Contributions to Linear Regression Diagnostics using the Singular Value Decomposition: Measures to Identify Outlying Observations, Influential Observations and Collinearity in Multivariate Data

Kutlwano K.K.M. Ramaboa

Thesis presented for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Statistical Sciences
Faculty of Science

University of Cape Town

February 2010

Acknowledgements

I am indebted to my supervisor, Professor Les Underhill, who contributed to the ideas presented in the thesis. I would also like to thank him for being patient, encouraging, and for the valuable advice, comments and professional guidance.

I am also grateful to Dr Trevor Wegner, former Associate Professor at the University of Cape Town, who has contributed significantly to my understanding and appreciation of Statistics.

I would also like to thank the following:

- Robert Burawundi, Bibo, Todd, my family and friends, whose continued encouragement and support is invaluable.
- Sue Kuyper.
- Staff from University of Cape Town's Interloans library.
- Colleagues who through informal discussions, I have learnt something.

Abstract

This thesis discusses the use of the singular value decomposition (SVD) in multiple regression with special reference to the problems of identifying outlying and/or influential observations, and the explanatory variables that are involved in collinear relationships.

Regression diagnostics are numerous and well-known, however most authors who have used the singular value decomposition in regression analysis have concentrated on the matrix of the right singular vectors (that is, the eigenvectors of $\mathbf{X}^T\mathbf{X}$). In this paper, we consider also the matrix of the left singular vectors (that is, the eigenvectors of $\mathbf{X}\mathbf{X}^T$).

We tap into the theory of correspondence analysis to demonstrate how the total variance in principal components analysis can be broken down along the principal axes, and further broken down into contributions of the rows and the columns of the data matrix. The algebraic expressions derived from decomposing the total variance of $\mathbf{X}$ are used extensively in the thesis to derive measures that aid in the identification of leverage points, and to identify the variables that are involved in collinear relationships and the thresholds to use in order to identify the variables that are involved in the collinearity.

The diagonal values of the hat matrix $\mathbf{H_x}$, where $\mathbf{H_x} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, are used in regression analysis to identify outlying observations in the explanatory variables that may alter the fit of the least squares regression line. The diagonal values however, are known to suffer from the effects of masking and swamping; thus we propose a procedure that is adapted in part from correspondence analysis, to identify leverage points in a data set. The procedure entails producing a leverage-distance (L-D) plot, which is useful in identifying any observations that may be masked or swamped. The procedure is also extended to the diagonal values of the $\mathbf{H}_z$ matrix, the matrix formed when we append the response vector, $\mathbf{y}$, to the matrix of explanatory variables, $\mathbf{X}$, to identify regression outliers. We also propose a residual measure, R_j, which provides insight into the role that each observation plays in determining the displacement of other observations from the least squares fit. The residuals, either in their raw or transformed form, are known to be a poor measure of fit since they may fail to identify the outlying observations when these observations are being accommodated by the least squares fit. Thus, R_j can be used in conjunction with existing measures that are based on the transformed residuals

to identify regression outliers.

The regression estimates such as the coefficients, are known to be easily affected by outlying observations, and by decomposing the regression coefficients, we illustrate how to determine the outlying observations that may have a disproportionate effect in the determination of the individual regression coefficients. An artificial data set and three examples from regression analysis literature are used to illustrate the procedure for outlying observations and proposed measures for outliers and influential observations.

A number of approaches have been proposed to identify the explanatory variables that are involved in collinear relationships, and to detect the coefficients that are most adversely affected. We propose an alternative measure for identifying variables that are involved in the collinearity that is based on examining the magnitude of the squared right singular vectors, which represent the proportion of variance due to an axis that is explained by a particular variable. We also develop a means of quantifying the meaning of 'large' for the magnitude of the eigenvectors of $\mathbf{X}\mathbf{X}^T$ that correspond to small singular values, and for the decomposed coefficient values that correspond to small singular values.

Principal components regression is a form of biased estimator that is used when there is collinearity among the explanatory variables, and we consider an alternative computational approach to principal components regression using the singular value decomposition, focusing particularly on employing values of the left singular vectors in expressing the principal components regression estimates where it is appropriate. We also demonstrate the usefulness of decomposing the multiple correlation coefficient, R^2, to determine the importance of the axes in explaining the amount of variation in $\mathbf{y}$, and a measure to determine the range of values in which prediction is reasonable when there is collinearity in the data is also proposed.

Table of Contents

Appendices

List of Figures

List of Tables

Chapter 1

Preliminaries

1.1 Notation

We consider the standard linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where the dimensions of the vectors and matrices are $n \times 1$, $n \times m$, $m \times 1$ and $n \times 1$ respectively, with $n \geq m$, and possibly including the intercept. We assume that the columns of $\mathbf{X}$ and the vector $\mathbf{y}$ have been standardised to have mean zero and variance one, and that $E(\boldsymbol{\varepsilon}) = 0$ and $\text{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$. In the discussion that follows, we assume that $\mathbf{X}$ does not include the intercept, although the results may be generalised to include the intercept.

The singular value decomposition (SVD) of the $n \times m$ matrix $\mathbf{X}$ of explanatory variables, of rank k, where $n \geq m$ and therefore $k \leq m$ is given by:

$$\mathbf{X} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^T \tag{1.1}$$

where
- $\mathbf{U}$ is an $n \times m$ matrix of left singular vectors ($\mathbf{u}_k$'s) of $\mathbf{X}$, or the eigenvectors of $\mathbf{X}\mathbf{X}^T$.
- $\mathbf{V}$ is an $m \times m$ matrix of right singular vectors ($\mathbf{v}_k$'s) of $\mathbf{X}$, or the eigenvectors of $\mathbf{X}^T\mathbf{X}$.
- $\mathbf{D}_\alpha$ is an $m \times m$ diagonal matrix with non-negative singular values (α_k) of $\mathbf{X}$, or the positive square roots of the eigenvalues of either $\mathbf{X}^T\mathbf{X}$ or $\mathbf{X}\mathbf{X}^T$.

For (1.1), we assume that $\mathbf{X}$ is of full column rank, that is, $k = m$ (cf. Belsley, Kuh and Welsch, 1980, p. 100), and that the singular values are arranged in order of decreasing magnitude (that is, $\alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_k > 0$) with their associated singular vectors arranged according to the order of the singular values.

The columns of $\mathbf{U}$ and $\mathbf{V}$ are orthonormal, hence $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_m$, but since $\mathbf{V}$ is square and $\mathbf{U}$ is rectangular, $\mathbf{V}\mathbf{V}^T = \mathbf{I}_m$ and $\mathbf{U}\mathbf{U}^T \neq \mathbf{I}_n$. The columns of $\mathbf{U}$ and $\mathbf{V}$ are also the row and column singular vectors of $\mathbf{X}$ respectively, since the columns of $\mathbf{U}$ form an orthonormal basis for the space spanned by the columns of $\mathbf{X}$, and the columns of $\mathbf{V}$ form an orthonormal basis for the space spanned by the rows of $\mathbf{X}$. Each $\mathbf{u}_k$ vector represents a linear combination of rows of $\mathbf{X}$ that tend to occur together in a consistent manner, hence most of the values of $\mathbf{X}$ are projected along the first few $\mathbf{v}_k$ axes, with the first singular vector of $\mathbf{V}$, $\mathbf{v}_1$, the major principal axis, representing the largest concentration of the $\mathbf{X}$ values (Mandel, 1982).

The SVD has been used in least squares problems (see amongst others Belsley *et al.* (1980), Mandel (1982), and Henshall and Smith (1996)) to decompose a matrix into several component matrices that are simpler geometrically than the original matrix (Green and Carroll, 1976). However most authors have concentrated on the matrix of the right singular vectors (that is, the eigenvectors of $\mathbf{X}^T\mathbf{X}$). In this thesis, we consider also the matrix of the left singular vectors (that is, the eigenvectors of $\mathbf{X}\mathbf{X}^T$).

1.1.1 Singular Value Decomposition and Regression Analysis

By substituting $\mathbf{X} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^T$ into $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, we can easily derive expressions for some of the quantities commonly estimated in regression using the least squares method:

The regression coefficients ($\hat{\boldsymbol{\beta}}$), and variance of $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{V}\mathbf{D}_\alpha^{-1}\mathbf{U}^T\mathbf{y}$$

$$\text{var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2 = (\mathbf{V}\mathbf{D}_\alpha^{-2}\mathbf{V}^T)\sigma^2$$

The residuals ($\hat{\boldsymbol{\varepsilon}}$), and variance of $\hat{\boldsymbol{\varepsilon}}$:

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{I} - \mathbf{U}\mathbf{U}^T)\mathbf{y}$$

$$\text{var}(\hat{\boldsymbol{\varepsilon}}) = \text{var}[(\mathbf{I} - \mathbf{U}\mathbf{U}^T)\mathbf{y}] = (\mathbf{I} - \mathbf{U}\mathbf{U}^T)\sigma^2$$

For any vector of explanatory variables, $\mathbf{x}_0$, the predicted values ($\hat{\mathbf{y}}_0$), and variance of the $\hat{\mathbf{y}}_0$:

$$\hat{\mathbf{y}}_0 = \hat{\boldsymbol{\beta}}^T \mathbf{x}_0 = \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0 \mathbf{y} = \mathbf{x}_0^T (\mathbf{V} \mathbf{D}_\alpha^{-2} \mathbf{V}^T) \mathbf{x}_0 \mathbf{y}$$

$$\text{var}(\hat{\mathbf{y}}_0) = \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0 \sigma^2 = \mathbf{x}_0^T (\mathbf{V} \mathbf{D}_\alpha^{-2} \mathbf{V}^T) \mathbf{x}_0 \sigma^2$$

The model variance:

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}}}{n-m} = \frac{\mathbf{y}^T (\mathbf{I} - \mathbf{U}\mathbf{U}^T) \mathbf{y}}{n-m}$$

The hat matrix:

$$\mathbf{H}_\mathbf{x} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T = \mathbf{U}\mathbf{U}^T$$

The notation used above, where boldface uppercase letters such as $\mathbf{X}$ are used to denote matrices; boldface lowercase letters such as $\mathbf{x}$ are used to denote vectors; the transpose of a vector $\mathbf{x}$ (or matrix $\mathbf{X}$) is written as $\mathbf{x}^T$ (or $\mathbf{X}^T$); the inverse of a matrix $\mathbf{X}$ is written as $\mathbf{X}^{-1}$, and a hat on a letter such as $\hat{\boldsymbol{\varepsilon}}$ is used to denote the estimate of its parameter $\boldsymbol{\varepsilon}$, will be employed throughout the thesis. Other notation not mentioned here will be introduced as the thesis develops.

1.2 Introduction

Regression analysis using the least squares approach is a widely used technique, and regression estimates are known to be easily affected by one or a few unusual observations and dependencies among the explanatory variables (a problem known as collinearity or multicollinearity). Unusual observations and collinearities when undetected, can cause problems in regression analysis, for example, by inflating the variance of the regression coefficients. This is often revealed when some of the regression assumptions are not satisfied (refer to Chatterjee and Hadi (1988)). The measures that have emerged as a result of trying to diagnose a failure of these assumptions, are commonly referred to as "regression diagnostics".

A lot of research has been done in the area of detecting unusual observations (refer to Barnett and Lewis (1994) for an extensive review of methods used to detect unusual observations), and most of the measures that have been developed work by measuring the

change in some parameter when one or more observations are deleted. The deletion statistics based on single observations may suffer from the masking effect, which occurs when some unusual observations are not detected because of the presence of another adjacent subset of unusual observations, and the swamping effect, which occurs when 'good' observations are incorrectly identified as unusual because of the presence of another adjacent subset of unusual observations (Hadi and Simonoff, 1993).

Group deletion statistics are rarely affected by these problems, although as discussed by Atkinson (1982), Kempthorne and Mendel (1990) and others, the difficulty associated with these methods is identifying the number of groups and number of observations to be considered for each group. For a detailed discussion of the methods that have been used to try and detect multiple outlying observations, refer to Ben-Gal (2005), Wang, Critchley and Smith (2003), Wisnowski, Montgomery and Simpson (2001) and Chiang (2008) and the authors cited therein.

Robust techniques are alternative diagnostics that have been developed to offset the effects of masking and swamping by minimising the impact of unusual observations in the estimation process (refer to Hadi and Simonoff (1993), Liang and Kvalheim (1996) for a survey of robust methods).

A lot of research has also been done in the area of understanding collinearities among the explanatory variables (refer to Belsley *et al.* (1980), Gunst and Mason (1977), and Hocking and Pendleton (1983)). Measures that are typically used to detect the presence of collinear relationships include *inter alia*: examining the magnitude of the condition indices, variance inflation factors, and the eigenvalues of $\mathbf{X}^T\mathbf{X}$ (Belsley *et al.*, 1980, Hocking and Pendleton, 1983, Mandel, 1982, Mansfield and Helms, 1982, Stewart, 1987). Estimators such as principal components regression and ridge regression, are typically used to compensate for the effects of collinearities (Mandel, 1982).

1.3 Objectives to the Research

The purpose of this thesis is to illustrate the advantages of using the singular value decomposition in multiple regression with special reference to problems of identifying unusual observations which may influence the regression coefficients and identifying the explanatory variables that are involved in collinear relationships. We focus specifically on the application of the matrix of the left singular vectors where it is appropriate.

The diagonal values of the hat matrix (that is, the h_i values) are used in regression analysis to identify outlying observations in the explanatory variables that may alter the fit

of the least squares line. The h_i values however, are known to suffer from the effects of masking and swamping, and in this thesis we demonstrate how decomposing a data matrix using the singular value decomposition technique can aid with the identification of observations that are being masked and swamped.

The residuals are also often examined to determine the observations that may have influenced the fit of the least squares regression line, because they take the response variable, $\mathbf{y}$, into account. The residuals, either in their raw or transformed form, are known to be a poor measure of fit since they may fail to identify the outlying observations when these observations are being accommodated by the least squares fit. Thus, we propose a measure that can be used in conjunction with the transformed residuals. The measure, which is based on the off-diagonal values of the hat ($\mathbf{H}_x$) matrix, defined in Section 1.1.1, determines the role that each observation plays in the displacement of other observations from the least squares fitted line.

The regression estimates such as the coefficients, are known to be easily affected by outlying observations, and measures such as DFBETAS (Belsley *et al.*, 1980), which are intended to measure the impact of an observation on the individual regression coefficients, are prone to the same problems as are the residuals and the diagonal values of the hat matrix since they are a function of the residuals, which are a poor measure of fit, and the diagonal values of the hat matrix, which may suffer from the masking and swamping effects. By decomposing the regression coefficients, we illustrate how to determine the outlying observations that may have a disproportionate effect in the determination of the individual regression coefficients.

A number of approaches have been proposed to identify the explanatory variables that are involved in collinear relationships, and to detect the coefficients that are most adversely affected. There are no thresholds to establish what a 'large' value is for the existing measures, and in this thesis we also develop a means of quantifying the meaning of 'large' for the magnitude of the eigenvectors of $\mathbf{X}^T\mathbf{X}$ that correspond to small singular values, and for the coefficient values that correspond to small singular values.

The last objective is to illustrate an alternative computational approach to principal components regression that is based on the SVD. Principal components regression is a form of biased estimator that is used when there is collinearity among the explanatory variables. Often only a subset of the principal axes are retained in the estimation of regression quantities, which may result in a decrease of the model's variance and/or decrease in the explanatory power of the model, and this may increase the bias of the regression quantities. We focus our attention particularly on employing values of the left singular vectors in expressing the principal components regression estimates where it is appropriate, and

to demonstrate the usefulness of decomposing the multiple correlation coefficient, R^2, to determine the importance of the axes in explaining the amount of variation in $\mathbf{y}$. We also propose a measure to determine the range of values for which prediction is reasonable when there is collinearity in the data.

Thus we consider two main areas of regression diagnostics. In the first part, we are concerned with the identification of outlying observations and determining which of the outlying observations influence the regression coefficients, and in the second part, we motivate for thresholds that may be used to identify the explanatory variables that are involved in collinear relationships, and regression coefficients that are most affected by the near dependencies.

1.4 Limitations

In this thesis, we focus only in the ways in which unusual data and the explanatory variables that are involved in collinear relationships can be detected, and identifying which of the unusual data and explanatory variables that are involved in collinear relationships have a disproportionate effect on the estimated regression coefficients. Thus, the extent to which the unusual data affects the regression coefficients is not assessed, and the remedial action that should be taken once influential observations have been identified is not considered. We also have not considered robust alternatives of the proposed measures.

1.5 Organization of the Thesis

The thesis is divided into eight chapters and three appendices. In the next chapter, we present the mathematical theory underlying the measures that are being introduced in the thesis. A procedure that aids with the identification of unusual observations in $\mathbf{X}$ is presented and exemplified in Chapter 3. In Chapter 4, we extend the procedure that is proposed in Chapter 3 to include the response variable, and also propose a measure which determines the role that each observation plays in the displacement of other observations from the least squares fitted line. The proposed measure should be used in conjunction with the Studentized residuals (or other transformed residuals) to identify outlying observations when we take the response variable into account. Chapter 5 then considers which of the outlying observations have a disproportionate effect on the regression coefficients. We make use of an artificial data set and three real data sets that have appeared in the literature on regression diagnostics in Chapters 3 to 5, to illustrate how the proposed measures operate.

In Chapter 6, we propose an alternative measure that may be used to identify the explanatory variables that are involved in collinear relationships, and thresholds for existing methods that are used for detecting the explanatory variables that are involved in collinear relationships are motivated and exemplified. We then illustrate the computational theory of principal components regression, focusing particularly on expressing the principal components regression estimates using the left singular vectors in Chapter 7, and conclude with a brief summary of the main contributions of this thesis in the final chapter.

Chapter 2

Graphical Display and Decomposition of the Principal Axes of X

In correspondence analysis (Greenacre, 1984), the key quantity known as inertia is a generalisation of the variance. Benzécri (1992) demonstrated how the inertia may be broken down along the principal axes, and further broken down into contributions of the rows and the columns of the data matrix. The same decomposition of the inertia can be applied to the total variance in principal components analysis (Greenacre, 1984). In this chapter, we apply the decomposition to an $n \times m$ data matrix ($\mathbf{X}$) with numeric data using the singular value decomposition of $\mathbf{X}$, to illustrate how the total variance of a matrix can be decomposed in numerous ways, that lead to an assortment of contributions towards the total variance of $\mathbf{X}$. The theory and notation introduced in this chapter will be used extensively in some of the chapters that follow.

2.1 Introduction

In correspondence analysis (Greenacre, 1984), the key quantity known as inertia is a generalisation of the variance. Benzécri (1992) demonstrated how the inertia may be broken down along the principal axes, and further broken down into contributions of the rows and the columns of a data matrix. The same decomposition of the inertia can be applied to the total variance in principal components analysis (Greenacre, 1984). In this chapter, we apply the same decomposition of the total variance to an $n \times m$ data matrix ($\mathbf{X}$) with numeric data using the singular value decomposition of $\mathbf{X}$, to illustrate how the total variance of the matrix can be decomposed in numerous ways, that lead to an assortment of contributions towards the total variance of $\mathbf{X}$.

This chapter is organised as follows. In the next section, we introduce the notation for plotting each of the rows (that is, the observations) and columns (that is, the variables) of $\mathbf{X}$, and some properties of the resulting graphical display. In section 2.3, we illustrate how to decompose the total variance of $\mathbf{X}$ into contributions from the observations and variables, and then further into contributions of the observations (or variables) to the axes and contributions of the axes to the observations (or variables).

2.2 Notation

Let $\mathbf{X} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^T$ be the SVD of the standardised matrix $\mathbf{X}$[1]. Put

$$\mathbf{F} = \mathbf{U}\mathbf{D}_\alpha \qquad \text{and} \qquad \mathbf{G} = \mathbf{V}\mathbf{D}_\alpha$$

The rows of $\mathbf{F} : n \times m$ and of $\mathbf{G} : m \times m$ provide coordinates for plotting each row of $\mathbf{X}$ (that is, each observation) and each column of $\mathbf{X}$ (that is, each variable) respectively on a new set of orthogonal axes, although the graphical display formed in this way is not a biplot (Gabriel, 1971), since $\mathbf{F}\mathbf{G}^T \neq \mathbf{X}$.

In consequence of the SVD,

$$\mathbf{X}\mathbf{V} = \mathbf{U}\mathbf{D}_\alpha = \mathbf{F}$$

so that

$$\mathbf{X}\mathbf{G}\mathbf{D}_\alpha^{-1} = \mathbf{F} \qquad \text{and} \qquad \mathbf{G} = \mathbf{X}^T\mathbf{F}\mathbf{D}_\alpha^{-\mathbf{1}} \tag{2.1}$$

These two results, (2.1), are known as the transition formulae, since the coordinates of the observations can be obtained from the coordinates of the variables, and conversely. The transition formulae provide a justification for simultaneously plotting the observations

[1]Throughout the thesis, we assume the matrix $\mathbf{X}$ to be standardised, unless otherwise stated.

and variables on the same set of axes (see Greenacre (1984) for further details in the context of correspondence analysis).

2.2.1 Properties of the Graphical Display

The properties of the above graphical display of **F** and **G** are (refer to Green and Carroll (1976) and Ludovic, Morineau and Warwick (1984) for algebraic details):

1. The distances between pairs of row points in the display (that is, the observations) are Euclidean:
   $\| \mathbf{f}_i - \mathbf{f}_{i'} \|^2 = (\mathbf{x}_i - \mathbf{x}_{i'})^T(\mathbf{x}_i - \mathbf{x}_{i'})$, where $\mathbf{f}_i$ is the ith row of **F**, and $\mathbf{x}_i$ is the ith row of **X**.

2. The Euclidean distance of the ith observation from the origin, $\| \mathbf{f}_i \|^2 = (\mathbf{x}_i^T\mathbf{x}_i)$ is the contribution of row i to the total variance. Thus if observation i is plotted far from the origin, it has, for one or more variables, coordinates that are far from the mean of that variable (or of those variables).

3. The distances between pairs of column points in the display (that is, the variables) are Euclidean:
   $\| \mathbf{g}_j - \mathbf{g}_{j'} \|^2 = (\mathbf{x}_j - \mathbf{x}_{j'})^T(\mathbf{x}_j - \mathbf{x}_{j'})$, where $\mathbf{g}_j$ is the jth row of **G**, and $\mathbf{x}_j$ is the jth column of **X**.

4. The Euclidean distance of the jth column point from the origin is proportional to the standard deviation of the jth variable, that is, $\| \mathbf{g}_j \|^2 = (\mathbf{x}_j^T\mathbf{x}_j) = (n-1)s_j^2$. But since **X** is standardised, the column points lie on a hypersphere.

5. The cosine of the angle between the vectors $\mathbf{g}_j$ and $\mathbf{g}_{j'}$ in the graphical display is the correlation between variables j and j':

   $$r_{jj'} = \frac{\mathbf{g}_j^T\mathbf{g}_{j'}}{\| \mathbf{g}_j \| \| \mathbf{g}_{j'} \|} = \cos\theta_{jj'}$$

   Thus variables that are highly correlated are either located close to one another ($\cos\theta_{jj'} \approx 1$ when $r_{jj'} \approx 1$), or far away from one another ($\cos\theta_{jj'} \approx -1$ when $r_{jj'} \approx -1$); whilst orthogonal (uncorrelated) variables are located a moderate distance from one another ($\cos\theta_{jj'} \approx 0$ when $r_{jj'} \approx 0$).

2.3 Decomposition of the Variance of X

To illustrate how to decompose the variance of **X**, we first observe the following equalities of the norm of **X**:

$$\begin{aligned} \| \mathbf{X} \|^2 &= \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}^2 = \text{tr}(\mathbf{X}^T\mathbf{X}) \\ &= \sum_{k=1}^{m} \alpha_k^2 \\ &= (n-1)\sum_{j=1}^{m} s_j^2 \end{aligned}$$

where α_k is the kth singular value, and $\sum_{j=1}^{m} s_j^2$ is the total variance of **X**.

We also note that

$$\mathbf{F}\mathbf{F}^T = \mathbf{U}\mathbf{D}_\alpha\mathbf{D}_\alpha\mathbf{U}^T = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^T\mathbf{V}\mathbf{D}_\alpha\mathbf{U}^T = \mathbf{X}\mathbf{X}^T \qquad \text{and}$$

$$\mathbf{G}\mathbf{G}^T = \mathbf{V}\mathbf{D}_\alpha\mathbf{D}_\alpha\mathbf{V}^T = \mathbf{V}\mathbf{D}_\alpha\mathbf{U}^T\mathbf{U}\mathbf{D}_\alpha\mathbf{V}^T = \mathbf{X}^T\mathbf{X}$$

So

$$\begin{aligned} \text{tr}(\mathbf{F}\mathbf{F}^T) = \text{tr}(\mathbf{G}\mathbf{G}^T) &= \| \mathbf{X} \|^2 \\ &= (n-1)\sum_{j=1}^{m} s_j^2 \end{aligned}$$

We thus have two ways of decomposing the total variance of **X**: **F** allows us to decompose the total variance of **X** in terms of the contributions of the observations, and **G** allows us to decompose the total variance of **X** in terms of the contributions of the variables.

The total variance of **X** may be decomposed further along the principal axes for each of the observations and variables (refer to Table 2.1):

Table 2.1
Decomposition of the variance. *Source*: Greenacre (1984)

		axes				
		1	2	$\cdots$	m	Total
observations	1	f_{11}^2	f_{12}^2	$\cdots$	f_{1k}^2	$\sum_{k=1}^m f_{1k}^2$
	2	f_{21}^2	f_{22}^2	$\cdots$	f_{2k}^2	$\sum_{k=1}^m f_{2k}^2$
	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
	n	f_{n1}^2	f_{n2}^2	$\cdots$	f_{nk}^2	$\sum_{k=1}^m f_{nk}^2$
	Total	α_1^2	α_2^2	$\cdots$	α_m^2	$\sum_{k=1}^m \alpha_k^2$
variables	1	g_{11}^2	g_{12}^2	$\cdots$	g_{1k}^2	$\sum_{k=1}^m g_{1k}^2$
	2	g_{21}^2	g_{22}^2	$\cdots$	g_{2k}^2	$\sum_{k=1}^m g_{2k}^2$
	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
	m	g_{m1}^2	g_{m2}^2	$\cdots$	g_{mk}^2	$\sum_{k=1}^m g_{mk}^2$

For the observations:

$$\begin{aligned}\text{Total variance} &= \frac{1}{n-1}\operatorname{tr}(\mathbf{F}\mathbf{F}^T) \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\left[\sum_{k=1}^{m} f_{ik}^2\right] \\ &= \frac{1}{n-1}\sum_{k=1}^{m}\left[\sum_{i=1}^{n} f_{ik}^2\right]\end{aligned}$$

where $\sum_{k=1}^{m} f_{ik}^2$ is the contribution of observation i to the total variance, and
$\sum_{i=1}^{n} f_{ik}^2$ is the contribution of the kth principal axis to the total variance.

Similarly, for the variables:

$$\begin{aligned}\text{Total variance} &= \frac{1}{n-1}\operatorname{tr}(\mathbf{G}\mathbf{G}^T) \\ &= \frac{1}{n-1}\sum_{j=1}^{m}\left[\sum_{k=1}^{m} g_{ik}^2\right] \\ &= \frac{1}{n-1}\sum_{k=1}^{m}\left[\sum_{j=1}^{m} g_{ik}^2\right]\end{aligned}$$

where $\sum_{k=1}^{m} g_{jk}^2$ is the contribution of variable j to the total variance, and
$\sum_{j=1}^{m} g_{jk}^2$ is the contribution of the kth principal axis to the total variance.

Below, we illustrate how the various decompositions of the total variance of $\mathbf{X}$ may be decomposed further, and it is the results generated here that will be used extensively in some of the chapters that follow.

2.3.1 Contributions of Observations or Variables to Axes

The quantity $\sum_{i=1}^{n} f_{ik}^2$, may be decomposed further into contributions of each observation to the variance of the kth principal axis. Thus

$$\frac{f_{ik}^2}{\sum_{i=1}^{n} f_{ik}^2} \qquad \text{for all} \quad i = 1, 2, \ldots, n$$

is interpreted as the proportion of the variance due to the kth principal axis that is explained by the ith observation.

Notice that

$$\frac{f_{ik}^2}{\sum_{i=1}^{n} f_{ik}^2} = u_{ik}^2 \qquad \text{for all} \quad i = 1, 2, \ldots, n$$

where u_{ik} is the value of the ith row and kth column of matrix $\mathbf{U}$ of the SVD of $\mathbf{X}$.

Therefore the squared values of matrix $\mathbf{U}$, that is the u_{ik}^2's, indicate the proportion of the variance due to the kth principal axis that is explained by the ith observation.

$\sum_{j=1}^{m} g_{jk}^2$ may also be decomposed further into contributions of each variable to the variance of the kth principal axis. Thus

$$\frac{g_{jk}^2}{\sum_{j=1}^{m} g_{jk}^2} \qquad \text{for all} \quad j = 1, 2, \ldots, m$$

is interpreted as the proportion of the variance due to the kth principal axis that is explained by variable j. Notice too that since $\sum_{j=1}^{m} g_{jk}^2 = \alpha_k^2$,

$$\frac{g_{jk}^2}{\sum_{j=1}^{m} g_{jk}^2} = v_{jk}^2$$

Therefore the squared values of matrix $\mathbf{V}$, that is the v_{jk}^2's, indicate the proportion of the variance due to the kth principal axis that is explained by the jth variable.

These contributions (of observations or variables to the axes), measure the importance of each point (observation or variable) in determining the direction of each axis. Thus large values imply that a point pulls or alters the direction of the axis away from where the majority of the other points are located.

2.3.2 Contributions of Axes to Observations or Variables

The same decomposition may be applied to the contributions of the axes to observation i or variable j.

Thus the quantity $\sum_{k=1}^{m} f_{ik}^2$ may be decomposed further into contributions of each of the m principal axes to the variance contributed by observation i. That is,

$$\frac{f_{ik}^2}{\sum_{k=1}^{m} f_{ik}^2} \qquad \text{for all} \quad k = 1, 2, \ldots, m$$

is interpreted as the proportion of the variance due to observation i that is explained by the kth principal axis.

Similarly, $\sum_{k=1}^{m} g_{jk}^2$ may be decomposed further into contributions of each of the m principal axes to the variance contributed by variable j. Thus

$$\frac{g_{jk}^2}{\sum_{k=1}^{m} g_{jk}^2} \qquad \text{for all} \quad k = 1, 2, \ldots, m$$

is interpreted as the proportion of the variance due to variable j that is explained by the

kth principal axis.

We observe that the contributions of the axes to an observation or variable are the squared cosines of the angles that the kth principal axis makes with the ith observation or the jth variable (refer to Figure 2.1). Large values of the squared cosine correspond to a small angle between the observation (or variable) and the axis, which indicates that the observation or variable lies in the direction of the axis or is highly correlated with the axis.

Notice too from the figure that $\cos^2\theta_{(kk')} = \dfrac{g_{jk}^2 + g_{jk'}^2}{\sum_{k=1}^{m} g_{jk}^2}$

is the squared cosine of the angle between the plane defined by the kth and the k'th principal axes and variable j. Clearly this result can be extended to the hyperplane defined by any $p(\leq m)$ principal axes.

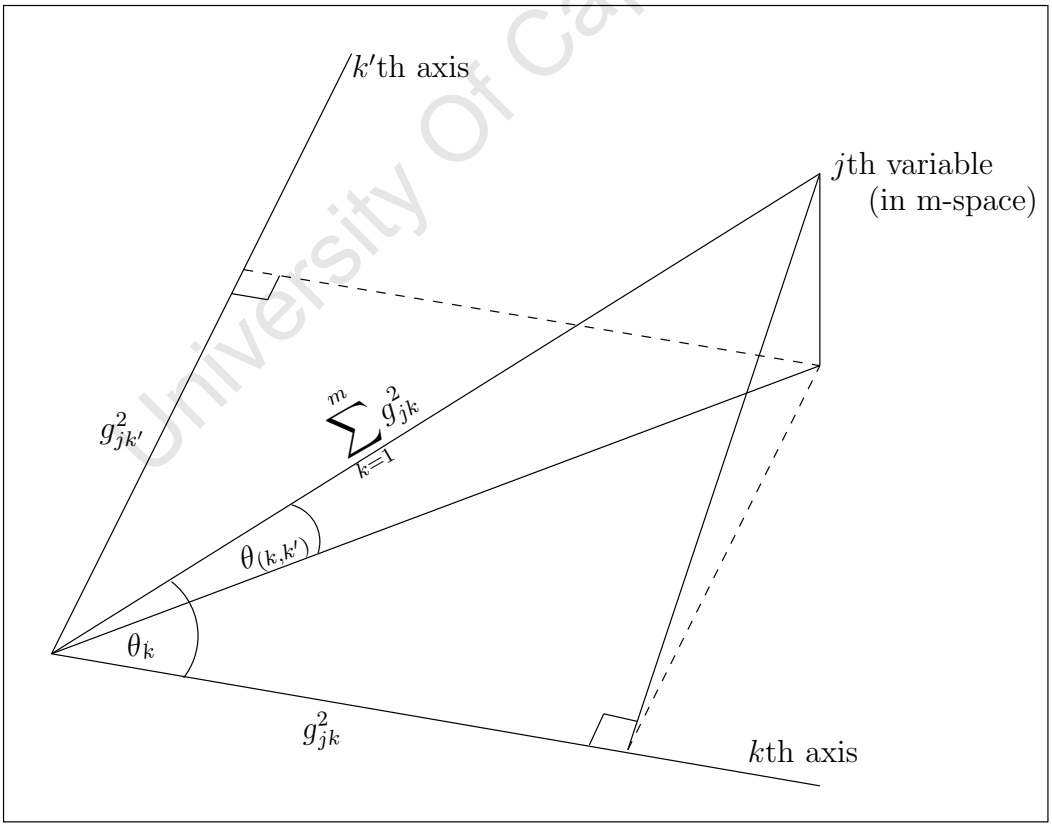


Figure 2.1: Illustration of squared cosine angle of the kth axis with the jth variable

2.4 Summary

In this chapter, we presented algebraic expressions of the decomposition of the total variance of **X** based on the singular value decomposition. We illustrated how the total variance of a data matrix, **X**, that contains numeric data may be decomposed into contributions of the rows and columns **X**, and further into contributions of the points (observations or variables) to the axes and contributions of the axes to the points. In some of the following chapters, we apply the various contributions of the total variance of **X** to propose measures that may be used in least squares regression to:

- aid with the identification of outlying observations in the explanatory variables.
- identify the explanatory variables that are involved in collinear relationships, and motivate for thresholds for existing methods that are used for detecting the explanatory variables that are involved collinear relationships.

In the next chapter, we consider the use of the diagonal values of the hat matrix (that is, the h_i values) in identifying unusual observations in **X**.

Chapter 3

Identifying Outlying Observations in X

Observations play an important role in determining the least squares line, and the regression quantities such as the coefficients, are known to be easily influenced by outlying observations. In this chapter, we consider a procedure to identify outlying observations in the explanatory variables (that is, based on the $\mathbf{X}$ matrix), using the diagonal values of the hat matrix (that is, the h_i values). The h_i values are known not to identify all leverage points correctly when there are multiple outliers in a data set, due to the effects of masking and swamping.

The procedure that we consider is adapted from a technique that is used in correspondence analysis to identify outlying points. We use the decompositions presented in Chapter 2, to identify the axis that an observation is outlying on. The advantage with this approach is that we are able to determine the type of outlier in the data, that is, whether the observation inflates variances because it is located in the first few axes, or whether the observation differs in multivariate structure because it is located in the last few axes or because the observation is not explained well by any axis. The masking and swamping effects that the h_i values are known to suffer from are minimized by examining the leverage-distance (L-D) plot.

An artificial data set, with various observations modified to be outliers, and three examples that have appeared in the literature on regression diagnostics are used to illustrate the proposed procedure.

3.1 Introduction

Observations play an important role in determining the least squares line, and the regression quantities such as the coefficients (refer to section 1.1, Chapter 1) are known to be easily affected by outlying observations. An observation is said to be an outlier if it appears to deviate markedly from the bulk of the data (Barnett and Lewis, 1994).

