapes detection is $\lesssim .05\ p$ if the 5% percentile ratios are chosen in establishing cutoff points for the bounds. Note that the bounds

B_1 and B_2 are exact for $m = 1$.

The recommended computational procedure for a particular m is as follows: firstly, order the v_{ii} and r_i from largest to smallest, making the calculation of r_*^2 trivial. For all subsets with $tr(V_I) < 1$, the three upper bounds for Δ_I are applied in the order B_3, then B_2, and finally B_1. Exact computation of Δ_I is necessary if $tr(V_I) \geq 1$ or all bounds are larger than their selected cutoff points.

Subsets should be considered according to the ordered v_{ii}. Hence, these subsets with high leverage are considered first. Once a point is reached where the bounds are sufficiently small, no further subsets comprising observations lower in the ordered lists of v_{ii} need be considered. Cook and Weisberg (1980) show that such a procedure can save considerable computational effort, particularly when n is large relative to q.

Using the data sets of chapter 13, an experiment was performed to determine the relative computational effort required for the calculation of the various bounds. The computer routine for the calculation of B_1 used an iterative procedure for calculating the largest eigenvalue given by Cooley and Lohnes (1971). It was shown empirically that calculation of the bound B_1 consumed more computer time (in c.p.u.) than exact calculation of the distance Δ_I. Hence, B_1 is of little practical value, unless a really efficient algorithm for the largest eigenvalue is available.

By following the procedure recommended above, without using B_1, real savings in computer effort were achieved. The c.p.u.'s consumed were overall 31.5% of the usage needed for exact calculation of all Δ_I. The bounds for B_2 and B_3 were determined using the ratios of the 5^{th} percentiles given in table 11.1. These proved good enough to allow all significantly large Δ_I to be detected.

When $m = 1$, the number of groups to be considered is relatively small, and the computational effort is also relatively small. There is little to be gained by applying the bounds procedure (which involves ordering the v_{ii} and r_i), and hence it is recommended that exact calculation of Δ_i be carried out when $m = 1$.

Note that values of Δ_I increase as m increases. This is because a large group of observations is generally more influential than a smaller group. We are not primarily concerned with the absolute magnitude of Δ_I, but rather its relative magnitude for various index sets I.

The purpose of studying sets of data points of size $m > 1$ is to find groups of observations which are not individually influential, but are influential when taken as a group. If the group includes one or more observations that are individually influential, then little information is gained, because the influence of the group will be partly due to those observations. Hence, when calculating the influence of data sets, further computational effort can be saved if we

exclude individual observations and smaller subsets already shown to be influential in their own right.

On the other hand, we may wish to examine the relationship between the observations in an influential group. In this case, all subsets of the data must be considered, even those with individually influential observations.

Upper bounds for $\tilde{\Delta}_I$, Δ_I^- and Δ_I^*:

To cut down computation, we would like upper bounds for the ridge measure of influence $\tilde{\Delta}_I$, the g-inverse measure Δ_I^- and the restricted least squares measures Δ_I^*. Unfortunately, neither $\tilde{\Delta}_I$ nor Δ_I^* can be simplified to expressions analagous to (3.18), and therefore bounds of the form of (11.3), (11.4) and (11.5) cannot be derived.

From (6.35),

$$\Delta_I^- = \frac{1}{qs_{-I}^2} \sum_{j=1}^{m} g_j^2 \frac{\lambda_j}{(1-\lambda_j)^2}$$

where $P_I = \Gamma'\Lambda\Gamma$ and $(g_j) = \Gamma r_I^-$.

It follows that

$$\Delta_I^- \leq \frac{1}{qs_{-I}^2} \frac{\lambda_m}{(1-\lambda_m)^2} \sum_{i\in I} (r_i^-)^2 \qquad (11.6)$$

$$\leq \frac{1}{qs_{-I}^2} \frac{\Sigma p_{ii}}{(1-\Sigma p_{ii})} \Sigma (r_i^-)^2 \text{ if } tr(P_I) < 1 \qquad (11.7)$$

$$\leq \frac{(r_*^-)^2}{qs_{-I}^2} \frac{\Sigma p_{ii}}{(1-\Sigma p_{ii})^2} \quad \text{where } (r_*^-)^2 = \max_I(\Sigma (r_i^-)^2) \qquad (11.8)$$

Hence, upper bounds for $\bar{\Delta}_I$ are established in the same way as those for Δ_I. The accuracy of these bounds can be determined empirically, as described in the previous section.

Computational effort in ridge and g-inverse regression:

In chapters 5 and 6 we suggested that, in the presence of collinearity, ridge or g-inverse regression might be more appropriate techniques than ordinary least squares regression. However, the use of ridge regression especially presents considerable computational difficulty with regard to the complexity and number of computations.

The computational formula for the ridge influence measure $\tilde{\Delta}_I$ was given by (5.27). It does not readily simplify to a form analogous to (3.18), and bounds for $\tilde{\Delta}_I$ are not easily derived. It seems that $\tilde{\Delta}_I$ must be computed exactly for all cases, which will involve considerable effort, since the computational formula (5.27) is itself complex and lengthy.

If a ridge trace is desired, then the analysis must be performed for several values of k. In preparing a ridge trace, the following computational procedure is suggested: Perform a complete analysis for a few selected values of k, say $k = 0.10$ to 0.50 in steps of 0.10. Record the ten most influential observations or groups for each value of k. Having isolated observations or groups that are potentially influential, a more detailed analysis can be performed on those observations/groups only, for $k = 0.01$ to 0.50 in steps of 0.01 say. In this way, a fairly complete ridge

trace may be prepared.

Computational effort increases considerably if either the sample size n or the number of variables q is large. If the sample is large, the number of potentially influential groups to be considered can be excessive. If the number of variables is large, the matrices involved in the computations are bigger, and the matrix manipulation consumes a great deal of computer time.

Table 11.2 below gives the amount of computer resource needed to analyse two of the smaller data sets of chapters 12 and 13 on the PRIME computer of the Graduate School of Business, University of Cape Town.

Table 11.2 : Computer resources needed for various regression techniques:

	Simulation data (chapter 12)	Data set number 1 (chapter 13)
Sample size	40	24
Number of variables	5	5
Computer resources (in c.p.u. seconds):		
Basic analysis		
m = 1,2	23	15
m = 3	233	129
G-inverse regression		
m = 1,2	58	26
m = 3	1660	372
Ridge regression		
m = 1,2	895	576
m = 3	11690	3078
Total	14559	4196
Approximate running time	8 hours	3 hours

It is seen that over 90% of the computational effort goes into analysing groups of size 3. In most practical situations, however, the influence of a triple is due to individually influential observations or pairs within the group. Considerable computational effort can therefore be saved by limiting the analysis to individual observations and pairs. For really large data sets, it is completely infeasible to even

attempt an analysis of groups of size 3.

The ridge regression technique requires over 85% of the computational effort in a complete analysis. Hence, there is serious doubt whether this enormous amount of computation can be justified. In the following two chapters we shall demonstrate that the ridge trace can be extremely valuable in the presence of collinearity. On the other hand, g-inverse regression, although computationally more efficient, yields very disappointing results overall.

Chapter 12

A SIMULATION STUDY

The effectiveness of the techniques discussed in this thesis may be tested using a contrived or simulated data set. In such a data set, all parameters are known, and the extent of collinearity between the variables can be controlled.

In this chapter, we shall examine in detail how the various techniques perform in the presence of known irregularities. In the following chapter, we shall apply the techniques to several real-life data sets from various disciplines in commerce.

A data set of 40 observations and 5 independent variables was generated as described below. The first independent variable X_1 is a trend variable, typical of many time series data sets : $X_1 = i$ for observation i, $i = 1,\ldots,n$. The variables X_2, X_3 and X_4 are independently normally distributed, with $X_2 \sim N(5,1)$ and X_3, $X_4 \sim N(10,1)$. An approximate linear relationship between the independent variables was introduced by making $X_5 = X_3 + X_4 + e$ where $e \sim N(0,0.1^2)$. Hence the linear relationship is not exact, but the correlation between X_5 and (X_3+X_4) is $>.999$. The dependent variable was generated as

$$Y = 100 + 2X_1 + 10X_2 + 10X_3 + 10X_4 + 10X_5 + e$$

where $e \sim N(0,5^2)$.

The following artificial extreme cases were introduced into the data set : Firstly, in observation 1, the variable X_2 was given a value of 10 (X_2 is usually centred on 5); also in generating Y, the random error term e was made to be zero. Hence, observation 1 has high leverage in an independent variable not affected by the collinearity, and the observation is not an outlier.

Secondly, in observations 2 and 3, X_3 and X_4 were centred on 15 instead of 10, and corresponding adjustments were made in the values of X_5 and Y. The result is that these observations have very high values of Y, X_3, X_4 and X_5. In addition, a possible outlier was introduced by subtracting 10 from the Y value of observation 2 only.

Thirdly, observation 4 is made to be a straightforward outlier by adding 20 to its Y value.

The complete data set and a summary of the rules used to generate it are given in appendix D. A printout of the first stage of the analysis is shown in appendix C, as a sample printout to illustrate the output of the computer routines.

The analysis initially warns us of the severe collinearity problem : a condition number $\kappa = 3962$ and smallest root $\lambda_1 < 0.001$. Observation 4 is identified as a significant outlier with $F = 14.03$, the critical value being 12.46 at the 5% significance level. The leverage of observation 4 is

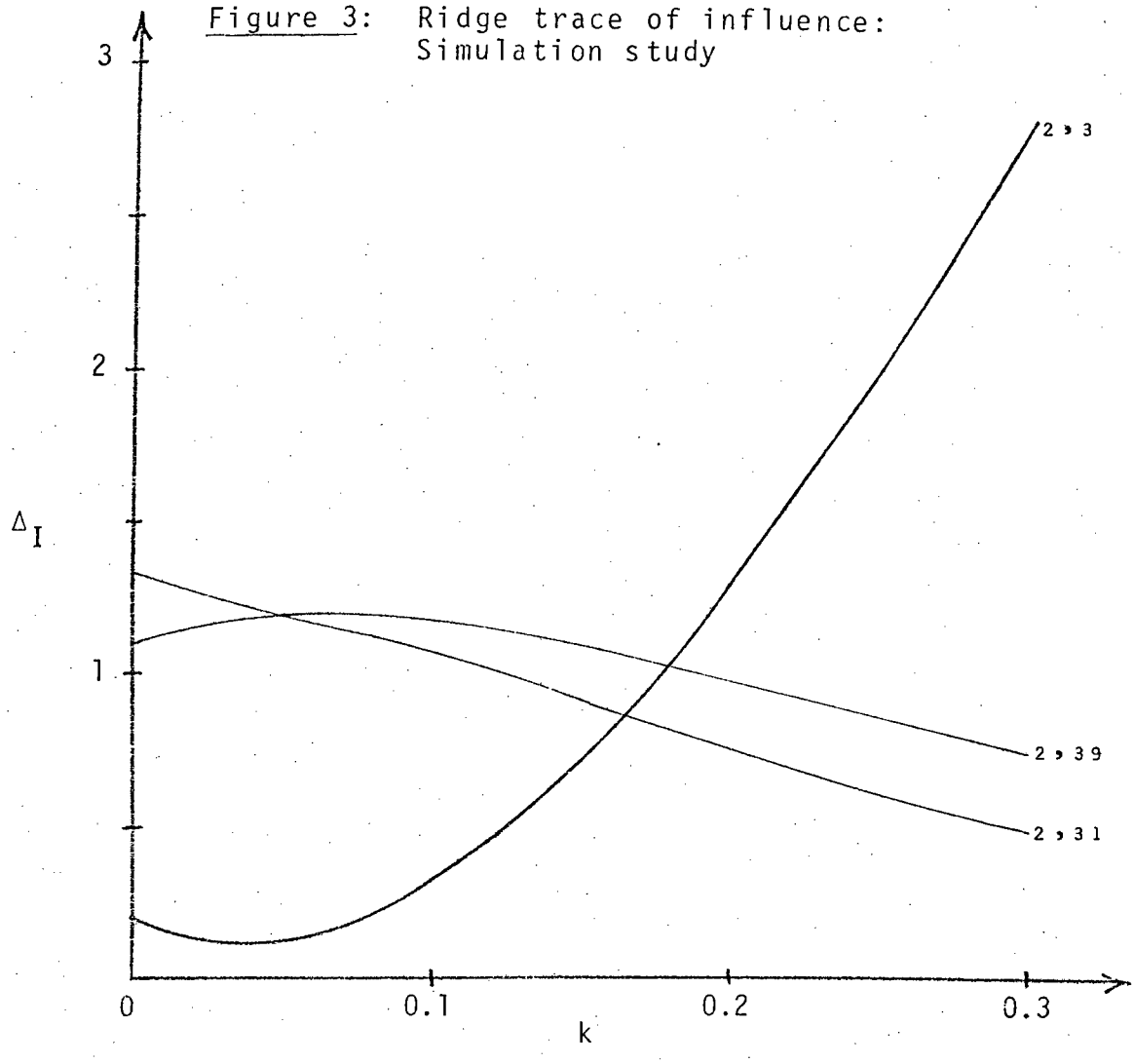
not unusual, yet it is still the most influential observation in the sample, with $\Delta_4 = 0.733$ corresponding to a 37.4% confidence ellipsoid. This gives sufficient evidence to eliminate observation 4 as an outlier of significant influence.

When observation 4 has been removed, further analysis reveals that there are no significant outliers in the sample. The highest value of t_i corresponds to observation 34, but it is not significant even at the 10% level; nor does this observation have any significant influence. Observations with high leverage are numbers 1, 2 and 3 (as we would hope) with leverage 0.90, 0.63 and 0.64 respectively. The influence measures are $\Delta_1 = 0.048$, $\Delta_2 = 0.844$ and $\Delta_3 = 0.609$. Observations 2 and 3 are by far most influential, corresponding to 45.5% and 27.9% confidence ellipsoids respectively.

Yet jointly observations 2 and 3 are uninfluential with $\Delta_{2,3} = 0.175$ only. This is apparently because one residual is positive and the other negative, the situation of figure 2 on page 16. Let us examine the ridge trace for the influence of selected groups of size 2 (figure 3). It is seen that for large k, the group of 2,3 dominates.

When groups of size 3 are examined, the pair 2,3 does not feature in the ordinary least squares analysis. However, when the ridge trace is applied, this pair dominates 8 of the ten most influential groups. It is noteworthy that the g-inverse regression analysis finds nothing out of the ordinary in the pair 2,3.

Figure 3: Ridge trace of influence: Simulation study



There is sufficient evidence that the analyst should pay particular attention to observations 2 and 3, which are individually influential, and also as a pair, although this latter fact is obscured by the collinearity in the data, and emerges only in the ridge trace. However, the least squares estimate of the regression coefficients would not change radically if this pair were to be removed, because $\Delta_{2,3}$ is small. The ridge estimate will change, as the pair becomes influential for larger values of k.

Whether or not the pair 2,3 is removed, all techniques (ordinary least squares, ridge and g-inverse regression) recognize observation 1 as a point of high leverage without significant influence. There is nothing to be gained from deleting the observation. If this were a real analysis, the researcher should note the presence of the high leverage point, and possibly investigate why such an observation is so different from the rest of the data set, yet consistent with the model.

Table 12.1 below lists the estimates of the regression parameters at various stages of the analysis. The only really significant impact on the parameter estimation was due to the removal of the outlier observation 4; prior to this, the least squares estimate was particularly poor, although other techniques appear to be more robust.

The ridge technique helped to identify an influential group of data, although the final ridge estimate of β is not very good. Nevertheless, identification of influential data is extremely valuable in its own right : this will be reinforced in the practical examples of chapter 13. The g-inverse technique contributed nothing new in the entire analysis.

Lack of accuracy in the parameter estimation is probably due to the severe collinearity in the data set. Recognizing that the collinearity is due to the relationship between X_5 and other variables, suppose we effectively remove X_5 by applying restricted least squares regression with the constraint $\beta_5 = 0$. Since the true values of β_3, β_4 and β_5 are 10, and $X_5 \simeq X_3 + X_4$, the new values of β_3 and β_4

Table 12.1: Regression parameters in simulation study:

	True value	Ordinary least squares	Ridge k = 0.1	G-inverse
Complete data set:				
Constant	100	113.47	145.36	116.45
x_1	2	2.04	1.76	2.01
x_2	10	9.93	8.04	9.94
x_3	10	22.88	10.58	11.36
x_4	10	24.03	12.52	12.23
x_5	10	-4.08	7.05	7.47
Outlier 4 deleted:				
Constant	100	111.02	139.87	110.79
x_1	2	2.07	1.81	2.08
x_2	10	9.60	7.87	9.60
x_3	10	10.68	10.83	11.78
x_4	10	10.84	12.21	11.97
x_5	10	8.76	7.30	7.65
2,3,4 deleted:				
Constant	100	108.84	126.62	108.98
x_1	2	2.04	1.79	2.03
x_2	10	9.23	7.62	9.22
x_3	10	11.67	9.91	10.31
x_4	10	12.77	11.69	11.37
x_5	10	7.54	8.75	8.92

are approximately 20 if β_5 is constrained to be 0.

If the variable X_5 is removed, the condition number of the matrix $X_1'X_1$ is $\kappa = 4.4$, and the collinearity problem is solved.

The restricted least squares analysis proceeds along similar lines to that described thus far in this chapter, and all the abnormalities in the data set are successfully identified. Estimates of the regression parameters are listed in table 12.2 below. Estimation is generally far more accurate than in the analysis of the unrestricted data set. Although we are able to develop workable techniques to overcome collinearity problems (ridge and g-inverse regression), it is obvious that the analysis is far simpler and more effective if the sources of collinearity can be eliminated.

Table 12.2: Restricted least squares estimates of regression parameters:

	True value	Complete data set	Outlier 4 deleted	Obsn. 2,3,4 deleted
Constant	100	114.49	109.27	108.27
x_1	2	2.03	2.09	2.05
x_2	10	9.94	9.62	9.26
x_3	20	18.81	19.37	19.10
x_4	20	19.86	19.80	20.43
x_5	0	0	0	0

Chapter 13

SOME PRACTICAL APPLICATIONS

In this chapter, various aspects of the theory developed in this thesis will be applied to real data sets drawn from different disciplines in South African commerce : economics, finance, market research, agriculture and personnel planning.

All calculations are performed by the original BASIC programs listed in appendix B. Sample printouts from the programs are given in appendix C. All data sets analysed in this chapter are listed in appendix D.

Example 1: Predicting changes in the price index.

In this data set, changes in the S.A. food consumer price index for the 24 months of 1980-1981 are predicted by changes in various consumer and production price indices in the previous month. Changes are expressed as the natural logarithm of the ratio of price indices in consecutive months; e.g. in January 1980, the food price index rose from 163.6 to 167.0, expressed in the data set as log (167.0/163.6) = 0.0206.

This is a reasonably small data set; 24 observations and 5 independent variables. Certain variables (x_1-x_3, x_1-x_4, x_3-x_4) are highly correlated and the condition number of $X'X$ is $\kappa = 238$, showing a potential collinearity problem.

There is one significant outlier (observation 9) corresponding to September 1980, when the food price index rose by 10.3 from 184.5 to 194.8. This observation also has fairly high leverage, and an influence of $\Delta_9 = 0.72$ which corresponds to a 36.0% confidence ellipsoid.

Observation 20 (August 1981) has very high leverage, but is not an outlier. In this month, the c.p.i. for non-food items increased abnormally, while two of the other independent variables actually fell. This unusual configuration of the independent variables led to the high leverage.

While the combination of observations 9 and 20 forms a fairly influential group, the most influential group of size 2 comprises observations 9 and 10, two very similar data points representing large rises in the dependent variable. The three points 9, 10 and 20 dominate all significantly influential groups of size 2 and 3.

The application of ridge and g-inverse regression gives regression coefficients $\tilde{b}$ and b^- substantially different from b, but no further influential observations or groups emerge. The collinearity is in fact due to a correlation of 0.99 between the variables x_3 and x_4. If x_3 (say) is removed, the condition number of the restricted $X'X$ falls to 8.2, indicating that the collinearity is removed. However, no further influential data emerges in the restricted least squares analysis. We would conclude, therefore, that in this data set the influential data are not obscured by the collinearity.

Data points deserving attention are the following: (1) the outlier at observation 9; (2) the jointly influential pair 9 and 10, both being large rises in the dependent variable; and (3) the unusual configuration of observation 20. In a small data set such as this, the presence of three "suspects" throws doubts on the validity of any conclusion that may be drawn from a conventional regression analysis.

Example 2A: Predicting stock market price changes.

In this data set, changes in the weekly price of the Western Deep gold share are predicted by changes in share prices, JSE indices and economic indicators of the previous week, over the last six months of 1981. Changes are again expressed as the logarithm of the ratio of prices in consecutive weeks.

The data set has relatively few observations (25) and a large number of predictor variables (10). The majority of the independent variables are highly correlated with each other, with the exception of the currency exchange rate variables x_9 and x_{10}. Of 21 correlations between the variables x_1 to x_7, 11 are over 0.70 and all are over 0.57. The condition number of $X'X$ is $\kappa = 426$. Hence the collinearity problem is severe.

There is an outlier at observation 13 (29/9/81) corresponding to the largest price drop of Western Deep in the sample. There are also potential outliers in observation 11 (15/9/81) corresponding to the largest price rise in the sample, and to a lesser extent in observation 15 (13/10/81), a strange case

of a 50 cent price rise where a massive price drop would be predicted.

The strangest case is observation 1 (7/7/81), where all the predictor variables except the Sterling exchange rate tumbled by record amounts, yet the price rise of the dependent variable is predicted quite accurately!! The small residual means that this observation cannot be considered an outlier. However, the leverage factor $\ell_1 = 4.80$ stands out, highlighting a very extreme design point.

The influence measure, which combines outlier and leverage effects, shows that the only influential observation is number 15, with $\Delta_{15} = 0.43$, corresponding to a 8.5% confidence ellipsoid. When considering influential groups, the pair of 11 and 13 have $\Delta_{11,13} = 2.76$ and stand out as a dominant influence. This pair feature in 8 of the 10 most influential groups of size 3, but the influence of these groups is obviously commanded by the pair 11 and 13.

Observations 11 and 13 are only moderately influential individually, but highly influential as a pair. This is a good illustration of the situation of figure 1 on page 16.

It is still rather disturbing that a design point as extreme as observation 1, with such high leverage, was missed as being potentially influential on its own. Let us examine the ridge trace for key observations in the data set.

It is seen from figure 4 that observation 15 appears highly

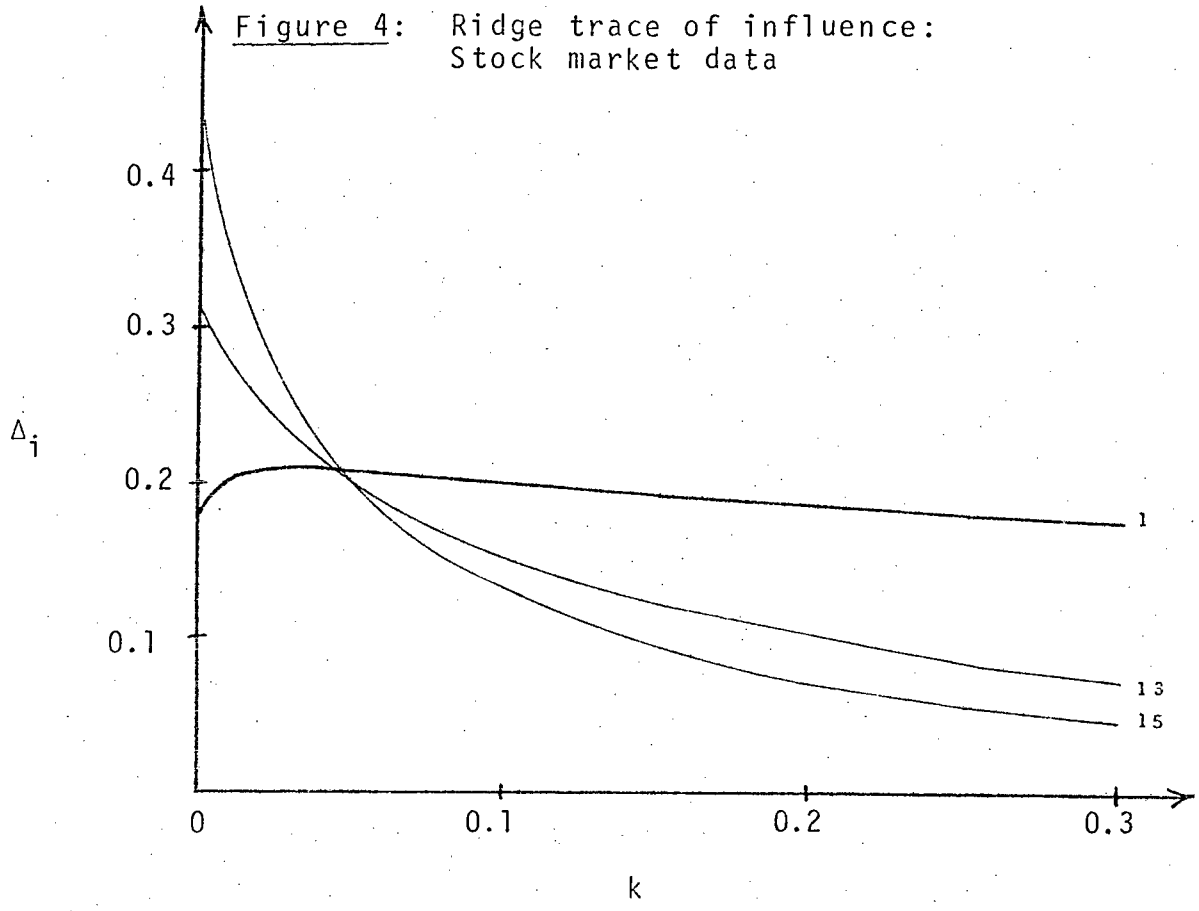


Figure 4: Ridge trace of influence: Stock market data

influential in ordinary least squares regression ($k = 0$), but that $\tilde{\Delta}_{15}$ declines rapidly. The same is true of observation 13 to a lesser extent. However, $\tilde{\Delta}_1$ rises and then declines very gradually, and for $k > 0.04$ observation 1 is the most influential data point. The conclusion is that some valuable information about the model was masked by the collinearity problem, and emerges only when ridge regression is applied. It should be noted that g-inverse regression finds observation 1 neither outlying nor at all influential.