In this chapter, we make use of the matrix of left singular vectors **U** (that is, the eigenvectors of $\mathbf{X}\mathbf{X}^T$), and consider a procedure for identifying outlying observations in the explanatory variables, the so-called leverage points, that could potentially alter the fit of the least squares regression line away from the direction of the majority of the observations. The proposed procedure is based on the diagonal (h_i) values of the hat ($\mathbf{H_x}$) matrix to identify the leverage points, and is an adaptation of a technique used to identify outliers in correspondence analysis.

In correspondence analysis, points that are outlying tend to dominate the interpretation of one or more of the axes. Bendixen (1996) defines outlying points in correspondence analysis as those points that contribute highly to an axis, are well explained by the axis (that is, a point with high absolute and relative contributions on a particular axis), and are also located far away from the centre of the plot, whilst Hoffman and Franke (1986) suggest considering a point with a large absolute contribution and a large principal coordinate on a major principal axis to be an outlier.

The h_i values are known to sometimes suffer from the effects of masking and swamping when there are multiple outliers in a data set. Masking occurs when some outlying observations are not detected because of the presence of another adjacent subset of outlying observations, or because the subset of outlying observations mask themselves. The h_i values of the masked observations are small because the sample mean and variance are skewed towards the masked observations, which results in the distance of the masked observations from the mean being small. Swamping occurs when 'good' observations are incorrectly identified as outlying because of the presence of another subset of outlying observations. The h_i values of the swamped observations are large because the sample mean and variance are skewed away from the swamped observations, which results in the distance of the swamped observations from the mean being large (Coleman, 1977, Hadi and Simonoff, 1993), and (Acuna and Rodriguez, 2004, cited in Ben-Gal, 2005).

In order to off-set the masking and swamping effects, we propose first identifying the axis that explains each observation well, and then using a simple graphical display that is based on the diagonal values of the hat matrix and the distance of the observations from the origin. This graphical display, which we have called a 'leverage-distance' or just 'L-D'

plot, will aid in revealing not only the true leverage points, but also those observations that are being masked (h_i value small, but observation located far from the origin) and swamped (h_i value large, but observation located close to the origin).

This chapter is organised as follows. In the next section, we briefly review the use of the diagonal values of the hat matrix in identifying outlying observations, using an artificial data set to illustrate some of the problems often encountered when using the h_i values to diagnose outlying observations in the explanatory variables. We then consider an alternative way of expressing the diagonal values of the hat matrix that is based on the decompositions presented in Chapter 2, and also describe how the decomposed h_i values can be adapted to a technique used in correspondence analysis to identify outlying points in section 3.3. A procedure to identify leverage points is proposed in section 3.4, and an artificial data set is used to illustrate the various steps of the procedure.

The proposed procedure is then applied to three real data sets that have appeared in the literature on regression diagnostics in section 3.5, and we then end the chapter by discussing and highlighting the importance of the findings of the results presented in the chapter.

3.2 The Diagonal Values of the Hat Matrix

The hat matrix, $\mathbf{H_x} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{U}\mathbf{U}^T$ (Hoaglin and Welsch, 1978, Welsch and Kuh, 1977), is a symmetric and idempotent matrix, which determines the predicted values by putting a 'hat' on $\mathbf{y}$ (that is, $\hat{\mathbf{y}} = \mathbf{H}_x\mathbf{y}$). The values of h_i, the diagonal values of the hat matrix, indicate the influence of y_i on the fitted value $\hat{y}_i$, and are used to identify outlying observations among the explanatory variables.

The magnitude of each h_i ($0 \leq h_i \leq 1$) is used to indicate whether an observation is outlying or not; with a small (large) value indicating that the observation lies close to (far from) the majority of the other observations, taking the correlation structure of the explanatory variables into account. As discussed in Hoaglin and Kempthorne (1986), there are various guidelines that are used to label an observation as having 'high leverage'. h_i values greater than $2m/n$ have been proposed by Hoaglin and Welsch (1978); Velleman and Welsh (1981) in addition, suggested labelling observations with h_i values greater than $3m/n$ when $m > 6$ and $n - m > 12$, and Huber (1981, cited in Hoaglin and Kempthorne, 1986, Velleman and Welsh, 1981) proposed examining observations with h_i values greater than 0.2. Hoaglin and Welsch (1978), and Hoaglin and Kempthorne (1986) further suggested a stem-and-leaf display of the h_i values to identify observations with

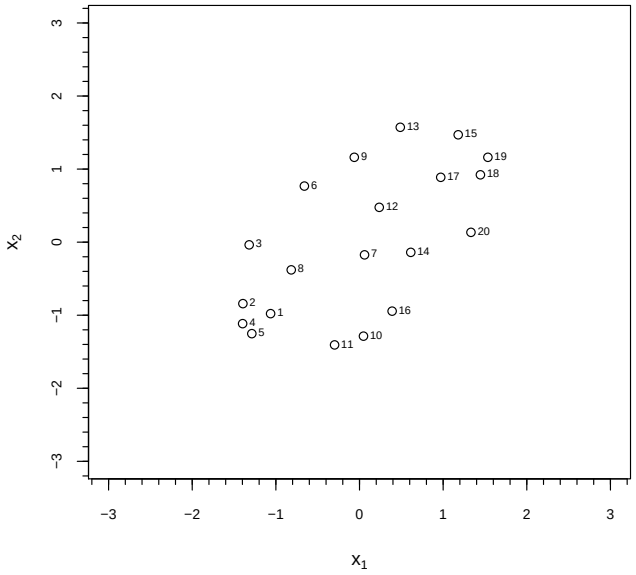
large h_i values, taking into account the various guidelines.

We illustrate the use of the diagonal values of the hat matrix with an artificial data set that is made up of two explanatory variables and twenty observations. Figure 3.1 illustrates various scatter plots of the artificial data set, and the stem-and-leaf displays of the corresponding h_i values are shown alongside the plots. In Figure 3.1(a), there are no leverage points, and in Figures 3.1(b) and 3.1(c), the positions of some observations have been modified, and the leverage points are indicated by black squares.

The h_i values can effectively identify individual outlying observations, but when a data set has multiple leverage points, the h_i values are known to sometimes fail to reveal outlying observations due to effects of masking and swamping.

Figure 3.2 illustrates more scatter plots of the artificial data set, and the stem-and-leaf displays of the corresponding h_i values are shown alongside the plots. In Figure 3.2(a), there are multiple leverage points indicated by black squares, although the leverage points are neither masked nor swamped. Figure 3.2(b) illustrates the masking effect. Observation 20 is outlying, but its h_i value is similar to those of other non-outlying observations in the sample. The observation is being masked by the presence of observations 18 and 19 that are close to it. In contrast, Figure 3.2(c) illustrates the swamping effect. Observation 13 is not outlying, although its h_i value is large relative to all other non-outlying observations in the sample. The observation is being swamped by the presence of observations 15 and 19 that are far from it.

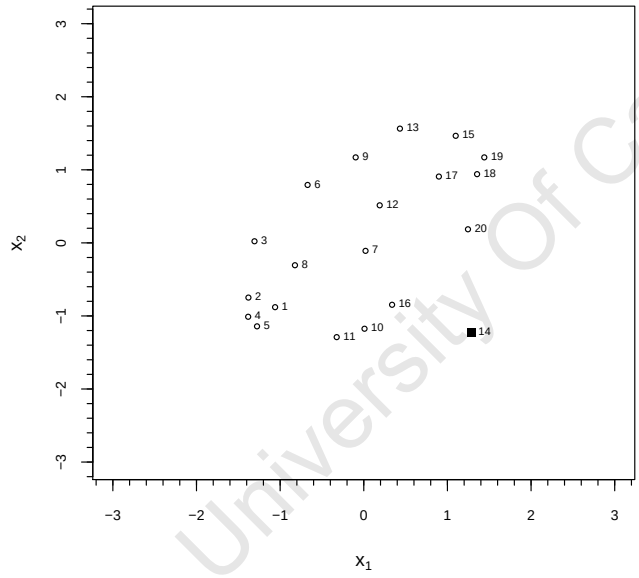
Figure 3.1: Scatter plots of two explanatory variables and the stem-and-leaf displays of the corresponding h_i values: Outliers are indicated by black square boxes.



(a)

stem	leaf
0	014467
1	00112233344555

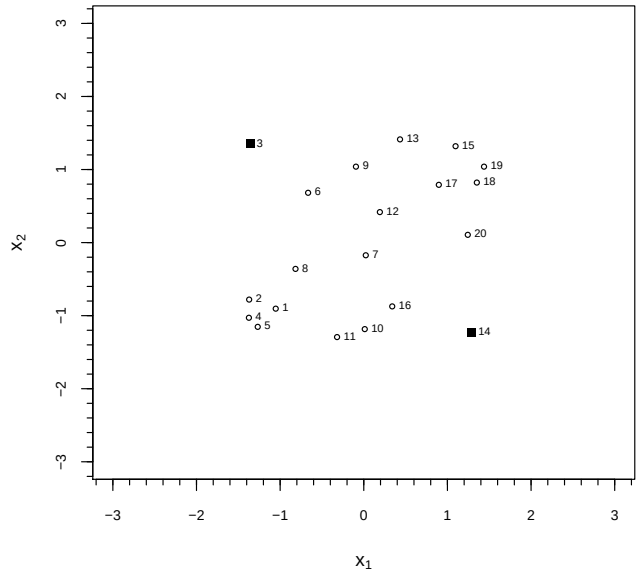
The decimal point is 1 digit to the left of the |



(b)

stem	leaf	observation
0	01467899	
1	00000112224	
2		
3	3	14

The decimal point is 1 digit to the left of the |



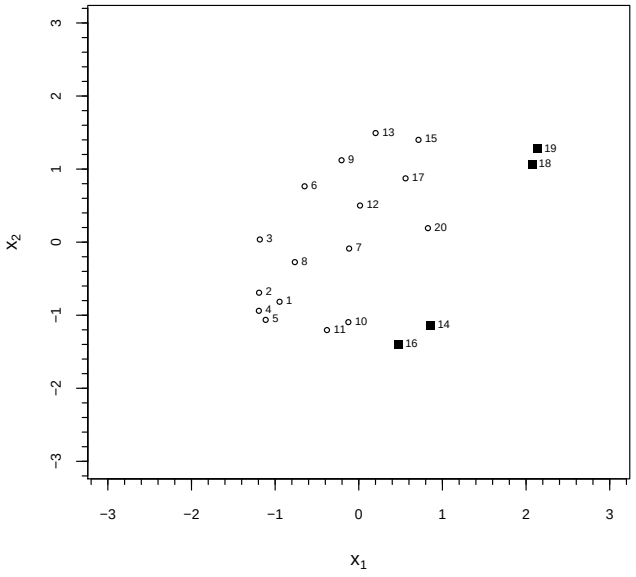
(c)

DATA1

stem	leaf	observation
0	01467777999	
1	0011223	
2	6	14
3	0	3

The decimal point is 1 digit to the left of the |

Figure 3.2: Scatter plots of two explanatory variables and the stem-and-leaf displays of the corresponding h_i values: Illustration of the effect of multiple outliers.

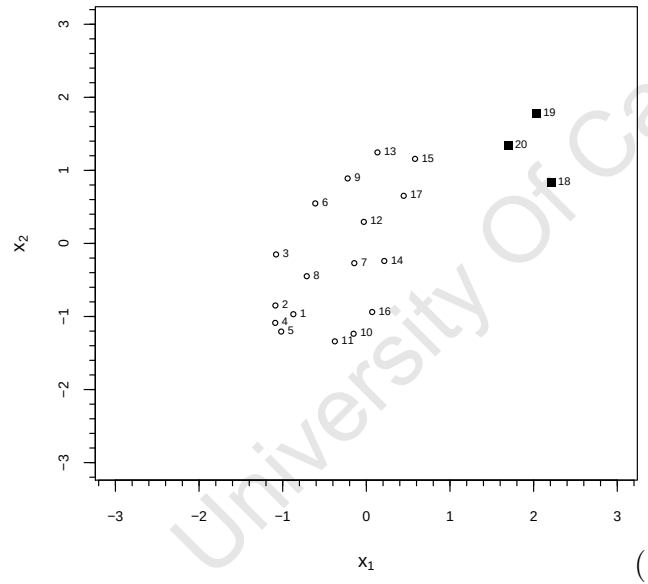


(a)

DATA2

stem	leaf	observation
0	02344678888	
1	00004	
2	0034	16,14,18,19

The decimal point is 1 digit to the left of the |

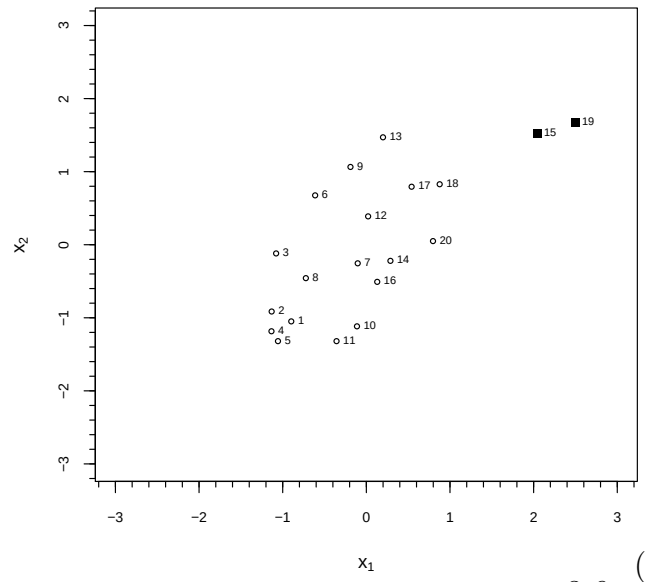


(b)

DATA3

stem	leaf	observation
0	0122356788	
1	01223445*	*20
2	3	19
3	2	18

The decimal point is 1 digit to the left of the |



(c)

DATA4

stem	leaf	observation
0	023334467789	
1	12366	
2	02	13,15
3	3	19

The decimal point is 1 digit to the left of the |

3.3 Decomposing the Diagonal Values of the Hat Matrix

In Chapter 2, section 2.3, we saw how the squared values of matrix **U**, that is, the u_{ik}^2's, indicate the proportion of the variance due to the kth principal axis that is explained by the ith observation, that is,

$$\frac{f_{ik}^2}{\sum_{i=1}^{n} f_{ik}^2} = u_{ik}^2 \qquad \text{for all} \quad i = 1, 2, \ldots, n$$

Note that

$$\sum_{k=1}^{m} u_{ik}^2 = \mathrm{h}_i \tag{3.1}$$

where h_i is the ith diagonal value of the hat matrix. Thus each h_i value represents the sum of the contributions of variance of each axis that is explained by an observation.

Recall that in correspondence analysis, points that are outlying contribute highly to the major axis, are well explained by the axis, and are also located far away from the centre of the plot. We adapt this technique that is used in correspondence analysis to identify outlying points and define a leverage point that is located in the direction of any kth axis, and not just the major axis. This is because when an axis is associated with a small singular value, the projection of the observations on that particular axis covers a smaller range than on an axis with a large singular value (refer to Mandel (1982)), and for the least squares fit, observations that are outlying on the minor axes could have a profound effect on the fit of the least squares regression line.

Due to the potential effects of masking and swamping, we start by defining an observation with a large h_i value to be outlying on a particular axis if it has a large value of

$$\frac{f_{ik}^2}{\sum_{i=1}^{n} f_{ik}^2} = u_{ik}^2 \equiv r_l \qquad \text{and} \qquad \frac{f_{ik}^2}{\sum_{k=1}^{m} f_{ik}^2} \equiv c_l$$

for some value k.

A large value of c_l implies that the ith observation is explained almost entirely by the kth principal axis, since the angle the observation makes with the axis is small. If this value

is less than 0.5, then the angle the observation makes with the axis is greater than 45°. Therefore we will consider any c_l value that is greater than 0.5 to be 'large'.

A large value of r_l, implies that the kth axis is, to a large extent, determined by (or dominated by) the ith observation. If an axis is determined equally by all the observations, then r_l will average $1/n$, therefore values greater than $2/n$ will be used to identify observations that dominate a particular axis.

Note that since a large value of r_l implies a large value of c_l, but the reverse is not true (Greenacre, 1984), and from (3.1), it is possible for an observation to have a large h_i value because the observation determines the direction of multiple axes, but may not be explained well by any of the axes. This type of observation is a leverage point that differs in structure from the bulk of data in the sample.

Notice also that the r_l values may also be masked when an observation is being masked, thus the proposed cut-off value of $2/n$ may not apply, although the masked observation should dominate the kth axis.

We illustrate in the next section how the three statistics together, that is h_i, r_l and c_l, can be used to provide a means of identifying leverage points in a data set where the effects of masking and swamping are minimised.

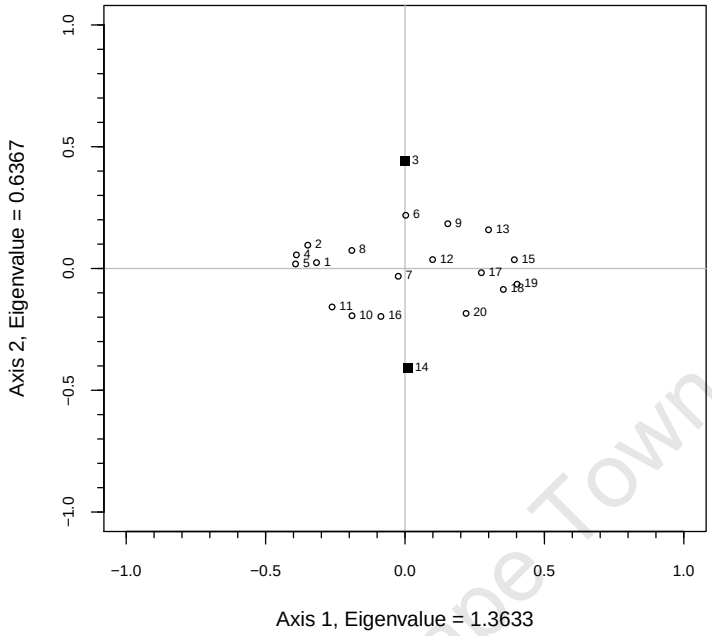
3.4 Procedure to Identify Leverage Points

The proposed procedure to identify leverage points, which we will illustrate by working through the artificial data from Figures 3.1(c) (p. 3-5), 3.2(a), 3.2(b) and 3.2(c) (p. 3-6), proceeds as follows (Note that in the explanation that follows, we will refer to the data from Figures 3.1(c), 3.2(a), 3.2(b) and 3.2(c) as DATA1, DATA2, DATA3 and DATA4 respectively. All computations on the data were performed in R (R Development Core Team, 2008), and the source codes written for the measures are included in Appendix C, section C.1 (p. C-1)):

1. Compute the r_l, c_l and h_i values.

   The values of h_i, r_l and c_l for DATA1 to DATA4 are shown in Figures 3.3(b) to 3.6(b). Also shown are the scatter plots of the data, produced using the row coordinates (**F** matrix) on an alternative set of orthogonal axes.

Figure 3.3: DATA1 - (a) Scatter plot of data from Figure 3.1(c) in the alternative orthogonal coordinate system. (b) Values of h_i, r_l, c_l and DIST for data from Figure 3.1(c).



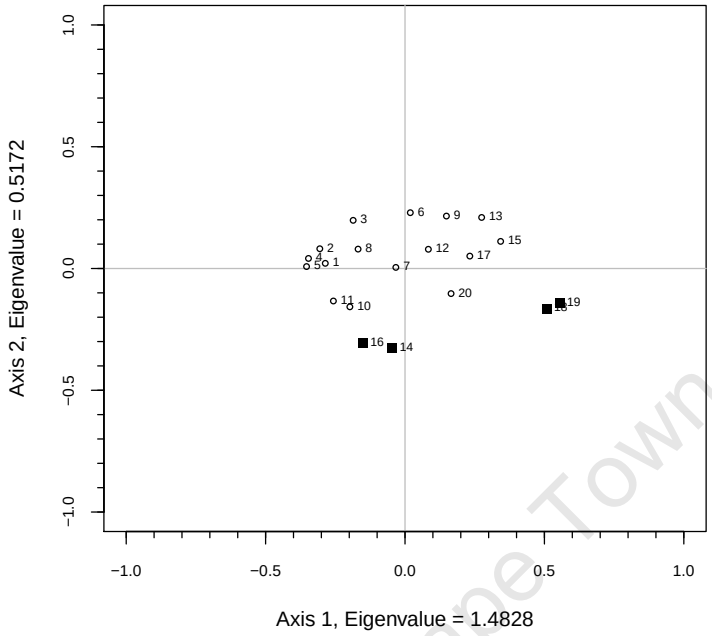
(a)

		r_l		c_l		
	h_i	Axis 1 (68.17%)	Axis 2 (31.83%)	Axis 1 (68.17%)	Axis 2 (31.83%)	DIST
1	0.075	0.074	0.001	0.994	0.006	0.101
2	0.103	0.089	0.014	0.930	0.070	0.131
3	0.303*	0.000	0.303*	0.000	1.000*	0.193
4	0.116	0.111	0.005	0.980	0.020	0.155
5	0.114	0.113	0.001	0.998	0.002	0.155
6	0.075	0.000	0.075	0.000	1.000	0.048
7	0.002	0.000	0.002	0.365	0.635	0.002
8	0.036	0.027	0.009	0.869	0.131	0.042
9	0.070	0.017	0.053	0.411	0.589	0.057
10	0.086	0.027	0.059	0.490	0.510	0.074
11	0.089	0.050	0.039	0.733	0.267	0.093
12	0.009	0.007	0.002	0.880	0.120	0.011
13	0.106	0.066	0.040	0.780	0.220	0.115
14	0.263*	0.000	0.263*	0.001	0.999*	0.167
15	0.115	0.113	0.002	0.992	0.008	0.155
16	0.066	0.005	0.061	0.162	0.838	0.046
17	0.055	0.055	0.000	0.996	0.004	0.075
18	0.103	0.091	0.012	0.944	0.056	0.132
19	0.125	0.118	0.007	0.975	0.025	0.166
20	0.088	0.035	0.053	0.585	0.415	0.082

* h_i large, $r_l > 0.1$ and $c_l > 0.5$.

(b)

Figure 3.4: DATA2 - (*a*) Scatter plot of data from Figure 3.2(*a*) in the alternative orthogonal coordinate system. (*b*) Values of h_i, r_l, c_l and DIST for data from Figure 3.2(*a*).



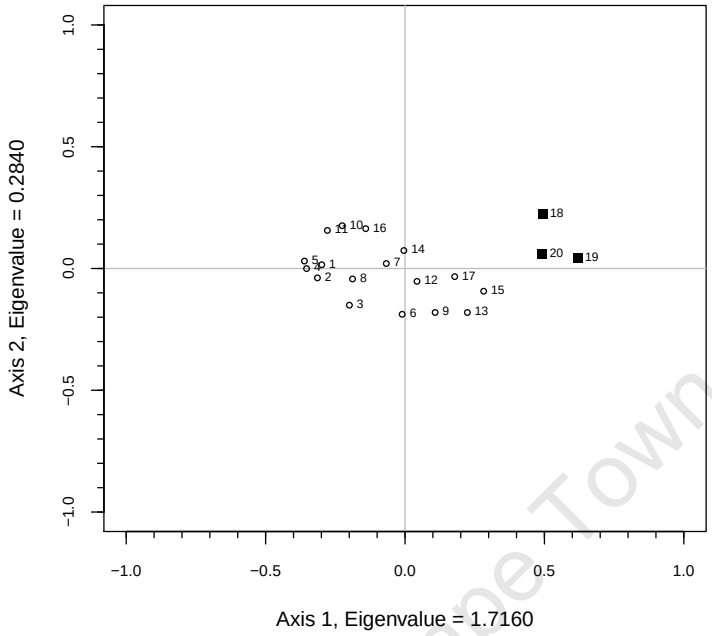
(*a*)

		r_l		c_l		
	h_i	Axis 1 (74.14%)	Axis 2 (25.86%)	Axis 1 (74.14%)	Axis 2 (25.86%)	DIST
1	0.056	0.055	0.001	0.995	0.005	0.082
2	0.076	0.063	0.013	0.934	0.066	0.100
3	0.099	0.023	0.076	0.469	0.531	0.074
4	0.084	0.081	0.003	0.986	0.014	0.122
5	0.084	0.084	0.000	0.999	0.001	0.125
6	0.102	0.000	0.102	0.007	0.993	0.053
7	0.001	0.001	0.000	0.982	0.018	0.001
8	0.031	0.019	0.012	0.817	0.183	0.035
9	0.105	0.015	0.090	0.322	0.678	0.068
10	0.074	0.026	0.048	0.612	0.388	0.064
11	0.079	0.045	0.034	0.787	0.213	0.084
12	0.017	0.005	0.012	0.530	0.470	0.013
13	0.136	0.051	0.085	0.632	0.368	0.119
14	0.205*	0.001	0.204*	0.019	0.981*	0.108
15	0.103	0.079	0.024	0.905	0.095	0.130
16	0.195*	0.015	0.180*	0.195	0.805*	0.116
17	0.041	0.036	0.005	0.954	0.046	0.057
18	0.226*	0.174*	0.052	0.905*	0.095	0.285
19	0.245*	0.207*	0.038	0.940*	0.060	0.327
20	0.039	0.018	0.021	0.720	0.280	0.038

* h_i large, $r_l > 0.1$ and $c_l > 0.5$.

(*b*)

Figure 3.5: DATA3 - (a) Scatter plot of data from Figure 3.2(b) in the alternative orthogonal coordinate system. (b) Values of h_i, r_l, c_l and DIST for data from Figure 3.2(b).



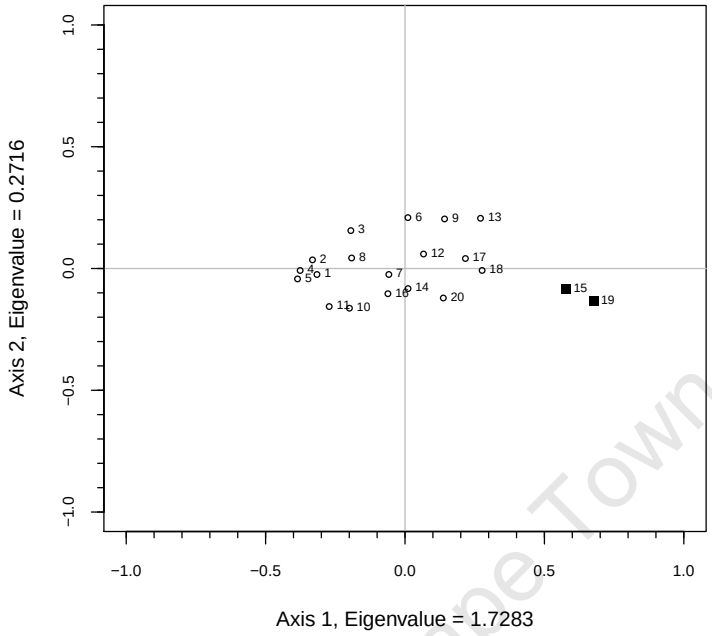
(a)

		r_l		c_l		
	h_i	Axis 1 (85.80%)	Axis 2 (14.20%) 2	Axis 1 (85.80%)	Axis 2 (14.20%)	DIST
1	0.053	0.052	0.001	0.997	0.003	0.089
2	0.063	0.058	0.005	0.985	0.015	0.100
3	0.103	0.023	0.080	0.637	0.363	0.063
4	0.073	0.073	0.000	1.000	0.000	0.125
5	0.079	0.076	0.003	0.993	0.007	0.131
6	0.124	0.000	0.124	0.003	0.997	0.035
7	0.004	0.003	0.001	0.916	0.084	0.005
8	0.027	0.021	0.006	0.951	0.049	0.037
9	0.122	0.007	0.115	0.264	0.736	0.044
10	0.139	0.030	0.109	0.621	0.379	0.082
11	0.131	0.045	0.086	0.760	0.240	0.102
12	0.011	0.001	0.010	0.399	0.601	0.005
13	0.144	0.029	0.115	0.605	0.395	0.083
14	0.019	0.000	0.019	0.003	0.997	0.005
15	0.077	0.046	0.031	0.901	0.099	0.088
16	0.106	0.012	0.094	0.426	0.574	0.047
17	0.023	0.019	0.004	0.966	0.034	0.033
18	0.320*	0.142*	0.178	0.829*	0.171	0.294
19	0.230*	0.224*	0.006	0.996*	0.004	0.386
20	0.153	0.141	0.012	0.986	0.014	0.246

* h_i large, $r_l > 0.1$ and $c_l > 0.5$.

(b)

Figure 3.6: DATA4 - (a) Scatter plot of data from Figure 3.2(c) in the alternative orthogonal coordinate system. (b) Values of h_i, r_l, c_l and DIST for data from Figure 3.2(c).



(a)

		r_l		c_l		
	h_i	Axis 1 (86.42%)	Axis 2 (13.58%)	Axis 1 (86.42%)	Axis 2 (13.58%)	DIST
1	0.060	0.058	0.002	0.994	0.006	0.100
2	0.069	0.064	0.005	0.989	0.011	0.111
3	0.111	0.022	0.089	0.609	0.391	0.062
4	0.082	0.082	0.000	1.000	0.000	0.142
5	0.093	0.086	0.007	0.988	0.012	0.150
6	0.161	0.000	0.161	0.002	0.998	0.044
7	0.004	0.002	0.002	0.848	0.152	0.004
8	0.028	0.021	0.007	0.952	0.048	0.039
9	0.164	0.012	0.152	0.327	0.673	0.062
10	0.121	0.023	0.098	0.599	0.401	0.066
11	0.133	0.043	0.090	0.752	0.248	0.098
12	0.016	0.003	0.013	0.555	0.445	0.008
13	0.199**	0.042	0.157	0.633	0.367	0.116
14	0.025	0.000	0.025	0.017	0.983	0.007
15	0.221*	0.194*	0.027	0.979*	0.021	0.342
16	0.041	0.002	0.039	0.260	0.740	0.014
17	0.033	0.027	0.006	0.965	0.035	0.049
18	0.044	0.044	0.000	0.999	0.001	0.077
19	0.331*	0.265*	0.066	0.962*	0.038	0.476
20	0.065	0.011	0.054	0.562	0.438	0.034

* h_i large, $r_l > 0.1$ and $c_l > 0.5$ on the kth axis.
** h_i large, but $r_l < 0.1$ and $c_l > 0.5$ (or the reverse) on the kth axis.

(b)

2. Identify observations with large h_i values using a stem-and-leaf display, and determine the direction that each observation is located on.

   The stem-and-leaf displays for DATA1 to DATA4 are shown in pages 3-5 and 3-6.

   In order to determine the direction of the axis that each observation is located on (or the axis that best explains each observation), we examine the values of c_l, since a large value of c_l implies that the angle between the observation and the axis is small, hence the observation is situated in the direction of that axis.

   In DATA1, observations 3 and 14 are outlying because they have large h_i values. Both leverage points determine the direction of and are well explained by (or lie in the direction of) the second axis. In DATA2, four observations are outlying, and they also have large h_i values. Observations 14 and 16 determine the direction of and lie in the direction of the second axis, whilst observations 18 and 19 determine the direction of and lie in the direction of the first axis. In DATA3 two observations, 18 and 19, have large h_i values. Both observations determine the direction of and are well explained by the first axis. In DATA4 three observations, 13, 15 and 19, have large h_i values. Observations 15 and 19 determine the direction of and lie in the direction of the first axis, whilst observation 13 appears to be responsible for determining the direction of the second axis, although it lies in the direction of the first axis. The axis that best explains each of the remaining observations for each data set can be seen by examining the c_l values.

   Note that examining the tables containing values of r_l and c_l not only enable us to see the axes that observations are outlying on, but also point towards the type of outliers in the data. As discussed in Gnanadesikan and Kettenring (1972), Hawkins and Fatti (1984), and Jolliffe (2002), the outlying observations that correspond to the first few axes are those that are "generally larger (or smaller) in overall size" compared to the rest of the observations, thus the observations inflate variances and covariances or correlations, since the first few axes explain most of the variation in **X**, whilst the outlying observations that correspond to the last few axes are those whose multivariate structure differs from the rest of the population. An alternative to identifying observations with large h_i values whose multivariate structure differs from the rest of the data in the sample is to examine the c_l values. These observations tend to be weakly correlated with all the axes (that is, $c_l < 0.5$ across all axes).

3. Compute the distance of each observation from the origin.

   In Chapter 2, section 2.2, we defined the distance of an observation from the origin

($\| \mathbf{f}_i^T \|^2$) to be equivalent to the contribution the observation makes to the total variance.

The last column ("DIST") of the tables in pages 3-9 to 3-12 shows the distance of each observation from the origin for each data set.

4. Identify observations that are located far from the origin.

   To aid with the identification of observations that are located far from the origin, we propose plotting a simple graphical display that is based on the diagonal values of the hat matrix and the distance of the observations from the origin. The graphical display, which we term the 'leverage-distance' or 'L-D' plot, is a plot of all observations with the h_i values plotted on the y-axis and the distance from the origin ($\| \mathbf{f}_i^T \|^2$) plotted on the x-axis, and recommend the use of different symbols for observations that are located in the direction of different axes. When n is large or when there is a concentration of scatter points from different axes around the same area, separate L-D plots for observations located on each axis may be easier to interpret.

   An advantage of the L-D plot is that it will aid not only in identifying the 'true' leverage points (where h_i and $\| \mathbf{f}_i \|^2$ are both large), but depending on the extent of deviation of an observation from the majority of data, the L-D plot will also aid in identifying observations that are being masked (h_i small but $\| \mathbf{f}_i \|^2$ large), as well as those observations that are being swamped (h_i large but $\| \mathbf{f}_i \|^2$ small).

   The L-D plots of DATA1 to DATA4 are shown in Figures 3.7 to 3.10. The leverage points of DATA1 and DATA2 can be seen easily from the plots. For DATA3, which has been modified to illustrate the masking effect, the L-D plot shows that the masked observation, 20, which has a low h_i value is also located far from the origin. From Figure 3.5(b), we see that observation 20 has large values of r_l and c_l on the first axis. For DATA4, which has been modified to illustrate the swamping effect, the L-D plot shows that the swamped observation, 13, which has a large h_i value is not located far from the origin, compared to other observations that are located in the direction of the same axis.

Figure 3.7: L-D plot of DATA1. Observation 3 and 4 are the leverage points, and are located on the second axis.

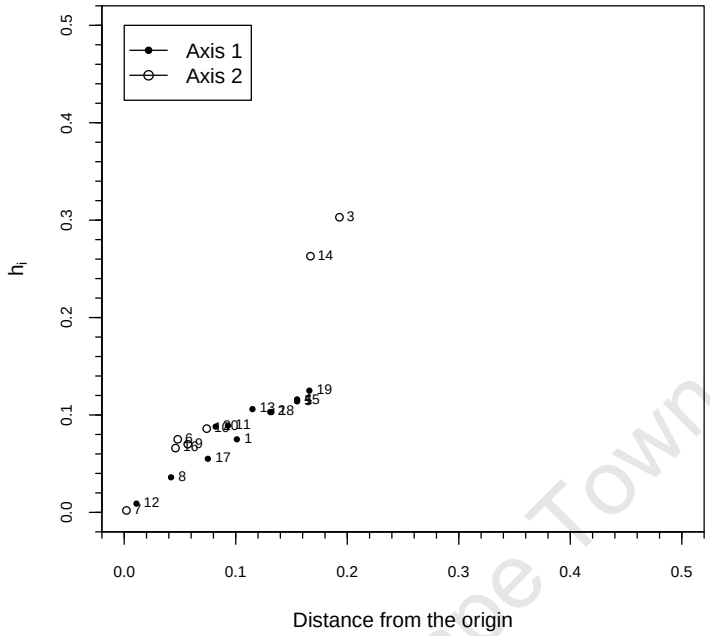


Figure 3.8: L-D plot of DATA2. Observations 14, 16, 18 and 19 are the leverage points, and two observations (18 and 19) are located on the first axis, whilst the other two observations, 14 and 16, are located on the second axis.