When the ridge trace for influence is examined for data pairs, the influence of 11 and 13 declines, and the group of 1 and

14 comes to dominate for larger values of k. Observation 14 is very similar to observation 1, where a row of declines in the independent variables successfully detect a rise in the dependent variable.

In conclusion, the regression coefficients for the complete data set and for data sets with influential observations and groups deleted are given in table 13.1 below. The impact of influential data points, particularly the influential group, is significant even by inspection, with large changes in many coefficients (e.g. b_3, b_8 and b_{10}).

Table 13.1: Regression coefficients for stock market data:

	Complete data set	Observation 15 deleted	Pair 11 and 13 deleted
Constant	.011	.013	.005
x_1	-.439	-.609	-.700
x_2	.261	.009	.723
x_3	.022	-.209	1.139
x_4	1.531	1.891	.816
x_5	-.545	-.083	-1.432
x_6	-3.573	-3.996	-2.549
x_7	-.246	-.618	-.956
x_8	.069	.399	-.276
x_9	.782	1.193	1.172
x_{10}	-.227	.491	1.398

Example 2B: Utilizing prior stock market information.

When performing the analysis of example 2A (predicting the

weekly changes of the Western Deep gold share in the second half of 1981), the results of an identical study conducted in 1980 were available. It is possible to combine this prior information with the sample data, as described in chapter 8.

In equation (8.1), the design matrix H will simply be the identity matrix I, and the parameters h and R will be taken to be the old estimates b and $(X'X)^{-1}$, with σ^2 regarded as unknown. Table 13.2 below shows how the regression coefficients change when prior information is incorporated.

Table 13.2: Regression coefficients for stock market data:

	1980 study	1981 study	1981 study with prior information
Constant	.016	.011	.009
x_1	-.197	-.439	-.427
x_2	-.373	.261	-.242
x_3	-3.135	.022	-1.589
x_4	3.969	1.531	2.727
x_5	-1.446	-.545	-.023
x_6	-.474	-3.573	-1.241
x_7	-.926	-.246	-.534
x_8	-.052	.069	-.030
x_9	-2.475	.782	-.597
x_{10}	3.204	-.227	.205

The restricted least squares analysis, incorporating the stochastic prior information of the 1980 study, reveals that there are now no significant outliers in the 1981 study.

Recall that observation 13 (29/9/81), corresponding to the largest price change in the sample, was regarded as a significant outlier in the (rather small) data set. However, in the 1980 study, there were three price changes of much larger magnitude, so their impact is reflected when the information is combined with further sample data, albeit indirectly.

Similarly, there are no really influential observations or even pairs in the combined model. Observation 1 remains a high leverage point with $\ell_1^* = 1.89$. There is a potentially outlying pair (observations 7 and 11) with $(t^*_{7,11})^2 = 6.59$ and $\Delta^*_{7,11} = 0.34$ corresponding to a 3.9% confidence ellipsoid.

The conclusion drawn from the above results is the following: although it may be intuitively attractive to utilize the results of past studies, to provide better parameter estimation overall, care must be exercised if the prior results are materially different from the current analysis, and/or they have large variance. In this example, the regression parameters of the 1980 and 1981 studies are significantly different; the prediction model appears to have altered over time. The effect of combining the studies is to completely swamp the detection of influential data, and to obscure the valuable insights into the data structure that can be gained from studying influential data for its own sakes.

Example 3: Relationships between media exposure in South Africa.

This is an interesting data set, because it is very different from the classical time series model to which regression analysis is most often applied. We examine the relationship between exposure to various media (cinema attendance, newspaper and magazine readership, television viewing and radio listenership) for several demographic groups in 1981.

The South African population is segmented according to several demographic variables. The following categories were included:

Race: White, Coloured, Asian, African.

Sex: Male, female.

Province: Cape, Natal, Transvaal, O.F.S.

Geographic density: City, town, village, rural.

Marital status: Married, single.

Age: 16-24, 25-34, 35-49.

A total of 104 combinations were considered, being 4 × race, 2 × sex and 13 × other variables. The population of some of these categories is zero (there are no Asians in the O.F.S. etc), so the total data set was of size 92. The total number of independent variables chosen was 11 (see appendix D for the raw data).

Cinema attendance was chosen to be the dependent variable, as it could be argued that exposure to advertising in the other media is likely to influence people's motivation to attend the cinema, and not vice versa. However, a worthwhile

analysis could be performed using other dependent variables. Market research analysts might be interested to note that 80% of variation in the dependent variable is explained by the regression. The variables with greatest impact on cinema attendance are exposure to magazines and Radio 5.

Several independent variables in the design are correlated, and the condition number of $X'X$ is $\kappa = 294$. Five eigenvalues of $X'X$ are <0.10.

The analysis shows that there is a substantial outlier at data point 50, Asian males in the Transvaal, where cinema attendance is almost double that which would be consistent with the rest of the model. This data point is also the most influential in the sample, with $\Delta_{50} = 0.30$, corresponding to a 1.3% confidence ellipsoid.

This illustrates a problem typical of large data sets. The larger the data set, the more unlikely it becomes that any individual observation, no matter how extreme, can exert significant influence upon the overall results. Thus a mechanical examination of the influence statistic, Δ_i, will fail to reveal anything out of the ordinary.

When the analysis is extended to groups of size 2, the pair of data points 50 and 59 (Asian males in the Transvaal and Asian females in the Transvaal) has $\Delta_{50,59} = 0.88$ corresponding to a 42.8% confidence ellipsoid, which is certainly significant. The connection between the data points is also interesting - we see that Asians in the Transvaal, males and

females, whose cinema attendance is much higher that would be expected, can influence the results of the entire study to a significant degree.

When a ridge trace is prepared for groups of size 2, only two pairs 50,54 and 50,55 stabilize. Data point 54 represents single Asians and point 55 represents Asians aged 16-24, both groups having very high cinema attendance. G-inverse regression (with 5 eigenvalues eliminated in calculating the g-inverse) finds only the pair 50 and 59 to be of significant influence. Hence, all evidence points to the conclusion that the Asian population behaves somewhat differently from the rest of the data set, and that the results of the study are significantly biased thereby.

Table 13.3 shows the regression coefficients for this data set. Changes induced by deleting data points are not as dramatic as those of data set 2, as expected due to the sample size.

Example 4: An attitudinal survey.

This very large data set was extracted from a survey of farmers' attitudes towards attributes of seed maize. Farmers in three climatic regions, one very dry, one wet and one in between, rated the importance of 17 attributes on a scale of 1 to 10. Of these, there were 9 attributes whose ratings differed significantly between regions.

The data set, another example of a non-time series application, has 184 data points and 9 independent variables. Correlations

Table 13.3: Regression coefficients for media exposure data:

	Complete data set	Pair 50 and 59 deleted	All Asians deleted
Constant	-14.81	-15.39	-16.73
x_1	4.05	0.54	-0.87
x_2	9.27	-0.01	0
x_3	9.21	11.03	11.71
x_4	9.52	8.94	8.74
x_5	0.07	0.03	0.01
x_6	-0.04	0.00	-0.01
x_7	-0.41	-0.28	-0.18
x_8	0.54	0.42	0.39
x_9	-0.02	-0.13	-0.20
x_{10}	0.75	0.91	0.92
x_{11}	0.11	0.32	0.38

between independent variables are all relatively low, and $\kappa = 3.2$ only. Hence there is no collinearity problem.

As the dependent variable is somewhat artificial, we are not primarily concerned with identifying outliers in y, although these may be of interest. By the nature of the data set, abnormalities in the X variables, i.e. high leverage points and influential observations, would be of more interest to the analyst.

As is to be expected with such a large data set, single data points exert only minimal influence on the overall results. The most influential data point is number 107, followed by

number 118, which is also a high leverage point. These two data points form the most influential pair with $\Delta_{107,118} = 0.33$, corresponding to a 2.8% confidence ellipsoid only, hardly an important influence. Hence we would conclude that, because of the size of the data set, the overall results are unchanged whether the marginally influential points are included or excluded.

But do these data points possibly hold some valuable information in their own right? Interestingly, the two points in question come from different climatic regions.

The farmer represented by data point 107, who comes from the very dry region, gave a rating of 5/10 to attribute C (drought resistance). Of 41 other farmers in the region, 38 rated this attribute 10/10 and 3 gave it 9/10; an average rating of 9.93. Why did only one particular farmer rate the attribute as low as 5/10?

In the medium climatic region, drought resistance is still regarded as very important, achieving an average rating of 9.62. Yet the farmer represented by data point 118 rated this attribute as 2/10 only!! The ratings of other attributes in data points 107 and 118 are reasonably consistent with the rest of the respondents.

The existence of such extreme values could have two possible origins. The particular respondents could have misunderstood the questionnaire or filled it in erroneously, in which case these data points should be discarded as errors. Alterna-

tively, the presence of a peculiar market segment may be indicated. This segment is interested in product attributes very different from the majority preference. The farmers involved probably have good irrigation. Although small, this segment may be valuable if products are available to meet its specialized needs. It would certainly be worthwhile to conduct follow up interviews with these respondents.

Example 5: Personnel planning in building society branches.

In this example, the number of clerical staff in each of 48 branches of a large building society is regressed on the annual volume of eight types of transactions. The resultant regression equation can be used for planning staff requirements in other branches. In addition, large residuals in the analysis would indicate that particular branches in the sample are over- or under-staffed.

Many of the eight predictor variables are highly correlated. The condition number is $\kappa = 347$, and 4 of the 8 eigenvalues of X'X are <0.10. Therefore, the collinearity problem is severe.

The ordinary least squares analysis reveals a massive outlier at observation 2, with $t_2^2 = 27.31$ ($F_{1,38}(1-.05/48) = 12.62$) and influence $\Delta_2 = 2.10$ corresponding to a 94.6% confidence ellipsoid! Observation 2 represents the second largest branch in the sample, and appears to be overstaffed by about 40 clerks. Its volume of transactions is uniformly low compared to other large branches in the sample. All the analysis is

dominated completely by this outlier, and its removal is recommended.

When observation 2 has been deleted, further analysis reveals that there are no other significant outliers in the data set, although several candidates are potential outliers. The most influential data point is observation 1, with $\Delta_1 = 0.94$ corresponding to a 49.4% confidence ellipsoid. This point represents the largest branch in the sample, with a clerical staff more than double that of the remaining branches; (the other really large branch, number 2, has been deleted). Observation 1 is also a high leverage point ($\ell_1 = 2.37$) because 6 of 8 values in the predictor variables are the largest in the sample by a long way. Observation 1 dominates the whole analysis, so should be eliminated as it is unduly influential.

While discussing high leverage points, it is worthwhile to note that observation 32 has a leverage $\ell_{32} = 6.31$. However, its residual is almost zero, so the data point is not an outlier, and its influence is negligible. This strange data point has several X values which are large for this branch size, offset by $x_4 = 0$, (this variable, loan transactions, has in fact the highest regression coefficient). The combination of high and low predictors balances out to be consistent with the overall model, despite the odd arrangement.

When observation 1 is deleted, the condition number of $X'X$ drops from 341 to 127, demonstrating again the dominance of

a single observation; however, the collinearity is still present. Least squares regression reveals no significant outliers, nor significantly influential points. The most influential point is observation 10 with $\Delta_{10} = 0.20$ corresponding to a 0.7% confidence ellipsoid. Observation 32 has very high leverage $\ell_{32} = 6.29$.

Figure 5 below shows the ridge trace of the influence of selected data points.

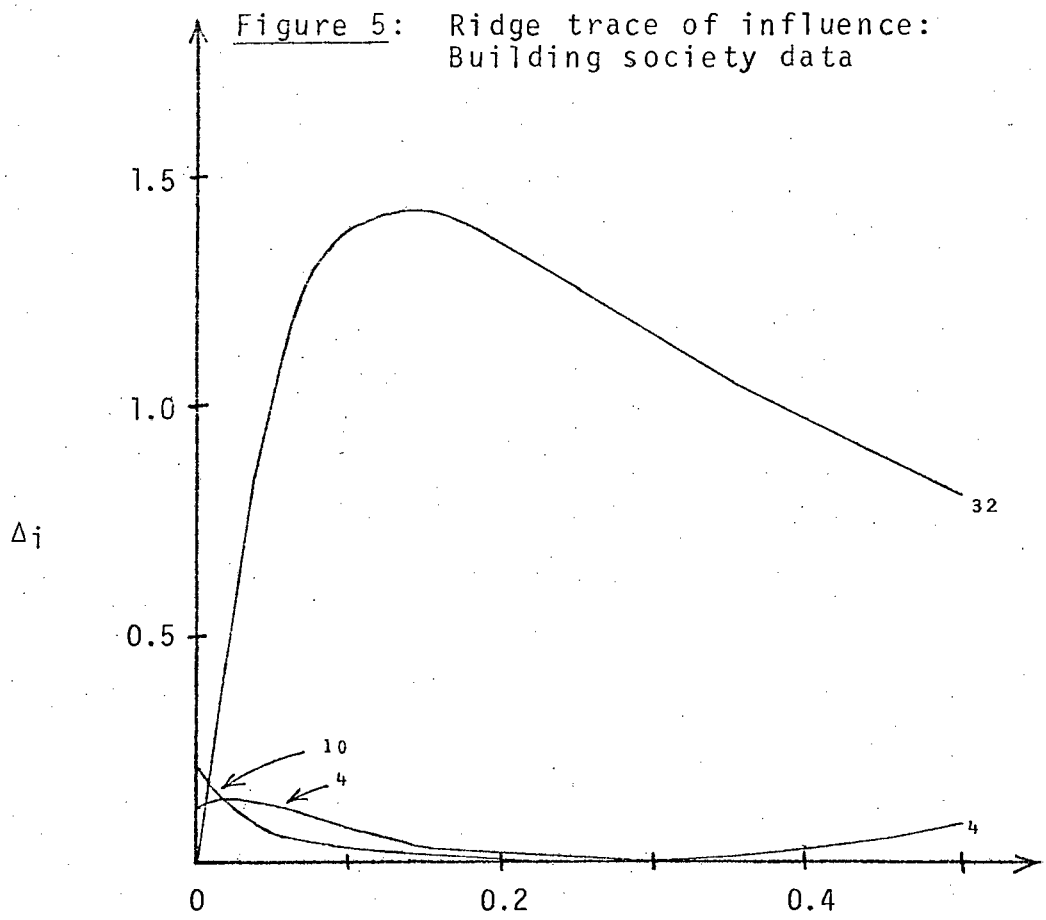


Figure 5: Ridge trace of influence: Building society data

The ridge trace for the outlier statistic $\tilde{t}_i$ follows a similar pattern, with observation 32 coming from almost zero to dominance for $k > 0.1$. This data point also dominates the ridge trace of all influential pairs. Hence we would conclude that observation 32, always a high leverage point, is in fact influential too, but its influence is hidden by the collinearity.

Generalized inverse regression, on the other hand, finds nothing abnormal about observation 32. This technique picks observation 4 as an influential point, with $\Delta_4^- = 0.51$, corresponding to a 14.3% confidence ellipsoid. This point also features in all influential pairs. Note the unusual behaviour of observation 4 in the ridge trace. Observation 4 is the third of the large branches in the original sample; the others have already been discarded as unduly influential. Yet the triple of large branches (1, 2 and 4) was not detected by any of the techniques to be particularly influential in the original analysis, apart from the influence of observations 1 and 2 in their own right. The analyst should be aware that large branches seem to have requirements different from the rest.

There is nothing to be gained from deleting observation 4 or observation 32. However, the very unusual mix of transactions performed in branch 32 warrants investigation.

Finally, table 13.4 below shows how the regression coefficients change with the deletion of influential data.

Table 13.4: Regression coefficients for building society data:

	Complete data set	Observation 2 deleted	Pair 1 and 2 deleted
Constant	3.20	3.21	5.87
x_1	.54	.49	.41
x_2	-5.27	-.49	.98
x_3	.82	7.99	9.28
x_4	11.53	15.94	17.00
x_5	-3.26	-3.07	-3.25
x_6	-4.58	-5.43	-5.12
x_7	-.18	.43	.36
x_8	4.16	3.00	2.70

Chapter 14

CONCLUSIONS

In this thesis, we have used variants of the studentized residual as a test statistic for detecting outliers, and a measure of influence similar to that proposed by Cook (1977). Of course, these are not the only forms of test statistics recommended in the current literature. For example, Hawkins (1980) suggests that outliers should be detected using recursive residuals; Andrews and Pregibon (1978) proposed a completely different measure of influence based on the determinants of various matrices in the analysis. Belsley, Kuh and Welsch (1980) analyse a host of statistics for identifying influential data; and Draper and John (1981) recommend that elements of both the Cook and Andrews/Pregibon statistics should be used.

The examples of chapters 12 and 13 demonstrate clearly that the techniques analysed in this study can successfully identify any influential data that might be of interest to the business analyst. The techniques are effective not only in a conventional, well-structured data set, but also in the presence of collinearity problems, and when restricted regression is more appropriate.

It is doubtful whether the use of other statistics would

provide more information. If several different statistics are used, the analysis is made more complex needlessly, and confusion may be caused if conflicting results are obtained.

It has been shown in several practical examples that the results of a regression analysis may be affected substantially by the presence of a small number of outlying or influential data points, especially in small samples. In such cases, the particular linear model does not apply to the bulk of the data points, but applies to the extreme group to an exagerated degree.

A simple scan of outlying observations, leverage points and influential data points and groups, such as that presented in this study, will provide information on

a. the homogeneity of the experimental design;
b. outlying observations, which may be errors or extreme results;
c. the real effect of extreme data points, and hence an indication of whether they should be excluded or not;
d. the identification of segments within the sample population;
e. the isolation of "interesting" data points that might deserve further investigation.

My recommendation is that no data analysis should be attempted without a preliminary scan of outliers and influential observations. If the data set contains outliers which are execution errors, then any results of the analysis would

necessarily be incorrect. The presence of a small, influential group of observations might bias the results significantly.

A tabulation of the studentized residuals, leverage and influence measure for each observation is an essential prerequisite. In addition, calculation of these statistics for small groups is recommended, plus a ridge trace when the X'X matrix is ill-conditioned.

The use of ridge regression consumes a great deal of computer time, but appears to be worthwhile when the ill-conditioning is severe. On the other hand, g-inverse regression, while computationally more efficient, yields disappointingly poor results.

To gain maximum benefit from any data study, some analysis beyond a mechanical examination of the statistics is required. The nature of the data set and the sample size must be considered. The appropriate regression technique must be chosen carefully. The analyst must question why particular observations are influential, and consider the implications of these findings, possibly beyond the scope of his current analysis.

In conclusion, there is often a mine of valuable information hidden beneath the surface of a superficial analysis. The study of influential observations is a powerful method of uncovering such information, and of measuring potential bias in conventional analyses.

Appendix A

TABLES FOR THE UPPER α/n POINTS OF t AND F DISTRIBUTIONS

In testing for a single outlier whose location is unknown beforehand, we require upper α/n points for the Student's t distribution (or alternatively the F distribution), to obtain critical points of the studentized residual.

In this appendix, tables are presented for various convenient values of n, q and α. The tables were generated using the Fortran subroutines of the Univac Statpack package.

UPPER α/n POINTS OF STUDENT'S t DISTRIBUTION WITH n-q-1 DEGREES OF FREEDOM

α = 0.10

n \ q	1	2	3	4	5	6	8	10	15	20	25	30
6	3.96											
7	3.68	4.15										
8	3.52	3.81	4.31									
9	3.42	3.62	3.93	4.47								
10	3.36	3.50	3.71	4.03	4.60							
12	3.28	3.36	3.48	3.64	3.86	4.22						
14	3.24	3.29	3.37	3.46	3.58	3.75	4.38					
16	3.21	3.26	3.31	3.37	3.45	3.55	3.86	4.53				
18	3.20	3.23	3.27	3.32	3.37	3.44	3.62	3.95				
20	3.20	3.22	3.25	3.29	3.33	3.37	3.50	3.69				
25	3.20	3.21	3.23	3.25	3.27	3.30	3.36	3.44	3.83			
30	3.21	3.22	3.23	3.24	3.26	3.27	3.31	3.35	3.53	3.95		
35	3.22	3.23	3.24	3.25	3.26	3.27	3.29	3.32	3.42	3.61	4.06	
40	3.24	3.24	3.25	3.26	3.27	3.27	3.29	3.31	3.38	3.48	3.67	4.15
45	3.25	3.26	3.26	3.27	3.27	3.28	3.29	3.31	3.36	3.42	3.53	3.73
50	3.27	3.27	3.28	3.28	3.29	3.29	3.30	3.31	3.35	3.40	3.47	3.58
60	3.30	3.30	3.30	3.31	3.31	3.31	3.32	3.33	3.35	3.38	3.42	3.47
70	3.32	3.33	3.33	3.33	3.33	3.34	3.34	3.35	3.36	3.38	3.40	3.43
80	3.35	3.35	3.35	3.35	3.36	3.36	3.36	3.37	3.38	3.39	3.41	3.43
90	3.37	3.37	3.37	3.38	3.38	3.38	3.38	3.38	3.39	3.40	3.42	3.43
100	3.39	3.39	3.39	3.40	3.40	3.40	3.40	3.40	3.41	3.42	3.43	3.44
150	3.48	3.48	3.48	3.48	3.48	3.48	3.48	3.48	3.48	3.49	3.49	3.50
200	3.54	3.54	3.54	3.54	3.54	3.54	3.54	3.54	3.54	3.55	3.55	3.55

UPPER α/n POINTS OF STUDENT'S t DISTRIBUTION WITH $n-q-1$ DEGREES OF FREEDOM

$\alpha = 0.05$

n \ q	1	2	3	4	5	6	8	10	15	20	25	30
6	4.85											
7	4.38	5.07										
8	4.12	4.53	5.26									
9	3.95	4.22	4.66	5.44								
10	3.83	4.03	4.32	4.77	5.60							
12	3.69	3.81	3.96	4.17	4.49	4.98						
14	3.61	3.69	3.79	3.91	4.07	4.30	5.16					
16	3.56	3.62	3.68	3.77	3.87	4.00	4.41	5.33				
18	3.53	3.57	3.62	3.68	3.75	3.83	4.08	4.51				
20	3.51	3.54	3.58	3.62	3.67	3.73	3.89	4.15				
25	3.48	3.51	3.53	3.55	3.58	3.61	3.69	3.79	4.30			
30	3.48	3.49	3.51	3.52	3.54	3.56	3.60	3.66	3.88	4.42		
35	3.48	3.49	3.50	3.51	3.52	3.54	3.57	3.60	3.73	3.96	4.53	
40	3.49	3.49	3.50	3.51	3.52	3.53	3.55	3.58	3.66	3.79	4.03	4.62
45	3.50	3.50	3.51	3.51	3.52	3.53	3.54	3.56	3.62	3.70	3.84	4.09
50	3.51	3.51	3.51	3.52	3.53	3.53	3.54	3.56	3.60	3.66	3.75	3.88
60	3.53	3.53	3.53	3.54	3.54	3.54	3.55	3.56	3.59	3.62	3.67	3.73
70	3.55	3.55	3.55	3.55	3.56	3.56	3.57	3.57	3.59	3.61	3.64	3.67
80	3.57	3.57	3.57	3.57	3.57	3.58	3.58	3.58	3.60	3.61	3.63	3.66
90	3.58	3.59	3.59	3.59	3.59	3.59	3.60	3.60	3.61	3.62	3.63	3.65
100	3.60	3.60	3.60	3.60	3.61	3.61	3.61	3.61	3.62	3.63	3.64	3.65
150	3.67	3.67	3.68	3.68	3.68	3.68	3.68	3.68	3.68	3.69	3.69	3.70
200	3.73	3.73	3.73	3.73	3.73	3.73	3.73	3.73	3.74	3.74	3.74	3.74

UPPER α/n POINTS OF STUDENT'S t DISTRIBUTION WITH n-q-1 DEGREES OF FREEDOM

α = 0.02

n \ q	1	2	3	4	5	6	8	10	15	20	25	30
6	6.25											
7	5.44	6.52										
8	4.98	5.60	6.76									
9	4.69	5.10	5.76	6.97								
10	4.50	4.79	5.21	5.89	7.17							
12	4.26	4.42	4.64	4.94	5.40	6.14						
14	4.12	4.22	4.36	4.53	4.76	5.08	6.35					
16	4.03	4.10	4.19	4.30	4.44	4.62	5.20	6.54				
18	3.96	4.02	4.09	4.16	4.26	4.37	4.71	5.31				
20	3.92	3.97	4.02	4.07	4.14	4.22	4.44	4.78				
25	3.86	3.88	3.91	3.95	3.98	4.02	4.12	4.26	4.94			
30	3.83	3.84	3.86	3.88	3.91	3.93	3.99	4.06	4.35	5.08		
35	3.81	3.82	3.84	3.85	3.87	3.89	3.92	3.97	4.13	4.43	5.19	
40	3.81	3.82	3.83	3.84	3.85	3.86	3.89	3.92	4.02	4.19	4.50	5.29
45	3.80	3.81	3.82	3.83	3.84	3.85	3.87	3.89	3.96	4.07	4.24	4.56
50	3.81	3.81	3.82	3.83	3.83	3.84	3.86	3.87	3.93	4.00	4.11	4.28
60	3.81	3.82	3.82	3.83	3.83	3.84	3.85	3.86	3.89	3.93	3.99	4.07
70	3.83	3.83	3.83	3.84	3.84	3.84	3.85	3.86	3.88	3.91	3.94	3.99
80	3.84	3.84	3.84	3.85	3.85	3.85	3.86	3.86	3.88	3.90	3.92	3.95
90	3.85	3.85	3.86	3.86	3.86	3.86	3.87	3.87	3.88	3.90	3.91	3.93
100	3.86	3.87	3.87	3.87	3.87	3.87	3.88	3.88	3.89	3.90	3.91	3.93
150	3.92	3.92	3.93	3.93	3.93	3.93	3.93	3.93	3.93	3.94	3.94	3.95
200	3.97	3.97	3.97	3.97	3.97	3.97	3.98	3.98	3.98	3.98	3.98	3.99