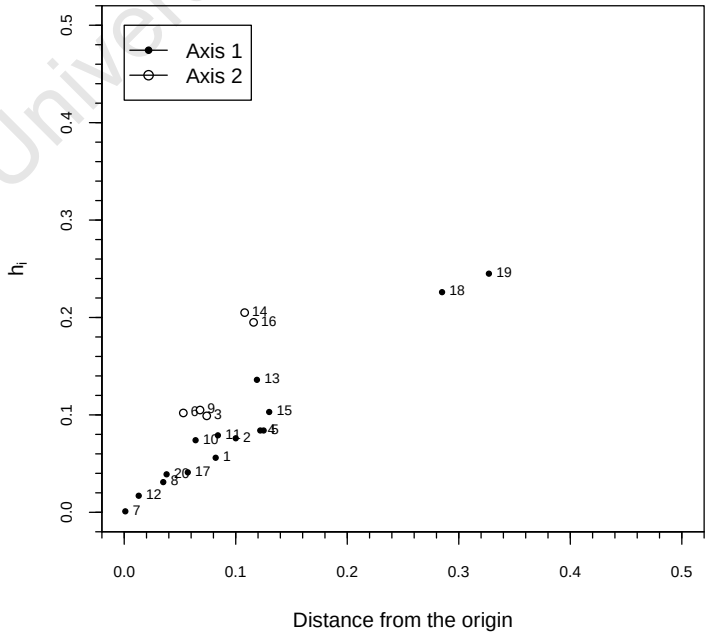


Figure 3.9: L-D plot of DATA3. Observations 18, 19, and 20 are the leverage points, although observation 20 is being masked since its h_i value is not large. All leverage points are located on the first axis.

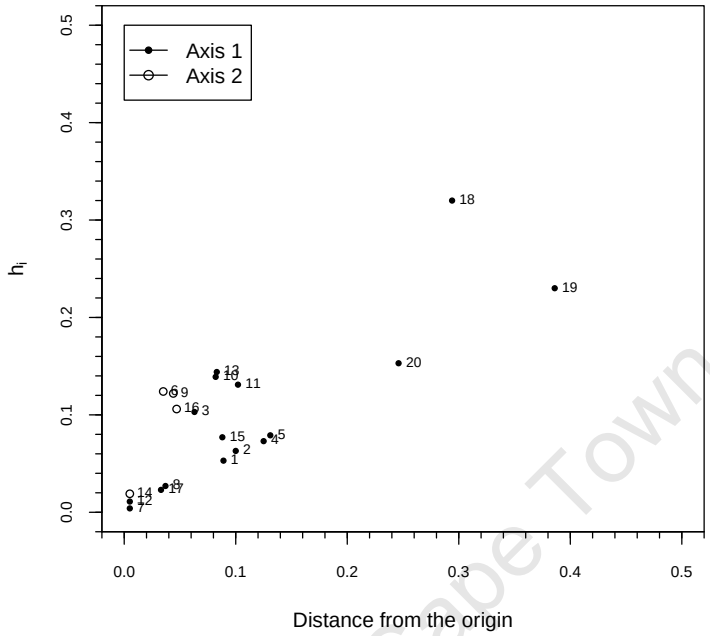
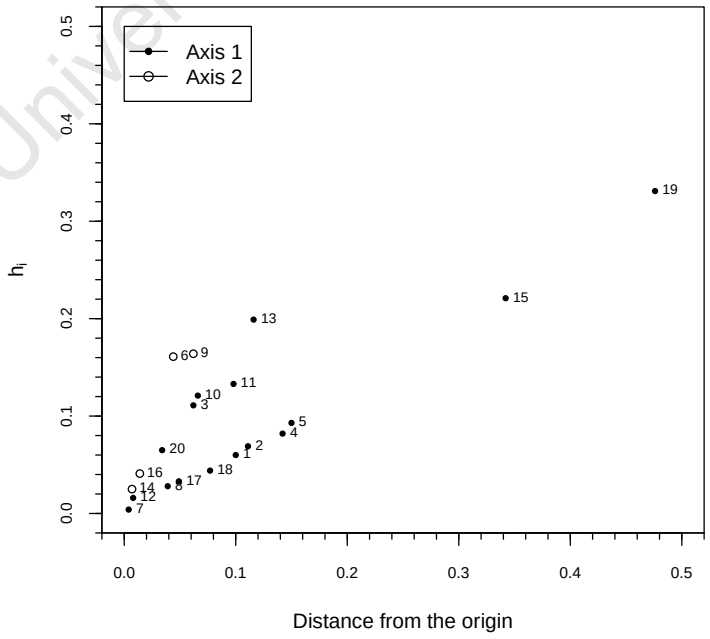


Figure 3.10: L-D plot of DATA4. Observations 15 and 19 are the leverage points, although observation 13 is being swamped since its h_i value is large. Both leverage points are located on the first axis.



Note that unless all the outlying observations are masking themselves, it may be difficult at times to differentiate between masked observations which do not deviate appreciably from the bulk of the data. In this instance, we recommend making use of the information contained in the off-diagonal values of the hat matrix (that is, the h_{ij}'s), since a large positive h_{ij} value indicates that observation i and observation j are situated on "the same side of the bulk of the cases nearly on the same line away from the centroid of the cases" (Gray and Ling, 1984). To illustrate, we consider the $\mathbf{H}_x$ matrix of DATA3 which is shown on page 3-18. Even though the L-D plot indicates that observation 20 is located far from the origin, since the masked observation is located on the same side as the two leverage points, examination of the h_{ij} values indicates a strong association between observations 18, 19 and 20.

Thus we see that the proposed procedure will identify all leverage points, and will minimise the effects of masking and swamping.

To summarize, the procedure discussed above entails the following steps:

Step 1 Compute the r_l, c_l and h_i values.

Step 2 Identify observations with large h_i values using a stem-and-leaf display, and determine the direction that each observation is located on (c_l).

Step 3 Compute the distance ($\| \mathbf{f}_i^T \|^2$) of each observation from the origin.

Step 4 Identify observations that are located far from the origin, using the L-D plot to identify the all 'true' leverage points, and if unsure about observations that are being masked, use the $\mathbf{H}_x$ matrix to determine their association with the identified leverage points.

Table 3.1

$\mathbf{H}_x$ matrix of DATA3. Entries are rounded values of $100 \times h_{ij}$.

Observations 18, 19 and 20 are the leverage points, and the h_{ij} values, together with the L-D plot (Figure 3.9), suggests that the three observations are located near each other. Observations 18 and 19 have a stronger association with each other compared to the association they have with observation 20.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	5	5	3	6	6	−1	1	3	−3	5	6	−1	−5	0	−5	3	−3	−7	−11	−8
2	5	6	6	6	6	3	1	4	0	2	3	−0	−2	−1	−4	0	−3	−12	−12	−10
3	3	6	10	4	3	10	−0	4	8	−7	−5	2	7	−4	2	−7	−0	−18	−9	−9
4	6	6	4	7	7	0	1	4	−2	5	6	−1	−5	0	−6	3	−4	−10	−13	−10
5	6	6	3	7	8	−2	2	3	−4	7	8	−1	−7	1	−7	5	−4	−8	−13	−10
6	−1	3	10	0	−2	12	−1	3	12	−12	−10	3	12	−5	6	−11	2	−15	−3	−4
7	1	1	−0	1	2	−1	0	0	−2	2	2	−1	−2	1	−2	2	−1	−0	−2	−2
8	3	4	4	4	3	3	0	3	2	−0	1	0	0	−1	−2	−1	−1	−9	−7	−6
9	−3	0	8	−2	−4	12	−2	2	12	−13	−12	4	13	−5	8	−11	3	−11	1	−1
10	5	2	−7	5	7	−12	2	−0	−13	14	13	−4	−14	5	−9	12	−4	7	−6	−3
11	6	3	−5	6	8	−10	2	1	−12	13	13	−4	−14	4	−10	11	−5	4	−8	−5
12	−1	−0	2	−1	−1	3	−1	0	4	−4	−4	1	4	−1	2	−3	1	−3	1	0
13	−5	−2	7	−5	−7	12	−2	0	13	−14	−14	4	14	−5	10	−12	4	−8	5	3
14	0	−1	−4	0	1	−5	1	−1	−5	5	4	−1	−5	2	−2	4	−1	6	1	1
15	−5	−4	2	−6	−7	6	−2	−2	8	−9	−10	2	10	−2	8	−8	4	1	9	6
16	3	0	−7	3	5	−11	2	−1	−11	12	11	−3	−12	4	−8	11	−3	9	−3	−1
17	−3	−3	−0	−4	−4	2	−1	−1	3	−4	−5	1	4	−1	4	−3	2	2	6	4
18	−7	−12	−18	−10	−8	−15	−0	−9	−11	7	4	−3	−8	6	1	9	2	32	**21**	**19**
19	−11	−12	−9	−13	−13	−3	−2	−7	1	−6	−8	1	5	1	9	−3	6	**21**	23	**19**
20	−8	−10	−9	−10	−10	−4	−2	−6	−1	−3	−5	0	3	1	6	−1	4	**19**	**19**	15

3.5 Illustrative Examples

In this section, we illustrate the procedure proposed above on three real data sets that have appeared in the literature on regression diagnostics.

Example 1: Hawkins, Bradu and Kass Data

The first data set that we consider is that of Hawkins, Bradu and Kass (1984) (the HBK artificial data set), which consists of one response variable, three explanatory variables and 75 observations. The data, which is available from the robustbase package found in R (Rousseeuw, P. and Croux, C. and Todorov, V. and Ruckstuhl, A. and Salibian-Barrera, M. and Maechler, M., 2007) has been reproduced in Appendix A (p. A-1), and is constructed so that the first fourteen observations are outliers, with the first ten observations "located at about the same place", whilst observations 11 to 14 make up the second group and are known to be good leverage points.

Table 3.3 shows the r_l, c_l, h_i and DIST values of the HBK data set for the first 20 observations (the complete table can be found in Appendix B (p. B-1)), and the stem-and-leaf display is shown in Table 3.2. From the tables, only observation 14, and possibly observations 12 and 13 are classified as leverage points because their h_i values are large compared to other observations in the data set. All three observations determine the direction of and are well explained by the first axis. The first 14 observations however, are all large contributors to the total variance (that is, they are located far from the origin).

Table 3.2
Hawkins, Bradu and Kass data: Stem-and-leaf display of the h_i values

stem	leaf	observation
0	000111111111111111222222222222222223333333333333444445555555555666677	
1	03	13, 12
2		
3		
4		
5	5	14

The decimal point is 1 digit to the left of the |.

Table 3.3
Hawkins, Bradu and Kass data: r_l, c_l, h_i and DIST values

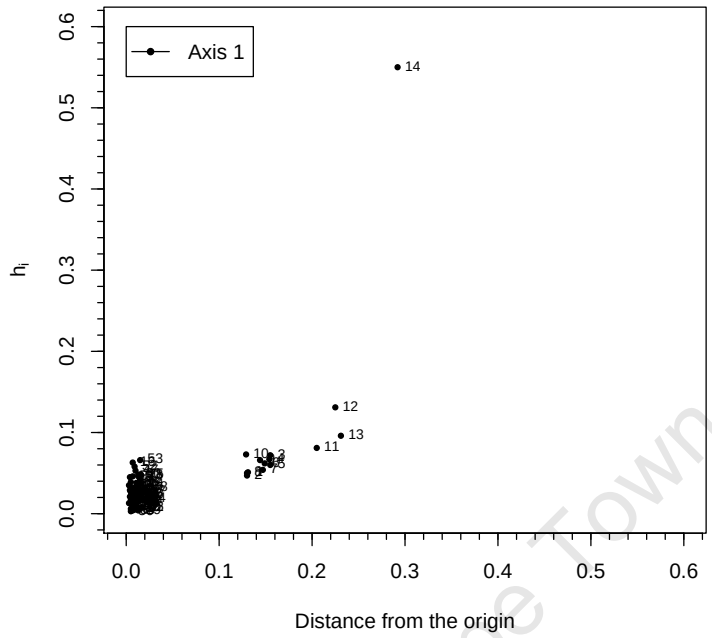
	r_l				c_l			
	Axis 1	Axis 2	Axis 3		Axis 1	Axis 2	Axis 3	
	(97.48%)	(1.89%)	(0.63%)	h_i	(97.48%)	(1.89%)	(0.63%)	DIST
1	0.045	0.005	0.001	0.051	0.998	0.002	0.000	0.131
2	0.045	0.001	0.001	0.047	0.999	0.000	0.000	0.130
3	0.053	0.007	0.012	0.072	0.996	0.003	0.001	0.155
4	0.053	0.002	0.012	0.067	0.998	0.001	0.001	0.154
5	0.053	0.000	0.007	0.060	0.999	0.000	0.001	0.155
6	0.051	0.011	0.000	0.062	0.996	0.004	0.000	0.149
7	0.050	0.003	0.001	0.054	0.999	0.001	0.000	0.147
8	0.044	0.002	0.004	0.050	0.999	0.001	0.001	0.130
9	0.049	0.001	0.016	0.066	0.997	0.001	0.002	0.144
10	0.044	0.001	0.028	0.073	0.995	0.001	0.004	0.129
11	0.070	0.002	0.009	0.081	0.999	0.001	0.001	0.205
12	0.077*	0.006	0.048	0.131*	0.995*	0.001	0.004	0.225
13	0.079*	0.000	0.017	0.096*	0.999*	0.000	0.001	0.231
14	0.095*	0.141	0.314	0.550*	0.952*	0.027	0.020	0.292
15	0.001	0.025	0.019	0.045	0.565	0.347	0.088	0.004
16	0.002	0.028	0.033	0.063	0.688	0.222	0.090	0.007
17	0.006	0.020	0.000	0.026	0.939	0.061	0.000	0.018
18	0.002	0.007	0.000	0.009	0.937	0.063	0.000	0.007
19	0.003	0.013	0.002	0.018	0.927	0.070	0.003	0.010
20	0.001	0.012	0.022	0.035	0.689	0.192	0.119	0.003
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

* h_i large, $r_l > 0.027$ and $c_l > 0.5$.

The L-D plot shown in Figure 3.11, indicates that the first fourteen observations deviate markedly from the other observations in the sample. Observations 1 to 10 which are clustered in the same area, and observation 11 are being masked. From Table 3.3, we see that all the eleven observations determine the direction of and are well explained by the first axis as well.

Thus all the leverage points in the Hawkins, Bradu and Kass (1984) data set are correctly identified when we use the L-D plot.

Figure 3.11: L-D plot for the Hawkins, Bradu and Kass data.



Example 2: Stack Loss Data

The second data set that we consider is the Stack Loss data set, which is taken from Brownlee (1965), p. 454, and is available from R (R Development Core Team, 2008), and has been reproduced in Appendix A (p. A-4). The data set comes from an experiment for the oxidation of ammonia to nitric acid conducted over 21 successive days, and consists of one response variable, the percentage of ammonia lost (stack. loss), and three explanatory variables, the flow of air to the plant (Air. Flow), the temperature of cooling water (Water. Temp), and the concentration of nitric acid in the absorbing liquid (Acid. Conc.).

Becker and Gather (1999) gave a summary of authors that have analysed this data set using various measures, and Meloun and Militký (2001) presented a survey using various diagnostic measures on the data set. Six observations, observations 1, 2, 3, 4, 17 and 21 (or combinations of), have been found to be outlying and/or influential.

We re-analysed the data using the procedure proposed above and the r_l, c_l, h_i and DIST values for the Stack Loss data are shown in Table 3.4, and the stem-and-leaf display is shown in Table 3.5.

Table 3.4
Stack Loss data: r_l, c_l, h_i and DIST values

	r_l				c_l			
	Axis 1 (71.10%)	Axis 2 (22.00%)	Axis 3 (6.90%)	h_i	Axis 1 (71.10%)	Axis 2 (22.00%)	Axis 3 (6.90%)	DIST
1	0.174*	0.062	0.018	0.254*	0.892*	0.099	0.009	0.415
2	0.162*	0.086	0.022	0.270*	0.849*	0.139	0.011	0.407
3	0.102	0.008	0.017	0.127	0.961	0.024	0.015	0.227
4	0.012	0.009	0.060	0.081	0.594	0.128	0.278	0.045
5	0.003	0.000	0.002	0.005	0.927	0.023	0.051	0.006
6	0.007	0.003	0.020	0.030	0.702	0.094	0.204	0.021
7	0.038	0.030	0.104	0.172	0.663	0.161	0.176	0.122
8	0.038	0.030	0.104	0.172	0.663	0.161	0.176	0.122
9	0.002	0.001	0.090	0.093	0.151	0.019	0.829	0.023
10	0.041	0.020	0.091	0.152	0.735	0.108	0.157	0.120
11	0.006	0.067	0.035	0.108	0.204	0.685	0.111	0.064
12	0.015	0.064	0.091	0.170	0.340	0.457	0.203	0.093
13	0.031	0.003	0.076	0.110	0.789	0.021	0.190	0.083
14	0.000	0.157	0.001	0.158	0.001	0.997	0.002	0.104
15	0.026	0.099	0.017	0.142	0.450	0.522	0.028	0.125
16	0.042	0.033	0.008	0.083	0.790	0.194	0.016	0.113
17	0.137*	0.227	0.000	0.364*	0.660*	0.339	0.000	0.442
18	0.074	0.028	0.011	0.113	0.885	0.103	0.013	0.179
19	0.053	0.026	0.048	0.127	0.805	0.124	0.071	0.139
20	0.019	0.013	0.000	0.032	0.822	0.178	0.000	0.050
21	0.018	0.034	0.185	0.237♯	0.389	0.225	0.386	0.099

* h_i large, $r_l > 0.095$ and $c_l > 0.5$.

♯ h_i large, and $r_l > 0.095$ on the third axis but $c_l < 0.5$ on all three axes.

Table 3.5
Stack Loss data: Stem-and-leaf display of the h_i values

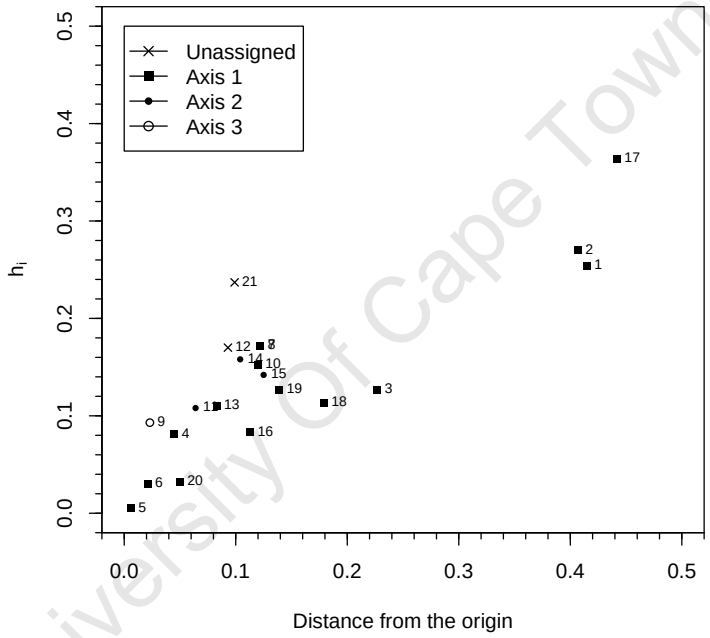
stem		leaf	observation
0	|	0 3 3 8 8 9	
1	|	1 1 1 3* 3 4 5 6 7 7 7	*3
2	|	4 5 7	21, 1, 2
3	|	6	17

The decimal point is 1 digit to the left of the |.

From the tables, observations 1, 2, 17 and 21 have large h_i values, and observations 1, 2 and 17 determine the direction of and are well explained by the first axis. Observation 21 on the other hand, is not explained well by any of the three axes, but is mostly responsible for determining the direction of the third axis (which explains only 6.9% of the total variance). This indicates that observation 21 does not fit the structure of the bulk of the data, and we will therefore treat the observation as a leverage point.

The DIST column of Table 3.4 indicates that observations 1, 2, 3 and 17 are also large contributors to the total variance. From Figure 3.12, observations 1, 2 and 17 which have large h_i values, are also located far from the origin. Observation 3 appears to be masked because it has a small h_i value but is located far from the origin relative to other observations. Since this observation does not deviate appreciably from the rest of the other observations on the L-D plot, we use the $\mathbf{H}_x$ matrix to confirm whether it is indeed being masked.

Figure 3.12: L-D plot for the Stack Loss data.



From the $\mathbf{H}_x$ matrix shown on the next page for the Stack Loss data (note that only the first five columns are shown, the complete $\mathbf{H}_x$ matrix can be found in Appendix A (p. B-9)), observations 1, 2 and 3 appear to be located on the same side of the axis, although the association is stronger between observations 1 and 2 compared to the association the two observations have with observation 3. Observation 17 is not strongly associated with any of these three observations since its h_{ij} value with all the three observations is very small and negative. This suggests that observation 17 is not located near observations 1, 2 and 3.

Thus the leverage points identified in the Stack Loss data are observations 1, 2, 3, 17 and 21.

Table 3.6
$\mathbf{H}_x$ matrix – Stack Loss data.
(Entries are rounded values of $100 \times h_{ij}$)

	1	2	3	4	5	⋯
1	25	**26**	**17**	4	2	⋯
2	**26**	27	**17**	4	2	⋯
3	**17**	**17**	13	1	1	⋯
4	4	4	1	8	2	⋯
5	2	2	1	2	0	⋯
6	3	3	1	5	1	⋯
7	−1	−2	0	8	2	⋯
8	−1	−2	0	8	2	⋯
9	−2	−2	−2	8	1	⋯
10	−1	0	−1	−8	−2	⋯
11	−7	−8	−2	−8	−2	⋯
12	−7	−8	−2	−11	−2	⋯
13	−2	−1	−2	−8	−2	⋯
14	−9	−11	−3	−4	−1	⋯
15	−16	−18	−10	−2	−1	⋯
16	−14	−15	−9	−2	−1	⋯
17	−4	−1	−8	1	−1	⋯
18	−9	−8	−9	1	−1	⋯
19	−8	−8	−9	4	−0	⋯
20	−3	−2	−3	−0	−1	⋯
21	7	6	8	−11	−1	⋯

Example 3: Health Club Data

The third data set that we consider is the Health Club data (refer to Appendix A, p. A-5), which appears in Chatterjee and Hadi (1988), p. 129, and originates from health records of 30 employees who were regular members of a company's health club. The data set consists of four explanatory variables: weight in pounds, resting pulse rate per minute, arm and leg strength, time (in seconds) in a quarter-mile trial run, and one response variable, time in seconds in a one-mile run.

Chatterjee and Hadi (1988) ran a model including the intercept, and found observation 23 to be outlying in the explanatory variables, observation 30 to be outlying in the residuals, and observation 28 to be influential on the coefficients based on the influence curve. According to the volume of confidence ellipsoids measure, observation 23 was the most influential, whilst observations 28 and 30 had the most influence on the likelihood function. Zhao, Lee and Hui (1994) on the other hand applied biased-corrected influence diagnostics on the data set, and found observation 23 to influence the intercept, whilst observation 28 exerted its influence on all the coefficients. Observations 8 and 30 appeared to be influential on the fourth coefficient.

We re-analysed the data using the procedure proposed above and the r_l, c_l, h_i and DIST values for the Health Club data are shown in Table 3.7, and the stem-and-leaf display is

shown in Table 3.8.

Table 3.7
Health Club data: r_l, c_l, h_i and DIST values

	r_l					c_l				
	Axis 1 (61.14%)	Axis 2 (25.33%)	Axis 3 (9.25%)	Axis 4 (4.28%)	h_i	Axis 1 (61.14%)	Axis 2 (25.33%)	Axis 3 (9.25%)	Axis 4 (4.28%)	DIST
1	0.101*	0.013	0.033	0.069	0.216*	0.869*	0.046	0.043	0.042	0.284
2	0.035	0.027	0.022	0.010	0.094	0.699	0.221	0.067	0.013	0.122
3	0.009	0.012	0.001	0.035	0.057	0.556	0.293	0.009	0.143	0.041
4	0.016	0.005	0.025	0.000	0.046	0.734	0.091	0.174	0.000	0.053
5	0.000	0.020	0.011	0.026	0.057	0.007	0.701	0.138	0.154	0.029
6	0.012	0.043	0.007	0.016	0.078	0.370	0.560	0.035	0.034	0.077
7	0.004	0.027	0.010	0.009	0.050	0.250	0.625	0.089	0.037	0.043
8	0.018	0.163*	0.000	0.006	0.187*	0.210	0.784*	0.000	0.005	0.210
9	0.036	0.070	0.003	0.006	0.115	0.545	0.443	0.006	0.006	0.159
10	0.002	0.087	0.000	0.006	0.095	0.052	0.937	0.000	0.011	0.094
11	0.023	0.014	0.033	0.110	0.180**	0.561	0.135	0.120	0.184	0.102
12	0.015	0.044	0.005	0.002	0.066	0.437	0.536	0.023	0.004	0.083
13	0.012	0.187*	0.053	0.012	0.264*	0.121	0.789*	0.082	0.009	0.240
14	0.015	0.001	0.091	0.021	0.128	0.480	0.020	0.452	0.048	0.075
15	0.023	0.003	0.002	0.009	0.037	0.908	0.053	0.014	0.024	0.061
16	0.017	0.023	0.082	0.076	0.198♯	0.379	0.216	0.283	0.121	0.108
17	0.029	0.001	0.154	0.030	0.214**	0.524	0.010	0.427	0.039	0.134
18	0.068*	0.022	0.001	0.103	0.194*	0.808*	0.106	0.001	0.085	0.207
19	0.006	0.002	0.004	0.001	0.013	0.777	0.129	0.089	0.006	0.018
20	0.053	0.003	0.013	0.012	0.081	0.932	0.019	0.034	0.015	0.140
21	0.016	0.029	0.035	0.006	0.086	0.482	0.353	0.154	0.012	0.084
22	0.054	0.005	0.024	0.006	0.089	0.902	0.032	0.059	0.007	0.147
23	0.244*	0.128	0.036	0.071	0.479*	0.794*	0.172	0.018	0.016	0.753
24	0.081	0.001	0.003	0.025	0.110	0.970	0.005	0.005	0.021	0.205
25	0.024	0.001	0.013	0.001	0.039	0.914	0.009	0.076	0.002	0.065
26	0.023	0.003	0.001	0.025	0.052	0.888	0.041	0.003	0.068	0.063
27	0.000	0.050	0.060	0.023	0.133	0.004	0.659	0.285	0.051	0.078
28	0.034	0.003	0.139	0.177	0.353♯	0.498	0.018	0.304	0.179	0.169
29	0.024	0.000	0.050	0.055	0.129	0.677	0.004	0.212	0.107	0.088
30	0.004	0.015	0.089	0.052	0.160♯	0.151	0.222	0.492	0.134	0.067

* h_i large, $r_l > 0.067$ and $c_l > 0.5$.

♯ h_i large, and $r_l > 0.067$ on one of the axis but $c_l < 0.5$ on all four axes.

** h_i large, but $r_l < 0.067$ and $c_l > 0.5$ (or the reverse) on the kth axis.

Table 3.8
Health data: Stem-and-leaf display of the h_i values

stem	leaf	observation
0	14455566788999	
1	0113336*8*9*9*	*30, *11, *8 ,*18
2	0226	16, 1, 17 ,13
3	5	28
4	8	23

The decimal point is 1 digit(s) to the left of the |

From the tables, observations 1, 13, 16, 17, 23 and 28 have large h_i values, while observations 8, 11, 18 and 30 are also suspect because they have relatively large h_i values compared to the rest of the data in the sample. Observations 1, 18 and 23 determine the direction of and are well explained by the first axis; observations 8 and 13 determine the direction of and are well explained by the second axis; observations 11 and 17 determine mostly the direction of the fourth axis, but are well explained by the first axis, and observations 16, 28 and 30 have large h_i values but are not explained well by any one of the four axes. Of the observations with large h_i values, observations 11, 17 and 30 do not appear to contribute much to the total variance.

From Figure 3.13, we see that observations 1, 8, 13, 18 and 23 are located far from the origin; observations 11 and 17, which are located in the direction of the first axis, are not located far from the origin, thus these observations appear to be swamped; observations 16, 28 and 30, do not lie in the direction of any axis, hence they do not fit the structure of the bulk of the data, and we will therefore treat these three observations as leverage points.

Notice that observation 24 has a low h_i value, but is located far from the origin, thus this observation appears to be masked. The $\mathbf{H}_x$ matrix for the Health Club data is shown on page 3-28. Observation 24 appears to have a strong association with observation 28, thus the presence of observation 28 may be masking observation 24.

Thus the identified leverage points for the Health Club data set are observations 1, 8, 13, 16, 18, 23, 24, 28 and 30.

Figure 3.13: L-D plot for the Health Club data.

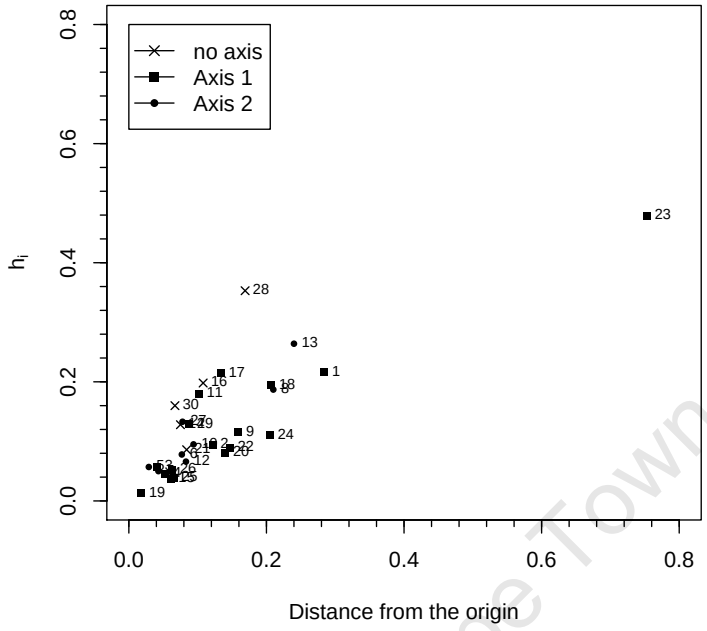


Table 3.9

$\mathbf{H}_x$ matrix – Health Club data. Entries are rounded values of $100 \times h_{ij}$.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	22	1	4	0	−4	−1	−5	2	0	0	−6	−6	0	−5	2	−4	−3	15	−1	−3	−1
2	1	9	6	6	−3	−5	−1	−8	−9	−5	−7	−2	7	0	−4	−2	7	−4	3	6	7
3	4	6	6	3	−4	−5	−3	−4	−6	−2	−7	0	7	3	−3	4	−1	2	1	4	2
4	0	6	3	5	0	−2	1	−5	−5	−2	−4	−1	1	−3	−2	−4	8	−4	2	4	6
5	−4	−3	−4	0	6	6	5	4	5	3	2	1	−10	−6	0	−5	7	−3	0	0	1
6	−1	−5	−5	−2	6	8	5	9	8	6	2	2	−13	−6	1	−4	4	2	−1	−2	−2
7	−5	−1	−3	1	5	5	5	5	3	4	−2	3	−10	−4	−2	−2	7	−2	1	2	1
8	2	−8	−4	−5	4	9	5	19	13	13	−5	7	−18	−2	−1	7	−3	12	−3	0	−9
9	0	−9	−6	−5	5	8	3	13	11	8	3	3	−13	−3	2	1	−3	6	−3	−4	−7
10	0	−5	−2	−2	3	6	4	13	8	10	−6	6	−13	−1	−2	6	−1	8	−2	1	−6
11	−6	−7	−7	−4	2	2	−2	−5	3	−6	18	−4	4	−1	7	−8	−4	−9	−1	−10	−1
12	−6	−2	0	−1	1	2	3	7	3	6	−4	7	−6	4	−3	8	−1	1	−1	4	−4
13	0	7	7	1	−10	−13	−10	−18	−13	−13	4	−6	26	12	1	5	−11	−6	1	−1	4
14	−5	0	3	−3	−6	−6	−4	−2	−3	−1	−1	4	12	13	−2	14	−12	0	−1	1	−5
15	2	−4	−3	−2	0	1	−2	−1	2	−2	7	−3	1	−2	4	−4	−3	0	−1	−5	−1
16	−4	−2	4	−4	−5	−4	−2	7	1	6	−8	8	5	14	−4	20	−13	7	−2	4	−8
17	−3	7	−1	8	7	4	7	−3	3	−1	−4	−1	−11	−12	−3	−13	21	−9	4	7	10
18	15	−4	2	−4	−3	2	−2	12	6	8	−9	1	−6	0	0	7	−9	19	−3	−1	−8
19	−1	3	1	2	0	−1	1	−3	−3	−2	−1	−1	1	−1	−1	−2	4	−3	1	2	3
20	−3	6	4	4	0	−2	2	0	−4	1	−10	4	−1	1	−5	4	7	−1	2	8	3
21	−1	7	2	6	1	−2	1	−9	−7	−6	−1	−4	4	−5	−1	−8	10	−8	3	3	9
22	−9	2	2	0	−2	−3	0	0	−2	1	−4	6	4	8	−4	11	−3	−3	0	5	−2
23	30	2	5	0	−7	−4	−10	−6	−3	−6	−1	−14	9	−6	6	−10	−7	17	−1	−8	1
24	4	−8	−6	−5	3	5	0	4	8	1	10	−3	−5	−4	6	−6	−4	3	−2	−9	−4
25	−3	5	2	4	1	−1	2	−3	−4	−1	−3	0	0	−2	−3	−2	8	−5	2	5	5
26	−8	2	−1	2	2	0	2	−5	−3	−3	3	0	2	−1	−1	−4	6	−10	2	2	4
27	−6	−1	−1	−2	−3	−5	−5	−10	−4	−8	12	−3	14	6	4	0	−8	−9	0	−5	1
28	−11	−12	−10	−8	2	3	−2	−3	7	−4	24	−3	4	3	9	−4	−11	−10	−3	−13	−5
29	−3	2	5	−1	−7	−7	−4	−1	−4	0	−6	4	10	12	−3	15	−10	3	−1	4	−4
30	−1	5	−1	6	5	2	4	−8	−4	−5	3	−5	−3	−11	0	−16	16	−10	4	2	10

Table 3.9
$\mathbf{H}_x$ matrix – Health Club data . . . continued.

	22	23	24	25	26	27	28	29	30
1	−9	30	4	−3	−8	−6	−11	−3	−1
2	2	2	−8	5	2	−1	−12	2	5
3	2	5	−6	2	−1	−1	−10	5	−1
4	0	−0	−5	4	2	−2	−8	−1	6
5	−2	−7	3	1	2	−3	2	−7	5
6	−3	−4	5	−1	−0	−5	3	−7	2
7	0	−10	−0	2	2	−5	−2	−4	4
8	0	−6	4	−3	−5	−10	−3	−1	−8
9	−2	−3	8	−4	−3	−4	7	−4	−4
10	1	−6	1	−1	−3	−8	−4	0	−5
11	−4	−1	10	−3	3	12	24	−6	3
12	6	−14	−3	0	−0	−3	−3	4	−5
13	4	9	−5	−0	2	14	4	10	−3
14	8	−6	−4	−2	−1	6	3	12	−11
15	−4	6	6	−3	−1	4	9	−3	0
16	11	−10	−6	−2	−4	−0	−4	15	−16
17	−3	−7	−4	8	6	−8	−11	−10	16
18	−3	17	3	−5	−10	−9	−10	3	−10
19	0	−1	−2	2	2	−0	−3	−1	4
20	5	−8	−9	5	2	−5	−13	4	2
21	−2	1	−4	5	4	1	−5	−4	10
22	9	−15	−7	2	2	1	−2	9	−6
23	−15	48	8	−5	−9	−1	−7	−5	1
24	−7	8	11	−5	−2	3	**14**	−7	−0
25	2	−5	−5	4	3	−2	−6	−1	5
26	2	−9	−2	3	5	3	3	−2	6
27	1	−1	3	−2	3	13	17	2	−1
28	−2	−7	**14**	−6	3	17	35	−4	−2
29	9	−5	−7	−1	−2	2	−4	13	−11
30	−6	1	−0	5	6	−1	−2	−11	16

3.6 Discussion and Summary

The regression quantities such as the coefficients, are known to be easily affected by outlying observations, and it is important to identify such observations, because they could potentially alter the fit of the least squares regression line away from the direction of the majority of the observations. In this chapter, we considered a procedure to identify leverage points that makes use of the matrix of the left singular vectors **U** (that is, the eigenvectors of $\mathbf{X}\mathbf{X}^T$).

The diagonal (h_i) values of the hat matrix are often used to identify observations that are outlying in the explanatory variables. Examining the h_i values by themselves may not correctly identify all leverage points as the h_i values are known to suffer from the masking and swamping effects when there are multiple outliers.

The proposed procedure, which is based on the h_i values, is an adaptation of a technique

used to identify outliers in correspondence analysis, and entails expressing the h_i values as sums of contributions of variance of each axis that is explained by each observation (r_l), and using these contributions, together with the contributions of variance of each observation that is explained by each axis (c_l), to identify the axis that observations are outlying on. Determining the axis that observations are outlying on is important because this tells us something about the type of outlier, that is, whether the observation inflates variances when located in the first few axes, or whether the observation differs in multivariate structure when located in the last few axes or because the observation is not explained well by any axis.

We suggested guidelines to determine large values of r_l and c_l ($r_l > 2/n$ and $c_l > 0.5$), and from the examples used, we notice also that when an observation is being masked, we are able to determine the axis that the observation is outlying on, since r_l and c_l will be large on the kth axis. When observations are being swamped, this pattern is not adhered to, that is, the observation is not well explained by the axis whose direction it determines the most.

We also suggested examining an easy to implement graphical display called a leverage-distance (L-D) plot that highlights any observations may be being masked or swamped.

The success of the proposed procedure was illustrated using an artificial data set, with various observations modified to be outliers, and computations were then carried on three real data sets that have appeared in the literature on regression diagnostics. Table 3.10 presents a summary of the observations that were identified as leverage points, in each of the real data sets using the proposed procedure.

Table 3.10: Summary of the identified leverage points

Data	Leverage Points
HBK	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14
Stack Loss	1, 2, 3, 17, 21
Health Club	1, 8, 13, 16, 18, 23, 24, 28, 30

Any differences between the results that we obtained and those obtained in the literature using the diagonal values of the hat matrix apart from not taking into account the effects of masking and swamping, may be due to the fact that the matrix **X** is standardised, and we have not included the intercept term in our analysis.

The procedure may also be applied to any $n \times m$ matrix where we assume the data to be

normal, for example, in discriminant analysis problems. Most measures typically used as diagnostics for discriminant analysis problems are a function of either the Mahalanobis distance of the observation from the group mean or the linear discriminant function (see Fung (1995) and Lachenbruch (1997)). The Mahalanobis distance however, can be written as a function of the h_i values (refer to Chatterjee and Hadi (1988)), thus any diagnostic measure based on the Mahalanobis distance will suffer from the same problems that the h_i values suffer from.

One drawback of using the proposed procedure is that observations with large h_i values that are not explained well by any axis are automatically treated as leverage points if they contribute highly to the determination of the direction of at least one of the axes. This may result in too many observations being declared incorrectly as leverage points.

We may also append the the vector **y** to the matrix of explanatory variables, **X**, and extend the procedure to consider the leverage points when we take the response variable into account. The results of this approach will be provided in the next chapter where we also consider another diagnostic measure that takes the response variable into account.

Chapter 4

Identifying Outlying Observations in the Residuals

In this chapter we consider observations that are outlying in the residuals. Observations that deviate from the bulk of the data can easily influence the fit of the least squares regression line, and the residuals, which take the response variable, $\mathbf{y}$, into account, are often examined to determine the observations that may have influenced the fit of the least squares regression line, hence affecting the regression coefficients.

Three different ways of identifying regression outliers are considered. The procedure that was proposed in Chapter 3 is extended to the diagonal values of the $\mathbf{H}_z$ matrix, where $\mathbf{H_z} = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T$, and $\mathbf{Z} = (\mathbf{X} : \mathbf{y})$. The L-D plot from this is informative in highlighting observations that may be masked or swamped, but we are not able to differentiate between leverage points and regression outliers, since h_z may be large because of the large h_i value or a large residual value.

One drawback of using transformed residuals such as the Studentized residuals is that they may fail to identify the regression outliers when these observations are being accommodated by the least squares fit, thus we propose a numerical measure to be used in conjunction with the Studentized residuals for identifying regression outliers. This measure is based on the role that each observation plays in the displacement of other observations from the least squares regression fitted line. The proposed measure is based on the off-diagonal values of the hat ($\mathbf{H}_x$) matrix, and computations on the examples based on the same real data sets that we considered in Chapter 3 are presented to illustrate how the proposed measure operates. The artificial examples used illustrate the success of using the measures together.

4.1 Introduction

In the preceding chapter we considered outlying observations in the explanatory variables only, and did not consider the outlyingness of the observations when we take the response variable, $\mathbf{y}$, into account. In this chapter, we consider the deviation of observations from the least squares regression line, that is, observations that are outlying in the residuals, also known as regression outliers.

The residuals (defined as the difference between the observed and predicted values by the regression model) are often examined because they not only take the response variable into account, but also aid in the investigation of regression assumptions. The residuals in their raw form ($\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{I} - \mathbf{U}\mathbf{U}^T)\mathbf{y}$) are often not used because they do not approximate the unobserved errors well, since the diagonal values of the hat matrix ($\mathbf{U}\mathbf{U}^T = \mathbf{H}_x$) need to be equal and the off-diagonal values need to be sufficiently small to ensure that the residuals are uncorrelated and have constant variance (Cook and Weisberg, 1982b). Thus, transformations of the residuals, *inter alia*: normalised residuals, standardised residuals, and the Studentized residuals, are often preferred over the raw residuals because they overcome some of the limitations of the raw residuals.

The transformed residuals are however not without any problems. The least squares method works by minimising the sum of squared deviations from the fitted line, and large deviations are often accommodated at the expense of less deviating observations. Thus observations that deviate from the bulk of the data may have small residuals because they may have pulled the fitted line in their direction. It is for this reason that diagnostics based on the least squares residuals often fail to reveal outlying observations (Hocking, 2003, Rousseeuw and Leroy, 1987).

In this chapter we extend the application of the procedure that was proposed in Chapter 3, and consider the outlying observations when we take the response variable into account. We also propose a measure to identify outlying observations in the residuals, or regression outliers, that is based on the residuals in their raw form, but differs from the diagnostics that are already in use in that we consider instead, the role that each observation plays in the displacement of other observations from the fitted least squares regression line.

This chapter is organised as follows. In the next section, we briefly review the use of the Studentized residuals in identifying regression outliers, and use some of the artificial data sets used in Chapter 3 to illustrate the problems encountered when using the Studentized residuals. We then extend the procedure that was proposed in Chapter 3 and consider the identification of outlying points when we take the response variable into account in section 4.3. In section 4.4, the second residual diagnostic measure is presented, using the

same artificial data sets to compare and contrast the differences between the results of the proposed measure and those of the Studentized residuals. We then consider the same three real data sets that were introduced in Chapter 3 in section 4.5, and end the chapter with a discussion of the main findings of the results presented in the chapter.

4.2 The Studentized Residuals

Let $\mathbf{X} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^T$ be the SVD of the standardised matrix $\mathbf{X}$. The regression residuals are given by

$$\begin{aligned}\hat{\boldsymbol{\epsilon}} &= \mathbf{y} - \hat{\mathbf{y}} \\ &= \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \left[\mathbf{I} - (\mathbf{U}\mathbf{D}_\alpha\mathbf{V}^T)(\mathbf{V}\mathbf{D}_\alpha^{-1}\mathbf{U}^T)\right]\mathbf{y} \\ &= (\mathbf{I} - \mathbf{U}\mathbf{U}^T)\mathbf{y}\end{aligned} \tag{4.1}$$

and the Studentized residuals are given by:

$$r^*_{(i)} = \frac{\hat{\epsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_i}} \tag{4.2}$$

where $\hat{\sigma}^2_{(i)} = \mathbf{y}^T_{(i)}(\mathbf{I} - \mathbf{H}_{x(i)})\mathbf{y}_{(i)}/(n-m-1)$ is the the estimate of σ when the ith observation is excluded, and $\mathbf{H}_{x(i)} = \mathbf{X}_{(i)}(\mathbf{X}^T_{(i)}\mathbf{X}_{(i)})^{-1}\mathbf{X}^T_{(i)}$ is the hat matrix excluding the ith observation.

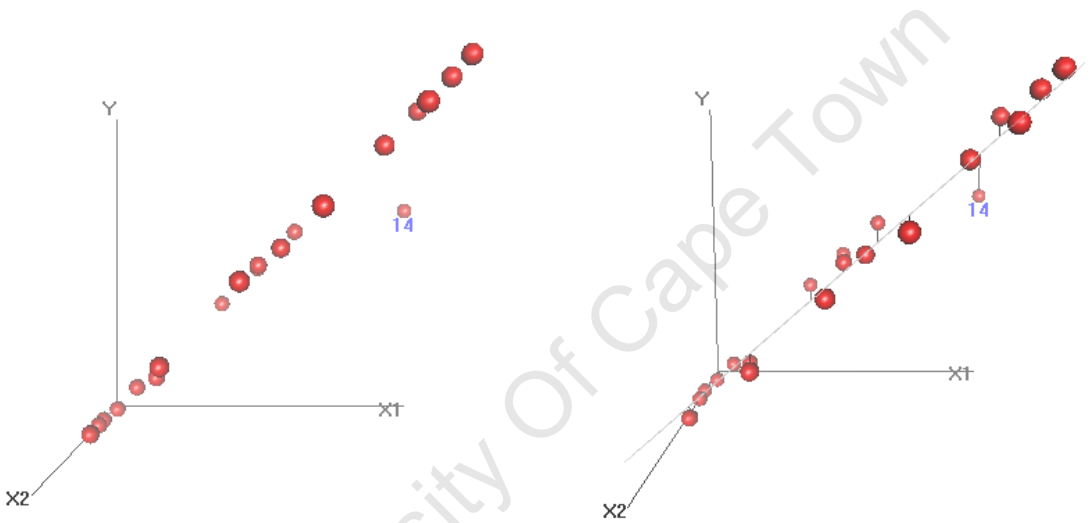
The Studentized residuals are the preferred residuals because they are easily related to the t-distribution, and have been standardised to have equal variances (Welsch and Kuh, 1977). Belsley *et al.* (1980) recommended a cut-off value of $\mid 2 \mid$ for $r^*_{(i)}$.

We have already mentioned one potential disadvantage of using the Studentized residuals; that is, large deviations may be accommodated at the expense of less deviating observations. Another potential drawback associated with the use of the Studentized residuals, is that they may suffer from the masking and swamping effects, since they are a function of the diagonal values of the hat matrix, $\mathbf{H}_x$. In the following three examples we illustrate these problems associated with using the Studentized residuals:

Example 1

The first example that we consider is the same one we used in Figure 3.1(b) (p. 3-5), where there was a single leverage point. The same leverage point, observation 14, has now been modified to be a regression outlier as illustrated in Figure 4.1(a). Figure 4.1(b) illustrates that this observation is not being accommodated excessively by the least squares fit at the expense of the other non-outlying observations, and Table 4.1 shows the Studentized residuals for this data set. Observation 14 has a large Studentized residual.

Figure 4.1: Example 1 – (a) 3-Dimensional scatter plot of two explanatory variables and the response variable. (b) The best linear least squares fit



(a) (b)

Table 4.1
Studentized residuals for Example 1

	$r^*_{(i)}$		$r^*_{(i)}$
1	0.218	11	1.139
2	−0.143	12	−0.092
3	−0.774	13	−0.763
4	0.112	14	−11.926 ♯
5	0.106	15	−0.126
6	−0.960	16	1.300
7	0.533	17	0.138
8	−0.579	18	0.569
9	−0.500	19	0.576
10	1.323	20	1.098

♯ $r^*_{(i)} > |2|$

Example 2

The second example that we consider is based on the data used in Figure 3.2(a) (p. 3-6), where four observations, observations 14, 16, 18 and 19 are leverage points. Observations 18 and 19 have been modified to be regression outliers as illustrated in Figure 4.2(a). Figure 4.2(b) illustrates that these observations are being accommodated by the least squares fit at the expense of other non-outlying observations, although their residuals are still large. Table 4.2 shows the Studentized residuals for the data set. From the table, we see that observations 18 and 19 have large Studentized residuals. Observation 20, whose distance from the regression plane is relatively large (refer to Figure 4.2(b)), also has a Studentized residual value that is greater than 2.

Figure 4.2: Example 2 – (a) 3-Dimensional scatter plot of two explanatory variables and the response variable. (b) The best linear least squares fit

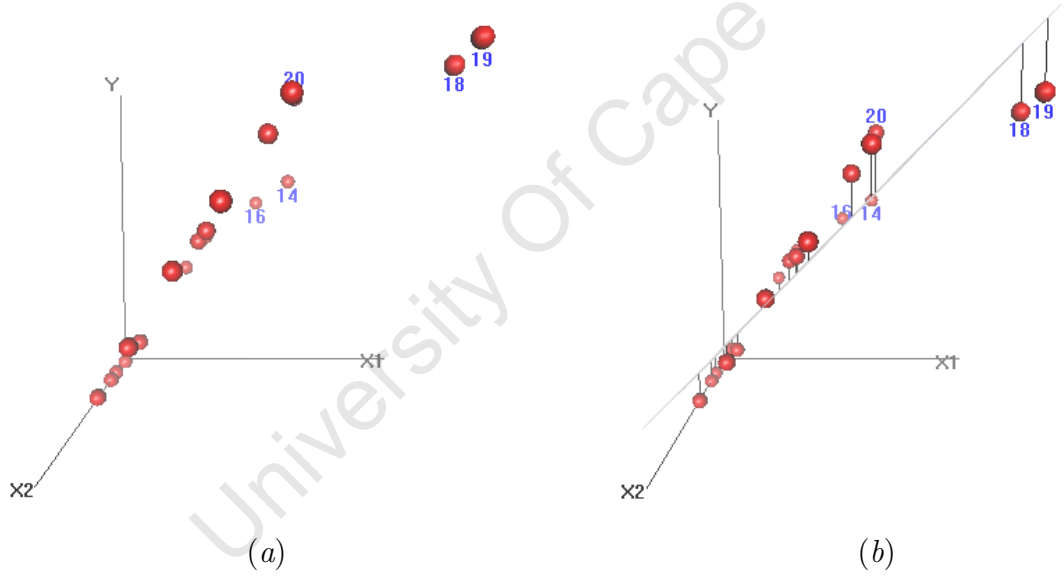


Table 4.2
Studentized residuals for Example 2

	$r^*_{(i)}$		$r^*_{(i)}$
1	−0.362	11	0.611
2	−0.791	12	0.512
3	−1.008	13	0.484
4	−0.684	14	−0.099
5	−0.616	15	1.356
6	−0.536	16	0.423
7	0.660	17	1.237
8	−0.545	18	$-2.225^\sharp$
9	0.161	19	$-2.228^\sharp$
10	0.970	20	$2.057^\sharp$

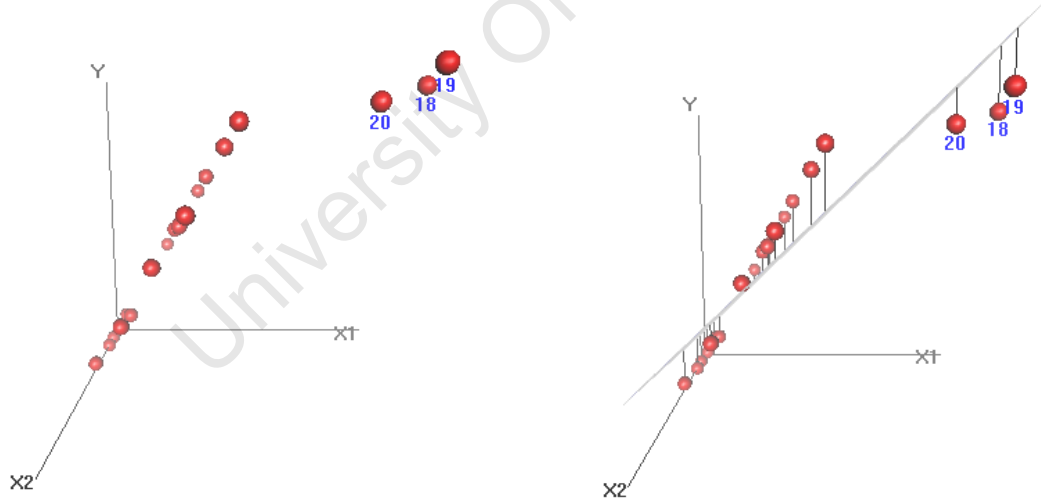
$^\sharp$ $r^*_{(i)} > |\,2\,|$

Example 3

The third example that we consider is based on the data used in Figure 3.3(a) (p. 3-9), where three observations, observations 18, 19 and 20 are leverage points, and observation 20 is a masked leverage point. All three observations have been modified to be regression outliers as illustrated in Figure 4.3(a). Figure 4.3(b) illustrates that these observations are being accommodated by the least square fit at the expense of other non-outlying observations.

Table 4.3 shows the Studentized residuals for this data set. From the table, we see that none of the three observations have Studentized residuals greater than 2, although the Studentized residual of observation 18 is large enough to be flagged as a regression outlier. From Figure 4.3(b), we see that the distance of observation 18 to the regression plane is relatively large compared to that of observations 19 and 20, hence the small Studentized residuals for these two observations.

Figure 4.3: Example 3 – (a) 3-Dimensional scatter plot of two explanatory variables and the response variable. (b) The best linear least squares fit



(a) (b)

Table 4.3
Studentized residuals for Example 3

	$r^*_{(i)}$		$r^*_{(i)}$
1	−0.566	11	0.480
2	−1.034	12	0.594
3	−1.081	13	0.699
4	−0.953	14	1.308
5	−0.865	15	1.663
6	−0.552	16	1.270
7	0.648	17	1.456
8	−0.623	18	−1.939
9	0.234	19	−1.229
10	0.907	20	−0.837

4.3 The Diagonal Values of $\mathbf{H}_z$

If we let $\mathbf{Z}$ be the matrix formed when we append the response variable, $\mathbf{y}$, to the matrix of explanatory variables $\mathbf{X}$, that is, $\mathbf{Z} = (\mathbf{X} : \mathbf{y})$, and $\mathbf{H_z} = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T$, then $\mathbf{H_z}$ can be written as

$$\mathbf{H_z} = \mathbf{H_x} + \frac{\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}^T}{\hat{\boldsymbol{\epsilon}}^T\hat{\boldsymbol{\epsilon}}}$$

which shows that $\mathbf{H_z}$ is a function of $\mathbf{H_x}$ and the residuals. Thus we see that $\mathbf{H_z}$ will be large whenever the first term, $\mathbf{H_x}$, is large and/or the second term is large.

The diagonal values of $\mathbf{H_z}$, that is, the h_z values, are often used to identify regression outliers. Values of h_z will however, suffer from the same problems of masking and swamping as the h_i values, as we have already seen in Chapter 3. Another potential problem with using the h_z values is that since they are a function of the residuals, and because of the possibility of the 'real' outlying observations being accommodated at the expense of the less deviating observations, the non-outlying observations may be swamped. Thus we advocate the same procedure that was proposed in Chapter 3 when dealing with the h_z values to identify regression outliers.

We illustrate the use of the procedure on the h_z values using the same three data sets used in the examples above for the Studentized residuals.

Example 1

Table 4.4 shows values of h_z, r_l, c_l and DIST for the first example, where observation 14 is a regression outlier. From the table, we see that observation 14 is outlying on the second

axis.

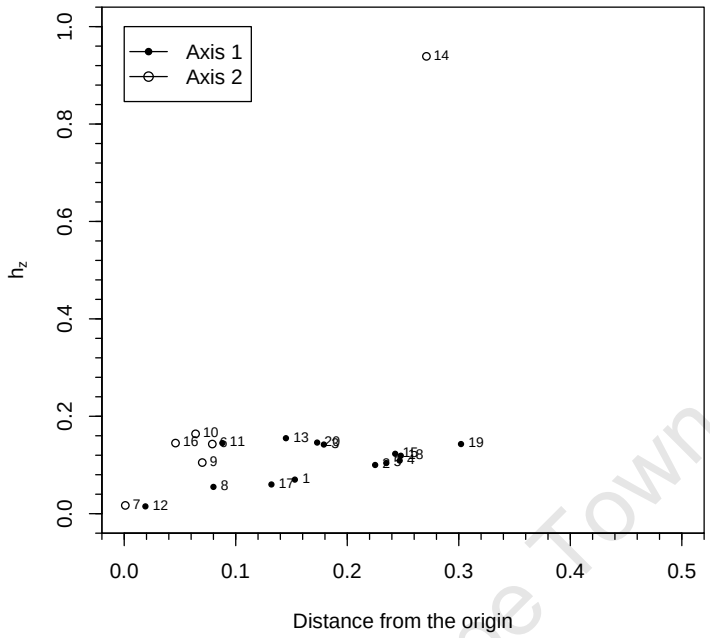
The L-D plot of the data is shown in Figure 4.4. From the plot, we see that even though some of the observations that are explained well by the first axis are located far from the origin (or make a large contribution to the total variance), the contribution of observation 14 to the total variance, which is located in the direction of the second axis, which explains only 22.25% of the total variance in the data, is comparatively large.

Table 4.4
Example 1 – r_l, c_l, h_z and DIST values

		r_l			c_l			
	h_z	Axis 1 (77.36%)	Axis 2 (22.25%)	Axis 3 (0.39%)	Axis 1 (77.36%)	Axis 2 (22.25%)	Axis 3 (0.39%)	DIST
1	0.070	0.066	0.000	0.004	0.999	0.001	0.000	0.153
2	0.100	0.095	0.005	0.000	0.984	0.016	0.000	0.225
3	0.142	0.060	0.060	0.022	0.774	0.224	0.001	0.179
4	0.109	0.107	0.000	0.002	1.000	0.000	0.000	0.247
5	0.104	0.101	0.001	0.002	0.996	0.004	0.000	0.235
6	0.143	0.005	0.099	0.039	0.152	0.842	0.006	0.079
7	0.017	0.000	0.001	0.016	0.050	0.684	0.266	0.001
8	0.055	0.033	0.007	0.015	0.938	0.060	0.002	0.080
9	0.105	0.005	0.089	0.011	0.152	0.846	0.002	0.070
10	0.164	0.005	0.077	0.082	0.185	0.800	0.015	0.064
11	0.145	0.020	0.060	0.065	0.537	0.454	0.009	0.088
12	0.015	0.005	0.009	0.001	0.684	0.316	0.000	0.019
13	0.155	0.036	0.090	0.029	0.583	0.415	0.002	0.145
14	0.939*	0.006	0.374*	0.559	0.054	0.922*	0.024	0.271
15	0.123	0.097	0.024	0.002	0.933	0.067	0.000	0.243
16	0.145	0.000	0.067	0.078	0.009	0.971	0.020	0.046
17	0.060	0.056	0.004	0.000	0.982	0.018	0.000	0.132
18	0.119	0.107	0.000	0.012	0.999	0.001	0.001	0.248
19	0.143	0.130	0.001	0.012	0.998	0.001	0.000	0.302
20	0.146	0.065	0.031	0.050	0.878	0.119	0.003	0.173

* h_z large, $r_l > 0.1$ and $c_l > 0.5$.

Figure 4.4: L-D plot for Example 1



Example 2

Table 4.5 shows values of h_z, r_l, c_l and DIST for the second example. Recall that for this data set, observations 14, 16, 18 and 19 are leverage points, and observations 18 and 19 have been modified to be regression outliers. Both observations 18 and 19 are located in the direction of the first axis.

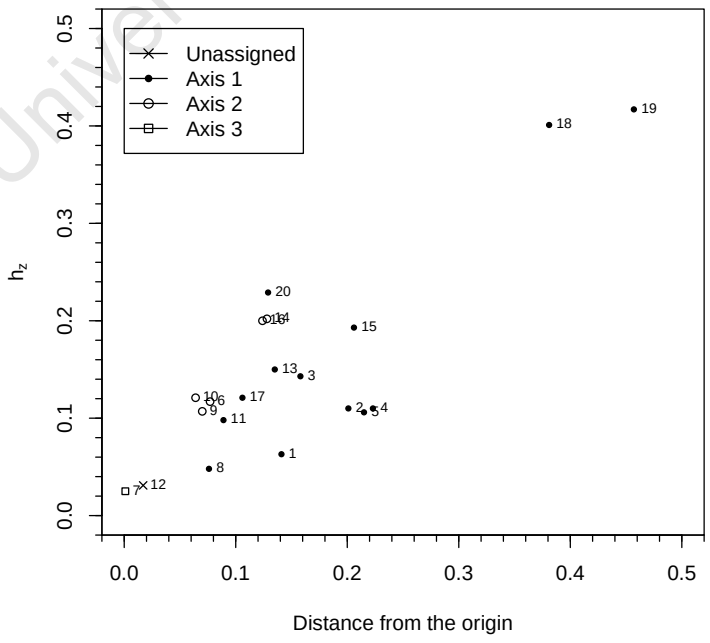
The L-D plot of the data is shown in Figure 4.5. From the plot, we see that observations 14, 16, 18 and 19 are located far from the origin (or make a large contribution to the total variance) relative to other observations that are located in the direction of the same axis. Observation 20, which has a large Studentized residual, appears to be swamped, since it has a large h_z value, but is not located too far from the origin relative to other non-outlying observations that are located in the direction of the first axis.

Table 4.5
Example 2 – r_l, c_l, h_z and DIST values

		r_l			c_l			
	h_z	Axis 1 (77.80%)	Axis 2 (20.70%)	Axis 3 (1.50%)	Axis 1 (77.80%)	Axis 2 (20.70%)	Axis 3 (1.50%)	DIST
1	0.060	0.000	0.003	0.063	0.998	0.001	0.001	0.141
2	0.085	0.005	0.020	0.110	0.980	0.016	0.004	0.201
3	0.051	0.061	0.031	0.143	0.753	0.238	0.009	0.158
4	0.095	0.000	0.015	0.110	0.997	0.000	0.003	0.223
5	0.092	0.002	0.012	0.106	0.993	0.005	0.003	0.215
6	0.004	0.106	0.007	0.117	0.135	0.861	0.004	0.077
7	0.000	0.000	0.025	0.025	0.121	0.087	0.792	0.001
8	0.030	0.008	0.010	0.048	0.928	0.066	0.006	0.076
9	0.004	0.099	0.004	0.107	0.121	0.876	0.003	0.070
10	0.007	0.072	0.042	0.121	0.269	0.702	0.030	0.064
11	0.022	0.058	0.018	0.098	0.589	0.402	0.009	0.089
12	0.003	0.012	0.016	0.031	0.489	0.467	0.045	0.017
13	0.030	0.105	0.015	0.150	0.511	0.484	0.005	0.135
14	0.003	0.193*	0.006	0.202*	0.059	0.939*	0.002	0.128
15	0.077	0.035	0.081	0.193	0.877	0.105	0.018	0.206
16	0.000	0.198*	0.002	0.200*	0.007	0.992*	0.001	0.124
17	0.042	0.007	0.072	0.121	0.926	0.043	0.031	0.106
18	0.155*	0.013	0.233	0.401*	0.951*	0.022	0.028	0.381
19	0.191*	0.001	0.225	0.417*	0.976*	0.002	0.022	0.457
20	0.046	0.022	0.161	0.229**	0.837	0.107	0.056	0.129

* h_z large, $r_l > 0.1$ and $c_l > 0.5$.
** h_i large, but $r_l < 0.1$ and $c_l > 0.5$ (or the reverse) on the kth axis.

Figure 4.5: L-D plot for Example 2



Example 3

Table 4.6 shows values of h_z, r_l, c_l and DIST for the third example. Recall that for this data set, observations 18, 19 and 20 are leverage points, and have been modified to be regression outliers as well.

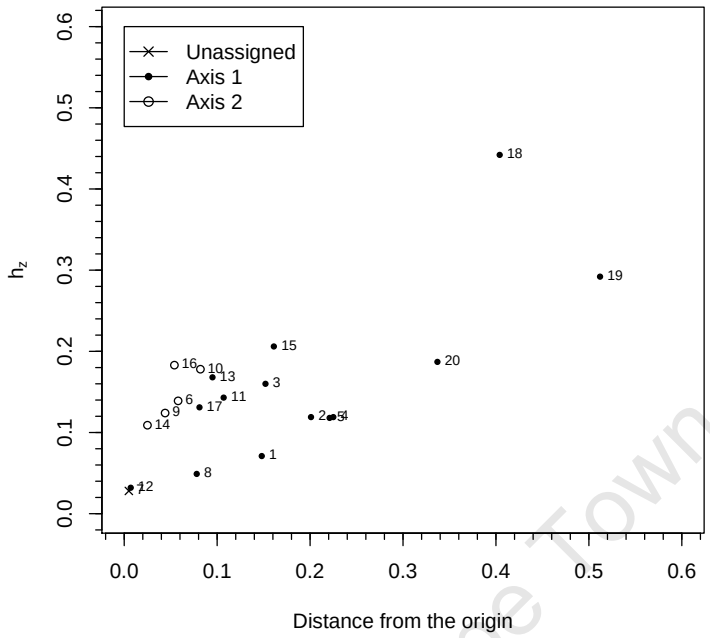
The L-D plot of the data is shown in Figure 4.6. From the plot, we see that observation 20 is being masked, whilst observation 15 appears to be swamped. All three observations (18, 19 and 20) are located in the direction of the first axis.

Table 4.6
Example 3 – r_l, c_l, h_z and DIST values

		r_l			c_l			
	h_z	Axis 1 (86.18%)	Axis 2 (11.58%)	Axis 3 (2.24%)	Axis 1 (86.18%)	Axis 2 (11.58%)	Axis 3 (2.24%)	DIST
1	0.071	0.057	0.001	0.013	0.992	0.002	0.006	0.148
2	0.119	0.076	0.008	0.035	0.975	0.013	0.012	0.201
3	0.160	0.046	0.092	0.022	0.779	0.211	0.010	0.152
4	0.119	0.086	0.000	0.033	0.990	0.000	0.010	0.225
5	0.118	0.084	0.003	0.031	0.986	0.004	0.009	0.221
6	0.139	0.004	0.134	0.001	0.188	0.810	0.001	0.058
7	0.028	0.001	0.006	0.021	0.326	0.405	0.269	0.005
8	0.049	0.029	0.008	0.012	0.954	0.036	0.010	0.078
9	0.124	0.002	0.107	0.015	0.139	0.838	0.023	0.044
10	0.178	0.011	0.150	0.017	0.349	0.637	0.014	0.082
11	0.143	0.026	0.114	0.003	0.628	0.370	0.002	0.107
12	0.032	0.002	0.005	0.025	0.547	0.225	0.228	0.007
13	0.168	0.022	0.101	0.045	0.602	0.366	0.032	0.095
14	0.109	0.003	0.040	0.066	0.275	0.550	0.175	0.025
15	0.206**	0.057	0.014	0.135	0.913	0.031	0.056	0.161
16	0.183	0.001	0.139	0.043	0.055	0.892	0.053	0.054
17	0.131	0.029	0.000	0.102	0.915	0.000	0.084	0.081
18	0.442*	0.140*	0.075	0.227	0.897*	0.065	0.038	0.404
19	0.292*	0.195*	0.002	0.095	0.986*	0.002	0.012	0.512
20	0.187	0.129	0.000	0.058	0.988	0.000	0.012	0.337

* h_z large, $r_l > 0.1$ and $c_l > 0.5$.
** h_i large, but $r_l < 0.1$ and $c_l > 0.5$ (or the reverse) on the kth axis.

Figure 4.6: L-D plot for Example 3



4.4 The Proposed Residual Diagnostic

From the examples used above, extending the procedure that was proposed in Chapter 3 to the diagonal values of the $\mathbf{H}_z$ matrix appears to work well in identifying the regression outliers. However, as already noted in the previous chapter, one potential disadvantage of using the proposed procedure is that observations with large h_z values that are not explained well by any axis are automatically treated as leverage points if they contribute highly to the determination of the direction of at least one of the axes. This may result in too many observations being declared incorrectly as regression outliers.

The Studentized residuals on the other hand, perform well when the outlying observations are not being accommodated by the least squares fit, but as we have already seen, the Studentized residuals may fail to highlight the regression outliers when these observations are being accommodated by the least squares fit.

In this section we propose a measure that should be used in conjunction with the Studentized residuals (or other transformed residuals), which will highlight those observations that are being accommodated by the least squares fit, and as a result, are inducing large residuals for other observations.

From (4.1), if we put

$$\mathbf{U} = \begin{bmatrix} \mathbf{u}_1^T \\ \mathbf{u}_2^T \\ \vdots \\ \mathbf{u}_n^T \end{bmatrix}$$

then the ith residual is given by

$$\begin{aligned} \hat{\epsilon}_i &= \left(1 - \mathbf{u}_i^T \mathbf{u}_i\right) y_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} \mathbf{u}_i^T \mathbf{u}_j y_j \\ &= r_{ii} + \sum_{\substack{j=1 \\ j \neq i}}^{n} r_{ij} \end{aligned}$$

Thus r_{ij} is the contribution of observation j to the ith residual. We can interpret

$$\sum_{i=1}^{n} r_{ij} \qquad \text{for all} \quad j = 1, 2, \ldots, n$$

as the sum of contributions of observation j to all the residuals. That is, it is the residual induced by observation j. But since some of the r_{ij}'s are positive and some negative, these contributions may cancel out. Thus of more interest is

$$r_j = \sum_{i=1}^{n} \mid r_{ij} \mid \qquad \text{for all} \quad j = 1, 2, \ldots, n$$

the sum of the absolute contributions of observation j to all the residuals, and also

$$\frac{r_j}{\sum_{j=1}^{n} r_j} \qquad \text{and} \qquad R_j = \frac{(n \times r_j)}{\sum_{j=1}^{n} r_j} \qquad \text{for all} \quad j = 1, 2, \ldots, n$$

The former indicates the proportion of the absolute contributions that is induced by observation j, and the latter is an easy-to-interpret statistic — if an observation contributes equally to all the residuals, then R_j will equal 1. Therefore values of R_j greater than 1 indicate that observation j makes a more-than-average contribution to the residuals of other observations. As a crude cut-off value, we will consider all observations which have

R_j values greater than 2 as observations that make a more-than-average contribution to the residuals of other observations.

R_j will be large when an observation is making a more-than-average contribution to the residuals of other observations, and the same data used in the examples above will be used to illustrate this. (Note that all computations on the examples were performed in R (R Development Core Team, 2008), and the source codes written for the measure are included in Appendix C, section C.1 (p. C-3).)

Example 1

Recall that for the first example, observation 14 has been modified to be a regression outlier as was illustrated in Figures 4.1(a) and 4.1(b) (p. 4-4). Since this observation is not being accommodated excessively by the least squares fit at the expense of the other non-outlying observations, the Studentized residuals correctly identify this observation as a regression outlier (refer to Table 4.7). The value of R_j for observation 14 is low, indicating that the observation is not making a more-than-average contribution to the residuals of other observations.

Table 4.7
R_j values and Studentized residuals for Example 1

	R_j	$r^*_{(i)}$
1	1.372	0.218
2	1.652	−0.143
3	1.412	−0.774
4	1.686	0.112
5	1.600	0.106
6	0.602	−0.960
7	0.002	0.533
8	1.030	−0.579
9	0.090	−0.500
10	0.210	1.323
11	0.604	1.139
12	0.318	−0.092
13	0.728	−0.763
14	1.098	$-11.926^{\sharp}$
15	1.418	−0.126
16	0.228	1.300
17	1.188	0.138
18	1.634	0.569
19	1.718	0.576
20	1.410	1.098

$^{\sharp}$ $r^*_{(i)} > |\,2\,|$

Example 2

For the second example, two observations, observations 18 and 19 have been modified to be regression outliers as was illustrated in Figures 4.2(a) and 4.2(b) (p. 4-5). These two observations are also being accommodated excessively by the least squares fit at the expense of the other non-outlying observations, and the Studentized residuals correctly identify these observations as regression outliers (refer to Table 4.8). The values of R_j for observations 18 and 19 are greater than 2, indicating that the observations are making a more-than-average contribution to the residuals of other observations. Observation 20, which has a large Studentized residual value, and was identified as being swamped using the L-D plot, has a low R_j value.