UPPER α/n POINTS OF STUDENT'S t DISTRIBUTION WITH $n-q-1$ DEGREES OF FREEDOM

$\alpha = 0.01$

n \ q	1	2	3	4	5	6	8	10	15	20	25	30
6	7.53											
7	6.35	7.84										
8	5.71	6.54	8.12									
9	5.31	5.84	6.71	8.38								
10	5.04	5.41	5.96	6.87	8.61							
12	4.71	4.91	5.19	5.58	6.17	7.15						
14	4.51	4.64	4.81	5.02	5.32	5.73	7.39					
16	4.38	4.48	4.59	4.72	4.90	5.12	5.86	7.60				
18	4.30	4.36	4.44	4.54	4.66	4.80	5.21	5.98				
20	4.23	4.29	4.35	4.42	4.50	4.60	4.86	5.29				
25	4.14	4.17	4.20	4.24	4.28	4.33	4.45	4.62	5.46			
30	4.09	4.11	4.13	4.15	4.18	4.21	4.28	4.36	4.71	5.60		
35	4.06	4.07	4.09	4.11	4.12	4.14	4.19	4.24	4.43	4.79	5.72	
40	4.04	4.05	4.06	4.08	4.09	4.10	4.14	4.17	4.29	4.49	4.87	5.83
45	4.03	4.04	4.05	4.06	4.07	4.08	4.10	4.13	4.22	4.34	4.54	4.93
50	4.03	4.03	4.04	4.05	4.06	4.07	4.08	4.10	4.17	4.25	4.38	4.59
60	4.03	4.03	4.04	4.04	4.05	4.05	4.06	4.08	4.12	4.17	4.23	4.32
70	4.03	4.03	4.04	4.04	4.05	4.05	4.06	4.07	4.09	4.13	4.17	4.22
80	4.04	4.04	4.04	4.05	4.05	4.05	4.06	4.07	4.09	4.11	4.13	4.17
90	4.05	4.05	4.05	4.05	4.06	4.06	4.06	4.07	4.08	4.10	4.12	4.14
100	4.06	4.06	4.06	4.06	4.06	4.07	4.07	4.07	4.09	4.10	4.11	4.13
150	4.11	4.11	4.11	4.11	4.11	4.11	4.11	4.11	4.12	4.12	4.13	4.14
200	4.15	4.15	4.15	4.15	4.15	4.15	4.15	4.15	4.16	4.16	4.16	4.16

UPPER α/n POINTS OF F DISTRIBUTION WITH 1 and n-q-1 DEGREES OF FREEDOM

α = 0.10

n \ q	1	2	3	4	5	6	8	10	15	20	25	30
6	15.69											
7	13.55	17.20										
8	12.40	14.52	18.62									
9	11.71	13.10	15.42	19.94								
10	11.26	12.25	13.75	16.26	21.20							
12	10.74	11.32	12.10	13.22	14.92	17.80						
14	10.47	10.86	11.34	11.98	12.85	14.08	19.20					
16	10.33	10.60	10.94	11.36	11.89	12.58	14.86	20.48				
18	10.25	10.46	10.71	11.01	11.37	11.81	13.12	15.58				
20	10.22	10.38	10.58	10.80	11.06	11.37	12.23	13.61				
25	10.23	10.33	10.44	10.57	10.71	10.88	11.28	11.82	14.70			
30	10.29	10.36	10.44	10.53	10.62	10.72	10.95	11.25	12.46	15.64		
35	10.39	10.44	10.49	10.55	10.62	10.69	10.84	11.03	11.71	13.01	16.45	
40	10.48	10.52	10.57	10.61	10.66	10.71	10.83	10.96	11.40	12.12	13.50	17.19
45	10.59	10.62	10.65	10.69	10.72	10.76	10.85	10.95	11.26	11.72	12.48	13.94
50	10.69	10.71	10.74	10.77	10.80	10.83	10.90	10.97	11.21	11.53	12.02	12.81
60	10.88	10.90	10.92	10.94	10.96	10.98	11.02	11.07	11.22	11.41	11.66	12.01
70	11.05	11.07	11.08	11.10	11.11	11.13	11.16	11.20	11.30	11.43	11.58	11.78
80	11.22	11.23	11.24	11.25	11.26	11.28	11.30	11.33	11.41	11.50	11.60	11.73
90	11.37	11.38	11.39	11.40	11.41	11.42	11.44	11.46	11.52	11.58	11.66	11.76
100	11.51	11.52	11.53	11.53	11.54	11.55	11.56	11.58	11.63	11.68	11.74	11.81
150	12.09	12.09	12.10	12.10	12.10	12.11	12.11	12.12	12.14	12.17	12.19	12.22
200	12.53	12.53	12.53	12.53	12.54	12.54	12.54	12.55	12.56	12.57	12.59	12.60

UPPER α/n POINTS OF F DISTRIBUTION WITH 1 and n-q-1 DEGREES OF FREEDOM

$\alpha = 0.05$

n \ q	1	2	3	4	5	6	8	10	15	30	25	30
6	23.53											
7	19.20	25.68										
8	16.93	20.48	27.68									
9	15.58	17.82	21.67	29.56								
10	14.69	16.24	18.64	22.78	31.33							
12	13.63	14.50	15.70	17.42	20.12	24.83						
14	13.04	13.61	14.33	15.28	16.58	18.48	26.67					
16	12.69	13.09	13.57	14.18	14.95	15.98	19.43	28.37				
18	12.47	12.76	13.11	13.54	14.05	14.70	16.61	20.31				
20	12.32	12.55	12.82	13.13	13.50	13.95	15.17	17.19				
25	12.15	12.29	12.44	12.62	12.81	13.04	13.59	14.34	18.46			
30	12.10	12.20	12.30	12.41	12.54	12.67	12.99	13.39	15.05	19.55		
35	12.12	12.18	12.26	12.34	12.42	12.52	12.73	12.98	13.89	15.67	20.51	
40	12.16	12.21	12.27	12.33	12.39	12.46	12.61	12.78	13.37	14.33	16.21	21.37
45	12.22	12.26	12.30	12.35	12.40	12.45	12.56	12.69	13.10	13.71	14.72	16.70
50	12.29	12.32	12.36	12.39	12.43	12.47	12.56	12.66	12.97	13.39	14.03	15.08
60	12.43	12.45	12.48	12.50	12.53	12.56	12.62	12.68	12.87	13.12	13.44	13.90
70	12.57	12.59	12.61	12.63	12.65	12.67	12.71	12.76	12.89	13.05	13.25	13.50
80	12.71	12.72	12.74	12.75	12.77	12.78	12.82	12.85	12.95	13.06	13.20	13.36
90	12.84	12.85	12.86	12.88	12.89	12.90	12.93	12.95	13.03	13.11	13.21	13.33
100	12.97	12.98	12.99	12.99	13.00	13.01	13.04	13.06	13.12	13.18	13.26	13.35
150	13.50	13.50	13.51	13.51	13.52	13.52	13.53	13.54	13.57	13.59	13.62	13.66
200	13.92	13.92	13.92	13.92	13.93	13.93	13.94	13.94	13.96	13.97	13.99	14.00

UPPER α/n POINTS OF F DISTRIBUTION WITH 1 AND n-q-1 DEGREES OF FREEDOM

α = 0.02

n \ q	1	2	3	4	5	6	8	10	15	20	25	30
6	39.11											
7	29.55	42.51										
8	24.81	31.41	45.67									
9	22.04	26.01	33.13	48.64								
10	20.26	22.90	27.12	34.73	51.45							
12	18.14	19.55	21.53	24.45	29.14	37.68						
14	16.95	17.84	18.98	20.51	22.65	25.82	40.35					
16	16.21	16.82	17.57	18.51	19.73	21.37	27.06	42.79				
18	15.72	16.17	16.70	17.34	18.13	19.13	22.14	28.20				
20	15.38	15.72	16.12	16.59	17.14	17.82	19.69	22.86				
25	14.88	15.09	15.31	15.56	15.85	16.18	16.99	18.11	24.42			
30	14.65	14.78	14.93	15.09	15.26	15.46	15.91	16.49	18.92	25.76		
35	14.53	14.63	14.73	14.84	14.96	15.09	15.39	15.74	17.04	19.62	26.94	
40	14.49	14.56	14.64	14.72	14.80	14.90	15.11	15.35	16.17	17.53	20.24	27.99
45	14.48	14.53	14.59	14.66	14.72	14.79	14.95	15.13	15.69	16.54	17.97	20.80
50	14.49	14.53	14.58	14.63	14.69	14.74	14.86	15.00	15.41	16.00	16.88	18.36
60	14.55	14.58	14.62	14.65	14.69	14.72	14.80	14.89	15.15	15.48	15.92	16.54
70	14.64	14.66	14.69	14.71	14.74	14.77	14.82	14.88	15.06	15.27	15.54	15.89
80	14.74	14.75	14.77	14.79	14.81	14.83	14.88	14.92	15.05	15.20	15.38	15.61
90	14.84	14.85	14.87	14.88	14.90	14.91	14.95	14.98	15.08	15.19	15.33	15.48
100	14.94	14.95	14.96	14.97	14.99	15.00	15.03	15.06	15.13	15.22	15.32	15.44
150	15.40	15.40	15.41	15.41	15.42	15.43	15.44	15.45	15.48	15.52	15.56	15.60
200	15.78	15.78	15.79	15.79	15.79	15.80	15.80	15.81	15.83	15.85	15.87	15.89

UPPER α/n POINTS OF F DISTRIBUTION WITH 1 AND n-q-1 DEGREES OF FREEDOM

α = 0.01

n \ q	1	2	3	4	5	6	8	10	15	20	25	30
6	56.68											
7	40.35	61.49										
8	32.59	42.79	65.96									
9	28.20	34.10	45.06	70.16								
10	25.41	29.25	35.51	47.18	74.14							
12	22.15	24.13	26.93	31.14	38.06	51.07						
14	20.34	21.55	23.12	25.25	28.26	32.81	54.59					
16	19.21	20.03	21.04	22.32	23.99	26.25	34.33	57.82				
18	18.45	19.05	19.75	20.61	21.67	23.02	27.16	35.71				
20	17.92	18.37	18.89	19.51	20.24	21.14	23.65	27.99				
25	17.12	17.38	17.66	17.99	18.36	18.78	19.84	21.31	29.82			
30	16.70	16.87	17.05	17.26	17.48	17.73	18.31	19.05	22.21	31.39		
35	16.46	16.58	16.71	16.85	17.01	17.17	17.54	17.99	19.65	22.98	32.76	
40	16.33	16.42	16.52	16.62	16.73	16.85	17.11	17.41	18.44	20.17	23.67	33.99
45	16.26	16.33	16.40	16.48	16.56	16.65	16.84	17.06	17.77	18.84	20.64	24.29
50	16.22	16.28	16.34	16.40	16.46	16.53	16.68	16.85	17.37	18.10	19.20	21.07
60	16.21	16.25	16.29	16.33	16.38	16.42	16.52	16.63	16.94	17.35	17.89	18.67
70	16.25	16.28	16.31	16.34	16.37	16.40	16.47	16.55	16.76	17.02	17.35	17.78
80	16.31	16.33	16.36	16.38	16.40	16.43	16.48	16.54	16.69	16.87	17.09	17.37
90	16.38	16.40	16.42	16.44	16.46	16.48	16.52	16.56	16.68	16.81	16.97	17.16
100	16.46	16.48	16.49	16.51	16.52	16.54	16.57	16.60	16.70	16.80	16.92	17.06
150	16.86	16.87	16.87	16.88	16.89	16.89	16.91	16.92	16.96	17.00	17.05	17.10
200	17.21	17.22	17.22	17.22	17.23	17.23	17.24	17.25	17.27	17.29	17.32	17.34

Appendix B

COMPUTER PROGRAMS

Original computer programs were developed to perform the analyses discussed in this thesis. The programs are written in fairly straightforward BASIC and are well documented, so no detailed explanations of programming details will be given here.

The first program performs ordinary least squares regression, ridge regression and generalized inverse regression. The program transforms its data as in (5.1) of chapter 5. The second program performs restricted least squares regression with either exact constraints or stochastic prior information. The source data are not transformed.

Both programs are available on the PRIME computer of the Graduate School of Business, University of Cape Town. Their relatively uncomplex structure makes them suitable for transcription to virtually any computer with a BASIC compiler, even a microcomputer. The programs have been written in such a way that memory storage is cut down at the expense of increased running time. This trade-off means that the programs may run on a microcomputer with a small memory capacity, although execution will be slow.

The programs are fairly general in that

a. there is technically no limit on the size of data sets;
b. each regression technique is purely optional;
c. the maximum size of groups to be examined is optional;

d. the significance level of the test statistics can be specified;

e. data files for source data are in free format.

In this appendix we shall list the general instructions for using the programs, and full listings of both programs.

Instructions for using computer routines for the detection of influential data

Two BASIC programs are available on the PRIME computer of the Graduate School of Business, University of Cape Town.

Program 1 : INF1 ordinary least squares regression
ridge regression
generalized inverse regression.

Program 2 : INF2 restricted regression with exact constraints
regression with stochastic prior information.

All options regarding regression techniques and size of data groups to be considered are entered on-line, in response to the relevant questions.

In order to run either program, a data file must be created beforehand. All data items in the file are in free field format, and data items on a line must be separated by commas.

The following structure must be followed for the data file:

1. Size of data set : 2 items
   - Item 1 : number of vector observations (n)
   - Item 2 : number of variables excluding constant (q-1)

2. Critical values for statistics : 8 items
   - Item 1 : significance level α (either 1,2,5 or 10)
   - Item 2 : upper α/n point of F(1,n-q-1) from Appendix A
   - Item 3 : upper α point of F(1,n-q-1)
   - Item 4 : upper α point of F(2,n-q-2)
   - Item 5 : upper α point of F(3,n-q-3)
   - Item 6 : lower 5% point of F(q,n-q-1)
   - Item 7 : lower 5% point of F(q,n-q-2)
   - Item 8 : lower 5% point of F(q,n-q-3)

3. Sample data : n lines
   One vector observation per line, in the form
   $y , x_1 , x_2 , \ldots , x_{q-1}$

4. Restricted least squares options : 2 items

   Item 1 : number of constraints (ℓ)

   Item 2 : option (0 = exact prior

   1 = stochastic prior)

5. Linear constraints : ℓ lines

   One vector constraint per line, in the form

   $h , H_1 , H_2 , \ldots , H_q$

6. Variance structure (stochastic prior only) : ℓ lines

   One row of R matrix per line, in the form

   $R_1 , R_2 , \ldots , R_\ell$

Note that numbers 4 to 6 above are required by program INF2 only, and that number 6 is required with stochastic prior information only.