Table 4.8
R_j values and Studentized residuals for Example 2

	R_j	$r^*_{(i)}$
1	1.340	−0.362
2	1.582	−0.791
3	1.356	−1.008
4	1.610	−0.684
5	1.530	−0.616
6	0.646	−0.536
7	0.174	0.660
8	1.044	−0.545
9	0.082	0.161
10	0.368	0.970
11	0.694	0.611
12	0.100	0.512
13	0.430	0.484
14	0.782	−0.099
15	1.024	1.356
16	0.290	0.423
17	0.848	1.237
18	2.476	$-2.225^\sharp$
19	2.546	$-2.228^\sharp$
20	1.078	$2.057^\sharp$

$^\sharp$ $r^*_{(i)} > |2|$

Example 3

For the third example, where observations 18, 19 and 20 are regression outliers as was illustrated in Figures 4.3(a) and 4.3(b) (p. 4-6), although Figure 4.3(b) illustrated that these observations were being accommodated by the least squares fit at the expense of other non-outlying observations.

Table 4.9 shows the R_j values and the Studentized residuals for this data set. From the table, we see the R_j values for observation 18, 19 and 20 are greater than 2, and none of the three observations have Studentized residuals greater than 2, although the Studentized residual of observation 18 is large enough to be flagged as a regression outlier.

Table 4.9
R_j values and Studentized residuals for Example 3

	R_j	$r^*_{(i)}$
1	1.304	−0.566
2	1.530	−1.034
3	1.312	−1.081
4	1.554	−0.953
5	1.472	−0.865
6	0.618	−0.552
7	0.252	0.648
8	1.036	−0.623
9	0.146	0.234
10	0.448	0.907
11	0.706	0.480
12	0.008	0.594
13	0.298	0.699
14	0.236	1.308
15	0.876	1.663
16	0.142	1.270
17	0.708	1.456
18	2.486*	−1.939
19	2.586*	−1.229
20	2.284*	−0.837

* $R_j > 2$

4.5 Illustrative Examples

We illustrate the extension of the procedure that was proposed in Chapter 3 using the diagonal values of the $\mathbf{H}_z$ matrix, the Studentized residuals and the residual measure, R_j, proposed in this chapter on the three real data sets that were introduced in the previous chapter.

Example 1: Hawkins, Bradu and Kass data

Recall that this data set has been constructed so that the first fourteen observations are outliers, with observations 11 to 14 known to be good leverage points.

Results Based on the Diagonal Values of $\mathbf{H}_z$

Table 4.10 shows the r_l, c_l, h_z and DIST values of the HBK data set for the first 20 observations (the complete table can be found in Appendix B (p. B-4)), and the stem-and-leaf display is shown in Table 4.11. From the tables, observations 11, 12, 13 and 14 are classified as regression outliers because their h_z values are large compared to other observations in the data set. All four observations determine the direction of and are well explained by the first axis. The first 14 observations however, are all large contributors to the total variance (that is, they are located far from the origin).

Table 4.10
Hawkins, Bradu and Kass data: r_l, c_l, h_z and DIST values

	r_l					c_l				
	Axis 1 (88.99%)	Axis 2 (9.19%)	Axis 3 (1.39%)	Axis 4 (0.43%)	h_z	Axis 1 (88.99%)	Axis 2 (9.19%)	Axis 3 (1.39%)	Axis 4 (0.43%)	DIST
1	0.056	0.022	0.002	0.001	0.081	0.960	0.039	0.000	0.000	0.209
2	0.058	0.028	0.004	0.001	0.091	0.952	0.047	0.001	0.000	0.217
3	0.066	0.025	0.003	0.003	0.097	0.962	0.037	0.001	0.000	0.245
4	0.063	0.013	0.005	0.004	0.085	0.977	0.021	0.001	0.000	0.229
5	0.065	0.020	0.000	0.001	0.086	0.970	0.030	0.000	0.000	0.239
6	0.063	0.021	0.007	0.004	0.095	0.964	0.034	0.002	0.000	0.234
7	0.066	0.034	0.001	0.010	0.111	0.948	0.051	0.000	0.001	0.248
8	0.059	0.032	0.000	0.000	0.091	0.946	0.054	0.000	0.000	0.220
9	0.060	0.017	0.004	0.006	0.087	0.970	0.029	0.001	0.000	0.220
10	0.057	0.026	0.005	0.012	0.100	0.953	0.045	0.001	0.001	0.212
11	0.041*	0.162	0.000	0.048	0.251*	0.707*	0.289	0.000	0.004	0.207
12	0.045*	0.180	0.017	0.133	0.375*	0.697*	0.289	0.004	0.010	0.229
13	0.050*	0.146	0.003	0.000	0.199*	0.767*	0.232	0.001	0.000	0.232
14	0.058*	0.213	0.093	0.227	0.591*	0.702*	0.267	0.018	0.013	0.294
15	0.001	0.002	0.029	0.014	0.046	0.639	0.105	0.223	0.032	0.007
16	0.001	0.000	0.029	0.034	0.064	0.693	0.017	0.212	0.078	0.008
17	0.005	0.000	0.021	0.000	0.026	0.937	0.008	0.056	0.000	0.021
18	0.002	0.000	0.008	0.000	0.010	0.948	0.000	0.051	0.000	0.009
19	0.003	0.000	0.013	0.003	0.019	0.932	0.004	0.060	0.004	0.012
20	0.001	0.000	0.013	0.021	0.035	0.746	0.004	0.168	0.082	0.004
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

* h_z large, $r_l > 0.027$ and $c_l > 0.5$.

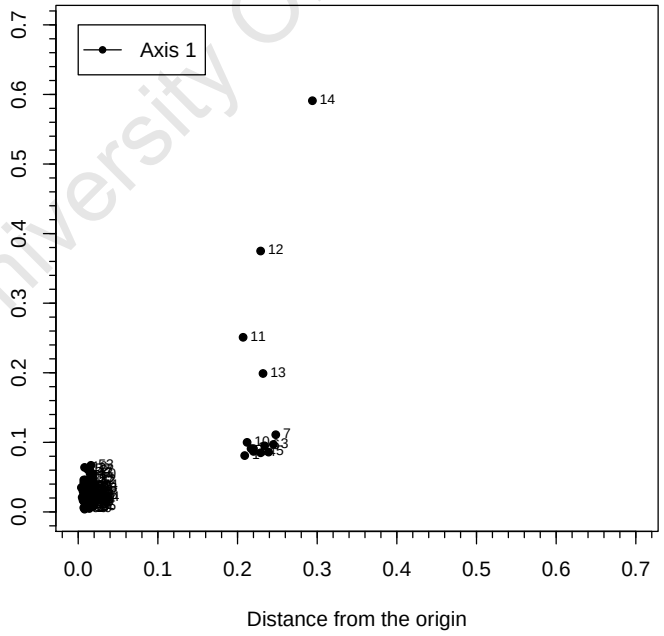
Table 4.11
Hawkins, Bradu and Kass data: Stem-and-leaf display of the h_z values

stem	leaf	observation
0	011111111111111122222222222222223333333333444444444455556666678999999	
1	0001	
2	05	13, 11
3	8	12
4		
5	9	14

The decimal point is 1 digit to the left of the |.

The L-D plot shown in Figure 4.7, indicates that the first fourteen observations deviate markedly from the other observations in the sample when we take the response variable into account. Observations 1 to 10 which are clustered in the same area, are being masked. From Table 4.10, we see that all the ten observations determine the direction of and are well explained by the first axis as well.

Figure 4.7: L-D plot for the Hawkins, Bradu and Kass data.



Results Based on the Studentized Residuals and Proposed Residual Measure (R_j)

The first column of Table 4.12 shows R_j values for the Hawkins, Bradu and Kass data.

From the table, observations 1 to 13 are labelled as regression outliers, and the R_j value for observation 14 is not greater than 2. The Studentized residuals indicate that the first ten observations are regression outliers.

Table 4.12
Hawkins, Bradu and Kass data: Residuals

	R_j	$r^*_{(i)}$		R_j	$r^*_{(i)}$		R_j	$r^*_{(i)}$
1	2.377*	2.427♯	26	0.801	-0.599	51	0.746	-0.167
2	2.270*	2.543♯	27	0.280	-0.570	52	0.264	-0.513
3	2.759*	2.600♯	28	0.374	-0.282	53	0.219	-0.167
4	2.627*	2.370♯	29	0.704	-0.282	54	1.170	-0.167
5	2.643*	2.514♯	30	0.132	-0.455	55	1.017	-0.369
6	2.439*	2.514♯	31	0.489	-0.369	56	0.994	-0.340
7	2.308*	2.744♯	32	1.129	-0.484	57	0.615	-0.167
8	2.435*	2.600♯	33	0.897	-0.541	58	0.832	-0.397
9	2.600*	2.398♯	34	0.651	-0.570	59	0.426	-0.455
10	2.544*	2.485♯	35	0.286	-0.282	60	0.373	-0.628
11	3.328*	-0.426	36	0.556	-0.657	61	0.770	-0.455
12	4.068*	-0.484	37	0.804	-0.541	62	0.701	-0.196
13	2.877*	-0.167	38	0.838	-0.109	63	1.128	-0.455
14	1.683	-0.340	39	0.340	-0.570	64	0.445	-0.513
15	0.485	-0.484	40	0.887	-0.513	65	0.193	-0.196
16	0.773	-0.196	41	0.115	-0.397	66	0.959	-0.628
17	1.073	-0.426	42	0.417	-0.570	67	0.762	-0.570
18	0.448	-0.369	43	1.218	-0.196	68	0.616	-0.196
19	0.931	-0.340	44	0.063	-0.570	69	0.646	-0.311
20	0.611	-0.253	45	0.508	-0.513	70	0.537	-0.167
21	0.538	-0.109	46	0.109	-0.484	71	0.560	-0.311
22	0.967	-0.282	47	0.289	-0.628	72	0.820	-0.426
23	0.832	-0.599	48	0.143	-0.340	73	0.665	-0.253
24	0.990	-0.167	49	0.525	-0.109	74	0.879	-0.628
25	0.648	-0.455	50	0.506	-0.484	75	0.349	-0.311

* $R_j > 2$.

♯ $r^*_{(i)} >\mid 2 \mid$.

Thus, using the procedure, we classify observations 1 to 14 as regression outliers, whilst the Studentized residuals and R_j suggest that only observation 1 to 13 are regression outliers.

Example 2: Stack Loss Data

Recall that for this data set, six observations, observations 1, 2, 3, 4, 17 and 21 (or combinations of) have been found to be outlying and/or influential.

Results Based on the Diagonal Values of $\mathbf{H}_z$

The r_l, c_l, h_z and DIST values for the Stack Loss data are shown in Table 4.13 and the stem-and-leaf display is shown in Table 4.14.

Table 4.13
Stack Loss data: r_l, c_l, h_z and DIST values

	r_l					c_l				
	Axis 1 (74.94%)	Axis 2 (18.35%)	Axis 3 (5.38%)	Axis 4 (1.33%)	h_i	Axis 1 (74.94%)	Axis 2 (18.35%)	Axis 3 (5.38%)	Axis 4 (1.33%)	DIST
1	0.222*	0.045	0.017	0.028	0.312*	0.945*	0.047	0.005	0.002	0.705
2	0.183*	0.051	0.010	0.047	0.291*	0.929*	0.063	0.004	0.004	0.591
3	0.132*	0.006	0.022	0.083	0.243*	0.966*	0.012	0.011	0.011	0.410
4	0.024	0.011	0.042	0.186	0.263**	0.725	0.080	0.093	0.101	0.098
5	0.002	0.000	0.003	0.016	0.021	0.761	0.004	0.104	0.131	0.007
6	0.004	0.000	0.030	0.047	0.081	0.563	0.008	0.308	0.121	0.021
7	0.020	0.052	0.113	0.019	0.204♯	0.485	0.309	0.197	0.008	0.123
8	0.022	0.049	0.108	0.004	0.183	0.526	0.287	0.185	0.002	0.125
9	0.000	0.000	0.107	0.041	0.148	0.013	0.006	0.897	0.085	0.026
10	0.026	0.036	0.096	0.004	0.162	0.628	0.207	0.163	0.002	0.126
11	0.008	0.047	0.050	0.041	0.146	0.321	0.494	0.154	0.031	0.070
12	0.015	0.040	0.116	0.041	0.212♯	0.447	0.287	0.244	0.021	0.102
13	0.028	0.006	0.071	0.017	0.122	0.802	0.043	0.146	0.009	0.104
14	0.002	0.152	0.003	0.001	0.158	0.053	0.942	0.005	0.000	0.119
15	0.036	0.077	0.007	0.054	0.174	0.642	0.332	0.009	0.017	0.169
16	0.050	0.022	0.004	0.012	0.088	0.894	0.097	0.006	0.004	0.166
17	0.098*	0.262	0.002	0.016	0.378*	0.602*	0.396	0.001	0.002	0.486
18	0.064	0.038	0.011	0.000	0.113	0.863	0.126	0.011	0.000	0.223
19	0.047	0.032	0.050	0.000	0.129	0.804	0.134	0.062	0.000	0.175
20	0.012	0.020	0.000	0.011	0.043	0.705	0.284	0.000	0.011	0.053
21	0.005	0.053	0.139	0.332	0.529♯	0.153	0.381	0.293	0.173	0.102

* h_z large, $r_l > 0.095$ and $c_l > 0.5$.

** h_z large, but $r_l < 0.067$ and $c_l > 0.5$ (or the reverse) on the kth axis.

♯ h_z large, and $r_l > 0.095$ on at least one axis but $c_l < 0.5$ on all four axes.

Table 4.14
Stack Loss data: Stem-and-leaf display of the h_z values

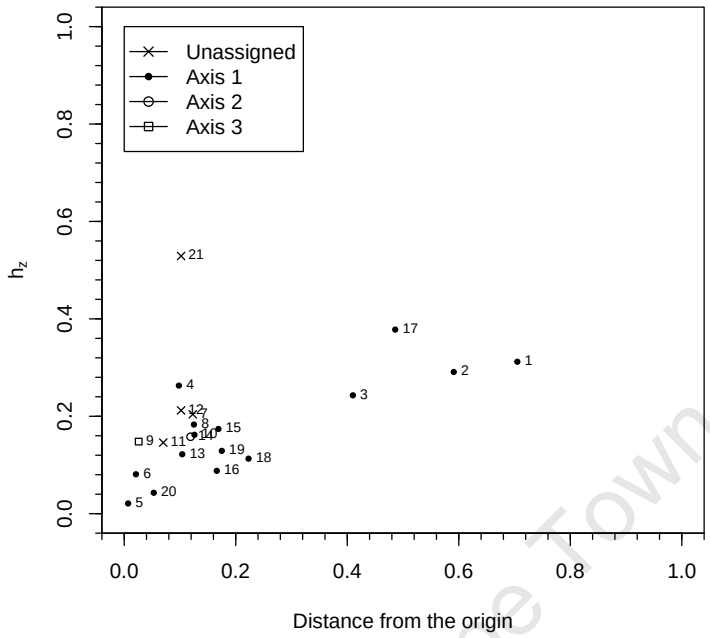
stem		leaf	observation
0	|	2 4 8 9	
1	|	1 2 3 5 5 6 6 7 8	
2	|	0 1 4 6 9	7, 12, 3, 4, 2
3	|	1 9	1 17
4	|		
5	|	3	21

The decimal point is 1 digit to the left of the |.

From the tables, observations 1, 2, 3, 4, 7, 12, 17 and 21 have large h_z values. Observations 1, 2, 3 and 17 determine the direction of and are well explained by the first axis; observation 4 determines the direction of the fourth axis, but is well explained by the first axis, and observations 7, 12 and 21 are not explained well by any of the four axes, but are responsible for determining the direction of the at least one axes. This indicates that the three observations do not fit the structure of the bulk of the data, and we will therefore treat the observations as outliers. The only observations that are large contributors to the total variance (that is, located far from the origin) are observations 1, 2, 3 and 17.

From Figure 4.8, observations 1, 2, 3 and 17 which have large h_z values, are also located far from the origin. Observation 4 appears to be swamped because it has a large h_z value but is not located far from the origin relative to other observations that are located on the first axis.

Figure 4.8: L-D plot for the Stack Loss data.



Results Based on the Studentized Residuals and Proposed Residual Measure (R_j)

The first column of Table 4.15 shows values of R_j and the Studentized residuals are shown on the second column. From the table, we see that observations 1 and 2 have R_j values that are greater that 2, and the value of R_j for observation 3 suggests that the observation is making a more-than-average contribution to the residuals of other observations, therefore we will consider the observation to be a regression outlier.

Values of the Studentized residuals, $r^*_{(i)}$, for the Stack Loss data are shown in the second column in Table 4.15. The Studentized residuals indicate the presence of only one outlier in the data set, observation 1, since it has a Studentized residual value greater than 2. However, the values of observations 2 and 3 suggest that the observations are possible regression outliers since their values are close to 2.

Table 4.15
Stack Loss data: Residuals

	R_j	$r^*_{(i)}$
1	2.316*	2.466♯
2	2.486*	1.962
3	1.919	1.962
4	0.607	1.055
5	0.336	0.048
6	0.536	0.048
7	0.557	0.149
8	0.531	0.250
9	0.132	−0.254
10	0.655	−0.355
11	0.890	−0.355
12	1.008	−0.456
13	0.674	−0.657
14	0.710	−0.557
15	1.621	−0.959
16	1.625	−1.060
17	0.699	−0.959
18	1.233	−0.959
19	1.054	−0.859
20	0.594	−0.254
21	0.819	−0.254

* $R_j > 2$.

♯ $r^*_{(i)} >\mid 2 \mid$.

Thus, when we use the procedure, we classify observations 1, 2, 3, 7, 12, 17 and 21 as regression outliers, whilst the Studentized residuals and R_j classify observation 1, 2 and 3 as the only regression outliers.

Example 3: Health Club Data

Results Based on the Diagonal Values of $\mathbf{H}_z$

The r_l, c_l, h_z and DIST values for the Health Club data are shown in Table 4.16, and the stem-and-leaf display is shown in Table 4.17.

Table 4.16
Health Club data: r_l, c_l, h_z and DIST values

	r_l						c_l					
	Axis 1 (61.14%)	Axis 2 (25.33%)	Axis 3 (25.33%)	Axis 4 (9.25%)	Axis 5 (4.28%)	h_i	Axis 1 (61.14%)	Axis 2 (25.33%)	Axis 3 (25.33%)	Axis 4 (9.25%)	Axis 5 (4.28%)	DIST
1	0.102*	0.016	0.012	0.069	0.037	0.236*	0.892*	0.045	0.015	0.041	0.007	0.371
2	0.040	0.023	0.014	0.024	0.006	0.107	0.785	0.143	0.037	0.032	0.003	0.166
3	0.010	0.011	0.001	0.016	0.033	0.071	0.653	0.220	0.008	0.072	0.047	0.049
4	0.011	0.003	0.035	0.000	0.008	0.057	0.649	0.049	0.292	0.001	0.010	0.054
5	0.001	0.022	0.012	0.007	0.018	0.060	0.086	0.662	0.165	0.048	0.039	0.033
6	0.017	0.042	0.006	0.003	0.011	0.079	0.535	0.424	0.026	0.007	0.007	0.101
7	0.004	0.027	0.004	0.000	0.060	0.095	0.287	0.584	0.039	0.001	0.088	0.048
8	0.033	0.162*	0.000	0.000	0.088	0.283*	0.380	0.597*	0.000	0.000	0.022	0.279
9	0.035	0.061	0.010	0.000	0.027	0.133	0.619	0.345	0.025	0.000	0.011	0.180
10	0.001	0.078	0.008	0.027	0.041	0.155	0.027	0.843	0.037	0.063	0.030	0.095
11	0.017	0.018	0.025	0.072	0.103	0.235**	0.506	0.171	0.103	0.151	0.069	0.106
12	0.011	0.047	0.003	0.000	0.015	0.076	0.425	0.548	0.014	0.001	0.012	0.087
13	0.026	0.192*	0.045	0.000	0.003	0.266*	0.278	0.655*	0.066	0.000	0.001	0.301
14	0.020	0.002	0.070	0.001	0.059	0.152	0.631	0.017	0.309	0.002	0.041	0.100
15	0.018	0.005	0.005	0.002	0.025	0.055	0.862	0.075	0.031	0.007	0.026	0.068
16	0.031	0.018	0.128	0.036	0.001	0.214**	0.547	0.101	0.309	0.043	0.000	0.185
17	0.009	0.006	0.230*	0.013	0.009	0.267*	0.212	0.041	0.722*	0.020	0.004	0.142
18	0.063	0.015	0.007	0.089	0.023	0.197**	0.835	0.065	0.012	0.081	0.007	0.243
19	0.004	0.001	0.011	0.002	0.007	0.025	0.617	0.072	0.264	0.021	0.026	0.019
20	0.052	0.004	0.009	0.018	0.000	0.083	0.933	0.024	0.021	0.022	0.000	0.181
21	0.007	0.020	0.085	0.015	0.056	0.183	0.246	0.234	0.435	0.038	0.046	0.086
22	0.065	0.005	0.027	0.001	0.001	0.099	0.924	0.022	0.052	0.001	0.000	0.229
23	0.204*	0.154	0.002	0.133	0.003	0.496*	0.778*	0.186	0.001	0.035	0.000	0.849
24	0.088	0.000	0.001	0.024	0.000	0.113	0.979	0.001	0.001	0.018	0.000	0.292
25	0.013	0.000	0.040	0.008	0.058	0.119	0.636	0.000	0.274	0.027	0.063	0.065
26	0.026	0.002	0.001	0.007	0.054	0.090	0.915	0.021	0.006	0.016	0.042	0.090
27	0.000	0.050	0.021	0.071	0.011	0.153	0.012	0.657	0.118	0.202	0.010	0.078
28	0.044	0.003	0.040	0.326	0.016	0.429**	0.602	0.015	0.075	0.304	0.005	0.238
29	0.044	0.001	0.084	0.034	0.001	0.164	0.756	0.005	0.198	0.040	0.000	0.189
30	0.005	0.013	0.065	0.000	0.225	0.308♯	0.221	0.176	0.390	0.000	0.213	0.074

* h_z large, $r_l > 0.067$ and $c_l > 0.5$.
♯ h_z large, and $r_l > 0.067$ on one of the axis but $c_l < 0.5$ on all five axes.
** h_z large, but $r_l < 0.067$ and $c_l > 0.5$ (or the reverse) on the kth axis.

Table 4.17
Health Club data: Stem-and-leaf display of the h_z values

stem		leaf	observation
0	|	3 6 6 6 7 8 8 8 9	
1	|	0 0 1 1 2 3 5 5 5 6 8	
2	|	0 1 3 4 7 7 8	18, 16, 11, 1, 13, 17, 8
3	|	1	30
4	|	3	28
5	|	0	23

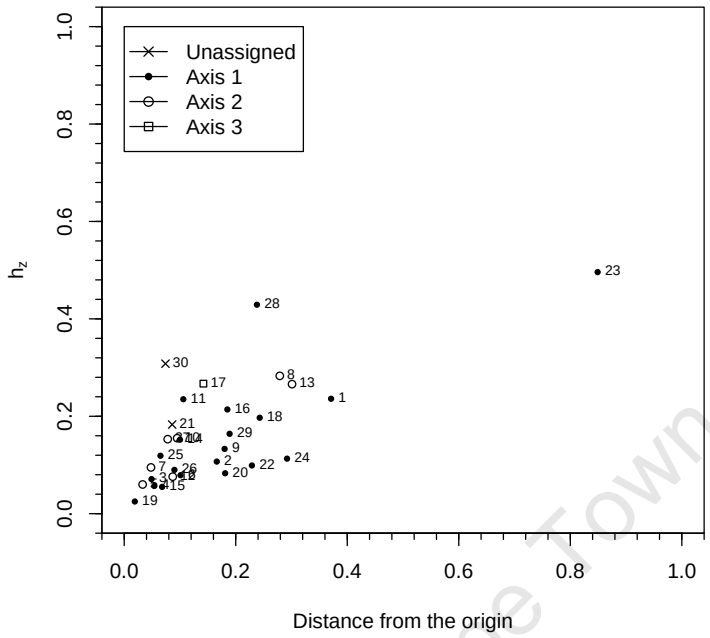
The decimal point is 1 digit to the left of the |.

From the tables, observations 1, 8, 11, 13, 16, 17, 18, 23, 28 and 30 have large h_z values. Observations 1 and 23 determine the direction of and are well explained by the first axis; observations 8 and 13 determine the direction of and are well explained by the second axis; observation 17 determines the direction of and is well explained by the third axis; observations 11, 16, 18 and 28 do not determine the direction of the axis whose direction they are located in. Observation 30 has a large h_z value but is not explained well by any one of the five axes. Of these observations with large h_z values, observations 1, 8, 13 and 23 contribute the most towards the total variance.

From Figure 4.9, we see that observations 1, 8, 13, 23 and 24 are located relatively far from the origin; observations 11 and 16, which are located in the direction of the first axis, are not located far from the origin, thus these observations appear to be swamped; observations 18 and 28, which are also located in the direction of the first axis, are borderline cases, thus will also treat these observations as outliers. Observation 30 does not lie in the direction of any axis, hence does not fit the structure of the bulk of the data, and we will therefore treat this observation as an outlier.

Notice that observation 24 has a low h_z value, but is located far from the origin, thus this observation appears to be masked.

Figure 4.9: L-D plot for the Health Club data.



Results Based on the Studentized Residuals and Proposed Residual Measure (R_j)

Table 4.18 shows R_j and $r^*_{(i)}$ values for the Health Club data. None of the observations are regression outliers according to these two measures, although the value of R_j for observation 24 suggests that the observation is making a more-than-average contribution to the residuals of other observations, therefore we will consider the observation to be a regression outlier.

Table 4.18
Health Club data: Residuals

	R_j	$r^*_{(i)}$
1	1.412	1.612
2	1.169	−1.156
3	1.026	−0.482
4	0.617	−0.204
5	0.695	0.367
6	1.262	0.850
7	0.130	−0.380
8	0.761	1.436
9	1.462	0.792
10	0.502	−0.189
11	1.108	0.323
12	0.819	−0.365
13	1.336	−1.346
14	1.591	−0.878
15	1.032	0.440
16	1.397	−1.522
17	0.145	0.484
18	1.076	1.041
19	0.390	−0.043
20	1.373	−1.112
21	0.564	0.294
22	1.778	−1.566
23	1.695	1.670
24	1.919	1.612
25	0.814	0.045
26	0.667	−0.907
27	0.500	−0.058
28	0.512	1.436
29	1.656	−1.742
30	0.595	−0.482

4.6 Discussion and Summary

In this chapter, we considered the response variable, $\mathbf{y}$, in determining the observations that deviate from the bulk of the data. Observations that deviate from the bulk of the data are known to easily influence the fit of the regression line, and the residuals, which take the response variable into account, are often examined to determine the observations

that may have influenced the fit of the regression line, hence affecting the regression coefficients.

The least squares method works by minimising the sum of squared deviations from the fitted line, and large deviations are often accommodated at the expense of less deviating observations. Thus, the existing diagnostic measures that are based on the residuals intended to detect regression outliers, may fail to reveal some outliers because observations that deviate from the bulk of the data may have small residuals since they may have pulled the fitted regression line in their direction.

We briefly reviewed the use of the Studentized residuals, which are transformed residuals which have been standardised to have equal variances. The Studentized residuals work well when the outlying observations are not being accommodated by the least squares fit, but may fail to highlight the regression outliers when these observations are being accommodated by the least squares fit. The Studentized residuals may also suffer from the effects of masking and swamping, as the artificial data set in example 2 demonstrated.

We also extended the procedure that was proposed in Chapter 3 to the diagonal values of the $\mathbf{H}_z$ matrix. Using the L-D plot has proven to be informative in highlighting observations that may be masked or swamped. The one drawback of using the proposed procedure which we have mentioned already in Chapter 3, is that too many observations may be declared incorrectly as regression outliers because observations with large h_z values that are not explained well by any axis are automatically treated as leverage points if they contribute highly to the determination of the direction of at least one of the axes. Another potential drawback with using the diagonal values of the $\mathbf{H}_z$ matrix is that we are not able to differentiate between leverage points and regression outliers, since h_z may be large because of a large h_i value and/or a large residual value.

The proposed measure, R_j, which should be used in conjunction with the Studentized residuals (or other transformed residuals), differs from existing measures in that it provides insight into the role that each observation plays in determining the displacement of other observations from the least squares fit.

We suggested a guideline to determine a large value of R_j ($R_j > 2$), and found the observations flagged as regression outliers by R_j and the Studentized residuals on the real data sets to be similar, except in the Health Club data set, where observation 24, which we found to be masked in Chapter 3, is flagged as a potential regression outlier by R_j, but the observation is not flagged by the Studentized residuals.

For the Hawkins, Bradu and Kass data set, the first 13 observations, which are known

leverage points were flagged as regression outliers using our proposed measure, R_j, even though only the first 10 observations are known to be regression outliers. This is a potential drawback with using R_j when observations are clustered nearly in the same area. As already pointed out in Chapter 3, a large h_{ij} value also indicates that observation i and observation j are situated on "the same side of the bulk of the cases nearly on the same line away from the centroid of the cases" (Gray and Ling, 1984). Thus leverage points that are located in the same area but far from the origin, and also lie in the direction of the fitted line may also be flagged as regression outliers when using R_j.

The two measures and the procedure based on the diagonal values of the $\mathbf{H}_z$ matrix appear to offer valuable insight into the nature of outlyingness of the observations, thus we recommend the use of all three measures in determining regression outliers. The procedure based on the diagonal values of the $\mathbf{H}_z$ matrix to determine the outliers and whether any observations are masked or swamped, and then R_j and the Studentized residuals to determine the real regression outliers.

Table 4.19 presents a summary of the observations that were identified as regression outliers, in each of the real data sets.

Table 4.19
Summary of regression outliers

Data	Regression Outliers
HBK	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13
Stack Loss	1, 2, 3
Health Club	24

In the next chapter, a diagnostic measure is proposed to determine which of the outlying observations identified in this chapter influence the regression coefficients.

Chapter 5

Identifying Influential Observations

The regression estimates such as the coefficients, are known to be easily affected by outlying observations, and in this chapter, we propose a measure which highlights the regression outliers identified in the preceding chapter that have a disproportionate effect on the individual regression coefficients.

Once again, we consider an existing measure, DFBETAS, that is intended to also measure the impact of an observation on the individual regression coefficients, and use the artificial data set used in Chapter 4 to illustrate the problems encountered when using the DFBETAS. The proposed measure appears to work well in identifying influential observations.

5.1 Introduction

The measures presented in the preceding two chapters to identify leverage points and regression outliers draw to our attention the observations that deviate from the bulk of the data, but fail to point out observations that are influential. Observations are said to be influential if they have a disproportionate effect on the estimated regression coefficients. Note that both leverage points and regression outliers need not be influential, and if they are influential, they need not influence all the regression quantities equally (Chatterjee and Hadi, 1986a).

In this chapter, an influence measure is proposed that measures the impact of an observation on the individual regression coefficients (refer to Chatterjee and Hadi (1986a) for a classification of 'influence' measures). In the next section, the diagnostic measure is presented. In section 5.3, we briefly review the use of the DFBETAS, an existing measure that determines the impact of an observation on the individual regression coefficients in identifying influential observations, and use the same artificial data set used in Chapter 4 to illustrate the problems encountered when using the DFBETAS. We then consider the same three real data sets that we have used in the preceding two chapters in section 5.4, and end the chapter with a discussion of the main findings of the results presented in the chapter.

5.2 The Diagnostic for Identifying Influential Observations

Let $\mathbf{X} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^T$ be the SVD of the standardised matrix $\mathbf{X}$. The regression coefficients are given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{V}\mathbf{D}_\alpha^{-1}\mathbf{U}^T\mathbf{y}$$

We can express the jth regression coefficient as

$$\begin{aligned}\hat{\beta}_j &= \sum_{i=1}^{n}\sum_{k=1}^{m}\frac{v_{jk}u_{ik}y_i}{\alpha_k} \\ &= \sum_{k=1}^{m}\frac{v_{jk}}{\alpha_k}\left[\sum_{i=1}^{n}u_{ik}y_i\right] \\ &= \sum_{i=1}^{n}y_i\left[\sum_{k=1}^{m}\frac{v_{jk}u_{ik}}{\alpha_k}\right]\end{aligned}$$

The sum $$\sum_{k=1}^{m} \frac{v_{jk}u_{ik}}{\alpha_k} \ (\equiv a_{ij})$$

depends on $\mathbf{X}$ alone. It may thus be interpreted as a measure of influence of the ith observation on the jth regression coefficient.

The quantity $y_i a_{ij}$ is the contribution which the ith observation makes to the jth regression coefficient. Note that the final value of $\hat{\beta}_j$ is the sum of the $y_i a_{ij}$ over all i observations (for $i = 1, 2 \ldots, n$). Thus each observation has an influence, small or large, in the determination of $\hat{\beta}_j$. This decomposition of $\hat{\beta}_j$ enables us to examine each observation in relation to each $\hat{\beta}_j$. Thus quantities helpful in this analysis are

$$\frac{\mid y_i a_{ij} \mid}{\sum_{i=1}^{n} \mid y_i a_{ij} \mid} \qquad \text{and} \qquad B_{ij} = \frac{n \times \mid y_i a_{ij} \mid}{\sum_{i=1}^{n} \mid y_i a_{ij} \mid}$$

The former measures the proportional absolute contribution of observation i to $\hat{\beta}_j$, and the latter is an easy-to-interpret statistic — for a given coefficient, large values of B_{ij} indicate a disproportionate contribution by an observation in the determination of the coefficient. Thus, if all observations contribute equally to the coefficient, then B_{ij} will equal 1. Values of B_{ij} greater than 1 therefore indicate that an observation makes a more-than-average contribution to the coefficient. As a crude cut-off value, we will consider all observations which have B_{ij} values greater than 2 to be influential on the jth coefficient.

5.3 DFBETAS

An existing measure that is intended to assess the impact of deleting the ith observation on the jth coefficient is DFBETAS (Belsley *et al.*, 1980) which is given by

$$\mathrm{D}_{ij}(i) = \frac{\hat{\beta}_j - \hat{\beta}_j(i)}{\mathrm{Var}(\hat{\beta}_j)} = \frac{\hat{\epsilon}_i}{\hat{\sigma}(i)(1 - \mathrm{h}_i)} \frac{c_{ji}}{\sqrt{\sum_{k=1}^{n} c_{jk}^2}}$$

where $\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

Values of $\mathrm{D}_{ij}(i)$ may be positive or negative, with a positive value indicating that the deletion of the ith observation results in a smaller coefficient, whilst a negative value

indicates that the deletion of the ith observation results in a larger coefficient.

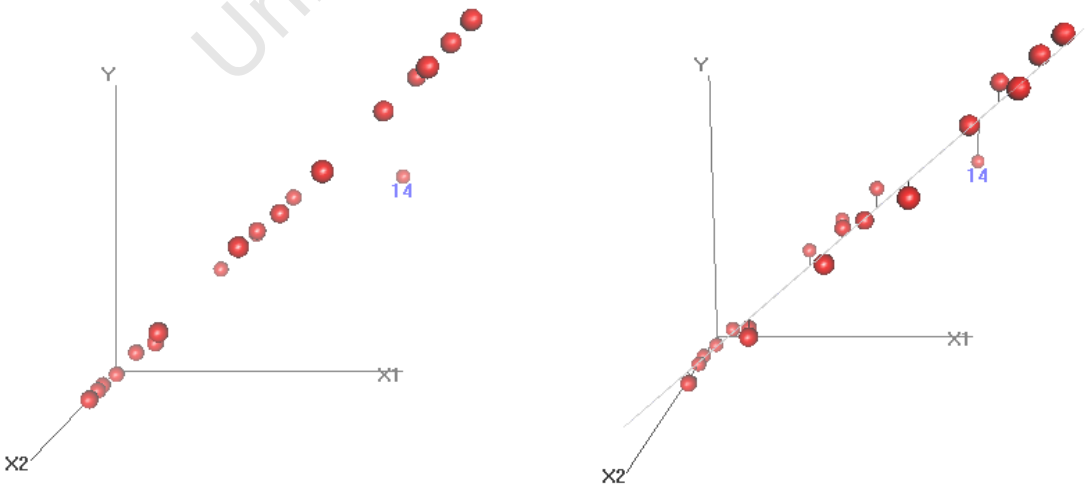
The drawback to using DFBETAS as diagnostics is that they are a function of the residuals which are a poor measure of fit, and the diagonal values of the hat matrix, h_i, which are affected by the masking and swapping effects. Therefore the DFBETAS are also prone to the same problems as the Studentized residuals and the diagonal values of the hat matrix (Rousseeuw and Leroy, 1987). Observations for which $D_{ij}(i)$ values exceed $| 2/\sqrt{n} |$ are generally said to influence the jth coefficient(s) (Belsley *et al.*, 1980).

The following examples, based on the same artificial data that we used in Chapter 4, will be used to illustrate the similarities and differences in the results obtained using the DFBETAS and the proposed measure. (Note that all computations were performed in R (R Development Core Team, 2008), and the source codes written for the measure are included in Appendix C, section C.3 (p. C-3).)

Example 1

Recall that for the first example, observation 14 is both a leverage point and a regression outlier that is not accommodated excessively by the least squares fit at the expense of the other non-outlying observations. Figures 4.1(a) and 4.1(b) have been reproduced in Figures 5.1(a) and 5.1(b).

Figure 5.1: Example 1 – (a) 3-Dimensional scatter plot of two explanatory variables and the response variable. (b) The best linear least squares fit



(a) (b)

Table 5.1 shows B_{ij} and DFBETAS values for this data set. From the table, we see that observation 14 has a large B_{ij} value on the second coefficient, although the B_{ij} value on the first coefficient is also considerably large. Observation 14 also has large DFBETAS values on both coefficients.

Table 5.1
B_{ij} and DFBETAS values for Example 1

	B_{ij}		DFBETAS	
	X1	X2	X1	X2
1	1.017	0.824	−0.041	−0.022
2	2.023	0.339	0.041	0.005
3	2.332	1.599	0.274	−0.123
4	1.830	1.058	−0.029	−0.011
5	1.448	1.457	−0.023	−0.015
6	0.882	1.437	0.256	−0.273
7	0.005	0.015	0.006	−0.011
8	0.848	0.139	0.106	−0.011
9	0.007	0.023	0.078	−0.158
10	0.024	0.086	0.162	−0.385
11	0.076	0.682	0.055	−0.323
12	0.011	0.206	0.001	−0.010
13	0.147	1.371	0.046	−0.279
14	1.762	2.867*	−8.458♯	9.005♯
15	0.761	2.388	−0.016	−0.033
16	0.340	0.730	0.223	−0.314
17	0.653	1.047	0.018	0.019
18	1.840	1.104	0.143	0.056
19	1.963	1.782	0.146	0.086
20	2.032	0.845	0.330	−0.090

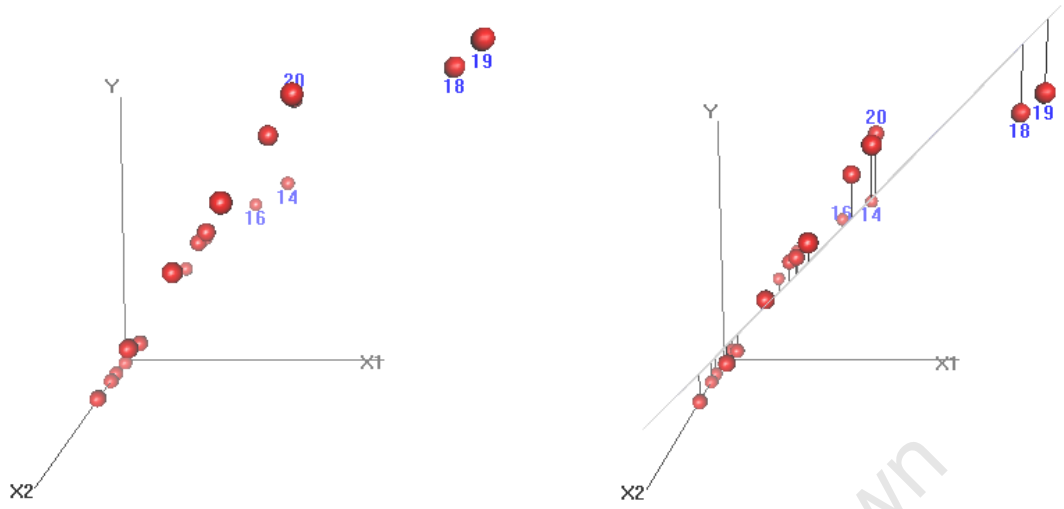
* $B_{ij} > 2$

♯ DFBETAS $> |\, 0.447 \,|$

Example 2

For the second example, observations 14, 16, 18 and 19 are leverage points, and observations 18 and 19 have been modified to be regression outliers. Figures 4.2(a) and 4.2(b), which illustrate that the two observations are being accommodated by the least squares fit at the expense of other non-outlying observations (although their residuals are still large), have been reproduced below in Figures 5.2(a) and 5.2(b).

Figure 5.2: Example 2 – (a) 3-Dimensional scatter plot of two explanatory variables and the response variable. (b) The best linear least squares fit



(a) (b)

Table 5.2 shows B_{ij} and DFBETAS values for this data set. From the table, we see that both observations 18 and 19 have large B_{ij} and DFBETAS values on the first coefficient.

Table 5.2
B_{ij} and DFBETAS values for Example 2

	B_{ij}		DFBETAS	
	X1	X2	X1	X2
1	0.857	0.886	0.055	0.035
2	1.726	0.397	0.187	0.027
3	2.169	1.742	0.320	−0.160
4	1.498	1.186	0.141	0.070
5	1.151	1.608	0.103	0.090
6	0.946	1.616	0.150	−0.160
7	0.010	0.005	−0.014	−0.004
8	0.811	0.191	0.093	−0.014
9	0.010	0.027	−0.033	0.055
10	0.021	0.089	0.102	−0.272
11	0.080	0.718	0.030	−0.170
12	0.074	0.261	−0.031	0.068
13	0.360	1.594	−0.070	0.193
14	1.232	2.184	−0.040	0.044
15	0.075	2.922	0.017	0.404
16	0.621	1.425	0.140	−0.200
17	0.192	1.373	0.046	0.206
18	3.389*	0.195	−1.082♯	0.039
19	3.401*	1.012	−1.024♯	−0.190
20	1.379	0.570	0.399	−0.103

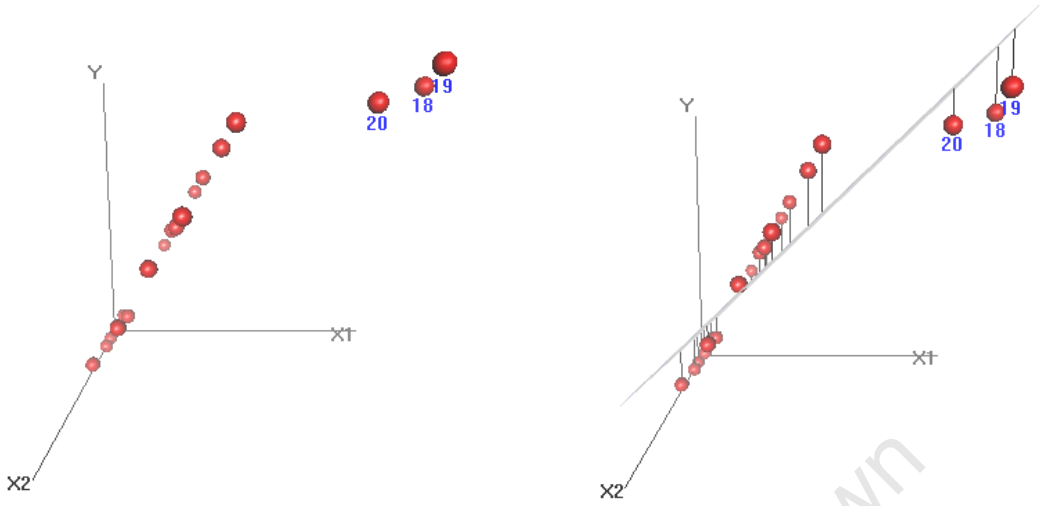
* $B_{ij} > 2$

♯ DFBETAS $>$ | 0.447 |

Example 3

For the third example, recall that observations 18, 19 and 20 have been modified to be regression outliers that are accommodated by the least squares fit at the expense of other non-outlying observations, and observations 19 and 20 have small Studentized residuals, whilst observation 20 is masked as seen from the L-D plot in Figure 4.6. Figures 4.3(a) and 4.3(b) have been reproduced in Figures 5.3(a) and 5.3(b).

Figure 5.3: Example 3 – (a) 3-Dimensional scatter plot of two explanatory variables and the response variable. (b) The best linear least squares fit



(a) (b)

Table 5.2 shows B_{ij} and DFBETAS values for this data set. From the table, we see that observation 18 has large B_{ij} and DFBETAS values on both coefficients, whilst observations 19 and 20 have large B_{ij} values on the first coefficient only. Both observations (19 and 20) have small DFBETAS values on both coefficients. Thus the two observations (19 and 20), which also have small Studentized residuals, do not have large DFBETAS.

Table 5.3
B_{ij} and DFBETAS values for Example 3

	B_{ij}		DFBETAS	
	X1	X2	X1	X2
1	0.389	0.955	0.034	0.066
2	1.363	0.258	0.168	0.025
3	2.608	2.136	0.364	−0.233
4	0.885	1.119	0.101	0.100
5	0.414	1.639	0.046	0.141
6	1.340	1.684	0.194	−0.191
7	0.008	0.037	0.010	−0.036
8	0.712	0.144	0.081	−0.013
9	0.017	0.026	−0.071	0.086
10	0.056	0.111	0.235	−0.362
11	0.367	0.862	0.099	−0.181
12	0.112	0.187	−0.047	0.062
13	0.766	1.485	−0.189	0.286
14	0.489	0.637	0.168	−0.171
15	0.597	2.294	−0.140	0.421
16	0.581	0.991	0.327	−0.436
17	0.040	0.838	−0.010	0.161
18	4.809*	2.868*	−1.252♯	0.584♯
19	2.432*	1.318	−0.351	−0.149
20	2.014*	0.412	−0.222	−0.036

* $B_{ij} > 2$
♯ DFBETAS $> |\, 0.447 \,|$

5.4 Illustrative Examples

Example 1: Hawkins, Bradu and Kass Data

Recall that for the Hawkins, Bradu and Kass data set, the first 14 observations are classified as leverage points, with the first 10 observations also constructed to be regression outliers. From Chapter 4, we classified the first 13 observations to be regression outliers using R_j and the Studentized residuals (observations 11 to 13 have large R_j values). Table 5.4 shows B_{ij} values for the HBK data. From the table, we see that the first 10 observations and observation 12 are flagged as having a disproportionate effect on at least one of the coefficients.

Table 5.5 shows the values for DFBETAS for the HBK data set. Only observations 11 to 14, which have large h_z values, have large DFBETAS values greater than $|\, 0.231 \,|$ on at least one coefficient. The coefficients affected can be seen from the table.

Table 5.4
Hawkins, Bradu and Kass data – Influential observations (B_{ij})

	X1	X2	X3
1	3.134*	1.987	1.025
2	1.028	0.178	2.151*
3	2.559*	6.852*	4.977*
4	2.917*	2.859*	5.412*
5	0.025	3.296*	4.052*
6	5.481*	1.608	0.501
7	4.257*	0.716	1.201
8	1.533	3.260*	2.985*
9	2.825*	4.034*	6.122*
10	3.790*	5.993*	8.029*
11	0.434	0.377	0.869
12	0.038	2.172*	1.880
13	0.240	0.471	0.327
14	0.513	5.063	2.845
15	1.722	0.416	1.378
16	0.784	0.314	0.718
17	1.029	0.625	0.098
18	0.528	0.273	0.187
19	0.559	0.641	0.160
20	0.714	0.398	0.731
⋮	⋮	⋮	⋮

* Regression outlier, and $B_{ij} > 2$.

Table 5.5
Hawkins, Bradu and Kass Data – Influential observations (DFBETAS)

	X1	X2	X3
1	0.115	−0.061	0.039
2	−0.043	0.006	0.092
3	0.079	−0.179	0.159
4	−0.084	−0.069	0.160
5	−0.001	−0.089	0.134
6	0.200	−0.049	−0.019
7	0.188	0.027	−0.054
8	0.060	−0.107	0.120
9	−0.085	−0.102	0.189
10	−0.125	−0.166	0.271
11	0.236	0.172	−0.486
12	−0.024	1.178	−1.247
13	−0.255	−0.420	0.356
14	0.329	−2.727	1.873
15	−0.061	−0.012	0.050
16	0.091	0.031	−0.086
17	−0.039	0.020	0.004
18	−0.014	0.006	0.005
19	−0.027	0.026	−0.008
20	0.024	0.011	−0.026
⋮	⋮	⋮	⋮

Thus, for the HBK data, we observe an instance when the DFBETAS fail to correctly identify the influential observations. B_{ij} correctly identifies all influential observations (including observation 12, which, because of the large R_j value, we would classify as having a disproportionate effect on the second coefficient).

Example 2: Stack Loss Data

Recall that for the Stack Loss data set, only observations 1, 2 and 3 are classified as regression outliers. Table 5.6 shows B_{ij} values for the Stack Loss data. From the table, we see that observations 1, 2, and 3 have a disproportionate effect on at least one coefficient.

Table 5.6
Stack Loss Data – Influential observations (B_{ij})

	Air. Flow	Water. Temp	Acid. Conc.
1	4.287*	1.887	2.703*
2	3.613*	1.501	2.866*
3	2.705*	0.055	0.390
4	1.253	3.185	0.098
5	0.007	0.028	0.004
6	0.032	0.086	0.004
7	0.268	0.449	0.339
8	0.450	0.753	0.569
9	0.434	0.791	0.132
10	0.580	1.054	0.722
11	0.252	1.054	0.443
12	0.610	1.907	0.403
13	0.939	1.951	0.857
14	0.125	0.975	1.506
15	1.324	0.331	2.016
16	1.137	0.366	1.068
17	0.153	0.834	3.930
18	0.843	0.835	1.482
19	1.293	1.792	1.013
20	0.036	0.030	0.277
21	0.659	1.136	0.177

* Regression outlier, and $B_{ij} > 2$.

Table 5.7
Stack Loss Data – Influential observations (DFBETAS)

	Air. Flow	Water. Temp	Acid. Conc.
1	0.385	0.099	−0.202
2	−0.241	−0.059	0.159
3	0.379	−0.005	−0.045
4	−0.403	0.601	0.026
5	0.012	−0.029	−0.007
6	0.103	−0.165	−0.012
7	0.260	−0.255	−0.273
8	0.149	−0.147	−0.157
9	0.300	−0.320	−0.076
10	0.121	−0.128	−0.125
11	0.105	−0.258	0.154
12	0.226	−0.414	0.124
13	−0.113	0.138	0.086
14	0.001	0.003	−0.007
15	−0.190	−0.028	0.240
16	−0.052	−0.010	0.041
17	0.019	−0.060	0.403
18	0.022	−0.013	0.032
19	0.051	−0.041	0.033
20	−0.010	0.005	−0.065
21	−1.537	1.554	−0.344

Table 5.7 shows the values for DFBETAS for the Stack Loss data set. None of the regression outliers are flagged as influential, and only observations 4 and 21, which have large h_z values, have $D_{ij}(i) >| 0.436 |$ on at least one coefficient.

Example 3: Health Club Data

The third example that we consider is the Health Club data set. Recall that for these data, observation 24 is the only regression outlier. Table 5.8 shows that the observation is most influential on the first coefficient using B_{ij}. The observation does not have a large DFBETAS value on any of the coefficients (refer to Table 5.9), instead, a majority of the observations with large h_z values have $D_{ij}(i) >| 0.365 |$ on at least one coefficient.

Table 5.8
Health Club Data – Influential observations (B_{ij})

	X1	X2	X3	X4
1	2.326	0.015	3.139	3.220
2	1.109	1.709	0.619	0.700
3	0.640	0.053	0.582	0.107
4	0.078	0.273	0.031	0.149
5	0.316	0.197	0.535	0.210
6	0.616	0.131	1.071	0.647
7	0.119	0.114	0.466	0.216
8	0.893	3.434	0.569	1.112
9	0.492	1.179	0.638	0.139
10	0.130	0.311	0.050	0.119
11	0.877	0.175	0.380	0.520
12	0.195	0.483	0.114	0.123
13	0.399	0.579	2.663	2.664
14	0.661	1.633	1.110	1.688
15	0.407	0.046	0.062	0.124
16	2.726	4.746	2.302	2.118
17	0.215	1.427	0.933	0.953
18	1.990	2.257	1.860	1.364
19	0.001	0.034	0.009	0.008
20	1.351	0.379	0.060	0.649
21	0.068	0.674	0.131	0.189
22	1.085	1.690	0.340	1.814
23	1.805	1.468	4.760	3.514
24	2.379*	0.448	0.931	0.132
25	0.008	0.049	0.020	0.018
26	0.780	1.088	0.928	0.396
27	0.084	0.015	0.002	0.121
28	5.071	0.725	2.191	4.096
29	2.613	3.141	2.747	2.283
30	0.567	1.527	0.755	0.606

* Regression outlier, and $B_{ij} > 2$.

Table 5.9
Health Club Data – Influential observations (DFBETAS)

	X1	X2	X3	X4
1	−0.196	−0.001	0.274	0.272
2	0.094	0.121	−0.054	−0.059
3	−0.118	−0.008	0.111	0.020
4	−0.031	−0.091	−0.013	0.059
5	−0.040	0.021	0.069	−0.026
6	−0.019	−0.004	0.035	−0.020
7	−0.052	0.042	0.211	−0.095
8	−0.182	0.589	−0.120	0.228
9	−0.071	−0.142	0.095	−0.020
10	0.138	−0.277	0.055	−0.127
11	−0.573	0.096	0.257	0.341
12	−0.043	0.088	−0.026	−0.027
13	0.012	0.014	−0.081	0.078
14	−0.095	0.196	0.164	−0.242
15	−0.092	0.009	0.015	0.028
16	0.212	−0.310	−0.185	0.165
17	0.096	−0.536	−0.432	0.428
18	−0.090	0.085	0.087	0.062
19	0.001	−0.053	−0.017	0.015
20	0.041	0.010	−0.002	−0.020
21	0.060	−0.499	−0.12	0.167
22	0.058	−0.076	−0.02	0.098
23	0.190	0.130	−0.518	−0.370
24	0.076	0.012	−0.031	−0.004
25	−0.038	−0.203	−0.103	0.090
26	−0.130	0.152	0.160	0.066
27	0.168	−0.026	0.005	−0.245
28	1.160	0.139	−0.519	−0.940
29	0.237	−0.239	−0.258	0.208
30	−0.429	0.970	0.591	−0.459

5.5 Discussion and Summary

The regression quantities such as the coefficients, are known to be easily affected by outlying observations. In this chapter, we proposed a measure, B_{ij}, that may be used

to identify the regression outliers that have a disproportionate effect on the individual regression coefficients. The B_{ij} values however do not tell us the actual impact the observation has on the estimation of the coefficient(s).

We compared B_{ij} to DFBETAS, an existing measure that determines the impact of an observation on the individual regression coefficients. The DFBETAS are a function of the residuals which are a poor measure of fit, and the diagonal values of the hat matrix, h_i, hence will suffer from the effects of masking and swamping like the Studentized residuals. The artificial data sets illustrated instances when the DFBETAS give accurate results, and instances when the DFBETAS give inaccurate results. For the real data sets, the DFBETAS failed to correctly identify any of the suspected regression outliers as being influential on the regression coefficients.

Table 5.10 presents a summary of the influential observations.

Table 5.10
Summary of influential observations

Data	Observations
HBK	1, 2, 3, 4, 5, 6, 7, 8, 9 ,10, 12
Stack Loss	1, 2, 3
Health Club	24

In the next chapter, measures for detecting the presence of collinear relationships among the explanatory variables are presented.

Chapter 6

Identifying Collinear Variables

Collinearity is a term used to denote the presence of linear relationships or 'near' linear relationships among the explanatory variables in regression analysis, and is known to have an adverse effect on the regression coefficients. A number of approaches have been proposed to identify the explanatory variables that are involved in collinear relationships, and to detect the coefficients that are most adversely affected. Two approaches that we consider in this chapter are examining the magnitude of the eigenvectors of $\mathbf{X}^T\mathbf{X}$ that correspond to small singular values, and decomposing the regression coefficients and examining the magnitude of the coefficient values that correspond to small singular values. The labelling of values as 'large' is arbitrary in both approaches, and in this chapter we develop a means of quantifying the meaning of 'large' for both methods.

Two real examples that have appeared in the literature are used to illustrate the operation of the proposed thresholds.

6.1 Introduction

Collinearity is a term used to denote the presence of linear relationships or 'near' linear relationships among the explanatory variables in regression analysis (Silvey, 1969). Collinearity is therefore a problem of the $\mathbf{X}$ matrix. Collinearity is known to have an adverse effect on the regression coefficients, in particular, the regression coefficients may be too large and have the wrong signs, and the variances of the coefficients tend to be much larger for the explanatory variables involved in collinear relationships than for uncorrelated explanatory variables (Gunst and Mason, 1980).

It is well known that collinearities reveal themselves by having singular values (α_k's) that are close to zero (refer to Belsley *et al.* (1980), Mandel (1982), and Hawkins and Fatti (1984)). Measures for detecting the presence of collinearity in the $\mathbf{X}$ matrix have been developed as a result of trying to understand the extent of collinearity in the $\mathbf{X}$ matrix, which variables are involved in the collinearity, and the extent to which the regression coefficients have been affected by the presence of collinearity. These include using *inter alia*:

Variance-inflation factors (VIFs): A variance-inflation factor indicates how much larger the variance of the coefficient estimate, var($\hat{\beta}_j$), will be for "collinear data than for orthogonal data – where each VIF is 1.0" (Mansfield and Helms, 1982).

If we let $\mathbf{X} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^T$ be the SVD of the $n \times m$ standardised matrix $\mathbf{X}$, of rank k, where $n \geq m$ and therefore $k = m$ (that is, we assume that $\mathbf{X}$ is of full column rank), then the variance-inflation factor for the jth variable is (Walker, 1989):

$$\text{VIF}(j) = \sum_{k=1}^{m} \frac{v_{jk}^2}{\alpha_k^2}$$

VIF(j) values greater that 10 are regarded as indicating the existence of collinearity among the explanatory variables.

Condition indices: A condition index ($\eta_k = \alpha_1/\alpha_k$) is used to identify the presence and strength of collinearity among the explanatory variables. Belsley *et al.* (1980) suggested that a condition index of between 5 and 10 indicates weak collinearity, and a condition index between 30 and 100 indicates moderate to strong collinearity. Thus, the stronger the collinearity among the explanatory variables, the higher the condition index.

Variance-decomposition proportions (VDPs): Variance-decomposition proportions measure the proportion of variance in the regression coefficients that is accounted for by each axis. The variance-decomposition proportion for the jth variable is given by:

$$\text{VDP}(j) = \frac{v_{jk}^2/\alpha_k^2}{\sum_{k=1}^{m}(v_{jk}^2/\alpha_k^2)}$$

Belsley *et al.* (1980) recommended that values greater than 0.5 be used as a threshold for identifying variables that are involved in the collinearity for an axis associated with a large condition index. However Belsley (1982) cautioned against the use of this threshold "when a given variate is involved in several coexisting near dependencies, for its variance-decomposition proportions can then be distributed across the VDP(j)'s so that all are relatively small".

Eigenvalues: A small eigenvalue (α_k^2) of the $\mathbf{X}^T\mathbf{X}$ matrix, is used to indicate the presence and strength of collinearity in a data set. Hocking (2003) suggested that if the smallest eigenvalue is less than 0.05, then this indicates serious collinearity, and if the smallest eigenvalue is less than 0.10, then this indicates moderate collinearity.

Large values of the eigenvectors of $\mathbf{X}^T\mathbf{X}$ (or the right singular vectors) that are associated with the small eigenvalues are used to indicate the number of explanatory variables that are involved in the collinearity and the nature of collinearity among the explanatory variables. The labelling of values as 'large' however is arbitrary, and in this chapter, we develop a means of quantifying the meaning of 'large' for values of the eigenvectors of $\mathbf{X}^T\mathbf{X}$ (or right singular vectors) that are associated with near-zero eigenvalues (or singular values).

Another approach that may be used to indicate the variables that are involved in the collinearity is that of decomposing the regression coefficients. Large coefficient values that are associated with near-zero singular values are used to indicate the variables that are involved in the collinearity. The labelling of values as 'large' however is also arbitrary, and we develop a means of quantifying the meaning of 'large' for the coefficient values that are associated with near-zero singular values.

This chapter is organised as follows. In section 6.2, we consider a measure that is based on the values of the eigenvectors of $\mathbf{X}^T\mathbf{X}$, and develop a means of quantifying the meaning of a 'large' eigenvector of $\mathbf{X}^T\mathbf{X}$ that is based on the values of the squared right singular vectors, using the decomposition of variance of the columns of the $\mathbf{X}$ matrix presented in

Chapter 2. In section 6.3, we also develop a means of quantifying the meaning of 'large' for coefficient values that are associated with near-zero singular values. In both sections 6.2 and 6.3, computations of the thresholds on two real data sets are also provided to illustrate their use. We then briefly mention collinear-influential observations in section 6.4, and end the chapter with a discussion of the main findings of the results presented in the chapter.

6.2 Using the Eigenvectors of $\mathbf{X}^T\mathbf{X}$ to Identify the Variables that are Involved in the Collinearity

In this section, we consider the values of the eigenvectors of $\mathbf{X}^T\mathbf{X}$, and develop a means of quantifying the meaning of a 'large' eigenvector that is based on the values of the squared right singular vectors, to identify the explanatory variables that are involved in collinear relationships.

Recall from Chapter 2 that,

$$\frac{g_{jk}^2}{\sum_{j=1}^{m} g_{jk}^2} = v_{jk}^2 \qquad \text{for all} \quad j = 1, 2, \ldots, m \tag{6.1}$$

is the proportion of the variance due to the kth principal axis that is explained by the jth variable, or the contribution the jth variable makes to the kth principal axis. The j variables involved in the collinearity will be those that have large values for (6.1) when α_k is near-zero. This is because

$$\mathbf{X}g_k = \mathbf{X}v_k\alpha_k = u_k\alpha_k^2$$
$$\text{i.e. } \mathbf{X}v_k = u_k\alpha_k$$

and since

$$\alpha_k \approx 0, \quad \mathbf{X}v_k \approx 0$$

That is, $\sum_{k=1}^{m} x_{ik}v_{jk} \approx 0 \qquad \text{for } i = 1, 2, \ldots, n$

There is thus a linear combination of columns of $\mathbf{X}$ which is near-zero, implying a collinearity. That is, collinearity involves those columns for which the v_{jk} (and hence g_{jk}) are relatively large when α_k is near-zero.

A large value for (6.1), a value of the squared right singular vector, implies that the kth axis is, to a large extent, determined by (or dominated by) the jth variable. If an axis is determined equally by all the variables, then this statistic will average $1/m$. Therefore we recommend that values greater than $2/m$ be used to identify variables that dominate a particular axis.

Notice that this method of identifying the explanatory variables that are involved in the collinearity is equivalent to that of examining the magnitude of the values of the eigenvectors of $\mathbf{X}^T\mathbf{X}$, that are associated with near-zero singular values or near-zero eigenvalues.

An advantage of using the eigenvectors of $\mathbf{X}^T\mathbf{X}$ to identify the explanatory variables that are involved in the collinearity is that the nature of relationship between the collinear variables can be determined, since $\mathbf{X}v_k \approx 0$, and writing this equation in terms of the unscaled variables identifies the relationships (Hocking, 2003).

A disadvantage of using the eigenvectors of $\mathbf{X}^T\mathbf{X}$ that are associated with near-zero singular values to identify the explanatory variables that are involved in collinear relationships is that the labelling of values as 'large' is arbitrary. Note though that as a consequence of using the proposed threshold of $2/m$ for values of the squared right singular vectors to identify the explanatory variables that are involved in the collinearity when α_k is near-zero, we are able to recommend a threshold that may be used for the eigenvectors of $\mathbf{X}^T\mathbf{X}$. Since the proposed threshold of $2/m$ is for values of the squared right singular vectors, we will consider values of the eigenvectors of $\mathbf{X}^T\mathbf{X}$ to be 'large' if they exceed $\mid \sqrt{2/m} \mid$.

Note that the threshold of $2/m$ (or $\mid \sqrt{2/m} \mid$ for the eigenvectors of $\mathbf{X}^T\mathbf{X}$) should be applied when the explanatory variables are involved in single dependencies. For explanatory variables involved in multiple dependencies, the threshold should be lowered as the proportion of the variance attributable to the jth explanatory variable may be distributed across multiple principal axes (that is, variable j determines the direction of multiple axes), or v_{jk} may be small because one or more of the explanatory variables that are involved in the multiple dependency may be orthogonal to another explanatory variable(s) involved in the same dependency (refer to Belsley and Klema (1974) for a discussion on this topic).

Condition indices ($\eta_k = \alpha_1/\alpha_k$'s) will be used to identify the presence and strength of collinearities in the $\mathbf{X}$ matrix. Thus the two or more explanatory variables that are involved in the collinearity will be those variables that have large values of the squared right singular vectors (or the eigenvectors of $\mathbf{X}^T\mathbf{X}$) on an axis (or axes) with a large condition index (or indices).

6.2.1 Illustrative Examples

In the examples that follow, all computations were performed in R (R Development Core Team, 2008), and the source codes written for each measure are included in Appendix C, section C.4 (p. C-4).

Example 1: Mason and Gunst Data

The first data set that we consider is the Mason and Gunst data set, found in Gunst and Mason (1980), Appendix A, p. 358, and has been reproduced in Appendix A, p. A-6. The data set is a compilation of entries from a much larger database from data set 31 of Loether, McTavish and Voxland (1975, cited in Gunst and Mason (1980)), and consists of one response variable, gross national product per capita, 1957, U.S. Dollars (GNP), and six explanatory variables: infant deaths per 1000 live births (INFD); number of inhabitants per physician (PHYS); population per square kilometre (DENS); population per 1000 hectares of agricultural land (AGDS); percentage literate of population aged 15 and over (LIT), and students enrolled in higher education per 100000 population (HIED).

Mason and Gunst (1985a) observed collinearity between DENS and AGDS, using the smallest eigenvalue (= 0.027) of the correlation matrix of the explanatory variables, the values on the corresponding eigenvector, and variance inflation factors. Walker (1989) and Hadi (1988) have also analysed the same data set and obtained results similar to those obtained by Mason and Gunst (1985a).

When we re-examined the data using the squared right singular vectors, the condition index corresponding to the smallest singular value indicated the presence of weak collinearity in the data set (refer to Table 6.1). Table 6.1 shows the singular values, condition indices, values of the squared right singular vectors for each axis, and the values of the eigenvectors of $\mathbf{X}^T\mathbf{X}$ (in square brackets). Since the last axis indicates the presence of collinearity, with a condition index of 9.651, we will only consider this axis in the interpretation of results that follow.

From Table 6.1, we identify collinearity between DENS and AGDS as the values of the squared right singular vector exceed $2/m = 0.333$, and the two explanatory variables together explain nearly all of the variance ($0.499 + 0.500 = 0.999$ or 99.9%) due to the last axis. The eigenvectors of $\mathbf{X}^T\mathbf{X}$ also identify collinearity between DENS and AGDS, since the values of the right singular vector exceed $\mid \sqrt{2/m} \mid = 0.577$.

Table 6.1
Mason and Gunst Data:
Identifying collinearity variables

	Axis 1	Axis 2	Axis 3	Axis 4	Axis 5	Axis 6
α_k	1.583	1.382	0.872	0.659	0.603	0.164
η_k	1.000	1.146	1.817	2.403	2.625	9.651
INFD	0.154	0.159	0.047	0.600	0.040	0.000
	[0.392]	[0.398]	[0.218]	[0.775]	[−0.200]	[−0.003]
PHYS	0.250	0.056	0.029	0.049	0.616	0.000
	[0.500]	[0.237]	[0.171]	[−0.221]	[0.785]	[0.017]
DENS	0.108	0.370	0.000	0.023	0.000	0.499*
	[0.328]	[−0.608]	[−0.010]	[0.153]	[0.004]	[0.706]
AGDS	0.107	0.371	0.001	0.020	0.001	0.500*
	[0.327]	[−0.609]	[0.031]	[0.143]	[0.024]	[−0.707]
LIT	0.260	0.017	0.079	0.307	0.337	0.000
	[−0.510]	[−0.131]	[−0.280]	[0.554]	[0.581]	[−0.003]
HIED	0.122	0.027	0.844	0.001	0.006	0.001
	[−0.349]	[−0.165]	[0.919]	[0.023]	[0.078]	[0.028]

* Variables that are involved in the collinearity

Example 2: Longley Data

The second data set that we consider is the Longley data set, found in the appendix of Longley (1967), Tables 1 and 2, p. 830-831. The data set is distributed with the R programme (R Development Core Team, 2008), and has been reproduced in Appendix A (p. A-8). The data set consists of six explanatory variables: Gross National Product Implicit Price Deflator (GNP.deflator), Gross National Product (GNP), Unemployment (Unemployed), Size of Armed Forces (Armed.Forces), Noninstitutional Population 14 Years of Age and Over (Population), the Time variable (Year), and one response variable, which was pieces of employment made additional to total employment (Employed).

The correlation matrix below (Table 6.2) shows that the response variable, Employed, is strongly positively correlated with four of the explanatory variables: GNP.deflator, GNP, Population and Year, and moderately correlated with Unemployed and Armed.Forces. The four explanatory variables that are strongly correlated with the response variable are also strongly correlated with each other.

Table 6.2
Longley Data: Correlation matrix

	Employed	GNP. deflator	GNP	Unem-ployed	Armed. Forces	Popula-tion	Year
Employed	1.000						
GNP.deflator	0.971	1.000					
GNP	0.984	0.992	1.000				
Unemployed	0.503	0.621	0.604	1.000			
Armed.Forces	0.457	0.465	0.446	−0.177	1.000		
Population	0.960	0.979	0.991	0.687	0.364	1.000	
Year	0.971	0.991	0.995	0.668	0.417	0.994	1.000

When we analysed the data, the fourth, fifth and sixth condition indices indicated the presence of weak to very strong collinearity in the data set (refer to Table 6.3, for the singular values, condition indices, values of the squared right singular vectors for each axis, and the values of the eigenvectors of $\mathbf{X}^T\mathbf{X}$ (in square brackets)).

From Table 6.3, we identify collinearity on the axes with large condition indices (fourth to sixth axes), between GNP.deflator, GNP, Population and Year, as they have large values of the squared right singular vectors. The values exceed $2/m = 0.333$ or are slightly below this cut-off since the explanatory variables are involved in multiple dependencies. GNP.deflator and Population account for 97.64% of the variation in the fourth axis, whilst Year and Population account for 86.27% of the variance in the fifth axis, and GNP and Year account for 87.28% of the variation in the sixth axis. The eigenvectors of $\mathbf{X}^T\mathbf{X}$ also identify collinearity between GNP.deflator, GNP, Population and Year, since the values exceed $\mid \sqrt{2/m} \mid = 0.577$ or are slightly below this cut-off because the explanatory variables are involved in multiple dependencies.

Table 6.3
Longley Data:
Identifying collinearity variables

	Axis 1	Axis 2	Axis 3	Axis 4	Axis 5	Axis 6
α_k	2.146	1.084	0.451	0.122	0.051	0.019
η_k	1.000	1.979	4.757	17.560	42.471	110.544
GNP.deflator	0.213	0.003	0.022	0.629*	0.114	0.018
	[0.462]	[−0.058]	[0.149]	[0.793]	[−0.338]	[0.135]
GNP	0.213	0.003	0.077	0.015	0.022	0.670*
	[0.462]	[−0.053]	[0.278]	[−0.122]	[0.150]	[−0.819]
Unemployed	0.103	0.355	0.530	0.000	0.000	0.012
	[0.321]	[0.596]	[−0.728]	[0.008]	[−0.009]	[−0.108]
Armed.Forces	0.041	0.637	0.315	0.006	0.001	0.000
	[0.202]	[−0.798]	[−0.562]	[−0.077]	[−0.024]	[−0.018]
Population	0.214	0.002	0.038	0.348*	0.301*	0.097
	[0.462]	[0.046]	[0.196]	[−0.590]	[−0.549]	[0.315]
Year	0.216	0.000	0.016	0.003	0.562*	0.203*
	[0.465]	[−0.001]	[0.128]	[−0.052]	[0.750]	[0.450]

* Variables that are involved in the collinearity

6.3 Decomposing the Regression Coefficients to Identify the Variables that are Involved in the Collinearity

The explanatory variables that are involved in collinear relationships are known to have large regression coefficients. In this section, we consider an alternative measure that may be used to identify the explanatory variables that are involved in the collinearity. The proposed measure highlights the explanatory variables that have large regression coefficient values on an axis (or axes) with near-zero singular value(s).