The output files are prepared in a Fortran-type format with printing control characters in column 1. The printer should be set accordingly.

```
1000   !   INFLUENTIAL OBSERVATIONS IN LINEAR REGRESSION
1005   !      Program 1 :  Ordinary least squares regression
1010   !                   Ridge regression
1015   !                   G-inverse regression
1020   !
1025   !=============================================================================
1030   !  READ DATA FROM FILE AND INITIALIZE ARRAYS
1035   !=============================================================================
1040   INPUT 'ENTER NAME OF DATA FILE  ',F1$
1045   INPUT 'DO YOU REQUIRE RIDGE REGRESSION AND RIDGE TRACE?  ',O2$
1050   O2$=LEFT(O2$,1)
1055   IF O2$<>'Y' AND O2$<>'N' THEN 1045
1060   INPUT 'DO YOU REQUIRE GENERALIZED INVERSE REGRESSION?    ',O3$
1065   O3$=LEFT(O3$,1)
1070   IF O3$<>'Y' AND O3$<>'N' THEN 1060
1075   IF O1$='N' AND O2$='N' AND O3$='N' THEN 1090
1080   INPUT 'ENTER MAXIMUM SIZE OF SUBSETS TO BE CONSIDERED    ',MO
1085  !
1090   DEFINE FILE #1 = F1$
1095   READ #1,N,Q                                  ! N = NUMBER OF ROW OBSERVATIONS
1100                                                ! Q = NUMBER OF VARIABLES  (EXCLUDING CONSTANT)
1105   Q1=Q+1
1110   READ #1,A1,FO,F1(1),F1(2),F1(3),F2(1),F2(2),F2(3)
1115                                                ! READ IN CRITICAL VALUES
1120   MAT X=ZER(N,Q)
1125   MAT Y=ZER(N)
1130   MAT P=ZER(N)
1135   MAT R=ZER(N)
1140   MAT V=ZER(N)
1145   MAT VO=ZER(N)
1150   MAT XO=ZER(Q)
1155   MAT B=ZER(Q)
1160   !
1165   FOR I=1 TO N
1170      READ* #1,Y(I)                             ! MAT Y = DEPENDENT VARIABLES
1175      FOR J=1 TO Q
1180         READ* #1,X(I,J)                        ! MAT X = INDEPENDENT VARIABLES
1185         NEXT J
1190      NEXT I
1195   CLOSE #1
1200   !
1205   INPUT 'ENTER NAME OF OUTPUT FILE   ',F2$
1210   DEFINE FILE #7 = F2$                         ! NOMINATE NAME OF OUTPUT FILE
1215   WRITE #7,'1MULTIPLE LINEAR REGRESSION PACKAGE   ---   M.JACOBS'
1220   WRITE #7,'+_____________________________________________________'
1225   WRITE #7,'ONAME OF DATA FILE : ':F1$
1230   WRITE #7,'ONUMBER OF VECTOR OBSERVATIONS          :':N
1235   WRITE #7,' NUMBER OF VARIABLES (EXCLUDING CONSTANT) :':Q
1240   !=============================================================================
1245   !   ROUTINE TO TRANSFORM VARIABLES
1250   !=============================================================================
1255   MAT M1=ZER(Q)                                ! MAT M1 = COLUMN MEANS
1260   MAT M2=ZER(Q)                                ! MAT M2 = COLUMN SUMS OF SQUARES
1265   FOR J=1 TO Q
1270      FOR I=1 TO N
1275         M1(J)=M1(J)+X(I,J)
1280         M2(J)=M2(J)+X(I,J)^2
1285         NEXT I
1290      M2(J)=SQR(M2(J)-M1(J)^2/N)
1295      M1(J)=M1(J)/N
1300      FOR I=1 TO N
1305         X(I,J)=(X(I,J)-M1(J))/M2(J)
1310         NEXT I
1315      NEXT J
1320   !
1325   FOR I=1 TO N
1330      M1(0)=M1(0)+Y(I)                          ! ELEMENT 0 FOR DEPENDENT VARIABLE
1335      M2(0)=M2(0)+Y(I)^2
1340      NEXT I
1345   M2(0)=SQR(M2(0)-M1(0)^2/N)
1350   M1(0)=M1(0)/N
1355   FOR I=1 TO N
1360      Y(I)=(Y(I)-M1(0))/M2(0)
1365      NEXT I
```

```
1370  !=========================================================================
1375  !   ROUTINE TO CALCULATE CORRELATION COEFFICIENTS
1380  !=========================================================================
1385  MAT T8=ZER(N,Q1)
1390  FOR I=1 TO N
1395     T8(I,1)=Y(I)                              ! MAT T8 (TEMP) = TRANSFORMED VARIABLES
1400     FOR J=2 TO Q1
1405        T8(I,J)=X(I,J-1)
1410        NEXT J
1415     NEXT I
1420  !
1425  MAT T9 = TRN(T8)                             ! MAT T9 = TEMPORARY MATRIX
1430  MAT T7 = T9*T8                               ! MAT T7 = CORRELATION MATRIX
1435  WRITE #7,'OCORRELATION COEFFICIENTS:'
1440  WRITE #7,'+_____________________________'
1445  WRITE #7,'O          Y ';
1450  FOR J=1 TO Q
1455     IF J<10 THEN L$='     X'+STR$(J)+' ' ELSE L$='    X'+STR$(J)+' '
1460     WRITE #7,L$;
1465     NEXT J
1470  WRITE #7
1475  WRITE #7
1480  FOR I=1 TO Q1
1485     IF I>1 THEN L$=' X'+STR$(I-1)  ELSE L$=' Y'
1490     WRITE #7 USING '#### ',L$;
1495     FOR J=1 TO Q1
1500        WRITE #7 USING '   -#.##',T7(I,J);
1505        NEXT J
1510     WRITE #7
1515     IF I=1 THEN WRITE #7
1520     NEXT I
1525  !=========================================================================
1530  !   CALCULATE X'X AND X'Y MATRICES
1535  !=========================================================================
1540  FOR I=1 TO Q
1545       FOR J=1 TO N
1550          XO(I)=XO(I) + X(J,I)*Y(J)           ! MAT XO = X'Y
1555          NEXT J
1560       NEXT I
1565  !
1570  MAT T9 = TRN(X)
1575  MAT X1 = T9*X                                ! MAT X1 = X'X
1580  DO=DET(X1)
1585  MAT W = INV(X1)                              ! MAT W = (X'X)^
1590  MAT W8 = W
1595  !
1600  MAT E0=X1
1605  N0=Q
1610  GOSUB 6090                                   ! EIGENVALUES OF X'X
1615  MAT E8=E1
1620  MAT E9=E2                                    ! STORE EIGENVALUES AND EIGENVECTORS
1625  !=========================================================================
1630  !   CALCULATE AND PRINT O.L.S. REGRESSION COEFFICIENTS
1635  !=========================================================================
1640  FOR I=1 TO Q
1645     FOR J=1 TO Q
1650        B(I)=B(I) + W(I,J)*XO(J)               ! MAT B = REGRESSION COEFFICIENTS  (BETA)
1655        NEXT J
1660     NEXT I
1665  !
1670  MAT B1=ZER(Q)                                ! CONVERT BETA TO RAW DATA FORM
1675  FOR I=1 TO Q
1680     B1(I)=B(I)*M2(0)/M2(I)
1685     B1(0)=B1(0)+B1(I)*M1(I)
1690     NEXT I
1695  B1(0)=M1(0)-B1(0)
1700  !
1705  WRITE #7,'-REGRESSION COEFFICIENTS:'
1710  WRITE #7,'+_____________________________'
1715  FOR I=0 TO Q
1720     IF I=0 THEN L$='OCONSTANT' ELSE L$=' X'+STR$(I)
1725     WRITE #7 USING '##########   --------#.####',L$,B1(I)
1730     NEXT I
```

```
1735  !=============================================================================
1740  !   ANALYSIS OF RESIDUALS
1745  !=============================================================================
1750  FOR I=1 TO N
1755     FOR J=1 TO Q
1760        P(I)=P(I) + X(I,J)*B(J)                    ! MAT P = PREDICTORS OF Y
1765        NEXT J
1770     NEXT I
1775  MAT R = Y - P                                    ! MAT R = RESIDUALS
1780  MAT R8 = R
1785  !
1790  FOR I=1 TO N                                     ! MAT V = DIAG ELEMENTS OF HAT MATRIX
1795     FOR J=1 TO Q
1800        FOR K=1 TO Q
1805           V(I)=V(I) + X(I,J)*X(I,K)*W(J,K)
1810           NEXT K
1815        NEXT J
1820     NEXT I
1825  MAT V8 = V
1830  !
1835  FOR I=1 TO N                                     ! S2 = SUM OF SQUARES
1840     S2=S2 + R(I)^2
1845     NEXT I
1850  S=SQR(S2/(N-Q1))                                 ! S = ESTIMATE OF STD DEVIATION OF ERRORS
1855  !
1860  R2 = 1 - S2*(N-1)/(N-Q1)                         ! R2 = MULTIPLE CORRELATION COEFF
1865  WRITE #7 USING '-UNCORRECTED R SQUARED :##.###',1-S2
1870  WRITE #7 USING ' CORRECTED R SQUARED   :##.###',R2
1875  WRITE #7 USING "ODET OF CORRELATION    :--.###^^^^",DO
1880  WRITE #7 USING ' CONDITION NUMBER      :########.#',E1(1)/E1(Q)
1885  WRITE #7,"OEIGENVALUES OF X'X"
1890  FOR I=1 TO Q
1895     WRITE #7 USING ' ----.###',E1(I);
1900     NEXT I
1905  WRITE #7
1910  !
1915  WRITE #7,'1ANALYSIS OF RESIDUALS:'
1920  WRITE #7,'+_____________________'
1925  WRITE #7,'0';TAB(47);'STANDARD';TAB(60);'STUDENT';TAB(86);'COOK'
1930  WRITE #7,'          OBSERVED   PREDICTED    RESIDUAL    RESIDUAL    RESIDUAL    LEVERAGE
'
1935  WRITE #7
1940  F1$=' ###.    -----#.##    -----#.##    -----#.##    -----#.##    -----#.## # -----#.##    -----#.
1945  !
1950  M9=0
1955  FOR I=1 TO N
1960     F$,D$=' '
1965     S9=SQR( (S2-R(I)^2/(1-V(I))) / (N-Q1-1) )    ! S9 = S(i)
1970     Z = R(I) / (SQR(1-V(I))*S)                   ! Z = STANDARDIZED RESIDUAL
1975     T = R(I) / (SQR(1-V(I))*S9)                  ! T = STUDENTIZED RESIDUAL
1980     F=T^2
1985     IF F>F1(1) THEN F$='*'
1990     L=V(I)/(1-V(I))                              ! L = LEVERAGE
1995     DO=Z^2 * L/Q1                                ! DO = COOK'S DISTANCE
2000     D1=T^2 * L/Q1                                ! D1 = INFLUENCE MEASURE
2005     IF D1>F2(1) THEN D$='#'
2010     IF ABS(T)<=M9 THEN 2025
2015        M9=ABS(T)                                 ! M9 = MAXIMUM STUDENTIZED RESIDUAL
2020        L9=I                                      ! L9 = ITS LOCATION
2025     WRITE #7 USING F1$,I,Y(I)*M2(O)+M1(O),P(I)*M2(O)+M1(O),R(I)*M2(O),Z,T,F$,L,DO,D1,D$
2030     NEXT I
2035  WRITE #7 USING '-MAX T =###.## CORRESPONDING TO OBSERVATION ###',M9,L9
2040  F=M9^2
2045  WRITE #7 USING 'OF =###.##',F
2050  WRITE #7,' CRITICAL POINT AT ';A1/100;'/';N;' LEVEL = ';FO
2055  WRITE #7,'-* : GREATER THAN ';100-A1;'
2060  WRITE #7,' # : GREATER THAN 5
```

```
2065  !=============================================================================
2070  !   ANALYSIS OF INFLUENTIAL GROUPS
2075  !=============================================================================
2080  !
2085  !-----------------------------------------------------------------------------
2090  !   SORT R INTO DESCENDING ORDER
2095  !-----------------------------------------------------------------------------
2100  MAT R1=ZER(N)                                  ! MAT R1 = SORTED RESIDUALS
2105  MAT R2=ZER(N)                                  ! MAT R2 = ORDER OF SORTED RESIDUALS
2110  M8=1E30
2115  FOR I=1 TO N
2120     M9=-1E30
2125     FOR J=1 TO N
2130        IF ABS(R(J))<=M9 OR ABS(R(J))>=M8 THEN 2145
2135          M9=ABS(R(J))
2140          R2(I)=J
2145        NEXT J
2150     M8=M9
2155     NEXT I
2160  FOR I=1 TO N
2165     R1(I)=R(R2(I))
2170     NEXT I
2175  !-----------------------------------------------------------------------------
2180  !   SORT DIAGONAL ELEMENTS OF HAT MATRIX INTO DESCENDING ORDER
2185  !-----------------------------------------------------------------------------
2190  MAT V1=ZER(N)                                  ! MAT V1 = SORTED Vii
2195  MAT V2=ZER(N)                                  ! MAT V2 = ORDER OF SORTED Vii
2200  M8=1E30
2205  FOR I=1 TO N
2210     M9=-1E30
2215     FOR J=1 TO N
2220        IF V(J)<=M9 OR V(J)>=M8 THEN 2235
2225          M9=V(J)
2230          V2(I)=J
2235        NEXT J
2240     M8=M9
2245     NEXT I
2250  FOR I=1 TO N
2255     V1(I)=V(V2(I))
2260     NEXT I
2265  !-----------------------------------------------------------------------------
2270  !   ANALYSIS OF GROUPS OF VARYING SIZES  (M)
2275  !-----------------------------------------------------------------------------
2280  FOR M=1 TO M0                                  ! M = NUMBER OF OBSERVATIONS IN GROUP
2285  !
2290  A0,C=0
2295  R0=R0+R1(M)^2                                  ! R0 = UPPER BOUND FOR R2
2300  IF M=2 THEN B0=F2(2)*4.310                     ! B0,B1 = CUTOFF POINTS FOR UPPER BOUNDS
2305  IF M=2 THEN B1=F2(2)*1.246
2310  IF M=3 THEN B0=F2(3)*9.846
2315  IF M=3 THEN B1=F2(3)*2.727
2320  !
2325  MAT R9=ZER(M)                                  ! LOCAL MATRICES R9,V9 ARE SUBMATRICES OF R,V
2330  MAT V9=ZER(M,M)
2335  MAT I9=IDN(M,M)
2340  !
2345  FOR I=1 TO M                                   ! SET UP INDEX SET K
2350     L(I)=I
2355     K(I)=V2(I)
2360     NEXT I
2365  !
2370  MAT G=ZER(10,M+3)                              ! MAT G = RECORD OF TEN MOST INFLUENTIAL GROUPS
2375  !
2380  R7,V7=0
2385  MAT V9=ZER
2390  FOR I=1 TO M
2395     R9(I)=R(K(I))
2400     V9(I,I)=V(K(I))
2405     FOR J=I+1 TO M
2410        FOR K=1 TO Q                             ! COMPUTE OFF-DIAG ELEMENTS OF HAT MATRIX
2415           FOR L=1 TO Q
2420              V9(I,J)=V9(I,J) + X(K(I),K)*X(K(J),L)*W(K,L)
2425              NEXT L
2430           NEXT K
2435        V9(J,I)=V9(I,J)
2440        NEXT J
2445     R7=R7+R9(I)^2
2450     V7=V7+V9(I,I)
2455     NEXT I
```

```
2460  !-------------------------------------------------------------------------
2465  !   CALCULATE UPPER BOUNDS AND VALUES OF DISTANCE MEASURE
2470  !-------------------------------------------------------------------------
2475  MAT T9 = I9-V9
2480  MAT T7 = INV(T9)                              ! MAT T7 (TEMP) = (I-Vi)~
2485  S9=0
2490  FOR I=1 TO M                                  ! S9 = ESTIMATE OF VARIANCE
2495     FOR J=1 TO M
2500        S9=S9 + R9(I)*R9(J)*T7(I,J)
2505        NEXT J
2510     NEXT I
2515  S9 = (S2-S9)/(N-G1-M)
2520  !
2525  IF M=1 OR V7>1 THEN 2575
2530  U1=R0*V7/(G1*S9*(1-V7)^2)                     ! U1 = FIRST UPPER BOUND FOR INFLUENCE
2535  IF U1>=B0 AND U1>=G(10,M+1) THEN 2550
2540    A0=1                                        ! A0 = QUICK ADVANCE INDICATOR
2545    GOTO 2815
2550  U2=R7*V7/(G1*S9*(1-V7)^2)                     ! U2 = SECOND UPPER BOUND
2555  IF U2>=B1 AND U2>=G(10,M+1) THEN 2575
2560    A0=1
2565    GOTO 2815
2570  !
2575  MAT T9=I9-V9
2580  MAT T7=INV(T9)                                ! MAT T7 (TEMP) = (I-Vi)~
2585  MAT T8=T7*V9
2590  MAT T9=T8*T7                                  ! MAT T9 (TEMP)~ = (I-Vi)~*Vi*(I-Vi)~
2595  !
2600  D1,T1=0                                       ! D1 = INFLUENCE
2605                                                ! T1 = GENERALIZED STUDENTIZED RESIDUAL
2610  FOR I=1 TO M
2615     FOR J=1 TO M
2620        D1=D1 + R9(I)*R9(J)*T9(I,J)
2625        T1=T1 + R9(I)*R9(J)*T7(I,J)
2630        NEXT J
2635     NEXT I
2640  D1 = D1/(G1*S9)
2645  T1 = T1/(M*S9)
2650  C=C+1                                         ! C = COUNTER
2655  !-------------------------------------------------------------------------
2660  !   CHECK FOR POSITION IN 'TOP TEN' GROUP
2665  !-------------------------------------------------------------------------
2670  IF D1<=G(10,M+1) THEN 2815
2675    FOR I=9 TO 1 STEP -1
2680       IF D1<G(I,M+1) THEN 2690
2685       NEXT I
2690    L=I+1
2695    FOR I=10 TO L+1 STEP -1                     ! INCORPORATE INTO 'TOP TEN' GROUP
2700       FOR J=1 TO M+3
2705          G(I,J)=G(I-1,J)
2710          NEXT J
2715       NEXT I
2720    FOR J=1 TO M
2725       G(L,J)=K(J)
2730       NEXT J
2735    G(L,M+1)=D1
2740    G(L,M+2)=T1
2745    !
2750    IF M>1 THEN 2765                            ! ROUTINE TO CALCULATE LEVERAGE
2755      G(L,M+3)=V9(1,1)/(1-V9(1,1))
2760      GOTO 2810
2765    MAT E0=V9
2770    N0=M
2775    GOSUB 6090                                  ! EIGENVALUES OF V9
2780    L1=0
2785    FOR J=1 TO M
2790       L1=L1 + E1(J)/(1-E1(J))
2795       NEXT J
2800    G(L,M+3)=L1
2805  !-------------------------------------------------------------------------
2810  !   ADVANCE INDEX SET
2815  !-------------------------------------------------------------------------
2820  U=M-A0
2825  L(U)=L(U)+1
2830  IF L(U)<=N-M+U THEN 2845
2835    U=U-1
2840    IF U>0 THEN 2825 ELSE 2895
2845  IF U=M THEN 2870
2850  U=U+1
2855  L(U)=L(U-1)
2860  GOTO 2825
2865  !
2870  FOR U=1 TO M
2875     K(U)=V2(L(U))
2880     NEXT U
2885  A0=0
2890  GOTO 2375
```

```
2895  !-------------------------------------------------------------------------------
2900  !   PRINT RESULTS OF ANALYSIS OF INFLUENTIAL GROUPS
2905  !-------------------------------------------------------------------------------
2910  FOR L=10 TO 1 STEP -1
2915     IF G(L,M+1)>0 THEN 2925                    ! DETERMINE HOW MANY GROUPS ARE INFLUENTIAL
2920     NEXT L
2925  !
2930  WRITE #7,'1ANALYSIS OF INFLUENTIAL GROUPS:'
2935  WRITE #7,'+_____________________________'
2940  WRITE #7,'ONUMBER OF OBSERVATIONS PER GROUP =':M
2945  WRITE #7 USING ' ## MOST INFLUENTIAL GROUPS SHOWN.',L
2950  G=1                                           ! G = MAXIMUM NUMBER OF SUBSETS
2955  FOR I=1 TO M
2960     G=G*(N-I+1)/I
2965     NEXT I
2970  WRITE #7,'ONUMBER OF SUBSETS CONSIDERED       =':G
2975  WRITE #7,' EXACT COMPUTATIONS OF INFLUENCE    =':C
2980  FO$='----#.## #  ----#.##     ----#.## #'
2985  WRITE #7,'O        OBSERVATIONS'
2990  WRITE #7,'         IN GROUP';TAB(22+4*M);' F STAT     LEVERAGE     INFLUENCE'
2995  WRITE #7
3000  FOR I=1 TO L
3005     F$,D$=' '
3010     IF G(I,M+2)>F1(M) THEN F$='*'
3015     IF G(I,M+1)>F2(M) THEN D$='#'
3020     WRITE #7 USING ' ##.    ',I;
3025     FOR J=1 TO M
3030        WRITE #7 USING ' ###',G(I,J);
3035        NEXT J
3040     WRITE #7,TAB(20+4*M);
3045     WRITE #7 USING FO$,G(I,M+2),F$,G(I,M+3),G(I,M+1),D$
3050     NEXT I
3055  WRITE #7,'-* : GREATER THAN ';100-A1;'
3060  WRITE #7,' # : GREATER THAN 5
3065  !
3070  NEXT M
```

```
3075  !=========================================================================
3080  !   RIDGE REGRESSION
3085  !=========================================================================
3090  IF O2$='N' THEN 5055                          ! PERFORM RIDGE REGRESSION ONLY ON REQUEST
3095  !
3100  MAT BO = ZER(Q,6)
3105  !
3110  FOR M=1 TO MO
3115  !
3120  AO,C,CO,KO=0
3125  MAT X9 = ZER(M,Q)                             ! MAT X9 = SUBMATRIX OF X
3130  MAT R9 = ZER(M)                               ! MAT R9 = SUBMATRIX OF R
3135  MAT R7 = ZER(M)                               ! MAT R7 = LEAST SQUARES EQUIVALENT
3140  MAT V9 = ZER(M,M)                             ! MAT V9 = SUBMATRIX OF V (HAT MATRIX)
3145  MAT V7 = ZER(M,M)                             ! MAT V7 = LEAST SQUARES EQUIVALENT
3150  MAT V6 = ZER(M,M)
3155  MAT I9 = IDN(M,M)
3160  MAT GO = ZER(60,M)                            ! MAT GO STORES INDICES OF POTENTIALLY INFLUENTIA
3165  !-------------------------------------------------------------------------
3170  !   STAGE 1:    k IN STEPS OF 0.10 TO LOCATE POTENTIAL GROUPS
3175  !-------------------------------------------------------------------------
3180  FOR K=0 TO 0.50 STEP 0.10
3185  !
3190  KO=KO+1
3195  MAT G=ZER(10,M)                               ! MAT G = 'TOP TEN' INFLUENTIAL GROUPS FOR PARTIC
3200  FOR I=1 TO 10
3205     G(I,0)=0
3210     NEXT I
3215  !
3220  MAT T9 = X1
3225  FOR I=1 TO Q
3230     T9(I,I)=T9(I,I) + K
3235     NEXT I
3240  MAT W = INV(T9)                               ! MAT W = (X'X+kI)~
3245  !
3250  MAT B=ZER
3255  FOR I=1 TO Q
3260     FOR J=1 TO Q
3265        B(I)=B(I) + W(I,J)*XO(J)                ! MAT B    = REGRESSION COEFFICIENTS
3270        NEXT J
3275     NEXT I
3280  FOR I=1 TO Q                                  ! COLUMNS 1-6 OF BO MATRIX CORRESPOND TO k=0 TO 0
3285     BO(I,KO)=B(I)*M2(0)/M2(I)
3290     BO(0,KO)=BO(0,KO)+BO(I,KO)*M1(I)
3295     NEXT I
3300  BO(0,KO)=M1(0)-BO(0,KO)
3305  !
3310  MAT P = ZER
3315   FOR I=1 TO N
3320      FOR J=1 TO Q
3325         P(I)=P(I) + X(I,J)*B(J)                ! MAT P = PREDICTORS OF Y
3330         NEXT J
3335      NEXT I
3340   MAT R = Y - P                                ! MAT R = RESIDUALS
3345  !
3350  MAT T8 = W*X1
3355  MAT W2 = T8 * W                               ! MAT W2 = WX'XW
3360  !
3365  FOR I=1 TO N
3370     V(I)=0
3375     FOR J=1 TO Q
3380        FOR L=1 TO Q
3385           V(I)=V(I) + X(I,J)*X(I,L)*W(J,L)     ! V(I) = Ith DIAG ELEMENT OF V=XWX'
3390           NEXT L
3395        NEXT J
3400     NEXT I
3405  !
3410  FOR I=1 TO M                                  ! SET UP INDEX SET K
3415     K(I)=I
3420     NEXT I
```

```
3425  !-------------------------------------------------------------------------
3430  !   ROUTINE TO CALCULATE INFLUENCE OF A GROUP OF OBSERVATIONS
3435  !-------------------------------------------------------------------------
3440  MAT V9=ZER
3445  MAT V7=ZER
3450  FOR I=1 TO M
3455     FOR J=1 TO Q
3460        X9(I,J)=X(K(I),J)                    ! MAT X9 = SUBMATRIX OF X
3465        NEXT J
3470     R9(I)=R(K(I))                           ! MAT R9 = SUBSET OF R
3475     R7(I)=R8(K(I))
3480     V9(I,I)=V(K(I))                         ! MAT V9 = SUBMATRIX OF V (HAT MATRIX)
3485     V7(I,I)=V8(K(I))
3490     FOR J=I+1 TO M
3495        FOR J1=1 TO Q                        ! COMPUTE OFF-DIAG ELEMENTS OF HAT MATRIX
3500           FOR J2=1 TO Q
3505              V9(I,J)=V9(I,J) + X(K(I),J1)*X(K(J),J2)*W(J1,J2)
3510              V7(I,J)=V7(I,J) + X(K(I),J1)*X(K(J),J2)*W8(J1,J2)
3515              NEXT J2
3520           NEXT J1
3525        V9(J,I)=V9(I,J)
3530        V7(J,I)=V7(I,J)
3535        NEXT J
3540     NEXT I
3545  !
3550  MAT T9=I9-V7
3555  MAT T7=INV(T9)
3560  S9=0                                       ! CALCULATE LEAST SQUARES ESTIMATE OF VARIANCE
3565  FOR I=1 TO M
3570     FOR J=1 TO M
3575        S9=S9 + R7(I)*R7(J)*T7(I,J)
3580        NEXT J
3585     NEXT I
3590  S9 = (S2-S9)/(N-Q1-M)
3595  !
3600  MAT T9 = I9-V9
3605  MAT T6 = INV(T9)                           ! MAT T6 = (I-Vi)^
3610  MAT T7 = X9*W2
3615  MAT T9 = TRN(X9)
3620  MAT T8 = T7*T9                             ! MAT T8 = XiWX'XWXi'
3625  MAT T7 = T6*T8
3630  MAT T9 = T7*T6
3635  !
3640  D1=0                                       ! D1 = INFLUENCE
3645  FOR I=1 TO M
3650     FOR J=1 TO M
3655        D1=D1 + R9(I)*R9(J)*T9(I,J)
3660        NEXT J
3665     NEXT I
3670  D1=D1/(Q1*S9)
3675  !-------------------------------------------------------------------------
3680  !   CHECK FOR POSITION IN 'TOP TEN' GROUP
3685  !-------------------------------------------------------------------------
3690  IF D1<=G(10,0) THEN 3765
3695    FOR I=9 TO 1 STEP -1
3700       IF D1<G(I,0) THEN 3710
3705       NEXT I
3710    L=I+1
3715    FOR I=10 TO L+1 STEP -1                  ! INCORPORATE INTO 'TOP TEN' GROUP
3720       FOR J=0 TO M
3725          G(I,J)=G(I-1,J)
3730          NEXT J
3735       NEXT I
3740    FOR J=1 TO M
3745       G(L,J)=K(J)
3750       NEXT J
3755    G(L,0)=D1
3760  !-------------------------------------------------------------------------
3765  !   ADVANCE INDEX SET
3770  !-------------------------------------------------------------------------
3775  U=M
3780  K(U)=K(U)+1
3785  IF K(U)<=N-M+U THEN 3800
3790    U=U-1
3795    IF U>0 THEN 3780 ELSE 3835
3800  IF U=M THEN 3430
3805  U=U+1
3810  K(U)=K(U-1)
3815  GOTO 3780
```

```
3820  !-------------------------------------------------------------------------
3825  !   ADD INDICES IN G TO MATRIX GO  (IF NOT ALREADY PRESENT)
3830  !-------------------------------------------------------------------------
3835  FOR I=1 TO 10
3840     FOR J=1 TO CO
3845        FOR L=1 TO M
3850           IF G(I,L)<>GO(J,L) THEN 3865
3855           NEXT L
3860        GOTO 3890
3865        NEXT J
3870     CO=CO+1                                  ! CO = NUMBER OF GROUPS RECORDED IN MAT GO
3875     FOR L=1 TO M
3880        GO(CO,L)=G(I,L)
3885        NEXT L
3890     NEXT I
3895  !
3900  NEXT K
3905  !-------------------------------------------------------------------------
3910  !   PRINT RIDGE REGRESSION COEFFICIENTS
3915  !-------------------------------------------------------------------------
3920  IF M>1 THEN 3985
3925  !
3930  WRITE #7,'1RIDGE REGRESSION: '
3935  WRITE #7,'+_____'
3940   WRITE #7,'-REGRESSION COEFFICIENTS: '
3945   WRITE #7,'+_____'
3950   WRITE #7,'0                k=0.00       k=0.10       k=0.20       k=0.30       k=0.40
3955   !
3960   F1$='##########  ------.####  ------.####  ------.####  ------.####  ------.####  ------.##+
3965   FOR I=0 TO Q
3970      IF I=0 THEN L$='0CONSTANT'  ELSE L$=' X'+STR$(I)
3975      WRITE #7 USING F1$,L$,BO(I,1),BO(I,2),BO(I,3),BO(I,4),BO(I,5),BO(I,6)
3980      NEXT I
3985  !-------------------------------------------------------------------------
3990  !   STAGE 2:
3995  !-------------------------------------------------------------------------
4000  MAT G1 = ZER(CO,51)                         ! MAT G1 = RIDGE TRACE OF INFLUENCE
4005  MAT G2 = ZER(CO,51)                         ! MAT G2 = RIDGE TRACE OF STUDENT RESIDUAL
4010  MAT G3 = ZER(CO,51)                         ! MAT G3 = RIDGE TRACE OF LEVERAGE
4015  KO=0
4020  !
4025  FOR K=0 TO 0.50 STEP 0.01
4030  KO=KO+1                                     ! KO = COUNTER 1 TO 51
4035  MAT T9 = X1
4040  FOR I=1 TO Q
4045     T9(I,I)=T9(I,I) + K
4050     NEXT I
4055  MAT W = INV(T9)                             ! MAT W = (X'X+kI)^
4060  MAT WO = W*W                                ! MAT WO = WW
4065  !
4070  MAT B=ZER
4075  FOR I=1 TO Q
4080     FOR J=1 TO Q
4085        B(I)=B(I) + W(I,J)*XO(J)              ! MAT B = REGRESSION COEFFICIENTS
4090        NEXT J
4095     NEXT I
4100  !
4105  MAT P = ZER
4110   FOR I=1 TO N
4115      FOR J=1 TO Q
4120         P(I)=P(I) + X(I,J)*B(J)              ! MAT P = PREDICTORS OF Y
4125         NEXT J
4130      NEXT I
4135   MAT R = Y - P                              ! MAT R = RESIDUALS
4140  !
4145  MAT T8 = W*X1
4150  MAT W2 = T8 * W                             ! MAT W2 = WX'XW
4155  !
4160  FOR I=1 TO N
4165     V(I),VO(I)=0
4170     FOR J=1 TO Q
4175        FOR L=1 TO Q
4180           V(I)=V(I) + X(I,J)*X(I,L)*W(J,L)  ! V(I) = Ith DIAG ELEMENT OF V=XWX'
4185           VO(I)=VO(I) + X(I,J)*X(I,L)*WO(J,L)          ! Ith DIAG ELEMENT OF XWWX'
4190           NEXT L
4195        NEXT J
4200     NEXT I
```

```
4205  !-------------------------------------------------------------------------
4210  !   ROUTINE TO CALCULATE INFLUENCE
4215  !-------------------------------------------------------------------------
4220  FOR G=1 TO CO
4225     FOR I=1 TO M
4230        K(I)=GO(G,I)                          ! INDEX SET K BASED ON GROUPS STORED IN GO
4235        NEXT I
4240     !
4245     MAT V9=ZER
4250     MAT V7=ZER
4255     MAT V6=ZER
4260     FOR I=1 TO M
4265        FOR J=1 TO Q
4270           X9(I,J)=X(K(I),J)
4275           NEXT J
4280        R9(I)=R(K(I))
4285        R7(I)=R8(K(I))
4290        V9(I,I)=V(K(I))                       ! MAT V9 = XWX'
4295        V7(I,I)=V8(K(I))                      ! MAT V7 = O.L.S. HAT MATRIX
4300        V6(I,I)=VO(K(I))                      ! MAT V6 = XWWX'
4305        FOR J=I+1 TO M
4310           FOR J1=1 TO Q                      ! COMPUTE OFF-DIAG ELEMENTS OF HAT MATRIX
4315              FOR J2=1 TO Q
4320                 V9(I,J)=V9(I,J) + X(K(I),J1)*X(K(J),J2)*W(J1,J2)
4325                 V7(I,J)=V7(I,J) + X(K(I),J1)*X(K(J),J2)*W8(J1,J2)
4330                 V6(I,J)=V6(I,J) + X(K(I),J1)*X(K(J),J2)*WO(J1,J2)
4335                 NEXT J2
4340              NEXT J1
4345           V9(J,I)=V9(I,J)
4350           V7(J,I)=V7(I,J)
4355           V6(J,I)=V6(I,J)
4360           NEXT J
4365        NEXT I
4370     !
4375     MAT T9=I9-V7
4380     MAT T7=INV(T9)
4385     S9=0                                     ! CALCULATE LEAST SQUARES ESTIMATE OF VARIANCE
4390     FOR I=1 TO M
4395        FOR J=1 TO M
4400           S9=S9 + R7(I)*R7(J)*T7(I,J)
4405           NEXT J
4410        NEXT I
4415     S9 = (S2-S9)/(N-Q1-M)
4420     !
4425     MAT T9 = I9-V9
4430     MAT T8 = (K)*V6
4435     MAT T8 = T9-T8
4440     MAT T5 = INV(T8)                         ! MAT T5 = (I-Qi)~
4445     MAT T6 = INV(T9)                         ! MAT T6 = (I-Vi)~
4450     MAT T7 = X9*W2
4455     MAT T9 = TRN(X9)
4460     MAT T8 = T7*T9                           ! MAT T8 = XiWX'XWXi'
4465     MAT T7 = T6*T8
4470     MAT T9 = T7*T6
4475     !
4480     D1,T1=0                                  ! D1 = INFLUENCE
4485                                              ! T1 = GENERALIZED STUDENTIZED RESIDUAL
4490     FOR I=1 TO M
4495        FOR J=1 TO M
4500           D1=D1 + R9(I)*R9(J)*T9(I,J)
4505           T1=T1 + R9(I)*R9(J)*T5(I,J)
4510           NEXT J
4515        NEXT I
4520     G1(G,KO)=D1/(Q1*S9)
4525     G2(G,KO)=T1/(M*S9)
4530     !
4535     IF M>1 THEN 4550                         ! ROUTINE TO CALCULATE LEVERAGE
4540       G3(G,KO)=(V9(1,1)-K*V6(1,1))/(1-V9(1,1)-K*V6(1,1))
4545       GOTO 4610
4550     MAT T9=(K)*V6
4555     MAT EO=V9-T9
4560     NO=M
4565     GOSUB 6090                               ! EIGENVALUES OF V9-kV6
4570     MAT LO=E1
4575     MAT EO=V9+T9
4580     GOSUB 6090                               ! EIGENVALUES OF V9+kV6
4585     L1=0
4590     FOR J=1 TO M
4595        L1=L1 + LO(J)/(1-E1(J))
4600        NEXT J
4605     G3(G,KO)=L1
4610     !
4615     NEXT G
4620  !
4625  NEXT K
4630  !
```

```
4635 !-------------------------------------------------------------------------
4640 !   PRINT SUMMARY RIDGE TRACE
4645 !-------------------------------------------------------------------------
4650 WRITE #7 USING "1SUMMARY RIDGE TRACE FOR M = #",M
4655 WRITE #7,'+_____'
4660 WRITE #7,'-OBSERVATIONS        k =      .00      .10      .20      .30      .40      .50'