In Chapter 5, section 5.2, we expressed the jth regression coefficient as

$$\begin{aligned}\hat{\beta}_j &= \sum_{i=1}^{n}\sum_{k=1}^{m}\frac{v_{jk}u_{ik}y_i}{\alpha_k} \\ &= \sum_{i=1}^{n} y_i\left[\sum_{k=1}^{m}\frac{v_{jk}u_{ik}}{\alpha_k}\right] \\ &= \sum_{k=1}^{m}\frac{v_{jk}}{\alpha_k}\left[\sum_{i=1}^{n} u_{ik}y_i\right]\end{aligned}$$

The quantity

$$\frac{v_{jk}}{\alpha_k}\sum_{i=1}^{n} u_{ik}y_i \equiv \mathrm{b}_{jk}$$

is the contribution of the kth principal axis to $\hat{\beta}_j$. Notice then that $\hat{\beta}_j$ is simply the sum of the contributions from each principal axis. For the axes that are associated with near-zero singular values, unless $\sum_{i=1}^{n} u_{ik}y_i$ is small, b_{jk} will be large. Clearly, a potential warning that all is not well in regression analysis occurs when an axis with a small singular value makes a large contribution to $\hat{\beta}_j$.

Gunst and Mason (1980) have proposed a similar measure (refer to Gunst and Mason (1980, p. 300)), but the notation used by Gunst and Mason differs from ours in that we have used values of the left singular vectors to express b_{jk}. As we have seen with the values of the eigenvectors of $\mathbf{X}^T\mathbf{X}$, Gunst and Mason did not quantify what a 'large' value of b_{jk} should be. Below, we propose a threshold that may be used to classify b_{jk} values as large or otherwise.

Quantities helpful in this analysis are:

$$\frac{\mid \mathrm{b}_{jk} \mid}{\sum_{k=1}^{m} \mid \mathrm{b}_{jk} \mid} \qquad \text{and} \qquad K_j = \frac{k\times \mid \mathrm{b}_{jk} \mid}{\sum_{k=1}^{m} \mid \mathrm{b}_{jk} \mid}$$

The former indicates the proportion of the absolute contributions of principal axis k to $\hat{\beta}_j$, and the latter is an easy-to-interpret statistic - values of K_j greater than 1 indicate that the axis makes a more-than-average contribution to the coefficient. As a crude cut-off value, we will investigate all axes with small singular values (using condition indices as a guideline), and with K_j values greater than 2 to identify the explanatory variables that have large coefficient estimates. Thus, two or more explanatory variables with K_j values greater than 2 on an axis (or axes) with a large condition index (or indices), are said to be involved in a collinearity.

Note once again that the threshold of 2 should be applied when the explanatory variables are involved in single dependencies. For explanatory variables involved in multiple dependencies, the threshold should be lowered as $\hat{\beta}_j$ will be contributing to multiple principal axes.

6.3.1 Illustrative Examples

Example 1: Mason and Gunst Data

Table 6.4 shows values of K_j. From Table 6.4, DENS and AGDS have large coefficient values on the smallest singular value, since $K_j > 2$ on the smallest singular value.

Table 6.4
Mason and Gunst Data: Identifying collinear variables

	Axis 1	Axis 2	Axis 3	Axis 4	Axis 5	Axis 6
α_k	1.583	1.382	0.872	0.659	0.603	0.164
η_k	1.000	1.146	1.817	2.403	2.625	9.651
INFD	2.340	0.906	1.032	1.200	0.516	0.012
PHYS	2.640	0.480	0.714	0.306	1.794	0.066
DENS	1.764	1.248	0.042	0.216	0.012	2.718*
AGDS	1.722	1.224	0.132	0.198	0.054	2.670*
LIT	2.592	0.252	1.128	0.732	1.278	0.012
HIED	1.746	0.312	3.642	0.030	0.168	0.102

* Variables with large regression coefficient values.

Example 2: Longley Data

For the Longley data set (Table 6.5), we identify collinearity between GNP.deflator, GNP, Population and Year as they have large values for K_j on the axes with relatively large condition indices (the threshold has been lowered since some $\hat{\beta}_j$'s are contributing towards multiple axes).

Table 6.5
Longley Data: Identifying collinear variables

	Axis 1	Axis 2	Axis 3	Axis 4	Axis 5	Axis 6
α_k	2.1455	1.0841	0.4510	0.1222	0.0505	0.0194
η_k	1.0000	1.9790	4.7570	17.5604	42.4710	110.5442
GNP.deflator	1.002	0.030	0.384	0.390	2.886*	1.302
GNP	0.546	0.018	0.390	0.030	0.696	4.314*
Unemployed	1.038	0.480	2.802	0.006	0.120	1.548*
Armed.Forces	0.960	0.948	3.174	0.084	0.456	0.378
Population	0.630	0.018	0.318	0.186	2.958*	1.896*
Year	0.498	0.000	0.162	0.012	3.174*	2.154*

* Variables with large regression coefficient values.

6.4 A Note about Collinearity-Influential Observations

Outlying observations in the explanatory variables have been shown to alter the collinearity structure of the $\mathbf{X}$ matrix (refer to Belsley *et al.* (1980), Mason and Gunst (1985a), Hadi (1988), Walker (1989) and Sengupta and Bhimasankaram (1997)). Such observations are known as collinearity-influential observations. Detecting the presence of the collinearity-influential observations is important for diagnostics of collinearity and for the type of estimation technique that is chosen as an alternative to the least squares method, since observations that mask the level of collinearity could adversely affect the results of robust estimators; whilst observations that induce the level of collinearity could affect the results of biased estimators.

Mason and Gunst (1985a) showed that collinearity can be increased without bounds when the leverage of an observation is increased, therefore the measures for detecting variables that are involved in the collinearity will be affected by outlying observations in the explanatory variables. We therefore recommend that measures introduced in Chapter 3 for detecting observations that are outlying in the explanatory variables be used to identify potential collinearity-influential observations. Diagnostics such as those proposed by Walker (1989), Hadi and Nyquist (1993), Sengupta and Bhimasankar (1997), and the authors cited therein, may be used to verify the nature and extent to which the observations are affecting the collinearity structure of the $\mathbf{X}$ matrix.

6.5 Discussion

Collinearity is known to have an adverse effect on the regression coefficients, in particular, the regression coefficients may be too large and have the wrong signs, and the variances of the coefficients tend to be much larger for the explanatory variables involved in collinear relationships than for uncorrelated explanatory variables.

In this chapter, we proposed an alternative measure for identifying variables that are involved in the collinearity that is based on examining the magnitude of the squared right singular vectors, which represent the proportion of variance due to an axis that is explained by a particular variable. As a result of the proposed threshold for identifying large values of the squared singular vectors for variables that are involved in the collinearity, we are able to propose a threshold for labelling large values of the eigenvectors of $\mathbf{X}^T\mathbf{X}$.

Two real data sets were used to illustrate the proposed measure and thresholds of $2/m$

and $|\sqrt{2/m}|$ for the squared singular vectors and the eigenvectors of $\mathbf{X}^T\mathbf{X}$ respectively, for values that correspond to near-zero singular values, which highlight the explanatory variables that are involved in collinear relationships.

We also proposed a threshold for an approach that is based on decomposing the regression coefficients into contributions from each axis to identify variables that are involved in collinear relationships. The results obtained using this approach mirror those obtained when observing 'large' values of the eigenvectors of $\mathbf{X}^T\mathbf{X}$ that correspond to near-zero singular values.

There are numerous approaches that may be employed to deal with collinear data, and these depend on the reason for the collinearity among the explanatory variables and the purpose of study:

Redundant variables may be eliminated: Redundant variables are eliminated if the collinearity is a function of the sample data and not a function of the population. Redundant variables will have large values of the squared right singular vectors for near-zero singular values. For a discussion on this topic, refer to Hocking (1983) and Hawkins and Fatti (1984).

Collect additional observations: Collection of additional observations has been proposed when the collinearity is a function of the sample data and the explanatory variables do not represent the range over which inferences are to be made. For a discussion on this topic, refer to Sengupta and Bhimasankaram (1997) and Hocking (2003).

Use biased estimators: Biased estimators such as principal components regression (discussed in the next chapter) are employed when the collinearity is a function of the population. Refer to Mason and Perreault Jr (1991) and the references cited therein, for a discussion on this and other methods that can be used for dealing with collinearity.

In the next chapter, we illustrate the computational theory of principal components regression, focusing particularly on expressing the principal components regression estimates using the left singular vectors when the purpose of analysis is predicting the response variable.

Chapter 7

Principal Components Regression

Principal components regression is a form of biased estimator that is used when there is collinearity among the explanatory variables, and often, only a subset of the principal axes are retained in the estimation of regression quantities. In this chapter we illustrate an alternative computational approach to principal components regression that is based on the SVD. We focus our attention particularly on employing values of the left singular vectors in expressing the principal components regression estimates where it is appropriate. Examples are used to demonstrate the usefulness of expressing and decomposing the multiple correlation coefficient, R^2, to determine the importance of the axes in explaining the amount of variation in $\mathbf{y}$, using the notation that has been introduced so far in the thesis. We also propose a measure to determine the range of values in which prediction is reasonable when there is collinearity in the data.

7.1 Introduction

In this chapter, we continue to consider the standard linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{7.1}$$

As already stated in Chapter 6, principal components regression is a form of biased estimator that is used when the collinearity among the explanatory variables is a function of the population. Principal components regression is often described using the transformed version of (7.1). That is,

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

where	$\mathbf{Z} = \mathbf{XV}$	is a matrix of principal components ($\mathbf{z}_k$),
	$\mathbf{V}$	is a matrix of right singular vectors ($\mathbf{v}_k$), and
	$\boldsymbol{\gamma} = \mathbf{V}^T\boldsymbol{\beta}$	is a vector of coefficients in the transformed model.

Thus the original explanatory variables are replaced by the uncorrelated components/axes (that is, the $\mathbf{z}_k$'s), which makes computations more stable (Jolliffe, 2002).

In principal components regression, often only a subset of the principal axes are retained in the estimation of regression quantities, using one of two criteria which will be described later in the chapter. Excluding a few axes does not result in the elimination of explanatory variables in the original model, but may result in a decrease of the model's variance and/or a decrease in the explanatory power of the model, and an increase in bias of the regression estimates.

Mandel (1982) illustrated how to carry out principal components regression by replacing the matrix $\mathbf{X}$ by its SVD (that is, $\mathbf{X} = \mathbf{UD}_\alpha\mathbf{V}^T$), and in this chapter, we illustrate an alternative computational approach to principal components regression that is based on the SVD when the goal of analysis is to predict values of the response variable. The computational approach illustrated in the chapter differs from Mandel's in that we include values of the left singular vectors in expressing the principal components regression estimates where it is appropriate.

This chapter is organised as follows. In the next section, we briefly review the computational theory of the bias introduced when selecting a subset of the axes in the estimation of regression coefficients in principal components regression. We also consider two strategies for selecting axes to retain in the estimation of regression estimates using principal components regression. Examples are used to demonstrate the usefulness of expressing

and decomposing the multiple correlation coefficient, R^2, to determine the importance of the axes in explaining the amount of variation in $\mathbf{y}$, using the notation that has been introduced so far in the thesis. In section 7.3, we consider extrapolation when collinearity is present. Here, we propose a measure to determine the range of values in which prediction is reasonable when there is collinearity in the data. We then end the chapter with a discussion of the main findings of the results presented in the chapter.

7.2 The Bias in Principal Components Regression

Selecting a subset of π (where $\pi = k_1, k_2, \ldots, k_p$) consisting of p of the m principal axes is equivalent to fitting the model:

$$\mathbf{y} = \mathbf{X}_\pi \boldsymbol{\beta}_\pi + \boldsymbol{\varepsilon}^*$$

where $\mathbf{X}_\pi = \mathbf{U}_\pi \mathbf{D}_{\alpha_\pi} \mathbf{V}_\pi^T$ is a rank p approximation to $\mathbf{X}$, with

$$\begin{aligned} \mathbf{U}_\pi &= (\mathbf{u}_{k_1}; \mathbf{u}_{k_2}, \cdots; \mathbf{u}_{k_p}) \\ \mathbf{V}_\pi &= (\mathbf{v}_{k_1}; \mathbf{v}_{k_2}, \cdots; \mathbf{v}_{k_p}) \\ \mathbf{D}_{\alpha_\pi} &= \text{diag}(\alpha_{k_1}; \alpha_{k_2}, \cdots; \alpha_{k_p}) \end{aligned}$$

and $\boldsymbol{\varepsilon}^*$ is the model's error term.

The estimator of $\boldsymbol{\beta}_\pi$ is

$$\hat{\boldsymbol{\beta}}_\pi = \mathbf{V}_\pi \mathbf{D}_{\alpha_\pi}^{-1} \mathbf{U}_\pi^T \mathbf{y}$$

and $\hat{\beta}_{\pi_j}$, the jth regression coefficient is given by

$$\hat{\beta}_{\pi_j} = \sum_{i=1}^{n} \sum_{k=1}^{p} \frac{v_{jk} u_{ik} y_i}{\alpha_k}$$

Assuming the full model to be unbiased,

$$E[\hat{\boldsymbol{\beta}}_\pi] = \mathbf{V}_\pi \mathbf{D}_{\alpha_\pi}^{-1} \mathbf{U}_\pi^T \mathbf{U} \mathbf{D}_\alpha \mathbf{V}^T \boldsymbol{\beta} = \mathbf{V}_\pi \mathbf{V}_\pi^T \boldsymbol{\beta}.$$

Thus $\hat{\boldsymbol{\beta}}_\pi$ is a biased estimator of $\boldsymbol{\beta}$. We examine the bias in more detail later.

The variance of $\hat{\boldsymbol{\beta}}_\pi$ is given by:

$$\begin{aligned}\text{var}(\hat{\boldsymbol{\beta}}_\pi) &= (\mathbf{X}_\pi^T\mathbf{X}_\pi)^{-1}\sigma^2 \\ &= \left(\mathbf{V}_\pi\mathbf{D}_{\alpha_\pi}^{-2}\mathbf{V}_\pi^T\right)\sigma^2\end{aligned}$$

So

$$\begin{aligned}\text{var}(\hat{\beta}_{\pi_j}) &= \left(\mathbf{V}_{\pi_j}^T\mathbf{D}_{\alpha_\pi}^{-2}\mathbf{V}_{\pi_j}\right)\sigma^2 \\ &= \sigma^2\sum_{k=1}^{p}\frac{v_{jk}^2}{\alpha_k^2}\end{aligned}$$

where $\mathbf{V}_{\pi_j}^T$ is the jth row of $\mathbf{V}_\pi$.

Since $\quad \text{var}(\hat{\beta}_j) = \sigma^2\sum_{k=1}^{m}\frac{v_{jk}^2}{\alpha_k^2}$

$$\text{var}(\hat{\beta}_{\pi_j}) \leq \text{var}(\hat{\beta}_j).$$

If at least one of the principal axes excluded from π has a small singular value, then $\text{var}(\hat{\beta}_{\pi_j})$ will in general be much smaller that $\text{var}(\hat{\boldsymbol{\beta}}_j)$. There is thus a trade-off between the fact that $\hat{\boldsymbol{\beta}}_\pi$ is biased, and the fact that it has a smaller variance.

7.2.1 Strategies for Retaining Axes in Principal Components Regression

There are two possible reasons for omitting axes from π:

(a) The axis is associated with a collinearity in $\mathbf{X}$.

(b) The axis is uncorrelated with $\mathbf{y}$.

We discuss each situation in turn.

Deleting the Axes that are Associated with a Collinearity in X

We assume that all the omitted axes represent collinearities. If axis k is involved in a collinearity, then

$$\mathbf{X}\mathbf{v}_k \approx 0.$$

The bias in prediction using $\hat{\boldsymbol{\beta}}_\pi$ and not $\hat{\boldsymbol{\beta}}$ is given by:

$$\hat{\mathbf{y}} - \hat{\mathbf{y}}_\pi = \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\hat{\boldsymbol{\beta}}_\pi = \mathbf{X}\mathbf{V}_{\bar{\pi}}\mathbf{D}_\alpha^{-1}\mathbf{U}^T\mathbf{y}$$

where $\mathbf{V}_{\bar{\pi}}$ consists of the right basic vectors omitted from π. But if $k \in \bar{\pi}$, then $\mathbf{X}\mathbf{v}_k \approx 0$, and hence

$$\mathbf{X}\mathbf{v}_{\bar{\pi}} \approx 0,$$

implying that

$$\hat{\mathbf{y}} - \hat{\mathbf{y}}_\pi \approx 0$$

There is in fact a wide range of $\hat{\boldsymbol{\beta}}$ values that will give essentially the same predicted values. Consider

$$\hat{\boldsymbol{\beta}}^* = \mathbf{V}_\pi\mathbf{D}_{\alpha_\pi}^{-1}\mathbf{U}_\pi^T\mathbf{y} + \mathbf{V}_{\bar{\pi}}\mathbf{D}^*\mathbf{D}_{\alpha_{\bar{\pi}}}^{-1}\mathbf{U}_{\bar{\pi}}^T\mathbf{y}$$

where $\mathbf{D}^* = \text{diag}(d_1, d_2, \cdots, d_{m-p})$ is an arbitrary diagonal matrix. Then

$$\hat{\mathbf{y}}^* = \mathbf{X}\hat{\boldsymbol{\beta}}^* = \hat{\mathbf{y}}_\pi + \mathbf{X}\mathbf{V}_{\bar{\pi}}\mathbf{D}^*\mathbf{D}_{\alpha_{\bar{\pi}}}^{-1}\mathbf{U}_{\bar{\pi}}^T\mathbf{y} \approx \hat{\mathbf{y}}_\pi$$

since $\mathbf{X}\mathbf{V}_{\bar{\pi}} \approx 0$

Thus if collinearity is present, $\hat{\boldsymbol{\beta}}$ can be decomposed into two parts, the part that expresses the real relationship, and the part that is near arbitrary, which simply expresses the collinearities.

This criterion of deleting an axis (or axes) associated with a collinearity in $\mathbf{X}$ will result in a model with little loss of the total variance in $\mathbf{X}$ and will help to control the inflation of variance in the $\hat{\boldsymbol{\beta}}_\pi$'s, which will also result in $\hat{\beta}_{\pi_j}$'s that are not inflated, resulting in more stable estimates of $\hat{\boldsymbol{\beta}}$ (Jolliffe, 2002). The estimator bias introduced however can be substantial if the deleted axes are closely correlated with the response variable, $\mathbf{y}$ (no bias is introduced if the singular value associated with the deleted axis is zero). Massy (1965) therefore recommended that this criterion be used only if the retained axes are interpretable, regardless of how the axes correlate with $\mathbf{y}$.

Deleting the Axes that are Uncorrelated with y

When the purpose of analysis is to predict future values, deleting the axes that are not important predictors of $\mathbf{y}$ is recommended (Massy, 1965). We illustrate below an approach that is based on decomposing the multiple correlation coefficient, R^2, that may be used to determine the importance of the axes in explaining the amount of variation in $\mathbf{y}$.

The transition formulae (Chapter 2, p. 2-2), enable us to add an extra row to the matrix $\mathbf{G}$ for the response variable, $\mathbf{y}$. Thus

$$\mathbf{g}_{m+1}^T = \mathbf{y}^T\mathbf{F}\mathbf{D}_\alpha^{-1}$$

Note that the m columns, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$, of $\mathbf{X}$ define an m dimensional subspace of n–space. The points, $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_m$ (the rows of $\mathbf{G}$) are the same points referred to a new set of m orthogonal axes, the principal axes, which form a basis for m–space.

The $m+1$ columns $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m, \mathbf{y}$ define an $m+1$ dimensional subspace of n–space. Thus the point $\mathbf{g}_{m+1}$, given by the transition formula is not the actual point $\mathbf{y}$ (in $m+1$ space), but its projection into the m–space defined by the principal axes. In fact $\mathbf{g}_{m+1}$ is simply the point $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ referred to the new coordinate system:

$$\begin{aligned}\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ &= \mathbf{U}\mathbf{U}^T\mathbf{y} \\ &= \mathbf{U}\mathbf{D}_\alpha^{-1}\mathbf{F}^T\mathbf{y} \\ &= \mathbf{U}\mathbf{g}_{m+1}\end{aligned}$$

The proof that $\mathbf{g}_{m+1}$ is the projection of $\mathbf{y}$ into the m–subspace now follows easily:

$$\hat{\mathbf{y}}^T(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{y}^T\mathbf{U}\mathbf{U}^T(\mathbf{y} - \mathbf{U}\mathbf{U}^T\mathbf{y}) = 0$$

so $\hat{\mathbf{y}}$ (and hence $\mathbf{g}_{m+1}$) is orthogonal to $\mathbf{y} - \hat{\mathbf{y}}$.

R^2, the multiple correlation coefficient is given by

$$R^2 = \frac{\| \hat{\mathbf{y}} \|^2}{\| \mathbf{y} \|^2} = \frac{\| \mathbf{g}_{m+1} \|^2}{\| \mathbf{y} \|^2} = \frac{\sum_{k=1}^{m} g_{m+1,k}^2}{\sum_{i=1}^{n} y_i^2} = (\cos\phi)^2$$

(Notice that the norm (length) of the vector is independent of the axis system)

where ϕ is the angle between $\mathbf{y}$ and the m–space defined by principal axes (that is, by the columns of $\mathbf{X}$).

Also

$$\frac{\mathbf{g}_{m+1,k}^2}{\| \mathbf{y} \|^2} = (\cos \phi_k)^2$$

where ϕ_k is the angle between $\mathbf{y}$ and the kth principal axis. This quantity may be interpreted as the contribution of the kth principal axis to R^2, since

$$(\cos \phi)^2 = \sum_{k=1}^{m} (\cos \phi_k)^2$$

The multiple correlation coefficient obtained by selecting a subset, π, of the principal axes is given by

$$R_\pi^2 = R_{\{k_1,k_2,\ldots,k_p\}}^2 = \sum_{i=1}^{p} \cos \phi_{k_i} = \frac{\sum_{i=1}^{p} \mathbf{g}_{m+1,k_i}^2}{\| \mathbf{y} \|^2}$$

Suppose an axis has been omitted from π because it is uncorrelated with $\mathbf{y}$ even though it accounts for a substantial proportion of the variability in $\mathbf{X}$. We recognise this situation when the axis has a substantial singular value, but makes a very small contribution to R^2. If this situation arises there must be one or more columns (variables) of $\mathbf{X}$ which are almost uncorrelated with $\mathbf{y}$. These variables will have near-zero values for their regression coefficients.

To illustrate, suppose variable $\mathbf{j}'$ has $\beta_{j'}$ close to zero. Then a substantial proportion of the variance induced by variable $\mathbf{j}'$ must be accounted for by one of the principal axes (the kth axis, say) which makes a small contribution to R^2 (that is, $g_{m+1,k}$ is small). This means that $\mathbf{v}_{jk}^T$ will be large, and since $\sum_{j=1}^{m} v_{jk}^2 = 1$, v_{jk} (or v_{jk}^2) must be small for $\mathbf{j} \neq \mathbf{j}'$.

The bias in $\hat{\beta}_{\pi_j}$ induced by omitting axis k is given by $\sum_{j=1}^{m} v_{kj} v_{jk} \beta_j$. This bias will be small, since when $\mathbf{j} \neq \mathbf{j}'$, v_{jk} is small or when $\mathbf{j} \neq \mathbf{j}'$, β_j is small. Thus omission of the axes which are nearly uncorrelated with $\mathbf{y}$ causes little bias in the regression coefficients, even though these axes may have large singular values, and account for a large proportion of the total variance in $\mathbf{X}$.

Appending a row of $(\cos\phi_k)^2$ values to the matrix of v_{jk}^2, which indicates the proportion of the variance due to the kth principal axis that is explained by variable j (refer to Chapter 6), we are able to determine the order of importance of the axes in explaining the variation in $\mathbf{y}$ and the variables that are almost uncorrelated with $\mathbf{y}$, in addition to identifying the variables that are involved in the collinearity as illustrated in Chapter 6.

Illustrative Examples

We revisit the examples used in Chapter 6 to demonstrate the use of appending a row of $(\cos\phi_k)^2$ values to the matrix of v_{jk}^2 values.

Example 1: Mason and Gunst Data

Table 7.1 shows the v_{jk}^2 values, the proportion of explained variation by the axes (Prop_Explained), $(\cos\phi_k)^2$ values, and the proportion contributed to R^2 by the axes (Prop_R^2), for the Mason and Gunst data set.

In Chapter 6, we identified the existence of collinearity between DENS and AGDS on the last axis. The order of importance of the axes in terms of the proportion contributed to R^2 by the axes is Axis 1, Axis 3, Axis 2, Axis 5, Axis 4 and Axis 6. The second axis explains 31.8% of the variation in $\mathbf{X}$, even though the axes makes a small contribution to R^2. The magnitude of the v_{jk}^2 values for DENS and AGDS on the second axis suggests that the two collinear variables are possibly not correlated with the response variable.

Table 7.1
Mason and Gunst Data: Contribution of axes to $R^2 = 0.581$

	Axis 1	Axis 2	Axis 3	Axis 4	Axis 5	Axis 6
α_k	1.583	1.382	0.872	0.659	0.603	0.164
Prop_Explained	0.418	0.318	0.127	0.072	0.061	0.004
η_k	1.000	1.146	1.817	2.403	2.625	9.651
INFD	0.154	0.159	0.047	0.600	0.040	0.000
PHYS	0.250	0.056	0.029	0.049	0.616	0.000
DENS	0.108	0.370	0.000	0.023	0.000	0.499
AGDS	0.107	0.371	0.001	0.020	0.001	0.500
LIT	0.260	0.017	0.079	0.307	0.337	0.000
HIED	0.122	0.027	0.844	0.001	0.006	0.001
$(\cos\phi_k)^2$	**0.431**	**0.048**	**0.082**	**0.005**	**0.012**	**0.002**
Prop_R^2	0.743	0.083	0.141	0.009	0.021	0.003

Example 2: Longley Data

Table 7.2 shows the v_{jk}^2 values, the proportion of explained variation by the axes (Prop_Explained), $(\cos\phi_k)^2$ values, and the proportion contributed to R^2 by the axes (Prop_R^2), for the Longley data set. The first axis accounts for most of the variability in $\mathbf{X}$ (76.7%), and contributes the most to R^2 (91.9%). The order of importance of the axes in terms of the proportion contributed to R^2 by the axes is Axis 1, Axis 3, Axis 2, Axis 5, Axis 6 and Axis 4. Thus for this data set, the bias introduced by omitting the axes that represent collinearities would be similar to the bias introduced when deleting the axes that are not strongly correlated with the response variable.

Table 7.2
Longley Data: Contribution of axes to $R^2 = 0.996$

	Axis 1	Axis 2	Axis 3	Axis 4	Axis 5	Axis 6
α_k	2.146	1.084	0.451	0.122	0.051	0.019
Prop_Explained	0.767	0.196	0.034	0.002	0.000	0.000
η_k	1.000	1.979	4.757	17.560	42.471	110.544
GNP.deflator	0.213	0.003	0.022	0.629*	0.114	0.018
GNP	0.213	0.003	0.077	0.015	0.022	0.670*
Unemployed	0.103	0.355	0.530	0.000	0.000	0.012
Armed.Forces	0.041	0.637	0.315	0.006	0.001	0.000
Population	0.214	0.002	0.038	0.348*	0.301*	0.097
Year	0.216	0.000	0.016	0.003	0.562*	0.203*
$(\cos\phi_k)^2$	**0.914**	**0.015**	**0.057**	**0.000**	**0.008**	**0.001**
Prop_R^2	0.919	0.015	0.057	0.000	0.008	0.001

The problem of selecting axes based on the size of the singular value and not considering how the axes correlate with the response variable if the purpose of analysis is prediction has been illustrated by Jolliffe (1982) and Hadi and Ling (1998), since the two strategies are not likely to produce the same results.

Sun (1995) proposed a similar approach for the second strategy that is based on looking at the correlation of each axes with the response variable in principal components regression, referred to as 'Correlated Principal Components Regression' (CPCR), and advocates the use of root mean square error of prediction criterion (RMSEP) to select the number of components to retain when prediction is the main purpose of analysis. There are various methods that exist for selecting the number of axes to retain in principal components regression depending of the strategy chosen, and the reader is referred to Jolliffe (2002).

7.3 Prediction when Collinearity is Present

Extrapolation, the prediction of the response variable for values of $\mathbf{x}_0 = (\mathbf{x}_{0_1}, \mathbf{x}_{0_2}, \ldots, \mathbf{x}_{0_m})$ outside the convex hull of the set of $\mathbf{x}$ values in the rows of $\mathbf{X}$ is dangerous, because the variance of the predicted value increases when a given value of $\mathbf{x}_0$ lies further away from the convex hull of the rows of $\mathbf{X}$. In the presence of collinearity, it is not always obvious whether a given value of $\mathbf{x}_0$ lies inside the convex hull of the rows of $\mathbf{X}$. Numerous authors (for example Hocking (2003), Mandel (1982)) have warned against this situation, since the predictions obtained using the full model compared to the reduced model will differ substantially.

The transition formulae, $\mathbf{F} = \mathbf{XGD}_\alpha^{-1}$, helps resolve this problem of describing the range of values in which prediction is reasonable. Given a vector $\mathbf{x}_0$, we find its coordinates with respect to the principal axes:

$$\mathbf{f}_0^T = \mathbf{x}_0^T \mathbf{GD}_\alpha^{-1}$$

We now compute

$$\frac{f_{0k}^2}{\sum_{k=1}^{m} f_{0k}^2} \qquad \text{for all} \quad k = 1, 2, \ldots, m$$

and interpret this as the squared cosine of the angle $\mathbf{x}_0$ makes with the kth principal axis. Thus

$$\frac{\sum_{k \in \pi} f_{0k}^2}{\sum_{k=1}^{m} f_{0k}^2}$$

is the squared cosine of the angle $\mathbf{x}_0$ makes with the p–subspace defined by π (refer to Figure 2.1, p. 2-8). If this value is less than 0.5, then the angle is greater than 45 °, and $\mathbf{x}_0$ lies more in the $m - p$ subspace defined by $\bar{\pi}$, and prediction is clearly unsatisfactory. A more stringent cut-off value could be $\sqrt{3}/2 = 0.866$ corresponding to the angle of 30 °.

Another important extrapolation check is that f_{0k} should follow the same pattern of collinearity as the bulk of the data. This is easily done by checking that f_{0k} lies within the limits of coordinates for the kth axis, that is

$$\min_{i=1\ldots n}\{f_{ik}\} \leq f_{0k} \leq \max_{i=1\ldots n}\{f_{ik}\} \qquad \text{for all} \quad k = 1, 2, \ldots, m$$

This can most easily be checked by computing

$$\frac{f_{0k}^2}{\sum_{i=1}^{n} f_{ik}^2}$$

and observing whether this value is greater than the

$$\frac{f_{ik}^2}{\sum_{i=1}^{n} f_{ik}^2} \qquad \text{for all} \quad i = 1, 2, \ldots, n$$

values which have already been computed.

7.4 Discussion

Principal components regression is a form of biased estimator that is used when there is collinearity among the explanatory variables, and in this chapter, we considered an alternative computational approach to principal components regression using the singular value decomposition.

We demonstrated the use of appending a row of $(\cos\phi_k)^2$ values (that is, the contribution of each axis to R^2) to the matrix of v_{jk}^2, which enables us to simultaneously determine the order of importance of the axes in explaining the variation in $\mathbf{y}$ and the variables that are almost uncorrelated with $\mathbf{y}$, in addition to identifying the variables that are involved in the collinearity, as was illustrated in Chapter 6.

We also saw the effectiveness of one of the measures proposed in Chapter 3, that is, the u_{ik}^2 values, to identify outlying observations in the explanatory variables, in determining the range of values for which prediction is reasonable when there is collinearity in the data.

In the next and final chapter of the thesis, we summarise the main findings of the thesis, and highlight our contributions and areas for further research.

Chapter 8

Concluding Remarks

Regression analysis using the least squares approach is a widely used technique, and regression estimates are known to be easily affected by one or a few unusual observations, and collinearity among the explanatory variables. In this thesis we used the singular value decomposition (SVD) in multiple regression, with special reference to problems of identifying unusual observations which may influence the regression coefficients and identifying the explanatory variables that are involved in collinear relationships.

The singular value decomposition (SVD) has been used in least squares problems, however most authors have concentrated on the matrix of right singular vectors (that is, the eigenvectors of $\mathbf{X}^T\mathbf{X}$). In this thesis, we considered also the matrix of left singular vectors (that is, the eigenvectors of $\mathbf{X}\mathbf{X}^T$).

8.1 Contributions to Research

A procedure to identify outliers which highlights observations that are masked or swamped

The diagonal values of the hat matrix (that is, the h_i values) are used in regression analysis to identify outlying observations in the explanatory variables that may alter the fit of the least squares line. The h_i values however, are known to suffer from the effects of masking and swamping, and in this thesis we proposed a procedure which is adapted in part from correspondence analysis, to identify the leverage points.

The procedure entails expressing the h_i values as sums of contributions of variance of each axis that is explained by each observation (r_l), and using these contributions, together with the contributions of variance of each observation that is explained by each axis (c_l), to identify the axis that observations are outlying on. We then produce a leverage-distance (L-D) plot which highlights any observations which may be masked or swamped.

The procedure for the h_i values was extended to the values of h_z, where the response variable is appended to the matrix of explanatory variables. Once again, the procedure was successful in identifying outlying observations that were masked or swamped.

Thus our contribution to research is a procedure which identifies outlying points and will suffer from very little (if any) effects of masking and swamping.

A drawback with using the procedure is that too many observations may be declared as leverage points, since observations with large h_i (or h_z) values that are not explained well by any axis are automatically treated as leverage points if they contribute highly to the determination of the direction of at least one of the axis. Another drawback that is associated with using the h_z values is that we are not able to differentiate between leverage points and regression outliers, since h_z may be large because of a large h_i value and/or a large residual value.

A measure to aid with the identification of observations that are being accommodated by the least squares fit

The residuals are also often examined to determine the observations that may have influenced the fit of the least squares regression line, because they take the response variable, $\mathbf{y}$, into account. The residuals, either in their raw or transformed form, are known to be a poor measure of fit since they may fail to identify the outlying observations when these observations are being accommodated by the least squares fit.

We proposed a measure, R_j, that can be used in conjunction with the transformed residuals. The measure, which is based on the off-diagonal values of the hat ($\mathbf{H}_x$) matrix, provides insight into the role that each observation plays in determining the displacement of other observations from the least squares fit. A drawback of using R_j however, is that since a large h_{ij} value also indicates that observation i and observation j are situated far from the bulk of the data, we may not be able to differentiate between observations with large R_j values because they are good leverage points or regression outliers.

A measure to identify the outlying observations that influence the regression coefficients

The regression estimates such as the coefficients, are known to be easily affected by outlying observations, and measures such as DFBETAS, which are intended to measure the impact of an observation on the individual regression coefficients, are prone to the same

problems as the residuals and the diagonal values of the hat matrix since they are a function of the residuals, which are a poor measure of fit, and the diagonal values of the hat matrix, which may suffer from the masking and swamping effects.

By decomposing the regression coefficients, we have proposed a measure, B_{ij}, which enables us to determine the outlying observations that may have a disproportionate effect in the determination of the individual regression coefficients, and do not suffer from the same problems as DFBETAS. A limitation with using the B_{ij} values is that although they highlight the regression coefficients affected by the outlying observations, they do not tell us the actual impact the observation has on the estimation of the coefficient(s).

Thus the proposed measures, even though they are exploratory, have value in revealing outlying and influential data. Another advantage of the proposed measures for identifying outlying and influential observations is that they do not entail deleting any observations from the analysis.