4665 WRITE #7
4670 !
4675 MAT G=ZER(CO)
4680 FOR I=1 TO 10                                  ! PRINT TOP TEN GROUPS ONLY
4685    M9=0
4690    FOR J=1 TO CO                               ! ROUTINE TO FIND LARGEST Di
4695       IF G(J)=1 THEN 4725
4700         FOR L=1 TO 51
4705            IF G1(J,L)<=M9 THEN 4720
4710              M9=G1(J,L)
4715              L9=J
4720            NEXT L
4725       NEXT J
4730    G(L9)=1
4735    !
4740    WRITE #7
4745    FOR J=1 TO M
4750       WRITE #7 USING ' ###',GO(L9,J):
4755       NEXT J
4760    WRITE #7,TAB(15);'Influence';               ! PRINT INFLUENCE
4765    FOR J=1 TO 51 STEP 10
4770       WRITE #7 USING ' ---.###',G1(L9,J);
4775       NEXT J
4780    WRITE #7
4785    WRITE #7,TAB(15);'Leverage ';               ! PRINT LEVERAGE
4790    FOR J=1 TO 51 STEP 10
4795       WRITE #7 USING ' ---.###',G3(L9,J);
4800       NEXT J
4805    WRITE #7
4810    WRITE #7,TAB(15);'F stat    ';              ! PRINT F STATISTIC
4815    FOR J=1 TO 51 STEP 10
4820       WRITE #7 USING ' ---.###',G2(L9,J);
4825       NEXT J
4830    WRITE #7
4835    NEXT I
4840 !-------------------------------------------------------------------------
4845 !   PRINT DETAILED RIDGE TRACE OF INFLUENCE
4850 !-------------------------------------------------------------------------
4855 WRITE #7 USING "1RIDGE TRACE OF INFLUENCE FOR M = #",M
4860 WRITE #7,'+_____'
4865 WRITE #7,'-OBSERVATIONS';TAB(23);
4870 FOR I=0 TO 0.09 STEP 0.01
4875    WRITE #7 USING '    +.##',I;
4880    NEXT I
4885 WRITE #7
4890 !
4895 MAT G=ZER(CO)
4900 FOR I=1 TO 10                                  ! PRINT TOP TEN GROUPS ONLY
4905    M9=0
4910    FOR J=1 TO CO                               ! ROUTINE TO FIND LARGEST Di
4915       IF G(J)=1 THEN 4945
4920         FOR L=1 TO 51
4925            IF G1(J,L)<=M9 THEN 4940
4930              M9=G1(J,L)
4935              L9=J
4940            NEXT L
4945       NEXT J
4950    G(L9)=1
4955    !
4960    WRITE #7
4965    FOR J=1 TO M
4970       WRITE #7 USING ' ###',GO(L9,J):
4975       NEXT J
4980    FOR J=0 TO 40 STEP 10
4985       WRITE #7,TAB(16);
4990       WRITE #7 USING 'k=.##   ',J/100;
4995       FOR L=1 TO 10
5000          WRITE #7 USING ' ---.###',G1(L9,J+L);
5005          NEXT L
5010       WRITE #7
5015       NEXT J
5020    WRITE #7,TAB(16);'k=.50   ';
5025    WRITE #7 USING ' ---.###',G1(L9,51)
5030    NEXT I
5035 !
5040 NEXT M
5045 !
```

```
5050  !=========================================================================
5055  !   GENERALIZED INVERSE REGRESSION
5060  !=========================================================================
5065  IF O3$='N' THEN 6070                          ! PERFORM G-INVERSE REGRESSION ONLY ON REQUEST
5070  !
5075  WRITE #7,'1GENERALIZED INVERSE REGRESSION:'
5080  WRITE #7,'+_____'
5085  FOR L=Q TO 1 STEP -1                          ! ELIMINATE EIGENVALUES UNTIL CONDITION NUMBER <=
5090     IF E8(1)/E8(L)<=30 THEN 5100
5095     NEXT L
5100  IF L<Q THEN 5120
5105    WRITE #7,'-No eigenvalues eliminated in calculating the generalized inverse.'
5110    WRITE #7,' See ordinary least squares analysis.'
5115    GOTO 6070
5120  !
5125  MAT X2=ZER(Q,Q)                               ! MAT X2 = G-INV(X'X)
5130  FOR I=1 TO Q
5135     FOR J=I TO Q
5140        FOR K=1 TO L
5145           X2(I,J)=X2(I,J) + E9(I,K)*E9(J,K)/E8(K)
5150           NEXT K
5155        IF J<>I THEN X2(J,I)=X2(I,J)
5160        NEXT J
5165     NEXT I
5170  !-------------------------------------------------------------------------
5175  !   CALCULATE AND PRINT REGRESSION COEFFICIENTS
5180  !-------------------------------------------------------------------------
5185  MAT B=ZER
5190  MAT W=X2
5195  !
5200   FOR I=1 TO Q
5205      FOR J=1 TO Q
5210         B(I)=B(I) + W(I,J)*XO(J)               ! MAT B = REGRESSION COEFFICIENTS  (BETA)
5215         NEXT J
5220      NEXT I
5225   !
5230   MAT B1=ZER(Q)                                ! CONVERT BETA TO RAW DATA FORM
5235   FOR I=1 TO Q
5240      B1(I)=B(I)*M2(0)/M2(I)
5245      B1(0)=B1(0)+B1(I)*M1(I)
5250      NEXT I
5255   B1(0)=M1(0)-B1(0)
5260   !
5265   WRITE #7,'0NUMBER OF EIGENVALUES ELIMINATED =':Q-L
5270   WRITE #7,'-REGRESSION COEFFICIENTS:'
5275   WRITE #7,'+_____'
5280   FOR I=0 TO Q
5285      IF I=0 THEN L$='0CONSTANT' ELSE L$=' X'+STR$(I)
5290      WRITE #7 USING '##########   --------#.####',L$,B1(I)
5295      NEXT I
5300  !
5305  MAT P=ZER
5310   FOR I=1 TO N
5315      FOR J=1 TO Q
5320         P(I)=P(I) + X(I,J)*B(J)                ! MAT P = PREDICTORS OF Y
5325         NEXT J
5330      NEXT I
5335   MAT R = Y - P                                ! MAT R = RESIDUALS
5340  !
5345  MAT V=ZER
5350  FOR I=1 TO N                                  ! MAT V = DIAG ELEMENTS OF HAT MATRIX
5355     FOR J=1 TO Q
5360        FOR K=1 TO Q
5365           V(I)=V(I) + X(I,J)*X(I,K)*W(J,K)
5370           NEXT K
5375        NEXT J
5380     NEXT I
5385  !
5390  FOR M=1 TO MO                                 ! M = NUMBER OF OBSERVATIONS IN GROUP
5395  !
5400  MAT X9=ZER(M,Q)                               ! MAT X9 = SUBMATRIX OF X
5405  MAT R9=ZER(M)                                 ! MAT R9 = SUBMATRIX OF R
5410  MAT R7=ZER(M)                                 ! MAT R7 = LEAST SQUARES EQUIVALENT
5415  MAT V9=ZER(M,M)                               ! MAT V9 = SUBMATRIX OF V
5420  MAT V7=ZER(M,M)                               ! MAT V7 = LEAST SQUARES EQUIVALENT
5425  MAT I9=IDN(M,M)
5430  !
5435  MAT G=ZER(10,M+3)                             ! MAT G = RECORD OF TEN MOST INFLUENTIAL GROUPS
5440  !
5445  FOR I=1 TO M                                  ! SET UP INDEX SET K
5450     K(I)=I
5455     NEXT I
```

```
5460 !-------------------------------------------------------------------------
5465 !   ROUTINE TO CALCULATE INFLUENCE OF A GROUP OF OBSERVATIONS
5470 !-------------------------------------------------------------------------
5475 MAT V9=ZER
5480 MAT V7=ZER
5485 FOR I=1 TO M
5490    FOR J=1 TO Q
5495       X9(I,J)=X(K(I),J)                    ! MAT X9 = SUBMATRIX OF X
5500       NEXT J
5505    R9(I)=R(K(I))                           ! MAT R9 = SUBSET OF R
5510    R7(I)=R8(K(I))
5515    V9(I,I)=V(K(I))                         ! MAT V9 = SUBMATRIX OF V (HAT MATRIX)
5520    V7(I,I)=V8(K(I))
5525    FOR J=I+1 TO M
5530       FOR J1=1 TO Q                        ! COMPUTE OFF-DIAG ELEMENTS OF HAT MATRIX
5535          FOR J2=1 TO Q
5540             V9(I,J)=V9(I,J) + X(K(I),J1)*X(K(J),J2)*W(J1,J2)
5545             V7(I,J)=V7(I,J) + X(K(I),J1)*X(K(J),J2)*W8(J1,J2)
5550             NEXT J2
5555          NEXT J1
5560       V9(J,I)=V9(I,J)
5565       V7(J,I)=V7(I,J)
5570       NEXT J
5575    NEXT I
5580 !
5585 MAT T9=I9-V7
5590 MAT T7=INV(T9)
5595 S9=0                                       ! CALCULATE LEAST SQUARES ESTIMATE OF VARIANCE
5600 FOR I=1 TO M
5605    FOR J=1 TO M
5610       S9=S9 + R7(I)*R7(J)*T7(I,J)
5615       NEXT J
5620    NEXT I
5625 S9 = (S2-S9)/(N-Q1-M)
5630 !
5635 MAT T9 = I9-V9
5640 MAT T7 = INV(T9)                           ! MAT T7 = (I-Vi)~
5645 MAT T8 = T7*V9
5650 MAT T9 = T8*T7                             ! MAT T9 = (I-Vi)~*Vi*(I-Vi)~
5655 !
5660 D1,T1=0                                    ! D1 = INFLUENCE
5665                                            ! T1 = GENERALIZED STUDENTIZED RESIDUAL
5670 FOR I=1 TO M
5675    FOR J=1 TO M
5680       D1=D1 + R9(I)*R9(J)*T9(I,J)
5685       T1=T1 + R9(I)*R9(J)*T7(I,J)
5690       NEXT J
5695    NEXT I
5700 D1=D1/(Q1*S9)
5705 T1=T1/(M*S9)
5710 !-------------------------------------------------------------------------
5715 !   CHECK FOR POSITION IN 'TOP TEN' GROUP
5720 !-------------------------------------------------------------------------
5725 IF D1<=G(10,M+1) THEN 5865
5730   FOR I=9 TO 1 STEP -1
5735      IF D1<G(I,M+1) THEN 5745
5740      NEXT I
5745   L=I+1
5750   FOR I=10 TO L+1 STEP -1                  ! INCORPORATE INTO 'TOP TEN' GROUP
5755      FOR J=1 TO M+3
5760         G(I,J)=G(I-1,J)
5765         NEXT J
5770      NEXT I
5775   FOR J=1 TO M
5780      G(L,J)=K(J)
5785      NEXT J
5790   G(L,M+1)=D1
5795   G(L,M+2)=T1
5800   !
5805   IF M>1 THEN 5820                         ! ROUTINE TO CALCULATE LEVERAGE
5810     G(L,M+3)=V9(1,1)/(1-V9(1,1))
5815     GOTO 5865
5820   MAT E0=V9
5825   N0=M
5830   GOSUB 6090                               ! EIGENVALUES OF V9
5835   L1=0
5840   FOR J=1 TO M
5845      L1=L1 + E1(J)/(1-E1(J))
5850      NEXT J
5855   G(L,M+3)=L1
5860 !-------------------------------------------------------------------------
5865 !   ADVANCE INDEX SET
5870 !-------------------------------------------------------------------------
5875 U=M
5880 K(U)=K(U)+1
5885 IF K(U)<=N-M+U THEN 5900
5890   U=U-1
5895   IF U>0 THEN 5880 ELSE 5925
5900 IF U=M THEN 5465
5905 U=U+1
5910 K(U)=K(U-1)
5915 GOTO 5880
```

```
5920  !-------------------------------------------------------------------------
5925  !   PRINT RESULTS OF ANALYSIS OF INFLUENTIAL GROUPS
5930  !-------------------------------------------------------------------------
5935  FOR L=10 TO 1 STEP -1
5940     IF G(L,M+1)>0 THEN 5950                  ! DETERMINE HOW MANY GROUPS ARE INFLUENTIAL
5945     NEXT L
5950  !
5955  IF M=1 THEN WRITE #7,'-ANALYSIS OF INFLUENTIAL GROUPS:'
5960  IF M=1 THEN WRITE #7,'+_____'
5965  IF M>1 THEN WRITE #7,'-'
5970  WRITE #7,'0NUMBER OF OBSERVATIONS PER GROUP =':M
5975  WRITE #7 USING ' ## MOST INFLUENTIAL GROUPS SHOWN.',L
5980  WRITE #7,'0       OBSERVATIONS'
5985  WRITE #7,'        IN GROUP';TAB(22+4*M);' F STAT     LEVERAGE     INFLUENCE'
5990  WRITE #7
5995  FOR I=1 TO L
6000     F$,D$=' '
6005     IF G(I,M+2)>F1(M) THEN F$='*'
6010     IF G(I,M+1)>F2(M) THEN D$='#'
6015     WRITE #7 USING ' ##.    ',I;
6020     FOR J=1 TO M
6025        WRITE #7 USING ' ###',G(I,J);
6030        NEXT J
6035     WRITE #7,TAB(20+4*M);
6040     WRITE #7 USING FO$,G(I,M+2),F$,G(I,M+3),G(I,M+1),D$
6045     NEXT I
6050  WRITE #7,'-* : GREATER THAN ';100-A1;'
6055  WRITE #7,' # : GREATER THAN 5
6060  !
6065  NEXT M
6070  !
6075  CLOSE #7
6080  PRINT ''
6085  END
6090  !=========================================================================
6095  !   ROUTINE TO CALCULATE EIGENVALUES OF X'X   (COOLEY-LOHNES)
6100  !=========================================================================
6105  MAT E1=ZER(NO)                                ! MAT E1 = EIGENVALUES
6110  MAT E2=ZER(NO,NO)                             ! MAT E2 = EIGENVECTORS
6115  MAT E3=ZER(NO)                                ! MAT E3,E4 = TEMP STORAGE
6120  MAT E4=ZER(NO)
6125  !
6130  LO=1
6135  TO=0.00001
6140  IO=0
6145  MAT E4=CON
6150  IO=IO+1
6155  FOR J1=1 TO NO
6160     E3(J1)=0
6165     FOR J2=1 TO NO
6170        E3(J1)=E3(J1)+E0(J1,J2)*E4(J2)
6175        NEXT J2
6180     NEXT J1
6185  E1(LO)=E3(1)
6190  S1=0
6195  FOR J1=1 TO NO
6200     E2(J1,LO)=E3(J1)/E3(1)
6205     S1=S1+ABS(E4(J1)-E2(J1,LO))
6210     E4(J1)=E2(J1,LO)
6215     NEXT J1
6220  IF IO<>200 THEN 6240
6225    IF SO>S1 THEN 6240
6230      PRINT ' ERROR STOP IN EIGENVALUE ROUTINE.'
6235      STOP
6240  SO=S1
6245  IF S1>TO THEN 6150
6250  S1=0
6255  FOR J1=1 TO NO
6260     S1=S1+E2(J1,LO)^2
6265     NEXT J1
6270  S1=SQR(S1)
6275  FOR J1=1 TO NO
6280     E2(J1,LO)=E2(J1,LO)/S1
6285     NEXT J1
6290  FOR J1=1 TO NO
6295     FOR J2=1 TO NO
6300        E0(J1,J2)=E0(J1,J2)-E2(J1,LO)*E2(J2,LO)*E1(LO)
6305        NEXT J2
6310     NEXT J1
6315  IF LO>=NO THEN 6330
6320    LO=LO+1
6325    GOTO 6140
6330  !
6335  RETURN
```

```
1000  !   INFLUENTIAL OBSERVATIONS IN LINEAR REGRESSION
1005  !      Program 2 :  Restricted least squares regression
1010  !                     exact prior restrictions
1015  !                     stochastic prior information
1020  !
1025  !=============================================================================
1030  !  READ DATA FROM FILE AND INITIALIZE ARRAYS
1035  !=============================================================================
1040  INPUT 'ENTER NAME OF DATA FILE   ',F1$
1045  INPUT 'ENTER MAXIMUM SIZE OF SUBSETS TO BE CONSIDERED     ',MO
1050  !
1055  DEFINE FILE #1 = F1$
1060  READ #1,N,Q                                  ! N = NUMBER OF ROW OBSERVATIONS
1065                                               ! Q = NUMBER OF VARIABLES  (EXCLUDING CONSTANT)
1070  Q1=Q+1
1075  READ #1,A1,F0,F1(1),F1(2),F1(3),F2(1),F2(2),F2(3)
1080                                               ! READ IN CRITICAL VALUES
1085  MAT X=CON(N,Q1)
1090  MAT Y=ZER(N)
1095  MAT P=ZER(N)
1100  MAT P1=ZER(N)
1105  MAT R=ZER(N)
1110  MAT V=ZER(N)
1115  MAT D=ZER(N)
1120  MAT XO=ZER(Q1)
1125  MAT B=ZER(Q1)
1130  MAT B1=ZER(Q1)
1135  !
1140  FOR I=1 TO N
1145     READ* #1,Y(I)                             ! MAT Y = DEPENDENT VARIABLES
1150     FOR J=2 TO Q1
1155        READ* #1,X(I,J)                        ! MAT X = INDEPENDENT VARIABLES
1160        NEXT J
1165     NEXT I
1170  !
1175  READ #1,HO,O1                                ! HO = NUMBER OF CONSTRAINTS
1180                                               ! O1 = OPTION  (0=EXACT   1=STOCHASTIC)
1185  MAT H1=ZER(HO)
1190  MAT H2=ZER(HO,Q1)
1195  MAT RO=ZER(HO,HO)
1200  FOR I=1 TO HO
1205     READ* #1,H1(I)                            ! MAT H1 = h
1210     FOR J=1 TO Q1
1215        READ* #1,H2(I,J)                       ! MAT H2 = H
1220        NEXT J
1225     NEXT I
1230  IF O1=1 THEN MAT READ #1,RO                  ! MAT RO = VARIANCE STRUCTURE OF PRIOR INFORMATI(
1235  CLOSE #1
1240  !
1245  INPUT 'ENTER NAME OF OUTPUT FILE    ',F2$
1250  DEFINE FILE #7 = F2$                         ! NOMINATE NAME OF OUTPUT FILE
1255  WRITE #7,'1MULTIPLE LINEAR REGRESSION PACKAGE   ---   M.JACOBS'
1260  WRITE #7,'+_____________________________________________________'
1265  WRITE #7,'ONAME OF DATA FILE : ':F1$
1270  WRITE #7,'ONUMBER OF VECTOR OBSERVATIONS          : ':N
1275  WRITE #7,' NUMBER OF VARIABLES (EXCLUDING CONSTANT) : ':Q
1280  !=============================================================================
1285  !    CALCULATE REGRESSION COEFFICIENTS FOR STOCHASTIC PRIOR
1290  !=============================================================================
1295  FOR I=1 TO Q1
1300       FOR J=1 TO N
1305          XO(I)=XO(I) + X(J,I)*Y(J)            ! MAT XO = X'Y
1310          NEXT J
1315       NEXT I
1320  !
1325  MAT T9 = TRN(X)
1330  MAT X1 = T9*X                                ! MAT X1 = X'X
1335  !
1340  IF O1=0 THEN 1490
1345    MAT Z=INV(RO)                              ! MAT Z = R~
1350    MAT T9=TRN(H2)
1355    MAT Z1=T9*Z
1360    MAT Z2=Z1*H2
1365    MAT T9=X1+Z2
1370    MAT A=INV(T9)                              ! MAT A = (X'X + H'R~H)~
1375    MAT Z3=XO
1380    FOR I=1 TO Q1                              ! MAT Z3 = X'y + H'R~h
1385       FOR J=1 TO HO
1390          Z3(I)=Z3(I) + Z1(I,J)*H1(J)
1395          NEXT J
1400       NEXT I
1405    FOR I=1 TO Q1                              ! MAT B1 = RESTRICTED REGRESSION COEFFS
1410       FOR J=1 TO Q1
1415          B1(I)=B1(I) + A(I,J)*Z3(J)
1420          NEXT J
1425       NEXT I
1430    !
1435    MAT H3=ZER(HO)
1440    FOR I=1 TO HO
1445       FOR J=1 TO Q1
1450          H3(I)=H3(I) + H2(I,J)*B1(J)
1455          NEXT J
1460       NEXT I
1465    MAT H3=H1-H3                               ! MAT H3 = h - Hb
1470    GOTO 1630
```

```
1475  !=========================================================================
1480  !   CALCULATE REGRESSION COEFFICIENTS FOR EXACT PRIOR
1485  !=========================================================================
1490  MAT W = INV(X1)                                  ! MAT W = (X'X)~
1495  FOR I=1 TO Q1
1500     FOR J=1 TO Q1
1505        B(I)=B(I) + W(I,J)*XO(J)                  ! MAT B = O.L.S. REGRESSION COEFFS
1510        NEXT J
1515     NEXT I
1520  !
1525  MAT T9=TRN(H2)
1530  MAT WO=W*T9                                      ! MAT WO = WH'
1535  MAT W1=TRN(WO)                                   ! MAT W1 = HW
1540  MAT W2=H2*WO                                     ! MAT W2 = HWH'
1545  MAT ZO=INV(W2)                                   ! MAT ZO = (HWH')~ = Z
1550  MAT Z1=WO*ZO                                     ! MAT Z1 = WH'Z
1555  MAT H3=ZER(HO)
1560  !
1565  FOR I=1 TO HO
1570     FOR J=1 TO Q1
1575        H3(I)=H3(I) + H2(I,J)*B(J)
1580        NEXT J
1585     NEXT I
1590  MAT H3=H1-H3                                     ! MAT H3 = h - Hb
1595  !
1600  FOR I=1 TO Q1
1605     FOR J=1 TO HO
1610        B1(I)=B1(I) +Z1(I,J)*H3(J)
1615        NEXT J
1620     NEXT I
1625  MAT B1=B+B1                                      ! MAT B1 = RESTRICTED LEAST SQUARES ESTIMATE
1630  !
1635  WRITE #7,'-RESTRICTED REGRESSION COEFFICIENTS:'
1640  WRITE #7,'+_____'
1645  FOR I=1 TO Q1
1650     IF I=1 THEN L$='OCONSTANT' ELSE L$=' X'+STR$(I-1)
1655     WRITE #7 USING '##########   --------#.####',L$,B1(I)
1660     NEXT I
1665  !=========================================================================
1670  !   ANALYSIS OF RESIDUALS
1675  !=========================================================================
1680  FOR I=1 TO N
1685     FOR J=1 TO Q1
1690        P(I)=P(I) + X(I,J)*B1(J)                  ! MAT P = PREDICTORS OF Y
1695        P1(I)=P1(I) + X(I,J)*B(J)
1700        NEXT J
1705     NEXT I
1710  MAT R = Y - P                                    ! MAT R = RESTRICTED RESIDUALS
1715  MAT R1 = Y - P1                                  ! MAT R1 = LEAST SQUARES RESIDUALS
1720  !
1725  IF O1=1 THEN 1765
1730    MAT TO = Z1*W1
1735    MAT TO = W-TO                                  ! MAT TO = W - WH'ZHW
1740    MAT T=ZER(N)
1745    FOR I=1 TO N
1750       S2=S2+R1(I)^2                               ! S2 = SUM OF SQUARES
1755       NEXT I
1760    GOTO 1815
1765  !
1770  FOR I=1 TO N
1775     S2=S2 + R(I)^2                               ! S2 = SUM OF SQUARES
1780     NEXT I
1785  FOR I=1 TO HO
1790     FOR J=1 TO HO
1795        S2=S2 + H3(I)*H3(J)*Z(I,J)
1800        NEXT J
1805     NEXT I
1810  !
1815  FOR I=1 TO N                                     ! MAT V = DIAG ELEMENTS OF XWX'
1820     FOR J=1 TO Q1                                 ! MAT T = DIAG ELEMENTS OF X(W-WH'ZHW)X'
1825        FOR K=1 TO Q1
1830           IF O1=0 THEN V(I)=V(I) + X(I,J)*X(I,K)*W(J,K)
1835           IF O1=0 THEN T(I)=T(I) + X(I,J)*X(I,K)*TO(J,K)
1840           IF O1=1 THEN V(I)=V(I) + X(I,J)*X(I,K)*A(J,K)
1845           NEXT K
1850        NEXT J
1855     NEXT I
1860  IF O1=1 THEN MAT R1=R
1865  IF O1=1 THEN MAT T=V
```

```
1870  !=========================================================================
1875  !   ANALYSIS OF INFLUENTIAL GROUPS
1880  !=========================================================================
1885  !
1890  !-------------------------------------------------------------------------
1895  !   SET UP SUBMATRICES
1900  !-------------------------------------------------------------------------
1905  FOR M=1 TO MO                                  ! M = NUMBER OF OBSERVATIONS IN GROUP
1910  !
1915  C=0
1920  !
1925  MAT R8=ZER(M)                                  ! MAT R8,R9 ARE SUBMATRICES OF R,R1
1930  MAT R9=ZER(M)
1935  MAT X9=ZER(M,Q1)                               ! MAT X9,V9 ARE SUBMATRICES OF X,V
1940  MAT V9=ZER(M,M)
1945  MAT T1=ZER(M,M)                                ! MAT T1 = SUBMATRIX OF TO
1950  MAT I9=IDN(M,M)
1955  !
1960  FOR I=1 TO M                                   ! SET UP INDEX SET K
1965     K(I)=I
1970     NEXT I
1975  !
1980  MAT G=ZER(10,M+3)                              ! MAT G = RECORD OF TEN MOST INFLUENTIAL GROUPS
1985  !
1990  MAT V9=ZER
1995  MAT T1=ZER
2000  FOR I=1 TO M
2005     R8(I)=R(K(I))
2010     R9(I)=R1(K(I))
2015     FOR J=1 TO Q1
2020        X9(I,J)=X(K(I),J)
2025        NEXT J
2030     FOR J=I TO M
2035        FOR K=1 TO Q1                           ! COMPUTE ELEMENTS OF HAT MATRICES
2040           FOR L=1 TO Q1
2045              IF O1=0 THEN V9(I,J)=V9(I,J) + X(K(I),K)*X(K(J),L)*W(K,L)
2050              IF O1=0 THEN T1(I,J)=T1(I,J) + X(K(I),K)*X(K(J),L)*TO(K,L)
2055              IF O1=1 THEN V9(I,J)=V9(I,J) + X(K(I),K)*X(K(J),L)*A(K,L)
2060              NEXT L
2065           NEXT K
2070        IF I<>J THEN V9(J,I)=V9(I,J)
2075        IF I<>J AND O1=0 THEN T1(J,I)=T1(I,J)
2080        NEXT J
2085     NEXT I
2090  MAT T9 = I9-V9
2095  MAT T7 = INV(T9)                               ! MAT T7 = (I-Vi)^
2100  !
2105  S9=0
2110  FOR I=1 TO M                                   ! S9 = ESTIMATE OF VARIANCE
2115     FOR J=1 TO M
2120        S9=S9 + R9(I)*R9(J)*T7(I,J)
2125        NEXT J
2130     NEXT I
2135  IF O1=0 THEN S9 = (S2-S9)/(N-Q1-M)
2140  IF O1=1 THEN S9 = (S2-S9)/(N-Q1-M+HO)
```

```
2145  !-------------------------------------------------------------------------
2150  !   CALCULATE REGRESSION COEFFS FOR TRUNCATED SAMPLE  (EXACT PRIOR)
2155  !-------------------------------------------------------------------------
2160  IF O1=1 THEN 2405
2165  !
2170  MAT T9 = I9-T1
2175  MAT T2 = INV(T9)                          ! MAT T2 = (I-Ti)^
2180  !
2185  MAT T8 = X9*W0
2190  MAT T6 = TRN(T8)
2195  MAT T5 = T7*T8
2200  MAT T4 = T6*T5
2205  MAT T4 = W2 + T4
2210  MAT Z9 = INV(T4)                          ! MAT Z9 = Z(i)
2215  !
2220  MAT T9 = TRN(X9)
2225  MAT T8 = W*T9
2230  MAT T6 = T8*T5
2235  MAT T6 = W0 + T6
2240  MAT C = T6*Z9                             ! MAT C
2245  !
2250  MAT T9 = C*H2
2255  MAT T6 = IDN(Q1,Q1)
2260  MAT T6 = T6 - T9                          ! MAT T6 = I-CH
2265  MAT T9 = T8*T7
2270  MAT T8 = T6*T9
2275  !
2280  MAT B2 = B
2285  FOR I=1 TO Q1
2290     FOR J=1 TO H0
2295        B2(I)=B2(I) + C(I,J)*H3(J)
2300        NEXT J
2305     FOR J=1 TO M
2310        B2(I)=B2(I) - T8(I,J)*R9(J)
2315        NEXT J
2320     NEXT I
2325  MAT B2 = B2 - B1
2330  !
2335  T1=0                                      ! T1 = GENERALIZATION OF STUDENTIZED RESIDUAL
2340  FOR I=1 TO M
2345     FOR J=1 TO M
2350        T1=T1 + R8(I)*R8(J)*T2(I,J)
2355        NEXT J
2360     NEXT I
2365  !
2370  D1=0                                      ! D1 = INFLUENCE
2375  FOR I=1 TO Q1
2380     FOR J=1 TO Q1
2385        D1=D1 + B2(I)*B2(J)*X1(I,J)
2390        NEXT J
2395     NEXT I
2400  GOTO 2495
2405  !-------------------------------------------------------------------------
2410  !   INFLUENCE AND F STATISTIC FOR STOCHASTIC PRIOR
2415  !-------------------------------------------------------------------------
2420  MAT T9 = X*A
2425  MAT T8 = TRN(X9)
2430  MAT T6 = T9*T8
2435  MAT T9 = T6*T7                            ! MAT T9 = XAXi'Nm^
2440  MAT T8 = TRN(T9)
2445  MAT T6 = T8*T9
2450  !
2455  D1,T1=0
2460  FOR I=1 TO M
2465     FOR J=1 TO M
2470        D1=D1 + R9(I)*R9(J)*T6(I,J)
2475        T1=T1 + R9(I)*R9(J)*T7(I,J)
2480        NEXT J
2485     NEXT I
2490  !
2495  D1=D1/(Q*S9)
2500  T1=T1/(M*S9)
2505  IF M=1 THEN D(K(1))=D1                    ! STORE D1
2510  C=C+1
```

```
2515  !-------------------------------------------------------------------------
2520  !   CHECK FOR POSITION IN 'TOP TEN' GROUP
2525  !-------------------------------------------------------------------------
2530  IF D1<=G(10,M+1) THEN 2680
2535    FOR I=9 TO 1 STEP -1
2540       IF D1<G(I,M+1) THEN 2550
2545       NEXT I
2550    L=I+1
2555    FOR I=10 TO L+1 STEP -1                    ! INCORPORATE INTO 'TOP TEN' GROUP
2560       FOR J=1 TO M+3
2565          G(I,J)=G(I-1,J)
2570          NEXT J
2575       NEXT I
2580    FOR J=1 TO M
2585       G(L,J)=K(J)
2590       NEXT J
2595    G(L,M+1)=D1
2600    G(L,M+2)=T1
2605    !
2610    IF O1=1 THEN MAT T1=V9
2615    IF M>1 THEN 2630                           ! ROUTINE TO CALCULATE LEVERAGE
2620      G(L,M+3)=T1(1,1)/(1-T1(1,1))
2625      GOTO 2675
2630    MAT E0=T1
2635    N0=M
2640    GOSUB 3070                                 ! EIGENVALUES OF T1
2645    L1=0
2650    FOR J=1 TO M
2655       L1=L1 + E1(J)/(1-E1(J))
2660       NEXT J
2665    G(L,M+3)=L1
2670  !-------------------------------------------------------------------------
2675  !   ADVANCE INDEX SET
2680  !-------------------------------------------------------------------------
2685  U=M
2690  K(U)=K(U)+1
2695  IF K(U)<=N-M+U THEN 2710
2700    U=U-1
2705    IF U>0 THEN 2690 ELSE 2735
2710  IF U=M THEN 1985
2715  U=U+1
2720  K(U)=K(U-1)
2725  GOTO 2690
2730  !-------------------------------------------------------------------------
2735  !   PRINT RESULTS OF RESIDUAL ANALYSIS
2740  !-------------------------------------------------------------------------
2745  IF M>1 THEN 2900
2750  WRITE #7,'1ANALYSIS OF RESIDUALS:'
2755  WRITE #7,'+_____'
2760  WRITE #7,'0';TAB(48);'STUDENT'
2765  WRITE #7,'          OBSERVED   PREDICTED    RESIDUAL    RESIDUAL    LEVERAGE    INFLUENCE'
2770  WRITE #7
2775  F1$=' ###.    -----#.##    -----#.##    -----#.##    -----#.## # -----#.##    -----#.## #'
2780  !
2785  M9=0
2790  FOR I=1 TO N
2795     F$,D$=' '
2800     IF O1=0 THEN S9=(S2-R1(I)^2)/(1-V(I))) / (N-Q1-1)    ! S9 = S(I)
2805     IF O1=1 THEN S9=(S2-R(I)^2/(1-V(I))) / (N-Q1-1+H0)
2810     T = R(I)/SQR((1-T(I))*S9)                  ! T = STUDENTIZED RESIDUAL
2815     F=T^2
2820     IF F>F1(1) THEN F$='*'
2825     L=T(I)/(1-T(I))                            ! L = LEVERAGE
2830     D1=D(I)
2835     IF D1>F2(M) THEN D$='#'
2840     IF ABS(T)<=M9 THEN 2855
2845       M9=ABS(T)                                ! M9 = MAXIMUM STUDENTIZED RESIDUAL
2850       L9=I                                     ! L9 = ITS LOCATION
2855     WRITE #7 USING F1$,I,Y(I),P(I),R(I),T,F$,L,D1,D$
2860     NEXT I
2865  WRITE #7 USING '-MAX T =###.## CORRESPONDING TO OBSERVATION ###',M9,L9
2870  F=M9^2
2875  WRITE #7 USING 'OF =###.##',F
2880  WRITE #7,' CRITICAL POINT AT ';A1/100;'/';N;' LEVEL = ';F0
2885  WRITE #7,'-* : GREATER THAN ';100-A1;'
2890  WRITE #7,' # : GREATER THAN 5
```

```
2895  !-------------------------------------------------------------------------
2900  !   PRINT RESULTS OF ANALYSIS OF INFLUENTIAL GROUPS
2905  !-------------------------------------------------------------------------
2910  FOR L=10 TO 1 STEP -1
2915     IF G(L,M+1)>0 THEN 2925                  ! DETERMINE HOW MANY GROUPS ARE INFLUENTIAL
2920     NEXT L
2925  !
2930  WRITE #7,'1ANALYSIS OF INFLUENTIAL GROUPS:'
2935  WRITE #7,'+_____________________________'
2940  WRITE #7,'ONUMBER OF OBSERVATIONS PER GROUP =':M
2945  WRITE #7 USING ' ## MOST INFLUENTIAL GROUPS SHOWN.',L
2950  WRITE #7,'ONUMBER OF SUBSETS CONSIDERED  =':C
2955  FO$='----#.## #  ----#.##    ----#.## #'
2960  WRITE #7,'O       OBSERVATIONS'
2965  WRITE #7,'        IN GROUP';TAB(22+4*M);' F STAT     LEVERAGE     INFLUENCE'
2970  WRITE #7
2975  FOR I=1 TO L
2980     F$,D$=' '
2985     IF G(I,M+2)>F1(M) THEN F$='*'
2990     IF G(I,M+1)>F2(M) THEN D$='#'
2995     WRITE #7 USING ' ##.    ',I;
3000     FOR J=1 TO M
3005        WRITE #7 USING ' ###',G(I,J);
3010        NEXT J
3015     WRITE #7,TAB(20+4*M);
3020     WRITE #7 USING FO$,G(I,M+2),F$,G(I,M+3),G(I,M+1),D$
3025     NEXT I
3030  WRITE #7,'-# : GREATER THAN ';100-A1;'
3035  WRITE #7,' # : GREATER THAN 5
3040  !
3045  NEXT M
3050  !
3055  CLOSE #7
3060  PRINT ''
3065  END
3070  !=========================================================================
3075  !   ROUTINE TO CALCULATE EIGENVALUES OF X'X   (COOLEY-LOHNES)
3080  !=========================================================================
3085  MAT E1=ZER(NO)                                 ! MAT E1 = EIGENVALUES
3090  MAT E2=ZER(NO,NO)                              ! MAT E2 = EIGENVECTORS
3095  MAT E3=ZER(NO)                                 ! MAT E3,E4 = TEMP STORAGE
3100  MAT E4=ZER(NO)
3105  !
3110  LO=1
3115  TO=0.00001
3120  IO=0
3125  MAT E4=CON
3130  IO=IO+1
3135  FOR J1=1 TO NO
3140     E3(J1)=0
3145     FOR J2=1 TO NO
3150        E3(J1)=E3(J1)+EO(J1,J2)*E4(J2)
3155        NEXT J2
3160     NEXT J1
3165  E1(LO)=E3(1)
3170  S1=0
3175  FOR J1=1 TO NO
3180     E2(J1,LO)=E3(J1)/E3(1)
3185     S1=S1+ABS(E4(J1)-E2(J1,LO))
3190     E4(J1)=E2(J1,LO)
3195     NEXT J1
3200  IF IO<>50 THEN 3220
3205    IF SO>S1 THEN 3220
3210      PRINT ' ERROR STOP IN EIGENVALUE ROUTINE.'
3215      STOP
3220  SO=S1
3225  IF S1>TO THEN 3130
3230  S1=0
3235  FOR J1=1 TO NO
3240     S1=S1+E2(J1,LO)^2
3245     NEXT J1
3250  S1=SQR(S1)
3255  FOR J1=1 TO NO
3260     E2(J1,LO)=E2(J1,LO)/S1
3265     NEXT J1
3270  FOR J1=1 TO NO
3275     FOR J2=1 TO NO
3280        EO(J1,J2)=EO(J1,J2)-E2(J1,LO)*E2(J2,LO)*E1(LO)
3285        NEXT J2
3290     NEXT J1
3295  IF LO>=NO THEN 3310
3300    LO=LO+1
3305    GOTO 3120
3310  !
3315  RETURN
```

Appendix C

SAMPLE COMPUTER PRINTOUT

A sample of the output of the computer programs is given in this appendix. The analysis shown is for the simulation study of chapter 12. The source data are listed in appendix D.

MULTIPLE LINEAR REGRESSION PACKAGE --- M. JACOBS

NAME OF DATA FILE : SIMDATA

NUMBER OF VECTOR OBSERVATIONS : 40
NUMBER OF VARIABLES (EXCLUDING CONSTANT) : 5

CORRELATION COEFFICIENTS:

	Y	X1	X2	X3	X4	X5
Y	1.00	.34	-.02	.82	.81	.91
X1	.34	1.00	-.38	-.05	.05	.00
X2	-.02	-.38	1.00	-.09	-.05	-.09
X3	.82	-.05	-.09	1.00	.60	.93
X4	.81	.05	-.05	.60	1.00	.85
X5	.91	.00	-.09	.93	.85	1.00

REGRESSION COEFFICIENTS:

CONSTANT	113.4729
X1	2.0391
X2	9.9340
X3	22.8808
X4	24.0325
X5	-4.0765

UNCORRECTED R SQUARED : .989
CORRECTED R SQUARED : .988

DET OF CORRELATION : 5.693E-04
CONDITION NUMBER : 3961.9

EIGENVALUES OF X'X

2.611	1.374	.637	.378	.001

ANALYSIS OF RESIDUALS:

	OBSERVED	PREDICTED	RESIDUAL	STANDARD RESIDUAL	STUDENT RESIDUAL	LEVERAGE	COOK DISTANCE	INFLUENCE
1.	587.23	588.94	-1.71	-0.37	-0.37	0.78	0.02	0.02
2.	715.93	725.51	-9.58	-1.98	-2.08 *	0.62	0.41	0.45 #
3.	766.81	755.55	11.25	2.32	2.49 *	0.61	0.55	0.63 #
4.	527.18	510.11	17.07	3.18	3.75 *	0.31	0.53	0.73 #
5.	512.06	510.88	1.17	0.20	0.20	0.13	0.00	0.00
6.	556.93	559.68	-2.75	-0.50	-0.50	0.27	0.01	0.01
7.	629.81	629.86	-0.05	-0.01	-0.01	0.12	0.00	0.00
8.	508.18	506.73	1.45	0.26	0.26	0.23	0.00	0.00
9.	569.06	577.68	-8.63	-1.45	-1.47	0.06	0.02	0.02
10.	541.93	546.04	-4.11	-0.71	-0.71	0.13	0.01	0.01
11.	558.81	562.42	-3.62	-0.60	-0.59	0.04	0.00	0.00
12.	562.18	563.75	-1.57	-0.26	-0.26	0.04	0.00	0.00
13.	616.06	609.20	6.86	1.14	1.14	0.04	0.01	0.01
14.	658.43	660.73	-2.30	-0.39	-0.38	0.07	0.00	0.00
15.	614.31	613.95	0.36	0.06	0.06	0.02	0.00	0.00
16.	549.68	554.61	-4.93	-0.83	-0.82	0.06	0.01	0.01
17.	545.06	549.45	-4.39	-0.78	-0.77	0.19	0.02	0.02
18.	579.93	582.32	-2.39	-0.40	-0.40	0.07	0.00	0.00
19.	612.81	616.43	-3.62	-0.65	-0.64	0.20	0.01	0.01
20.	571.18	577.31	-6.13	-1.04	-1.04	0.08	0.01	0.01
21.	584.56	583.75	0.81	0.14	0.13	0.06	0.00	0.00
22.	586.93	580.48	6.45	1.07	1.07	0.04	0.01	0.01
23.	568.81	566.99	1.82	0.34	0.33	0.29	0.01	0.01
24.	638.18	639.16	-0.98	-0.18	-0.18	0.25	0.00	0.00
25.	670.56	669.76	0.79	0.13	0.13	0.08	0.00	0.00
26.	589.43	591.24	-1.80	-0.31	-0.30	0.08	0.00	0.00
27.	585.81	587.47	-1.66	-0.29	-0.28	0.12	0.00	0.00
28.	630.68	628.98	1.70	0.28	0.28	0.05	0.00	0.00
29.	643.56	639.98	3.58	0.60	0.59	0.05	0.00	0.00
30.	628.93	618.27	10.66	1.81	1.87	0.08	0.05	0.05
31.	645.31	639.80	5.51	0.92	0.91	0.05	0.01	0.01
32.	602.68	597.83	4.85	0.82	0.81	0.07	0.01	0.01
33.	661.06	657.25	3.80	0.66	0.65	0.13	0.01	0.01
34.	651.93	665.19	-13.26	-2.32	-2.50 *	0.16	0.14	0.17
35.	623.81	629.15	-5.34	-0.90	-0.90	0.07	0.01	0.01
36.	619.68	622.23	-2.55	-0.46	-0.45	0.22	0.01	0.01
37.	659.06	657.89	1.17	0.20	0.20	0.09	0.00	0.00
38.	706.43	702.91	3.52	0.62	0.62	0.19	0.01	0.01
39.	627.81	631.97	-4.16	-0.72	-0.72	0.14	0.01	0.01
40.	642.18	639.48	2.70	0.47	0.47	0.15	0.01	0.01

MAX T = 3.75 CORRESPONDING TO OBSERVATION 4

F = 14.03
CRITICAL POINT AT .05/40 LEVEL = 12.46

* : GREATER THAN 95% POINT OF F(0,34)
\# : GREATER THAN 5% POINT OF F(6,34)

ANALYSIS OF INFLUENTIAL GROUPS:

NUMBER OF OBSERVATIONS PER GROUP = 1
10 MOST INFLUENTIAL GROUPS SHOWN.

NUMBER OF SUBSETS CONSIDERED = 40
EXACT COMPUTATIONS OF INFLUENCE = 40

	OBSERVATIONS IN GROUP	F STAT	LEVERAGE	INFLUENCE
1.	4	14.03 *	0.31	0.73 #
2.	3	6.22 *	0.61	0.63 #
3.	2	4.32 *	0.62	0.45 #
4.	34	6.23 *	0.16	0.17
5.	30	3.51	0.08	0.05
6.	9	2.17	0.06	0.02
7.	17	0.60	0.19	0.02
8.	1	0.13	0.78	0.02
9.	20	1.08	0.08	0.01
10.	19	0.41	0.20	0.01

* : GREATER THAN 95% POINT OF F(1,33)
# : GREATER THAN 5% POINT OF F(6,33)

ANALYSIS OF INFLUENTIAL GROUPS:

NUMBER OF OBSERVATIONS PER GROUP = 2
10 MOST INFLUENTIAL GROUPS SHOWN.

NUMBER OF SUBSETS CONSIDERED = 780
EXACT COMPUTATIONS OF INFLUENCE = 111

	OBSERVATIONS IN GROUP	F STAT	LEVERAGE	INFLUENCE
1.	2 4	13.28 *	0.96	2.42 #
2.	4 34	14.22 *	0.47	1.40 #
3.	4 19	7.83 *	0.55	1.02 #
4.	4 39	7.80 *	0.47	0.95 #
5.	3 34	7.41 *	0.77	0.94 #
6.	4 35	8.08 *	0.40	0.93 #
7.	4 30	10.51 *	0.40	0.92 #
8.	4 5	7.15 *	0.48	0.87 #
9.	1 4	6.90 *	1.24	0.87 #
10.	3 13	4.28 *	0.66	0.84 #

* : GREATER THAN 95% POINT OF F(2,32)
# : GREATER THAN 5% POINT OF F(6,32)

RIDGE REGRESSION:

REGRESSION COEFFICIENTS:

	k=0.00	k=0.10	k=0.20	k=0.30	k=0.40	k=0.50
CONSTANT	113.4729	145.3630	169.8642	191.0619	209.7224	226.3714
X1	2.0391	1.7557	1.5601	1.4064	1.2819	1.1787
X2	9.9340	8.0363	6.6622	5.6282	4.8260	4.1886
X3	22.8808	10.5759	9.9113	9.4004	8.9767	8.6109
X4	24.0325	12.5170	12.4030	12.1880	11.9213	11.6318
X5	-4.0765	7.0512	6.7952	6.5456	6.3140	6.0996

SUMMARY RIDGE TRACE FOR M = 1

OBSERVATIONS	k =	.00	.10	.20	.30	.40	.50
3	Influence	.630	.942	1.318	1.684	2.029	2.344
	Leverage	.607	.538	.488	.445	.408	.376
	F stat	6.220	11.095	17.824	25.758	34.680	44.380
4	Influence	.733	.204	.085	.028	.004	.000
	Leverage	.314	.149	.132	.118	.107	.098
	F stat	14.027	8.389	4.017	1.486	.265	.015
2	Influence	.448	.364	.244	.139	.063	.019
	Leverage	.621	.529	.461	.410	.368	.334
	F stat	4.324	4.482	3.605	2.370	1.224	.406
38	Influence	.012	.034	.072	.111	.146	.177
	Leverage	.186	.132	.113	.100	.089	.080
	F stat	.382	1.601	3.972	6.988	10.422	14.118
6	Influence	.011	.062	.118	.153	.170	.177
	Leverage	.275	.176	.135	.106	.086	.071
	F stat	.249	2.225	5.689	9.499	13.236	16.735
8	Influence	.003	.025	.065	.104	.138	.166
	Leverage	.230	.096	.081	.070	.062	.055
	F stat	.067	1.600	4.999	9.279	13.988	18.881
34	Influence	.165	.019	.000	.004	.012	.019
	Leverage	.159	.105	.078	.060	.047	.038
	F stat	6.232	1.146	.008	.450	1.624	3.128
5	Influence	.001	.017	.054	.091	.122	.148
	Leverage	.127	.101	.084	.071	.062	.055
	F stat	.040	1.057	4.040	7.963	12.335	16.902
17	Influence	.019	.039	.056	.069	.077	.083
	Leverage	.188	.134	.102	.081	.066	.056
	F stat	.600	1.834	3.547	5.509	7.588	9.706
40	Influence	.005	.024	.044	.057	.063	.065
	Leverage	.147	.092	.072	.057	.047	.040
	F stat	.217	1.634	3.886	6.222	8.417	10.404

RIDGE TRACE OF INFLUENCE FOR M = 1

OBSERVATIONS		+.00	+.01	+.02	+.03	+.04	+.05	+.06	+.07	+.08	+.09
3	k=.00	.630	.614	.649	.684	.720	.757	.793	.830	.867	.905
	k=.10	.942	.979	1.017	1.055	1.092	1.130	1.168	1.205	1.243	1.280
	k=.20	1.318	1.355	1.392	1.429	1.466	1.503	1.540	1.576	1.612	1.649
	k=.30	1.684	1.720	1.755	1.791	1.825	1.860	1.894	1.928	1.962	1.996
	k=.40	2.029	2.062	2.095	2.127	2.159	2.190	2.222	2.253	2.284	2.314
	k=.50	2.344									
4	k=.00	.733	.413	.382	.353	.327	.303	.280	.259	.239	.221
	k=.10	.204	.188	.173	.159	.146	.134	.123	.112	.103	.094
	k=.20	.085	.077	.070	.063	.057	.051	.046	.041	.036	.032
	k=.30	.028	.024	.021	.018	.016	.013	.011	.009	.007	.006
	k=.40	.004	.003	.002	.002	.001	.001	.000	.000	.000	.000
	k=.50	.000									
2	k=.00	.448	.439	.434	.429	.422	.414	.406	.396	.386	.375
	k=.10	.364	.353	.341	.329	.317	.305	.292	.280	.268	.256
	k=.20	.244	.233	.221	.210	.199	.188	.178	.167	.157	.148
	k=.30	.139	.130	.121	.112	.104	.097	.089	.082	.076	.069
	k=.40	.063	.057	.052	.047	.042	.037	.033	.029	.025	.022
	k=.50	.019									
38	k=.00	.012	.008	.010	.012	.015	.018	.021	.024	.027	.031
	k=.10	.034	.038	.042	.045	.049	.053	.057	.061	.064	.068
	k=.20	.072	.076	.080	.084	.088	.092	.096	.099	.103	.107
	k=.30	.111	.114	.118	.122	.125	.129	.132	.136	.139	.143
	k=.40	.146	.149	.153	.156	.159	.162	.165	.168	.171	.174
	k=.50	.177									
6	k=.00	.011	.008	.012	.018	.023	.029	.036	.042	.049	.055

SUMMARY RIDGE TRACE FOR M = 2

OBSERVATIONS		k =	.00	.10	.20	.30	.40	.50
2	3	Influence	.134	.466	1.231	2.259	3.400	4.549
		Leverage	2.330	2.050	1.823	1.627	1.460	1.315
		F stat	3.625	5.718	8.933	13.426	19.222	26.181
3	8	Influence	.605	1.278	1.994	2.687	3.326	3.897
		Leverage	.863	.667	.600	.545	.497	.456
		F stat	3.017	6.887	12.797	19.905	27.802	36.220
3	5	Influence	.615	1.171	1.826	2.461	3.048	3.576
		Leverage	.763	.664	.595	.539	.491	.450
		F stat	3.016	6.410	11.965	18.726	26.273	34.336
3	4	Influence	.723	.843	1.416	2.093	2.793	3.473
		Leverage	1.011	.737	.665	.604	.552	.507
		F stat	10.376	10.738	13.070	17.452	23.387	30.503
3	38	Influence	.727	1.159	1.749	2.346	2.917	3.443
		Leverage	.808	.680	.613	.557	.509	.467
		F stat	3.438	6.721	11.742	17.821	24.703	32.178
3	16	Influence	.835	1.280	1.794	2.280	2.726	3.127
		Leverage	.698	.607	.545	.494	.450	.413
		F stat	3.840	7.348	12.040	17.411	23.301	29.579
3	17	Influence	.702	1.125	1.633	2.130	2.595	3.016
		Leverage	.799	.678	.597	.533	.482	.439
		F stat	3.531	6.791	11.323	16.657	22.595	28.975
3	25	Influence	.655	1.143	1.664	2.158	2.610	3.014
		Leverage	.709	.608	.546	.494	.451	.413
		F stat	3.090	6.499	11.211	16.695	22.720	29.122
3	14	Influence	.598	1.037	1.541	2.034	2.493	2.909
		Leverage	.726	.643	.583	.531	.486	.446
		F stat	3.019	5.546	9.236	13.692	18.750	24.267
3	10	Influence	.738	1.101	1.593	2.066	2.504	2.899
		Leverage	.750	.597	.536	.485	.442	.405
		F stat	3.542	6.940	12.036	17.929	24.364	31.168

RIDGE TRACE OF INFLUENCE FOR M = 2

OBSERVATIONS			+.00	+.01	+.02	+.03	+.04	+.05	+.06	+.07	+.08	+.09
2	3	k=.00	.134	.125	.145	.169	.198	.232	.270	.312	.359	.410
		k=.10	.466	.526	.590	.658	.730	.805	.884	.967	1.052	1.140
		k=.20	1.231	1.325	1.421	1.520	1.620	1.723	1.827	1.933	2.040	2.149
		k=.30	2.259	2.370	2.482	2.595	2.709	2.823	2.938	3.053	3.169	3.285
		k=.40	3.400	3.516	3.632	3.748	3.863	3.979	4.094	4.208	4.322	4.436
		k=.50	4.549									
3	8	k=.00	.605	.670	.733	.798	.864	.931	.999	1.068	1.137	1.208
		k=.10	1.278	1.349	1.421	1.492	1.564	1.636	1.708	1.779	1.851	1.923
		k=.20	1.994	2.065	2.136	2.206	2.276	2.346	2.415	2.484	2.552	2.620
		k=.30	2.687	2.753	2.819	2.885	2.950	3.014	3.078	3.141	3.203	3.265
		k=.40	3.326	3.386	3.446	3.504	3.563	3.620	3.677	3.733	3.788	3.843
		k=.50	3.897									
3	5	k=.00	.615	.619	.675	.733	.793	.854	.916	.978	1.042	1.106
		k=.10	1.171	1.236	1.301	1.367	1.432	1.498	1.564	1.629	1.695	1.761
		k=.20	1.826	1.891	1.956	2.020	2.084	2.148	2.212	2.275	2.337	2.399
		k=.30	2.461	2.522	2.583	2.643	2.703	2.762	2.820	2.878	2.935	2.992
		k=.40	3.048	3.104	3.159	3.213	3.267	3.320	3.372	3.424	3.475	3.526
		k=.50	3.576									
3	4	k=.00	.723	.545	.561	.583	.609	.640	.674	.712	.753	.797
		k=.10	.843	.892	.944	.997	1.052	1.109	1.168	1.228	1.290	1.352
		k=.20	1.416	1.481	1.546	1.613	1.680	1.747	1.816	1.884	1.954	2.023
		k=.30	2.093	2.163	2.233	2.303	2.373	2.443	2.514	2.584	2.654	2.724
		k=.40	2.793	2.863	2.932	3.001	3.070	3.138	3.206	3.273	3.340	3.407
		k=.50	3.473									
3	38	k=.00	.727	.681	.728	.778	.830	.882	.936	.991	1.046	1.102
		k=.10	1.159	1.217	1.275	1.333	1.392	1.451	1.510	1.570	1.629	1.689
		k=.20	1.749	1.809	1.869	1.929	1.989	2.049	2.109	2.168	2.228	2.287
		k=.30	2.346	2.404	2.463	2.521	2.578	2.636	2.693	2.749	2.806	2.861
		k=.40	2.917	2.971	3.026	3.080	3.133	3.186	3.239	3.291	3.342	3.393
		k=.50	3.443									

GENERALIZED INVERSE REGRESSION:

NUMBER OF EIGENVALUES ELIMINATED = 1

REGRESSION COEFFICIENTS:

CONSTANT	116.4467
X1	2.0146
X2	9.9394
X3	11.3590
X4	12.2315
X5	7.4652

ANALYSIS OF INFLUENTIAL GROUPS:

NUMBER OF OBSERVATIONS PER GROUP = 1
10 MOST INFLUENTIAL GROUPS SHOWN.

	OBSERVATIONS IN GROUP	F STAT	LEVERAGE	INFLUENCE
1.	3	5.77 *	0.60	0.58 #
2.	4	15.52 *	0.17	0.44 #
3.	2	4.26 *	0.62	0.44 #
4.	34	5.78 *	0.15	0.15
5.	30	4.06	0.06	0.04
6.	1	0.24	0.75	0.03
7.	17	0.58	0.19	0.02
8.	19	0.72	0.14	0.02
9.	33	0.56	0.11	0.01
10.	36	0.32	0.20	0.01

* : GREATER THAN 95% POINT OF F(1,33)
# : GREATER THAN 5% POINT OF F(6,33)

NUMBER OF OBSERVATIONS PER GROUP = 2
10 MOST INFLUENTIAL GROUPS SHOWN.

	OBSERVATIONS IN GROUP	F STAT	LEVERAGE	INFLUENCE
1.	2 4	14.20 *	0.82	2.02 #
2.	4 34	15.18 *	0.33	1.15 #
3.	3 34	6.83 *	0.75	0.83 #
4.	3 36	3.29	0.85	0.76 #
5.	3 16	3.54 *	0.69	0.76 #
6.	3 35	3.77 *	0.68	0.75 #
7.	3 13	3.79 *	0.63	0.73 #
8.	4 39	8.56 *	0.33	0.70 #
9.	3 26	3.05	0.73	0.68 #
10.	2 32	2.97	0.73	0.68 #

* : GREATER THAN 95% POINT OF F(2,32)
# : GREATER THAN 5% POINT OF F(6,32)

MULTIPLE LINEAR REGRESSION PACKAGE --- M. JACOBS

NAME OF DATA FILE : SIMDATA

NUMBER OF VECTOR OBSERVATIONS : 40
NUMBER OF VARIABLES (EXCLUDING CONSTANT) : 5

RESTRICTED REGRESSION COEFFICIENTS:

CONSTANT	114.4882
X1	2.0305
X2	9.9364
X3	18.8139
X4	19.8649
X5	0.0000

ANALYSIS OF RESIDUALS:

	OBSERVED	PREDICTED	RESIDUAL	STUDENT RESIDUAL	LEVERAGE	INFLUENCE
1.	587.23	589.14	-1.91	-0.42	0.83	0.03
2.	715.93	725.49	-9.56	-2.13 *	0.69	0.62 #
3.	766.81	755.71	11.10	2.51 *	0.66	0.84 #
4.	527.18	509.42	17.76	3.76 *	0.20	0.58 #
5.	512.06	510.86	1.20	0.21	0.16	0.00
6.	556.93	559.33	-2.40	-0.44	0.28	0.01
7.	629.81	630.31	-0.50	-0.08	0.10	0.00
8.	508.18	507.37	0.81	0.14	0.15	0.00
9.	569.06	577.29	-8.23	-1.40	0.06	0.02
10.	541.93	545.54	-3.61	-0.61	0.10	0.01
11.	558.81	562.43	-3.62	-0.60	0.07	0.00
12.	562.18	563.87	-1.69	-0.28	0.07	0.00
13.	616.06	609.49	6.57	1.10	0.04	0.01
14.	658.43	660.60	-2.16	-0.36	0.09	0.00
15.	614.31	613.94	0.36	0.06	0.05	0.00
16.	549.68	554.52	-4.84	-0.82	0.08	0.01
17.	545.06	549.42	-4.36	-0.78	0.22	0.03
18.	579.93	582.76	-2.83	-0.47	0.05	0.00
19.	612.81	616.89	-4.08	-0.71	0.18	0.02
20.	571.18	576.78	-5.60	-0.93	0.05	0.01
21.	584.56	584.25	0.31	0.05	0.04	0.00
22.	586.93	580.72	6.21	1.04	0.05	0.01
23.	568.81	567.79	1.02	0.17	0.14	0.00
24.	638.18	638.26	-0.07	-0.01	0.07	0.00
25.	670.56	669.49	1.06	0.18	0.09	0.00
26.	589.43	591.24	-1.81	-0.31	0.11	0.00
27.	585.81	587.72	-1.92	-0.33	0.14	0.00
28.	630.68	628.90	1.78	0.30	0.07	0.00
29.	643.56	639.75	3.80	0.63	0.07	0.01
30.	628.93	617.94	10.99	1.94	0.09	0.07
31.	645.31	639.78	5.53	0.93	0.07	0.01
32.	602.68	597.59	5.10	0.86	0.09	0.01
33.	661.06	657.05	4.01	0.69	0.15	0.01
34.	651.93	665.04	-13.10	-2.50 *	0.19	0.23
35.	623.81	629.39	-5.58	-0.95	0.09	0.02
36.	619.68	622.47	-2.79	-0.50	0.24	0.01
37.	659.06	657.63	1.42	0.24	0.10	0.00
38.	706.43	703.24	3.19	0.56	0.19	0.01
39.	627.81	631.76	-3.96	-0.69	0.16	0.02
40.	642.18	639.78	2.41	0.42	0.16	0.01

MAX T = 3.76 CORRESPONDING TO OBSERVATION 4

F = 14.13
CRITICAL POINT AT .05/40 LEVEL = 12.46

* : GREATER THAN 95% POINT OF F(1,33)
# : GREATER THAN 5% POINT OF F(6,33)

ANALYSIS OF INFLUENTIAL GROUPS:

NUMBER OF OBSERVATIONS PER GROUP = 1
10 MOST INFLUENTIAL GROUPS SHOWN.

NUMBER OF SUBSETS CONSIDERED = 40

	OBSERVATIONS IN GROUP	F STAT	LEVERAGE	INFLUENCE
1.	3	6.31 *	0.66	0.84 #
2.	2	4.52 *	0.69	0.62 #
3.	4	14.13 *	0.20	0.58 #
4.	34	6.27 *	0.19	0.23
5.	30	3.76	0.09	0.07
6.	1	0.17	0.83	0.03
7.	17	0.61	0.22	0.03
8.	9	1.97	0.06	0.02
9.	19	0.51	0.18	0.02
10.	35	0.90	0.09	0.02

* : GREATER THAN 95% POINT OF F(1,33)
# : GREATER THAN 5% POINT OF F(6,33)

ANALYSIS OF INFLUENTIAL GROUPS:

NUMBER OF OBSERVATIONS PER GROUP = 2
10 MOST INFLUENTIAL GROUPS SHOWN.

NUMBER OF SUBSETS CONSIDERED = 780

	OBSERVATIONS IN GROUP	F STAT	LEVERAGE	INFLUENCE
1.	2 4	12.92 *	0.90	2.24 #
2.	4 34	14.11 *	0.39	1.28 #
3.	3 13	4.41 *	0.74	1.18 #
4.	3 4	11.41 *	0.90	1.05 #
5.	3 34	7.14 *	0.85	1.04 #
6.	3 16	3.77 *	0.76	1.01 #
7.	3 38	3.51 *	0.88	1.00 #
8.	3 33	3.60 *	0.83	0.99 #
9.	3 36	3.43 *	0.93	0.99 #
10.	3 35	3.87 *	0.76	0.98 #

* : GREATER THAN 95% POINT OF F(2,32)
# : GREATER THAN 5% POINT OF F(6,32)

Appendix D

DATA SETS

Listings of the source data used in the simulation study of chapter 12 and the examples of chapter 13 are given in this appendix.

SIMULATION DATA:

Y = 100 + (2 * X1) + (10 * X2) + (10 * X3) + (10 * X4) + (10 * X5) + E
where E distributed N(0,25)
X1 = trend variable
X2 distributed N(5,1)
X3 distributed N(10,1)
X4 distributed N(10,1)
X5 = X3 + X4 + E where E distributed N(0,0.1)

OUTLIERS INSERTED:
Observation 1 : X2=10
Observation 2 : X3,X4 centred on 15
Y = Y-10
Observation 3 : X3,X4 centred on 15
Observation 4 : Y = Y+20

	Y	X1	X2	X3	X4	X5
1.	587.231	1	10.000	8.808	10.448	19.266
2.	715.932	2	4.370	15.011	14.151	29.192
3.	766.807	3	5.573	16.714	13.355	30.119
4.	527.182	4	6.776	7.417	9.058	16.245
5.	512.057	5	4.980	8.120	9.261	17.322
6.	556.932	6	3.183	11.323	9.464	20.648
7.	629.807	7	6.386	12.026	10.667	22.775
8.	508.182	8	4.589	8.230	8.870	17.201
9.	569.057	9	5.792	9.933	10.073	19.878
10.	541.932	10	4.495	9.136	9.776	18.755
11.	558.807	11	6.198	9.339	9.480	18.781
12.	562.182	12	5.401	10.042	9.183	19.208
13.	616.057	13	5.105	11.245	10.386	21.684
14.	658.432	14	6.308	11.948	11.589	23.511
15.	614.307	15	4.511	11.151	10.792	21.937
16.	549.682	16	5.214	8.355	9.995	18.314
17.	545.057	17	3.917	10.558	8.198	18.691
18.	579.932	18	5.120	9.261	10.401	19.767
19.	612.807	19	5.823	11.964	9.105	21.144
20.	571.182	20	5.526	9.667	9.308	18.820
21.	584.557	21	5.230	9.370	10.011	19.497
22.	586.932	22	4.933	10.073	9.214	19.323
23.	568.807	23	4.636	10.276	8.417	18.850
24.	638.182	24	5.839	11.480	10.120	21.376
25.	670.557	25	6.042	12.183	10.823	22.953
26.	589.432	26	5.745	8.386	10.526	18.930
27.	585.807	27	3.448	9.089	10.730	19.906
28.	630.682	28	5.651	9.792	10.933	20.733
29.	643.557	29	3.855	10.995	11.136	22.109
30.	628.932	30	3.558	10.198	10.839	20.986
31.	645.307	31	5.761	10.401	10.542	20.962
32.	602.682	32	4.464	10.105	9.245	19.289
33.	661.057	33	4.667	12.308	9.948	22.216
34.	651.932	34	6.370	11.511	10.151	21.642
35.	623.807	35	4.573	9.714	10.855	20.669
36.	619.682	36	3.776	8.917	11.558	20.595
37.	659.057	37	4.480	10.620	11.261	21.872
38.	706.432	38	4.183	11.823	12.464	24.448
39.	627.807	39	2.886	11.026	10.167	21.175
40.	642.182	40	5.589	10.230	9.870	20.201

DATA SET 1:

COMPARISON OF CHANGES IN PRICE INDICES 1980-1981

KEY : Changes represented by LOG[x(t)/x(t-1)]
c.p.i. = consumer price index
p.p.i. = production price index

Y = Change in c.p.i. for FOOD in month i
X1 = Change in c.p.i. for FOOD in month i-1
X2 = Change in c.p.i. for ITEMS EXCLUDING FOOD in month i-1
X3 = Change in p.p.i. for FOOD in month i-1
X4 = Change in p.p.i. for AGRICULTURE in month i-1
X5 = Change in p.p.i. for S.A. MANUFACTURING in month i-1

	Y	X1	X2	X3	X4	X5
1.	.0206	.0055	.0049	.0013	.0025	.0082
2.	.0012	.0206	.0043	-.0286	-.0215	.0075
3.	.0030	.0012	.0121	.0072	.0102	.0092
4.	.0224	.0030	.0066	.0117	.0057	.0097
5.	.0104	.0224	.0083	.0180	.0075	.0146
6.	.0155	.0104	.0118	.0350	.0271	.0128
7.	.0302	.0155	.0156	.0303	.0234	.0077
8.	.0169	.0302	.0092	.0454	.0321	.0206
9.	.0543	.0169	.0096	.0319	.0300	.0228
10.	.0466	.0543	.0079	.0871	.0783	.0130
11.	.0199	.0466	.0145	.0717	.0610	.0118
12.	.0171	.0199	.0039	.0126	.0183	.0082
13.	.0127	.0171	.0082	.0148	.0128	.0041
14.	.0139	.0127	.0038	-.0166	-.0152	.0055
15.	.0014	.0139	.0146	.0152	.0152	.0095
16.	.0000	.0014	.0096	-.0321	-.0262	.0050
17.	.0005	.0000	.0026	.0094	.0096	.0104
18.	.0159	.0005	.0089	-.0237	-.0145	.0093
19.	.0036	.0159	.0094	.0391	.0325	.0097
20.	.0218	.0036	.0297	-.0009	-.0014	.0186
21.	.0196	.0218	.0145	.0028	-.0009	.0229
22.	.0099	.0196	.0143	.0210	.0251	.0092
23.	.0102	.0099	.0102	.0201	.0200	.0100
24.	.0101	.0102	.0053	.0110	.0076	.0077

DATA SET 2:

PREDICTING STOCK EXCHANGE PRICE CHANGES Last half of 1981

KEY : Changes represented by LOG[x(t)/x(t-1)]
Y relates to week i
All X relate to week i-1

Y = Change in share price WESTERN DEEP (in cents)
X1 = Change in share price Western Deep previous week
X2 = Change in share price Randfontein
X3 = Change in JSE index West Wits
X4 = Change in JSE index All Gold
X5 = Change in JSE index Financial
X6 = Change in JSE index Industrial
X7 = Change in gold price (in $/oz)
X8 = Change in London silver price (in p/oz)
X9 = Change in Sterling/Rand exchange rate
X10 = Change in US $/Rand exchange rate

	Y	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
1.	.103	-.137	-.118	-.155	-.128	-.075	-.050	-.087	-.194	.019	-.021
2.	.088	.103	.070	.061	.066	.014	-.004	-.048	.030	-.009	-.022
3.	-.011	.088	-.050	.100	.071	.009	.017	-.015	.055	-.003	-.008
4.	-.055	-.011	.059	-.010	.006	.007	.013	.015	-.009	-.003	-.020
5.	-.011	-.055	-.039	.007	.001	-.002	.008	-.007	-.020	-.019	-.012
6.	.000	-.011	.000	-.009	-.008	-.003	.008	-.035	.004	.012	-.015
7.	.087	.000	.068	.053	.047	.004	.017	.040	.058	.021	.005
8.	.010	.087	.037	.069	.079	.010	.031	.034	.054	-.015	.012
9.	-.010	.010	.009	-.031	-.027	.018	-.002	-.024	-.071	-.003	-.002
10.	.016	-.010	.044	.045	.049	.010	.011	.036	.049	-.005	.000
11.	.102	.016	.038	.014	.029	.018	.016	.032	.106	.021	-.012
12.	-.028	.102	.098	.046	.065	.018	.017	.045	.091	-.014	.017
13.	-.116	-.028	.044	.008	.003	.012	.010	.011	-.079	.007	.003
14.	.087	-.116	-.113	-.035	-.038	-.015	-.018	-.074	-.109	.005	-.016
15.	.010	.087	.141	.020	.025	.015	.012	.043	-.118	-.026	.007
16.	-.030	.010	.070	.003	.017	-.005	.025	-.016	.106	-.005	-.007
17.	-.083	-.030	.010	-.031	-.014	.010	.017	.000	.010	.009	-.006
18.	.011	-.083	-.019	-.073	-.069	-.008	-.021	-.034	-.022	-.007	-.011
19.	-.011	.011	-.013	-.013	-.024	-.008	-.013	.000	-.021	-.020	.013
20.	-.033	-.011	-.027	-.022	-.038	-.009	.004	-.031	-.017	-.018	-.006
21.	-.094	-.033	-.031	-.063	-.060	.003	-.003	-.037	-.114	-.002	-.003
22.	-.006	-.094	-.018	-.018	-.026	-.004	-.001	.000	.002	-.013	.001
23.	.089	-.006	.062	.057	.041	.019	.008	.007	-.026	-.015	-.002
24.	-.023	.089	.046	.051	.050	-.006	.009	.039	.095	.000	-.009
25.	.034	-.023	.025	-.004	.003	.006	.006	-.002	.030	.021	-.020

DATA SET 3:

RELATIONSHIPS BETWEEN MEDIA EXPOSURES FOR VARIOUS POPULATION GROUPS

Y = percent CINEMA ATTENDANCE in past two weeks
X1 = Coloured
X2 = Asian
X3 = African
X4 = Male
X5 = percent reading English daily newspaper
X6 = percent reading Afrikaans daily newspaper
X7 = percent reading any weekly newspaper
X8 = percent reading any magazine
X9 = percent watching prime time television (2000 - 2130)
X10 = percent listening to Radio 5
X11 = percent listening to Springbok radio

	Y	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
1.	20.3	0	0	0	1	45.4	34.7	86.0	89.8	60.1	15.8	24.9
2.	26.4	0	0	0	1	69.8	6.3	85.8	94.9	57.8	20.1	31.8
3.	23.5	0	0	0	1	40.6	33.2	79.2	91.9	56.2	17.1	27.2
4.	22.2	0	0	0	1	15.9	51.2	76.3	89.8	69.2	12.4	33.6
5.	23.9	0	0	0	1	56.4	27.3	83.2	90.9	56.9	21.0	27.7
6.	24.5	0	0	0	1	25.6	37.5	78.6	91.0	62.4	11.8	29.0
7.	17.8	0	0	0	1	18.0	41.5	81.8	93.7	58.4	7.5	29.2
8.	14.5	0	0	0	1	19.4	39.0	77.3	93.7	62.2	7.6	24.1
9.	15.3	0	0	0	1	42.1	33.5	81.0	91.7	64.0	11.6	28.4
10.	48.2	0	0	0	1	48.3	27.0	83.8	92.7	42.8	35.5	25.3
11.	46.1	0	0	0	1	39.3	32.1	82.3	95.4	48.3	37.0	30.5
12.	28.3	0	0	0	1	45.5	30.7	84.0	94.2	55.1	21.2	25.7
13.	12.9	0	0	0	1	48.5	34.0	83.2	92.1	63.0	8.9	23.8
14.	18.6	0	0	0	0	38.7	26.6	81.0	91.3	63.8	11.4	32.5
15.	20.4	0	0	0	0	64.8	2.7	83.5	94.4	61.5	15.5	42.3
16.	23.4	0	0	0	0	34.4	25.6	74.2	93.2	63.9	11.5	33.9
17.	16.3	0	0	0	0	16.8	53.0	67.0	95.3	61.8	6.5	42.0
18.	25.2	0	0	0	0	50.6	22.0	78.0	92.5	63.5	15.6	35.8
19.	20.3	0	0	0	0	17.1	28.2	75.6	91.9	64.0	7.4	32.9
20.	9.4	0	0	0	0	14.6	33.0	80.9	93.3	66.1	3.5	34.2
21.	10.1	0	0	0	0	17.6	29.2	67.9	96.5	59.3	2.6	35.3
22.	16.1	0	0	0	0	36.5	26.1	77.7	93.8	67.4	8.5	35.7
23.	50.0	0	0	0	0	46.9	27.4	79.1	96.9	45.5	33.8	31.3
24.	42.5	0	0	0	0	38.9	23.9	79.2	97.1	54.4	29.4	36.8
25.	22.6	0	0	0	0	38.2	27.0	77.1	95.9	63.6	13.1	32.9
26.	15.7	0	0	0	0	35.8	26.9	80.1	95.3	67.8	6.3	33.1
27.	12.8	1	0	0	1	29.1	13.0	58.8	48.7	27.1	9.8	21.6
28.	19.9	1	0	0	1	55.0	5.1	67.1	48.6	51.4	7.5	41.7
29.	16.2	1	0	0	1	55.1	8.0	77.6	60.4	45.5	13.3	28.4
30.	22.0	1	0	0	1	25.8	31.3	67.1	58.8	30.5	17.5	23.7
31.	13.0	1	0	0	1	4.3	16.9	51.9	43.2	13.9	4.7	21.2
32.	3.7	1	0	0	1	.2	10.4	23.3	23.8	3.5	1.7	13.1
33.	6.2	1	0	0	1	32.8	11.5	59.3	44.2	33.0	7.2	23.9
34.	26.1	1	0	0	1	29.6	15.3	63.1	59.3	23.4	14.4	23.2
35.	29.3	1	0	0	1	31.8	14.7	65.9	60.8	26.5	15.8	24.0
36.	8.4	1	0	0	1	35.5	11.5	60.5	54.2	30.2	11.7	17.9
37.	5.1	1	0	0	1	32.3	15.2	62.5	44.0	32.8	5.1	26.2
38.	7.5	1	0	0	0	26.5	10.0	57.3	60.9	28.9	9.4	31.1
39.	14.4	1	0	0	0	36.7	3.1	57.5	62.9	49.7	8.6	52.9
40.	10.1	1	0	0	0	47.6	6.2	72.1	72.7	47.0	13.9	37.5
41.	6.5	1	0	0	0	13.0	19.5	63.2	69.0	31.8	7.9	40.3
42.	8.5	1	0	0	0	2.0	19.8	46.4	55.0	10.5	2.1	26.0
43.	2.7	1	0	0	0	1.5	4.0	22.8	29.9	2.0	3.8	21.4
44.	4.8	1	0	0	0	28.4	8.1	55.0	58.2	36.1	5.8	34.9
45.	15.4	1	0	0	0	29.8	12.1	65.6	72.9	24.2	15.7	29.2
46.	15.1	1	0	0	0	30.2	10.8	63.7	77.5	26.5	16.8	28.3
47.	8.9	1	0	0	0	28.8	12.2	58.3	64.4	33.9	6.7	29.9
48.	1.4	1	0	0	0	28.2	8.8	56.6	54.5	33.7	5.7	38.2
49.	31.1	0	1	0	1	71.9	1.2	86.2	63.7	38.2	31.3	37.3

50.	41.7	0	1	0	1	69.4	1.0	78.8	60.2	50.3	14.8	24.9
51.	35.7	0	1	0	1	77.0	1.3	88.7	64.2	42.7	31.5	33.3
52.	21.6	0	1	0	1	49.4	1.1	68.5	46.7	22.4	23.8	37.4
53.	18.7	0	1	0	1	71.2	.7	85.2	56.9	42.6	21.7	35.4
54.	59.5	0	1	0	1	73.4	2.0	86.3	77.2	36.8	43.5	36.4
55.	58.1	0	1	0	1	72.2	3.1	85.7	76.0	34.5	41.9	35.0
56.	28.7	0	1	0	1	73.6	.3	90.3	70.6	39.2	35.2	42.6
57.	16.1	0	1	0	1	73.9	.0	88.2	59.1	47.4	17.8	28.7
58.	17.9	0	1	0	0	46.6	.0	69.5	54.2	42.9	25.9	37.9
59.	32.2	0	1	0	0	43.2	.5	65.8	59.5	51.7	13.4	38.5
60.	22.4	0	1	0	0	50.7	.1	72.3	59.1	48.5	27.1	38.4
61.	10.7	0	1	0	0	25.6	.0	49.0	33.4	20.7	15.5	31.7
62.	16.3	0	1	0	0	45.7	.0	70.0	53.5	46.3	18.8	42.5
63.	37.9	0	1	0	0	59.2	.3	82.6	76.1	42.8	46.7	27.3
64.	30.7	0	1	0	0	51.8	.2	80.9	75.7	37.8	40.6	32.2
65.	16.6	0	1	0	0	52.8	.0	75.4	58.4	49.2	22.1	49.1
66.	13.3	0	1	0	0	43.0	.0	63.2	46.4	51.8	12.4	39.6
67.	8.3	0	0	1	1	11.9	.8	21.3	24.4	4.8	2.4	2.5
68.	10.5	0	0	1	1	9.9	.2	46.6	42.0	3.5	3.1	10.6
69.	21.4	0	0	1	1	17.6	1.1	22.7	35.2	3.9	2.5	1.9
70.	10.0	0	0	1	1	9.0	4.8	14.9	34.2	3.2	.4	.9
71.	25.3	0	0	1	1	28.0	1.7	41.8	44.7	6.5	4.1	3.3
72.	35.0	0	0	1	1	18.7	2.9	29.2	47.7	6.4	2.8	3.2
73.	25.8	0	0	1	1	17.9	2.9	29.2	50.0	7.6	2.3	2.3
74.	3.8	0	0	1	1	3.8	.3	15.9	22.6	1.2	1.3	4.1
75.	8.2	0	0	1	1	13.1	1.2	21.5	24.8	3.2	1.3	2.4
76.	24.5	0	0	1	1	15.6	1.1	32.7	46.3	5.1	3.7	5.3
77.	26.0	0	0	1	1	10.1	1.4	29.6	46.9	4.0	4.4	4.6
78.	22.7	0	0	1	1	25.0	1.7	37.3	47.9	6.7	2.7	4.9
79.	5.9	0	0	1	1	14.8	1.3	23.4	24.7	3.3	1.6	2.2
80.	2.6	0	0	1	0	3.6	.5	10.7	16.1	1.7	.6	.8
81.	3.0	0	0	1	0	4.3	.0	33.4	33.8	.4	2.2	.9
82.	3.0	0	0	1	0	5.3	.6	11.7	24.7	2.4	.6	.8
83.	.3	0	0	1	0	4.0	1.5	5.4	22.2	1.8	1.2	.2
84.	7.1	0	0	1	0	17.4	1.5	34.4	44.7	5.2	2.3	2.6
85.	4.6	0	0	1	0	6.0	2.6	26.1	42.3	10.0	.9	1.6
86.	.8	0	0	1	0	4.5	.0	15.2	36.1	1.7	2.5	1.2
87.	1.4	0	0	1	0	1.0	.2	10.5	16.8	.2	.6	.2
88.	1.3	0	0	1	0	3.9	.3	11.5	16.0	1.8	.7	.7
89.	6.3	0	0	1	0	6.3	1.1	24.7	44.3	2.2	2.1	1.1
90.	6.4	0	0	1	0	4.9	.9	20.2	42.4	1.9	2.3	1.2
91.	2.0	0	0	1	0	5.7	.6	18.0	23.1	2.6	1.4	.9
92.	1.2	0	0	1	0	5.7	.4	17.7	21.7	1.7	.1	.9

DATA SET 4:

RELATIONSHIP BETWEEN CLIMATIC REGION AND ATTRIBUTES OF HYBRID MAIZE

Y = Bioclimatic region (1=dry 2=medium 3=wet)
X1 = Rating for attribute A
X2 = Rating for attribute B
X3 = Rating for attribute C
X4 = Rating for attribute D
X5 = Rating for attribute E
X6 = Rating for attribute F
X7 = Rating for attribute G
X8 = Rating for attribute H
X9 = Rating for attribute I

	Y	X1	X2	X3	X4	X5	X6	X7	X8	X9
1.	2	7	8	9	6	5	8	8	6	4
2.	2	8	7	10	8	6	8	7	7	6
3.	2	8	7	9	8	7	8	7	8	6
4.	2	8	7	9	6	6	9	7	8	6
5.	1	8	9	10	8	6	9	8	8	5
6.	1	8	9	10	1	6	8	8	8	2
7.	1	8	7	9	9	7	9	7	6	4
8.	2	7	8	10	9	7	8	7	8	6
9.	1	8	9	9	8	7	9	8	8	5
10.	2	7	8	10	8	5	9	8	7	6
11.	1	7	8	10	7	6	8	7	9	5
12.	2	8	7	8	7	6	8	7	7	6
13.	1	6	7	10	8	5	8	6	5	3
14.	2	7	7	10	8	6	7	6	7	4
15.	2	9	8	10	8	8	9	9	9	8
16.	2	10	8	10	4	8	10	6	5	5
17.	2	6	7	10	5	6	10	6	6	6
18.	2	10	9	10	4	9	9	6	7	4
19.	2	10	6	10	8	8	9	5	5	5
20.	2	6	9	10	6	5	8	4	6	4
21.	2	10	7	10	8	5	10	10	8	10
22.	2	10	5	10	10	6	10	8	7	6
23.	2	9	7	10	8	4	10	7	8	9
24.	2	10	5	10	10	6	10	8	8	9
25.	2	10	6	10	9	7	10	9	10	10
26.	2	10	5	10	9	10	10	8	8	8
27.	2	10	10	10	10	10	7	3	5	1
28.	2	10	9	10	8	6	10	5	8	10
29.	2	10	4	10	5	7	5	5	6	7
30.	2	9	6	10	6	3	7	5	6	8
31.	2	10	5	10	1	1	10	8	8	8
32.	2	7	9	10	7	4	10	7	9	7
33.	2	10	9	10	9	10	10	8	9	10
34.	2	7	6	9	9	1	1	7	8	7
35.	2	7	6	10	5	3	8	6	6	6
36.	2	7	6	10	3	3	8	5	6	4
37.	2	9	5	10	8	8	9	7	8	5
38.	2	9	6	10	10	10	10	4	8	1
39.	2	10	6	9	4	2	8	7	6	9
40.	2	10	5	7	4	4	10	7	5	5
41.	2	10	8	10	3	1	10	5	5	10
42.	2	10	8	10	9	4	8	2	6	2
43.	2	8	7	7	4	5	6	6	4	7
44.	2	8	8	9	8	6	9	8	8	7
45.	2	4	9	10	8	3	7	9	8	6
46.	2	6	9	10	10	5	8	5	8	4
47.	2	6	8	10	10	3	8	8	9	4
48.	2	7	9	10	10	6	10	6	9	3
49.	2	7	9	10	9	8	9	7	5	9
50.	2	5	9	9	10	6	10	8	10	4
51.	1	2	6	10	5	10	10	9	10	9

52.	1	2	5	10	7	9	10	8	7	8
53.	1	1	1	10	9	8	10	7	9	9
54.	1	7	1	10	7	6	7	5	1	1
55.	1	9	3	10	5	4	6	4	9	10
56.	1	6	5	10	9	9	8	6	5	8
57.	2	6	8	10	4	6	9	3	6	3
58.	2	7	5	10	7	8	10	10	5	2
59.	2	6	5	9	8	7	9	9	8	7
60.	2	8	9	10	7	8	8	8	8	8
61.	2	8	8	10	10	10	10	8	9	6
62.	2	10	6	8	10	8	8	10	7	3
63.	2	7	3	10	8	9	5	3	7	4
64.	1	10	8	10	8	10	5	9	8	7
65.	2	10	10	10	9	10	9	7	10	7
66.	2	4	6	10	9	5	8	8	9	5
67.	2	8	8	10	8	8	8	6	7	6
68.	2	10	10	10	10	10	10	5	10	5
69.	2	10	5	10	7	10	9	8	9	8
70.	2	3	4	10	10	10	10	8	9	8
71.	2	7	6	10	7	8	8	6	6	5
72.	2	5	6	10	7	8	10	4	7	6
73.	2	8	8	10	7	8	9	9	9	9
74.	2	9	8	10	9	7	9	5	4	1
75.	2	10	9	10	10	10	7	4	3	2
76.	2	8	6	10	10	10	9	5	8	4
77.	2	10	9	10	10	10	8	6	8	4
78.	2	10	5	10	9	10	6	7	6	4
79.	2	10	5	10	10	10	5	5	10	5
80.	2	8	8	10	7	8	8	6	8	6
81.	2	7	4	10	9	8	8	3	5	8
82.	2	8	6	10	9	9	10	5	8	5
83.	3	4	6	3	10	5	8	2	4	4
84.	3	5	8	5	8	7	6	8	9	8
85.	3	6	10	9	9	7	9	4	9	6
86.	3	5	9	3	8	1	2	10	10	5
87.	3	9	9	4	3	8	2	8	8	3
88.	3	7	7	8	10	5	5	8	6	5
89.	3	8	10	10	8	1	8	9	10	5
90.	3	6	7	4	6	1	2	8	9	1
91.	3	9	5	7	8	9	5	7	5	8
92.	3	5	7	10	7	7	3	1	7	7
93.	3	1	10	9	9	1	5	9	10	1
94.	3	8	8	8	8	8	5	5	5	5
95.	3	2	6	8	10	1	7	7	7	3
96.	3	8	9	10	9	6	7	2	10	1
97.	3	8	6	10	8	9	8	4	9	1
98.	3	5	3	9	8	6	8	3	8	1
99.	3	8	8	9	9	4	3	7	6	1
100.	3	7	8	10	9	8	10	5	6	3
101.	3	7	5	8	9	4	9	2	7	1
102.	3	2	10	10	3	1	4	3	8	1
103.	3	2	6	7	9	4	8	4	9	1
104.	3	4	8	9	9	5	5	6	2	1
105.	3	8	5	9	5	8	2	5	5	5
106.	3	6	7	10	6	5	8	6	6	3
107.	1	3	9	5	8	10	9	7	9	6
108.	1	2	9	10	6	9	9	4	10	5
109.	1	1	8	10	10	3	10	9	7	10
110.	1	2	5	10	5	9	9	9	3	7
111.	1	1	2	10	8	8	9	8	7	8
112.	1	3	4	10	8	8	10	9	3	9
113.	1	1	6	10	8	10	8	8	7	10
114.	1	1	10	10	6	10	8	9	8	8
115.	1	4	4	10	6	9	8	1	7	7
116.	1	4	4	10	4	10	5	8	6	10
117.	1	2	9	10	3	1	9	9	7	5
118.	2	7	4	2	6	6	10	6	8	5
119.	2	10	7	10	5	5	7	4	6	6

120.	2	4	6	10	9	5	9	7	6	5
121.	2	5	6	8	10	4	9	6	8	4
122.	2	5	7	10	8	5	8	6	9	6
123.	2	8	9	10	8	7	9	4	10	8
124.	2	8	9	10	7	6	9	4	9	2
125.	1	7	8	10	7	8	7	8	6	5
126.	1	6	7	10	8	4	9	7	5	3
127.	1	3	7	10	5	8	8	8	4	5
128.	1	5	9	10	8	5	9	8	9	9
129.	2	2	5	10	8	6	8	5	5	3
130.	2	1	5	10	10	1	7	3	7	1
131.	1	3	5	10	5	3	8	7	8	2
132.	2	5	10	10	8	5	9	9	10	10
133.	1	3	7	10	5	5	8	8	7	2
134.	1	6	7	9	10	10	9	7	3	5
135.	1	6	9	10	8	2	10	5	5	4
136.	1	7	10	10	10	6	10	8	8	4
137.	2	10	5	10	8	9	5	9	5	1
138.	2	10	7	10	10	10	10	7	8	1
139.	2	9	5	10	8	9	9	6	5	2
140.	2	10	3	8	10	10	8	9	10	3
141.	2	10	8	10	8	8	8	6	6	6
142.	2	10	6	8	10	6	8	8	5	3
143.	2	8	4	10	8	6	9	7	6	2
144.	2	6	10	8	9	5	7	5	8	4
145.	2	9	5	10	8	6	8	8	8	1
146.	2	8	9	10	9	8	9	9	9	3
147.	2	6	8	10	5	10	10	8	5	2
148.	2	6	4	10	10	3	8	8	4	2
149.	2	8	7	10	7	7	6	8	7	2
150.	2	5	8	10	8	5	10	9	8	5
151.	1	10	5	10	6	10	10	5	5	5
152.	2	6	8	9	8	3	7	8	7	7
153.	2	8	6	10	8	8	9	7	3	2
154.	2	10	6	10	9	8	8	6	6	1
155.	2	10	8	10	5	5	6	8	10	3
156.	1	2	8	10	9	10	8	7	8	5
157.	2	2	5	10	6	9	10	10	8	5
158.	2	8	9	9	7	7	5	9	8	5
159.	2	5	8	9	9	4	9	5	9	5
160.	2	9	7	9	9	3	9	7	7	6
161.	2	6	4	10	7	1	5	8	7	9
162.	2	7	6	9	9	3	8	8	7	6
163.	2	3	7	9	9	2	8	7	5	7
164.	2	8	5	9	9	2	7	6	8	7
165.	2	5	7	8	9	2	9	8	5	4
166.	2	8	7	10	7	2	7	8	8	7
167.	2	10	6	10	8	9	10	7	6	9
168.	1	10	7	10	10	6	7	7	8	6
169.	1	10	7	10	6	9	8	9	6	8
170.	1	6	4	10	5	7	10	10	9	10
171.	1	10	6	10	9	7	10	8	10	10
172.	1	10	5	10	5	9	9	7	8	8
173.	2	9	6	10	9	9	10	10	9	10
174.	2	10	8	10	10	10	10	8	8	8
175.	2	9	8	10	7	9	9	8	8	9
176.	2	10	7	10	6	8	10	8	8	7
177.	1	10	7	10	8	7	6	8	8	9
178.	1	10	6	10	7	6	8	10	6	8
179.	3	7	5	6	5	5	7	2	10	1
180.	3	7	8	9	7	2	7	4	10	4
181.	3	10	8	10	10	9	6	4	10	7
182.	3	7	8	9	6	5	8	6	6	4
183.	3	9	10	10	8	5	6	5	5	10
184.	3	7	5	9	7	6	3	2	7	4

DATA SET 5:

ANALYSIS OF BUILDING SOCIETY STAFF REQUIREMENTS

Y = Clerical staff
X1 = Banking hall transactions (in thousands)
X2 = Number of savings cheques drawn ('000)
X3 = Investor transactions ('000)
X4 = Loans transactions ('000)
X5 = Insurance transactions ('000)
X6 = Loans information transactions ('000)
X7 = Name and address capture ('000)
X8 = Other general information capture ('000)

	Y	X1	X2	X3	X4	X5	X6	X7	X8
1.	497.0	379.604	14.984	8.107	.099	.000	22.406	35.008	114.871
2.	329.0	196.433	5.302	3.456	.133	.000	11.232	15.875	61.681
3.	146.0	105.975	3.837	1.686	.666	.860	3.838	8.509	27.333
4.	249.0	209.451	10.621	4.913	.052	.000	14.099	21.368	63.052
5.	82.0	46.641	2.310	1.554	.322	.397	3.701	5.601	17.713
6.	98.0	57.725	1.226	1.615	.648	.822	4.408	.770	21.410
7.	32.0	14.870	.237	1.021	.219	.622	1.164	.022	5.687
8.	23.0	15.601	.393	.703	.050	.069	1.330	1.592	4.980
9.	36.0	16.826	.405	.441	.276	.208	1.012	1.844	5.739
10.	109.0	80.640	1.512	1.369	1.840	1.252	3.799	8.726	23.455
11.	124.0	86.852	3.345	2.269	.732	.398	5.088	8.405	24.547
12.	97.0	40.114	2.652	1.644	1.099	.634	3.730	7.199	21.496
13.	87.0	43.165	1.544	1.560	.603	.661	4.225	6.647	15.769
14.	65.0	33.714	1.463	1.196	.434	.501	2.580	5.468	18.354
15.	46.0	24.735	.484	.659	.400	.275	1.359	1.763	6.728
16.	78.0	48.252	.995	.917	.662	.771	3.120	7.005	15.839
17.	64.0	26.543	.734	1.583	.585	.499	1.906	4.181	12.678
18.	62.0	35.609	2.026	.943	.423	.579	2.419	3.052	12.150
19.	43.0	23.206	.697	.902	.311	.507	1.832	1.468	9.305
20.	25.0	18.481	.266	.323	.183	.119	.632	1.678	5.753
21.	37.0	21.287	.651	.818	.435	.379	1.508	2.751	7.275
22.	26.0	12.469	.766	.395	.278	.406	.806	.038	5.547
23.	175.0	110.420	5.540	5.030	.011	.000	7.923	.000	42.429
24.	42.0	20.472	1.306	.722	.255	.362	1.646	2.765	10.097
25.	126.0	54.363	3.494	2.233	1.455	2.172	4.777	7.638	32.259
26.	84.0	37.196	1.392	.878	.700	.891	3.499	5.996	20.131
27.	160.0	50.162	4.086	1.690	2.133	1.644	4.233	9.987	30.808
28.	60.0	41.921	3.457	.851	.023	.000	2.061	5.406	13.163
29.	71.0	31.187	2.411	.756	1.278	.826	2.792	5.404	21.218
30.	89.0	50.144	2.006	.995	1.310	1.238	4.554	5.475	21.487
31.	96.0	50.299	3.856	.990	1.202	1.351	2.987	.287	19.488
32.	64.0	25.574	4.057	1.099	.000	3.907	5.740	18.624	25.574
33.	90.0	48.554	4.412	1.515	.055	.000	2.282	5.543	18.321
34.	93.0	42.784	1.117	.500	.281	.288	1.567	5.060	20.122
35.	45.0	23.720	.741	.492	.380	.307	1.252	2.909	9.628
36.	37.0	19.295	.674	.380	.383	.389	2.579	2.378	7.179
37.	18.0	11.038	.297	.160	.064	.036	.985	.952	3.369
38.	130.0	5.607	.071	.025	3.762	3.559	.248	2.165	18.882
39.	54.0	4.035	.092	.000	2.613	2.597	.194	2.612	9.783
40.	129.0	5.380	.094	.058	5.068	4.226	.509	3.842	19.208
41.	7.0	7.880	.120	.074	.021	.000	.138	.000	1.624
42.	8.0	4.729	.153	.119	.027	.000	.195	.000	.831
43.	55.0	2.874	.016	.012	2.240	3.017	.297	1.175	11.029
44.	86.0	49.364	4.121	.792	.017	.000	3.057	6.665	25.349
45.	57.0	15.977	1.523	.429	.887	.652	.984	3.149	8.706
46.	61.0	22.383	1.970	.546	.877	.846	3.522	3.417	13.195
47.	78.0	42.041	1.979	.881	.002	.000	1.755	5.374	15.549
48.	95.0	83.133	4.326	1.363	.006	.000	5.348	7.616	26.262

BIBLIOGRAPHY

A. ACADEMIC BACKGROUND:

1. ANDREWS, D.F. and PREGIBON, D. (1978) : Finding the outliers that matter. Journal of the Royal Statistical Society, B40, 85-93.

2. BARNETT, V. and LEWIS, T. (1978) : Outliers in statistical data. John Wiley & Sons.

3. BECKMAN, R.J. and TRUSSELL, H.J. (1974) : The distribution of an arbitrary studentized residual and the effects of updating in multiple regression. Journal of the American Statistical Association, 69, 199-201.

4. BELSLEY, D.A., KUH, E. and WELSCH, R.E. (1980) : Regression diagnostics : identifying influential data and sources of collinearity. John Wiley & Sons.

5. BINGHAM, C. (1977) : Some identities useful in the analysis of residuals from linear regression. University of Minnesota, School of Statistics, technical report no. 300.

6. COOK, R.D. (1977) : Detection of influential observations in linear regression. Technometrics, 19, 15-18.

7. COOK, R.D. (1979) : Influential observations in linear regression. Journal of the American Statistical Association, 74, 169-174.

8. COOK, R.D. and PRESCOTT, P. (1981) : On the accuracy of Bonferroni significance levels for detecting outliers in linear models. Technometrics, 23, 59-63.

9. COOK, R.D. and WEISBERG, S. (1980) : Characterizations of an empirical influence function for detecting influential cases in regression. Technometrics, 22, 495-508.

10. COOLEY, W.W. and LOHNES, P.R. (1971) : Multivariate data analysis. John Wiley & Sons.

11. DOORNBOS, R. (1981) : Testing for a single outlier in a linear model. Biometrics, 37, 705-711.

12. DRAPER, N.R. and JOHN, J.A. (1981) : Influential observations and outliers in regression. Technometrics, 23, 21-26.

13. ELLENBERG, J.H. (1973) : The joint distribution of the standardized least squares residuals from a general linear regression. Journal of the American Statistical Association, 68, 941-943.

14. ELLENBERG, J.H. (1976) : Testing for a single outlier from a general linear regression. Biometrics, 32, 637-645.

15. FORSYTHE, G. and MOLER, C.B. (1967) : Computer solution of linear algebraic systems. Prentice-Hall.

16. GENTLEMAN, J.F. and WILK, M.B. (1975) : Detecting outliers II: supplementing the direct analysis of residuals. Biometrics, 31, 387-410.

17. GOLDBERGER, A.S. (1964) : Econometric theory. John Wiley & Sons.

18. GRAYBILL, F.A. (1961) : An introduction to linear statistical models, vol. 1. McGraw-Hill.

19. GRAYBILL, F.A. (1976) : Theory and application of the linear model. Duxbury Press.

20. GUNST, R.F. and MASON, R.L. (1976) : Generalized mean square error properties of regression estimators. Communication in Statistics, A5, 1501-1508.

21. HAWKINS, D.M. (1980) : Identification of outliers. Chapman and Hall.

22. HOAGLIN, D.C. and WELSCH, R.E. (1978) : The Hat matrix in regression and ANOVA. The American Statistician, 32, 17-22.

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

24. HOERL, A.E. and KENNARD, R.W. (1970) : Ridge regression: applications to non-orthogonal problems. Technometrics, 12, 69-82.

25. JACOBS, M. (1982) : Influential observations in linear regression. Graduate School of Business, University of Cape Town, unpublished technical report.

26. JOHNSON, S.R., REIMER, S.C. and ROTHROCK, T.P. (1972) : Principal components and the problem of multicollinearity. Metroeconomica, 25, 306-314.

27. JUDGE, G.G. and BOCK, M.E. (1978) : The statistical implications of pre-test and Stein-rule estimators in econometrics. North-Holland.

28. JUDGE, G.G., GRIFFITHS, W.E., HILL, R.C. and LEE, T-C. (1980) : The theory and practice of econometrics. John Wiley & Sons.

29. KHATRI, C.G. (1962) : Conditions for Wishartness and independence of second degree polynomials in a normal vector. Annals of Mathematical Statistics, 33, 1002-1007.

30. LUND, R.E. (1975) : Tables for an approximate test for outliers in linear models. Technometrics, 17, 473-476.

31. MARQUARDT, D.W. (1974) : Generalized inverse, ridge regression, biased linear estimation and nonlinear estimation. Technometrics, 12, 591-612.

32. MARSHALL, A.W. and OLKIN, I. (1971) : Norms and inequalities for condition numbers, III. Stanford University, technical report no. 53.

33. NAGAR, A.L. and KAKWANI, N.C. (1964) : The bias and moment matrix of a mixed regression estimator. Econometrica, 32, 389-402.

34. PRESCOTT, P. (1975) : An approximate test for outliers in linear models. Technometrics, 17, 129-132.

35. PRICE, J.M. (1982) : Comparisons among regression estimators under the generalized mean square error criterion. Communication in Statistics, A11, 1965-1984.

36. SEARLE, S.R. (1971) : Linear models. John Wiley & Sons.

37. THEIL, H. (1978) : Introduction to econometrics. Prentice-Hall.

38. THEOBALD, C.M. (1974) : Generalization of mean square error applied to ridge regression. Journal of the Royal Statistical Society, B36, 103-106.

39. 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.

40. TROSKIE, C.G., COUTSOURIDES, D., JACOBS, M. and DUNNE, T.T. (1982) : Detection of outliers in the presence of multicollinearity. University of Cape Town, Department of Mathematical Statistics, technical report no. 7.

B. SOURCES OF ILLUSTRATIVE EXAMPLES:

DATA SET

1. Bulletin of Statistics (1981): Vol. 15, Department of Statistics, Pretoria.

2A. Financial Mail (1981): Vol. 80-82.

2B. Financial Mail (1980): Vol. 76-78.

3. All Media and Products Survey (1981): S.A. Advertising and Research Foundation.

4. HEDLEY, S.M. (1982): A study of product attribute analysis in the hybrid maize seed industry with specific reference to Poineer Seed Company (Pty) Ltd. Graduate School of Business, University of Cape Town, unpublished technical report.

5. BURGER, A.P. (1982): Practical applications of regression diagonistics. Presentation to the South African Statistical Association, Cape Town, 1982.

17 APR 1984