Thresholds for variables that are involved in collinear relationships

A number of approaches have been proposed to identify the explanatory variables that are involved in collinear relationships, and to detect the coefficients that are most adversely affected. We proposed an alternative measure for identifying variables that are involved in the collinearity that is based on examining the magnitude of the squared right singular vectors, which represent the proportion of variance due to an axis that is explained by a particular variable.

As a result of the threshold proposed for the squared right singular vectors (that is, $v_{jk}^2 > 2/m$), we were able to recommend a threshold for a measure that is based on examining the magnitude of the eigenvectors of $\mathbf{X}^T\mathbf{X}$ ($v_{jk} >\mid \sqrt{2/m} \mid$) that correspond to small singular values to determine the explanatory variables that are involved in the collinearity. We also motivated for a threshold to use for the decomposed regression coefficients that correspond to small singular values to determine the coefficients that are affected by the collinearity.

Decomposing R^2 to determine the importance of the axes in explaining the variation in y

Principal components regression is a form of biased estimator that is used when there is collinearity among the explanatory variables, and often only a subset of the principal axes

are retained in the estimation of regression quantities.

In this thesis, we focused particularly on employing values of the left singular vectors in expressing the principal components regression estimates where it is appropriate, and demonstrated the usefulness of decomposing the multiple correlation coefficient, R^2, to determine the importance of the axes in explaining the amount of variation in $\mathbf{y}$. We also illustrated the use of appending a row of $(\cos\phi_k)^2$ values (that is, the contribution of each axis to R^2) to the matrix of v_{jk}^2, which enables use to simultaneously determine the order of importance of the axes in explaining the variation in $\mathbf{y}$ and the variables that are almost uncorrelated with $\mathbf{y}$, in addition to identifying the variables that are involved in the collinearity.

We also saw the effectiveness of one using the u_{ik}^2 values in determining the range of values for which prediction is reasonable when there is collinearity in the data.

8.2 Further Research Directions

Throughout the thesis, we have assumed that the $\mathbf{X}$ matrix is standardised to have zero mean and unit variance. As already mentioned in Chapter 3, depending on the extent of deviation of the outlying observations from the bulk of the data, the L-D plot may fail to reveal any observations that are being masked or swamped, thus there is a need to consider robust estimates of the mean and variance.

When examining the distance of the observations from the origin using the L-D plot, there is no guideline that may be used in order to determine what should constitute a 'large' distance from the origin for the observations that are located on the various axes. Thus, research to establish thresholds to use for the observations that are located on the various axes is worth pursuing.

In Chapter 4, the Hawkins, Bradu and Kass data set presented a challenge in that the three good leverage points (observations 11 to 13) also had large R_j values, simply because they cluster together far away from the bulk of the other observations. The L-D plot may be used to get a sense of the proximity of the observations relative to each other by looking at the distance of the observations from the origin, however, there is a need for an approach that will enable us to differentiate between observations that have large R_j values because they are regression outliers that may be responsible for inducing large residuals for other observations, and those observations that are located far from the bulk of the data but in the general direction of the least squares fitted line.

Throughout the thesis, we have also recommended thresholds to use for the measures proposed which are only intended to provide a rough approximation. Since the measures introduced are exploratory, it may be worthwhile considering their statistical properties. We have also not attempted to make suggestions about what to do with observations that are influential, thus this would be something to consider next.

Bibliography

Afifi, A.A. and Clark, V. (1990) *Computer-Aided Multivariate Analysis*. Chapman & Hall, New York, 2nd edition.

Anderson, T.W. (1958) *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, Inc., New York.

Andrews, D.F. and Pregibon, D. (1978) Finding the Outliers that Matter. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(1):85–93.

Atkinson, A.C. (1982) Regression Diagnostics, Transformations and Constructed Variables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(1):1–36.

Atkinson, A.C. (1985) *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*. Clarendon Press Oxford University Press, Oxford, New York.

Atkinson, A.C. (1986) [Influential Observations, High Leverage Points, and Outliers in Linear Regression]: Comment: Aspects of Diagnostic Regression Analysis. *Statistical Science*, 1(3):397–402.

Barnett, V. and Lewis, T. (1994) *Outliers in Statistical Data*. Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics. Wiley, Chichester, West Sussex, 3rd edition.

Becker, C. and Gather, U. (1999) The Masking Breakdown Point of Multivariate Outlier Identification Rules. *Journal of the American Statistical Association*, 94(447):947–955.

Belsley, D.A. (1982) Assessing the Presence of Harmful Collinearity and Other Forms of Weak Data through a Test for Signal-to-Noise. *Journal of Econometrics*, 20(2):211–253.

Belsley, D.A. (1984) Demeaning Conditioning Diagnostics through Centering. *The American Statistician*, 38(2):73–77.

Belsley, D.A. (1987) [Collinearity and Least Squares Regression]: Comment: Well-Conditioned Collinearity Indices. *Statistical Science*, 2(1):86–91.

Belsley, D.A. (1991) *Conditioning Diagnostics: Collinearity and Weak Data in Regression*. John Wiley & Sons, Inc., USA.

Belsley, D.A. and Hocking, R.R. (1984) Eigenvector Weaknesses and Other Topics for Assessing Conditioning Diagnostics. *Technometrics*, 26(3):297–301.

Belsley, D.A. and Klema, V. (1974) Detecting and Assessing the Problems Caused by Multi-Collinearity: A Use of the Singular-Value Decomposition. *SSRN eLibrary*.

Belsley, D.A., Kuh, E. and Welsch, R.E. (1980) *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, New York.

Ben-Gal, I. (2005) Outlier Detection. In O. Maimon and L. Rokach, eds., *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*, 131–146. Kluwer Academic Publishers.

Bendixen, M. (1996) A Practical Guide to the Use of Correspondence Analysis in Marketing Research. Technical Report 2, Marketing Research On-Line.

Benzécri, J.P. (1992) *Correspondence Analysis Handbook*. Marcel Dekker, New York.

Brant, R. (1986) [Influential Observations, High Leverage Points, and Outliers in Linear Regression]: Comment. *Statistical Science*, 1(3):405–407.

Brownlee, K.A. (1965) *Statistical Theory and Methodology in Science and Engineering*. Wiley, New York, 2nd edition.

Chatterjee, S. and Hadi, A.S. (1986a) Influential Observations, High Leverage Points, and Outliers in Linear Regression. *Statistical Science*, 1(3):379–393.

Chatterjee, S. and Hadi, A.S. (1986b) Influential Observations, High Leverage Points, and Outliers in Linear Regression: Rejoinder. *Statistical Science*, 1(3):415–416.

Chatterjee, S. and Hadi, A.S. (1988) *Sensitivity Analysis in Linear Regression*. Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics. Wiley, USA.

Chiang, J. (2008) The Algorithm for Multiple Outliers Detection Against Masking and Swamping Effects. *International Journal of Contemporary Mathematical Sciences*, 3(17):839–859.

Coleman, D.E. (1977) Finding Leverage Groups. *SSRN eLibrary*.

Cook, R.D. (1979) Influential Observations in Linear Regression. *Journal of the American Statistical Association*, 74(365):169–174.

Cook, R.D. (1986) [Influential Observations, High Leverage Points, and Outliers in Linear Regression]: Comment. *Statistical Science*, 1(3):393–397.

Cook, R.D. and Weisberg, S. (1980) Characterizations of an Empirical Influence Function for Detecting Influential Cases in Regression. *Technometrics*, 22(4):495–508.

Cook, R.D. and Weisberg, S. (1982a) Criticism and Influence Analysis in Regression. *Sociological Methodology*, 13:313–361.

Cook, R.D. and Weisberg, S. (1982b) *Residual and Influence in Regression*. Monograms on Statistics and Applied Probability. Chapman & Hall, New York.

Cook, R.D., Weisberg, S., Carr, D.B., Pregibon, D., Atkinson, A.C., Tierney, L. and Welsch, R.E. (1989) Regression Diagnostics with Dynamic Graphics: [With Discussions and Response]. *Technometrics*, 31(3):277–311.

Davies, L. and Gather, U. (1993) The Identification of Multiple Outliers. *Journal of the American Statistical Association*, 88(423):782–792.

Dunteman, G.H. (1989) *Principal Components Analysis*. Sage University Papers Series. Quantitative Applications in the Social Sciences. Sage Publications, Newbury Park, California.

Everitt, B.S. (1978) *Graphical Techniques for Multivariate Data*. Heinemann Educational Books Ltd, London.

Everitt, B.S. and Dunn, G. (1991) *Applied Multivariate Data Analysis*. Edward Arnold, London.

Fatti, P.L., Hawkins, D.H. and Raath, L.E. (1982) Discriminant Analysis. In D.H. Hawkins, ed., *Topics in Applied Multivariate Analysis*, chapter 1, 1–71. Cambridge University Press, Cambridge.

Flury, B. (2007) *Flury: Data Sets from Flury, 1997*. R package version 0.1-2.

Fomby, T.B. and Hill, R.C. (1978) Deletion Criteria for Principal Components Regression Analysis. *American Journal of Agricultural Economics*, 60(3):524–527.

Fox, J. (1991) *Regression Diagnostics*. Sage University Papers Series. Quantitative Applications in the Social Sciences. Sage Publications, Newbury Park, California.

Fung, W.K. (1993) Unmasking Outliers and Leverage Points: A Confirmation. *Journal of the American Statistical Association*, 88(422):515–519.

Fung, W.K. (1995) Diagnostics in Linear Discriminant Analysis. *Journal of the American Statistical Association*, 90(431):952–956.

Fung, W.K. and Kwan, C.W. (1997) A Note on Local Influence Based on Normal Curvature. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(4):839–843.

Gabriel, K.R. (1971) The Biplot Graphic Display of Matrices with Application to Principal Component Analysis. *Biometrika*, 58(3):453–467.

Galmacci, G. (1996) Collinearity Detection in Linear Regression Models. *Computational Economics*, 9(3):215–227.

Gnanadesikan, R. and Kettenring, J.R. (1972) Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data. *Biometrics*, 28(1):81–124.

Golub, G.H. and Reinsch, C. (1970) Singular Value Decomposition and Least Squares Solutions. *Numerische Mathematik*, 14(5):403–420.

Golub, G.H. and Van Loan, C.F. (1996) *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, 3rd edition.

Good, I.J. (1969) Some Applications of the Singular Value Decomposition of a Matrix. *Technometrics*, 11(4):823–831.

Gower, J.C. and Hand, D.J. (1996) *Biplots*. Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Gray, J.B. (1989) On the Use of Regression Diagnostics. *The Statistician*, 38(2):97–105.

Gray, J.B. and Ling, R.F. (1984) K-Clustering as a Detection Tool for Influential Subsets in Regression. *Technometrics*, 26(4):305–318.

Green, P.E. and Carroll, J.D. (1976) *Mathematical Tools for Applied Multivariate Analysis*. Academic Press, New York.

Greenacre, M.J. (1984) *Theory and Applications of Correspondence Analysis*. Academic Press, London.

Greenacre, M.J. (2007) *Correspondence Analysis in Practice*. Chapman & Hall/CRC, London, 2nd edition.

Gunst, R.F. and Mason, R.L. (1977) Advantages of Examining Multicollinearities in Regression Analysis. *Biometrics*, 33(1):249–260.

Gunst, R.F. and Mason, R.L. (1980) *Regression Analysis and its Application: A Data-Oriented Approach*. Marcel Dekker, Inc. New York.

Hadi, A.S. (1988) Diagnosing Collinearity-Influential Observations. *Computational Statistics & Data Analysis*, 7(2):143–159.

Hadi, A.S. (1992) Identifying Multiple Outliers in Multivariate Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(3):761–771.

Hadi, A.S. and Ling, R.F. (1998) Some Cautionary Notes on the Use of Principal Components Regression. *The American Statistician*, 52(1):15–19.

Hadi, A.S. and Nyquist, H. (1993) Further Theoretical Results and a Comparison between Two Methods for Approximating Eigenvalues of Perturbed Covariance Matrices. *Statistics and Computing*, 3(3):113–123.

Hadi, A.S. and Simonoff, J.S. (1993) Procedures for the Identification of Multiple Outliers in Linear Models. *Journal of American Statistical Association*, 88(424):1264–1272.

Hadi, A.S. and Wells, M.T. (1990) Assessing the Effects of Multiple Rows on the Condition Number of a Matrix. *Journal of the American Statistical Association*, 85(411):786–792.

Hair, J.F., Anderson, R.E., Tatham, R.L. and Black, W.C. (2006) *Multivariate Data Analysis*. Pearson Prentice Hall, Upper Saddle, New Jersey, sixth edition.

Hammarling, S. (1985) The Singular Value Decomposition in Multivariate Statistics. *ACM SIGNUM Newsletter*, 20(3):2–25.

Hawkins, D.M., Bradu, D. and Kass, G.V. (1984) Location of Several Outliers in Multiple-Regression Data Using Elemental Sets. *Technometrics*, 26(3):197–208.

Hawkins, D.M. and Fatti, L.P. (1984) Exploring Multivariate Analysis Using the Minor Principal Components. *The Statistician*, 33(4):325–338.

Henshall, J.M. and Smith, D.M. (1996) Iterative Refinement of the Singular-Value Decomposition Solution to Regression Equations. *Computational Statistics & Data Analysis*, 22(6):573–582.

Hines, O.R.J. and Hines, W.G.S. (1995) Exploring Cooks Statistic Graphically. *The American Statistician*, 49(4):389–394.

Hoaglin, D.C. and Kempthorne, P.J. (1986) [Influential Observations, High Leverage Points, and Outliers in Linear Regression]: Comment. *Statistical Science*, 1(3):408–412.

Hoaglin, D.C. and Welsch, R.E. (1978) The Hat Matrix in Regression and ANOVA. *The American Statistician*, 32(1):17–22.

Hocking, R.R. (1983) Developments in Linear Regression Methodology: 1959-1982. *Technometrics*, 25(3):219–230.

Hocking, R.R. (2003) *Methods and Applications of Linear Models: Regression and the Analysis of Variance*. Wiley Interscience, Hoboken, N.J., 2nd edition.

Hocking, R.R. and Pendleton, O.J. (1983) The Regression Dilemma. *Communications in Statistics - Theory and Methods*, 12(5):497–527.

Hoffman, D.L. and Franke, G.R. (1986) Correspondence Analysis: Graphical Representation of Categorical Data in Marketing Research. *Journal of Marketing Research*, 23(3):213–227.

Jolliffe, I.T. (1982) A Note on the Use of Principal Components in Regression. *Applied Statistics*, 31(3):300–303.

Jolliffe, I.T. (2002) *Principal Components Analysis*. Springer, USA, 2nd edition.

Kempthorne, P.J. and Mendel, M.B. (1990) Unmasking Multivariate Outliers and Leverage Points: Comment. *Journal of the American Statistical Association*, 85(411):647–648.

Lachenbruch, P.A. (1997) Discriminant Diagnostics. *Biometrics*, 53(4):1284–1292.

Liang, Y. and Kvalheim, O.M. (1996) Robust Methods for Multivariate Analysis – A Tutorial Review. *Chemometrics and Intelligent Laboratory Systems*, 32(1):1–10.

Longley, J.W. (1967) An Appraisal of Least Squares Programs for the Electronic Computer from the Point of View of the User. *Journal of the American Statistical Association*, 62(319):819–841.

Ludovic, L., Morineau, A. and Warwick, K.M. (1984) *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York.

Mandel, J. (1982) Use of the Singular Decomposition in Regression Analysis. *The American Statistician*, 36(1):15–24.

Mansfield, E.R. and Helms, B.P. (1982) Detecting Multicollinearity. *The American Statistician*, 36(3):158–160.

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979) *Multivariate Analysis*. Probability and Mathematical Statistics. Academic Press, London.

Mason, C.H. and Perreault Jr, W.D. (1991) Collinearity, Power, and Interpretation of Multiple Regression Analysis. *Journal of Marketing Research*, 28(3):268–280.

Mason, R.L. and Gunst, R.F. (1985a) Outlier-Induced Collinearities. *Technometrics*, 27(4):401–407.

Mason, R.L. and Gunst, R.F. (1985b) Selecting Principal Components in Regression. *Statistics & Probability Letters*, 3(6):299–301.

Massy, W.F. (1965) Principal Components Regression in Exploratory Statistical Research. *Journal of the American Statistical Association*, 60(309):234–256.

Meloun, M. and Militký, J. (2001) Detection of Single Influential Points in OLS Regression Model Building. *Analytica Chimica Acta*, 439(2):169–191.

Montgomery, D.C., Peck, E.A. and Vining, G.G. (2001) *Methods of Multivariate Analysis*. Wiley Series in Probability and Statistics. Wiley, New York, 3rd edition.

Morrison, D.F. (1976) *Multivariate Statistical Methods*. McGraw-Hill Series in Probability and Statistics. McGraw-Hill, New York, 2nd edition.

Morrison, D.G. (1969) On the Interpretation of Discriminant Analysis. *Journal of Marketing Research*, 6(2):156–163.

Pena, D. and Yohai, V.J. (1995) The Detection of Influential Subsets in Linear Regression by using an Influence Matrix. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):145–156.

Pregibon, D. (1981) Logistic Regression Diagnostics. *The Annals of Statistics*, 9(4):705–724.

R Development Core Team (2008) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Austria.

Rencher, A.C. (2002) *Methods of Multivariate Analysis*. Wiley, USA, 2nd edition.

Riani, M. and Atkinson, A.C. (2001) A Unified Approach to Outliers, Influence, and Transformations in Discriminant Analysis. *Journal of Computational and Graphical Statistics*, 10(3):513–544.

Rousseeuw, P.J. and Leroy, A.M. (1987) *Robust Regression and Outlier Detection*. Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics. Wiley, New York.

Rousseeuw, P.J. and van Zomeren, B.C. (1990) Unmasking Multivariate Outliers and Leverage Points. *Journal of American Statistical Association*, 85(411):633–639.

Rousseeuw, P. and Croux, C. and Todorov, V. and Ruckstuhl, A. and Salibian-Barrera, M. and Maechler, M. (2007) *robustbase: Basic Robust Statistics*. R package version 0.2-8.

Sengupta, D. and Bhimasankaram, P. (1997) On the Roles of Observations in Collinearity in the Linear Model. *Journal of American Statistical Association*, 92(439):1024–1032.

Silvey, S.D. (1969) Multicollinearity and Imprecise Estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 31(3):539–552.

Snee, R.D. and Marquardt, D.W. (1984) Comment: Collinearity Diagnostics Depend on the Domain of Prediction, the Model, and the Data. *The American Statistician*, 38(2):83–87.

Spector, P. (1994) *An Introduction to S and S-Plus*. Duxbury Press, USA.

Stewart, G.W. (1987) Collinearity and Least Squares Regression. *Statistical Science*, 2(1):68–84.

Stewart, G.W. (1993) On the Early History of the Singular Value Decomposition. *SIAM Review*, 35(4):551–556.

Stine, R.A. (1995) Graphical Interpretation of Variance Inflation Factors. *The American Statistician*, 49(1):53–56.

Sun, J. (1995) A Correlation Principal Component Regression Analysis of NIR Data. *Journal of Chemometrics*, 9(1):21–29.

Velleman, P.F. (1986) [Influential Observations, High Leverage Points, and Outliers in Linear Regression]: Comment. *Statistical Science*, 1(3):412–413.

Velleman, P.F. and Welsh, R.E. (1981) Efficient Computing of Regression Diagnostics. *The American Statistician*, 35(4):234–242.

Walker, E. (1989) Detection of Collinearity-Influential Observations. *Communications in Statistics: Theory and Methods*, 18(5):1675–1690.

Walker, E. and Birch, J.B. (1988) Influence Measures in Ridge Regression. *Technometrics*, 30(2):221–227.

Wang, D.Q., Critchley, F. and Smith, P.J. (2003) The Multiple Sets of Deletion Measures and Masking in Regression. *Communications in Statistics: Theory & Methods*, 32(2):407–413.

Wang, S. and Nyquist, H. (1991) Effects on the Eigenstructure of a Data Matrix when Deleting an Observation. *Computational Statistics & Data Analysis*, 11(2):179–188.

Weisberg, S. (1986) [Influential Observations, High Leverage Points, and Outliers in Linear Regression]: Comment. *Statistical Science*, 1(3):414–415.

Welsch, R. and Kuh, E. (1977) Linear Regression Diagnostics. Working Paper Series 173, National Bureau of Economic Research, Inc.

Welsh, R.E. (1986) [Influential Observations, High Leverage Points, and Outliers in Linear Regression]: Comment. *Statistical Science*, 1(3):403–405.

Wisnowski, J.W., Montgomery, D.C. and Simpson, J.R. (2001) A Comparative Analysis of Multiple Outlier Detection Procedures in the Linear Regression Model. *Computational Statistics & Data Analysis*, 36(3):351–382.

Zhao, Y., Lee, A.H. and Hui, Y. (1994) Influence Diagnostics for Generalized Linear Measurement Error Models. *Biometrics*, 50(4):1117–1128.

Appendix A

Data Sets

A.1 Data Used in Chapter 3

A.1.1 Data for Example 1: Hawkins, Bradu and Kass data (Hawkins *et al.*, 1984)

Table A.1
Hawkins, Bradu and Kass data

	X1	X2	X3	Y
1	10.1	19.6	28.3	9.7
2	9.5	20.5	28.9	10.1
3	10.7	20.2	31	10.3
4	9.9	21.5	31.7	9.5
5	10.3	21.1	31.1	10
6	10.8	20.4	29.2	10
7	10.5	20.9	29.1	10.8
8	9.9	19.6	28.8	10.3
9	9.7	20.7	31	9.6
10	9.3	19.7	30.3	9.9
11	11	24	35	-0.2
12	12	23	37	-0.4
13	12	26	34	0.7
14	11	34	34	0.1
15	3.4	2.9	2.1	-0.4
16	3.1	2.2	0.3	0.6
17	0	1.6	0.2	-0.2
18	2.3	1.6	2	0
19	0.8	2.9	1.6	0.1
⋮	⋮	⋮	⋮	⋮

Table A.1
continued

	X1	X2	X3	Y
⋮	⋮	⋮	⋮	⋮
20	3.1	3.4	2.2	0.4
21	2.6	2.2	1.9	0.9
22	0.4	3.2	1.9	0.3
23	2	2.3	0.8	-0.8
24	1.3	2.3	0.5	0.7
25	1	0	0.4	-0.3
26	0.9	3.3	2.5	-0.8
27	3.3	2.5	2.9	-0.7
28	1.8	0.8	2	0.3
29	1.2	0.9	0.8	0.3
30	1.2	0.7	3.4	-0.3
31	3.1	1.4	1	0
32	0.5	2.4	0.3	-0.4
33	1.5	3.1	1.5	-0.6
34	0.4	0	0.7	-0.7
35	3.1	2.4	3	0.3
36	1.1	2.2	2.7	-1
37	0.1	3	2.6	-0.6
38	1.5	1.2	0.2	0.9
39	2.1	0	1.2	-0.7
40	0.5	2	1.2	-0.5
41	3.4	1.6	2.9	-0.1
42	0.3	1	2.7	-0.7
43	0.1	3.3	0.9	0.6
44	1.8	0.5	3.2	-0.7
45	1.9	0.1	0.6	-0.5
46	1.8	0.5	3	-0.4
47	3	0.1	0.8	-0.9
48	3.1	1.6	3	0.1
49	3.1	2.5	1.9	0.9
50	2.1	2.8	2.9	-0.4
51	2.3	1.5	0.4	0.7
52	3.3	0.6	1.2	-0.5
53	0.3	0.4	3.3	0.7
⋮	⋮	⋮	⋮	⋮

Table A.1
continued

	X1	X2	X3	Y
⋮	⋮	⋮	⋮	⋮
54	1.1	3	0.3	0.7
55	0.5	2.4	0.9	0
56	1.8	3.2	0.9	0.1
57	1.8	0.7	0.7	0.7
58	2.4	3.4	1.5	-0.1
59	1.6	2.1	3	-0.3
60	0.3	1.5	3.3	-0.9
61	0.4	3.4	3	-0.3
62	0.9	0.1	0.3	0.6
63	1.1	2.7	0.2	-0.3
64	2.8	3	2.9	-0.5
65	2	0.7	2.7	0.6
66	0.2	1.8	0.8	-0.9
67	1.6	2	1.2	-0.7
68	0.1	0	1.1	0.6
69	2	0.6	0.3	0.2
70	1	2.2	2.9	0.7
71	2.2	2.5	2.3	0.2
72	0.6	2	1.5	-0.2
73	0.3	1.7	2.2	0.4
74	0	2.2	1.6	-0.9
75	0.3	0.4	2.6	0.2

A.1.2 Data for Example 2: Stack Loss data (Brownlee, 1965)

Table A.2
Stack Loss data

	Air.Flow	Water.Temp	Acid.Conc.	stack.loss
1	80	27	89	42
2	80	27	88	37
3	75	25	90	37
4	62	24	87	28
5	62	22	87	18
6	62	23	87	18
7	62	24	93	19
8	62	24	93	20
9	58	23	87	15
10	58	18	80	14
11	58	18	89	14
12	58	17	88	13
13	58	18	82	11
14	58	19	93	12
15	50	18	89	8
16	50	18	86	7
17	50	19	72	8
18	50	19	79	8
19	50	20	80	9
20	56	20	82	15
21	70	20	91	15

A.1.3 Example 3: Health Club data

Table A.3
Health Club data (Chatterjee and Hadi, 1988)

	weight	pulse	arm.leg	time(0.25)	time(1)
1	217	67	260	91	481
2	141	52	190	66	292
3	152	58	203	68	338
4	153	56	183	70	357
5	180	66	170	77	396
6	193	71	178	82	429
7	162	65	160	74	345
8	180	80	170	84	469
9	205	77	188	83	425
10	168	74	170	79	358
11	232	65	220	72	393
12	146	68	158	68	346
13	173	51	243	56	279
14	155	64	198	59	311
15	212	66	220	77	401
16	138	70	180	62	267
17	147	54	150	75	404
18	197	76	228	88	442
19	165	59	188	70	368
20	125	58	160	66	295
21	161	52	190	69	391
22	132	62	163	59	264
23	257	64	313	96	487
24	236	72	225	84	481
25	149	57	173	68	374
26	161	57	173	65	309
27	198	59	220	62	367
28	245	70	218	69	469
29	141	63	193	60	252
30	177	53	183	75	338

A.2 Data Used in Chapter 6

A.2.1 Example 1: Mason and Gunst's data (Gunst and Mason, 1980)

Table A.4
Mason and Gunst's data

	INFD	PHYS	DENS	AGDS	LIT	HIED	GNP
Australia	19.5	806	1	21	98.5	856	1316
Austria	37.5	695	84	1720	98.5	546	670
Barbados	60.4	3000	548	7121	91.1	24	200
Belgium	35.4	819	301	5257	96.7	536	1196
Brit. Guiana	67.1	3900	3	192	74.0	27	235
Bulgaria	45.1	740	72	1380	85.0	456	365
Canada	27.3	900	2	257	97.5	645	1947
Chile	127.9	1700	11	1164	80.1	257	379
Costa Rica	78.9	2600	24	948	79.4	326	357
Cyprus	29.9	1400	62	1042	60.5	78	467
Czechoslovakia	31.0	620	108	1821	97.5	398	680
Denmark	23.7	830	107	1434	98.5	570	1057
El Salvador	76.3	5400	127	1497	96.4	89	219
Finland	21.0	1600	13	1512	29.4	529	794
France	27.4	1014	83	1288	57.5	667	943
Guatemala	91.9	6400	36	1365	29.4	135	189
Hong Kong	41.5	3300	3082	98143	57.5	176	272
Hungary	47.6	650	108	1370	97.5	258	490
Iceland	22.4	840	2	79	98.5	445	572
India	225.0	5200	138	2279	19.3	220	73
Ireland	30.5	1000	40	598	98.5	362	550
Itay	48.7	746	164	2323	87.5	362	516
Jamaica	58.7	4300	143	3410	77.0	42	316
Japan	37.7	930	254	7563	98.0	750	306
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table A.4
continued

⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Luxemborg	31.5	910	123	2286	96.5	36	1388
Malaya	68.9	6400	54	2980	38.4	475	356
Malta	38.3	980	1041	8050	57.6	142	377
Mauritius	69.5	4500	352	4711	51.8	14	225
Mexico	77.7	1700	18	296	50.0	258	262
Netherlands	16.5	900	346	4855	98.5	923	836
New Zealand	22.8	700	9	170	98.5	839	1310
Nicaragua	71.7	2800	10	824	38.4	110	160
Norway	20.2	946	11	3420	98.5	258	1130
Panama	54.8	3200	15	838	65.7	371	329
Poland	74.7	1100	96	1411	95.0	351	475
Portugal	77.5	1394	100	1087	55.9	272	224
Puerto Rico	52.4	2200	271	4030	81.0	1192	563
Romania	75.7	788	78	1248	89.0	226	360
Singapore	32.3	2400	2904	108214	50.0	437	400
Spain	43.5	1000	61	1347	87.0	258	293
Sweden	16.6	1089	17	1705	98.5	401	1380
Switzerland	21.1	765	133	2320	98.5	398	1428
Taiwan	30.5	1500	305	10446	54.0	329	161
Trinidad	45.4	2300	168	4383	73.8	61	423
United Kingdom	24.1	935	217	2677	98.5	460	1189
United States	26.4	780	20	399	98.0	1983	2577
USSR	35.0	578	10	339	95.0	539	600
West Germany	33.8	798	217	3631	98.5	528	927
Yugoslavia	100.0	1637	73	1215	77.0	524	265

A.2.2 Example 2: Longley data

Table A.5
Longley data (Longley, 1967)

	GNP.deflator	GNP	Unemployed	Armed.Forces	Population	Year	Employed
1947	83.0	234289	2356	1590	107608	1947	60323
1948	88.5	259426	2325	1456	108632	1948	61122
1949	88.2	258054	3682	1616	109773	1949	60171
1950	89.5	284599	3351	1650	110929	1950	61187
1951	96.2	328975	2099	3099	112075	1951	63221
1952	98.1	346999	1932	3594	113270	1952	63639
1953	99.0	365385	1870	3547	115094	1953	64989
1954	100.0	363112	3578	3350	116219	1954	63761
1955	101.2	397469	2904	3048	117388	1955	66019
1956	104.6	419180	2822	2857	118734	1956	67857
1957	108.4	442769	2936	2798	120445	1957	68169
1958	110.8	444546	4681	2637	121950	1958	66513
1959	112.6	482704	3813	2552	123366	1959	68655
1960	114.2	502601	3931	2514	125368	1960	69564
1961	115.7	518173	4806	2572	127852	1961	69331
1962	116.9	554894	4007	2827	130081	1962	70551

Appendix B

Complete Tables for Examples

B.1 Example 1: Hawkins, Bradu and Kass data (1984) — Chapter 3

Table B.1
Hawkins, Bradu and Kass data: r_l, c_l, h_i and DIST values

	r_l				c_l			
	Axis 1 (97.48%)	Axis 2 (1.89%)	Axis 3 (0.63%)	h_i	Axis 1 (97.48%)	Axis 2 (1.89%)	Axis 3 (0.63%)	DIST
1	0.045	0.005	0.001	0.051	0.998	0.002	0.000	0.131
2	0.045	0.001	0.001	0.047	0.999	0.000	0.000	0.130
3	0.053	0.007	0.012	0.072	0.996	0.003	0.001	0.155
4	0.053	0.002	0.012	0.067	0.998	0.001	0.001	0.154
5	0.053	0.000	0.007	0.060	0.999	0.000	0.001	0.155
6	0.051	0.011	0.000	0.062	0.996	0.004	0.000	0.149
7	0.050	0.003	0.001	0.054	0.999	0.001	0.000	0.147
8	0.044	0.002	0.004	0.050	0.999	0.001	0.001	0.130
9	0.049	0.001	0.016	0.066	0.997	0.001	0.002	0.144
10	0.044	0.001	0.028	0.073	0.995	0.001	0.004	0.129
11	0.070	0.002	0.009	0.081	0.999	0.001	0.001	0.205
12	0.077	0.006	0.048	0.131	0.995	0.001	0.004	0.225
13	0.079	0.000	0.017	0.096	0.999	0.000	0.001	0.231
14	0.095	0.141	0.314	0.550	0.952	0.027	0.020	0.292
15	0.001	0.025	0.019	0.045	0.565	0.347	0.088	0.004
16	0.002	0.028	0.033	0.063	0.688	0.222	0.090	0.007
17	0.006	0.020	0.000	0.026	0.939	0.061	0.000	0.018
18	0.002	0.007	0.000	0.009	0.937	0.063	0.000	0.007
19	0.003	0.013	0.002	0.018	0.927	0.070	0.003	0.010
20	0.001	0.012	0.022	0.035	0.689	0.192	0.119	0.003
21	0.002	0.010	0.005	0.017	0.883	0.101	0.016	0.005
22	0.004	0.028	0.001	0.033	0.864	0.135	0.001	0.012
23	0.003	0.002	0.012	0.017	0.959	0.013	0.029	0.008
24	0.003	0.001	0.008	0.012	0.980	0.005	0.015	0.010
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table B.1
continued

	r_l				c_l			
	Axis 1	Axis 2	Axis 3		Axis 1	Axis 2	Axis 3	
	(97.48%)	(1.89%)	(0.63%)	h_i	(97.48%)	(1.89%)	(0.63%)	DIST
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
25	0.005	0.000	0.003	0.008	0.995	0.001	0.004	0.016
26	0.003	0.015	0.000	0.018	0.898	0.101	0.001	0.009
27	0.001	0.024	0.004	0.029	0.624	0.356	0.020	0.004
28	0.003	0.004	0.003	0.010	0.971	0.022	0.007	0.009
29	0.004	0.000	0.000	0.004	1.000	0.000	0.000	0.013
30	0.003	0.000	0.030	0.033	0.944	0.001	0.055	0.010
31	0.002	0.035	0.009	0.046	0.707	0.269	0.024	0.007
32	0.005	0.014	0.005	0.024	0.939	0.054	0.007	0.014
33	0.002	0.002	0.009	0.013	0.963	0.015	0.022	0.007
34	0.006	0.002	0.010	0.018	0.982	0.008	0.010	0.018
35	0.001	0.018	0.002	0.021	0.720	0.271	0.009	0.004
36	0.003	0.005	0.002	0.010	0.966	0.029	0.005	0.009
37	0.004	0.040	0.002	0.046	0.827	0.171	0.003	0.013
38	0.004	0.001	0.003	0.008	0.991	0.005	0.004	0.012
39	0.003	0.016	0.002	0.021	0.914	0.082	0.004	0.011
40	0.004	0.012	0.000	0.016	0.949	0.051	0.000	0.014
41	0.001	0.038	0.000	0.039	0.575	0.424	0.001	0.005
42	0.005	0.012	0.025	0.042	0.920	0.048	0.032	0.015
43	0.004	0.038	0.005	0.047	0.848	0.145	0.007	0.015
44	0.003	0.004	0.021	0.028	0.932	0.023	0.045	0.009
45	0.004	0.011	0.000	0.015	0.946	0.054	0.000	0.012
46	0.003	0.004	0.018	0.025	0.938	0.024	0.038	0.009
47	0.003	0.049	0.001	0.053	0.723	0.276	0.001	0.010
48	0.001	0.026	0.000	0.027	0.700	0.300	0.000	0.005
49	0.001	0.020	0.012	0.033	0.710	0.241	0.049	0.005
50	0.002	0.000	0.001	0.003	0.994	0.003	0.002	0.005
51	0.003	0.011	0.009	0.023	0.910	0.070	0.020	0.009
52	0.002	0.054	0.002	0.058	0.635	0.360	0.005	0.009
53	0.005	0.009	0.052	0.066	0.902	0.034	0.064	0.015
54	0.003	0.005	0.018	0.026	0.938	0.029	0.033	0.011
55	0.004	0.015	0.001	0.020	0.935	0.063	0.002	0.013
56	0.002	0.000	0.022	0.024	0.941	0.001	0.058	0.007
57	0.004	0.006	0.000	0.010	0.971	0.029	0.000	0.011
58	0.001	0.002	0.024	0.027	0.889	0.019	0.092	0.005
59	0.002	0.000	0.002	0.004	0.992	0.002	0.006	0.007
60	0.004	0.018	0.026	0.048	0.887	0.076	0.037	0.013
61	0.003	0.034	0.001	0.038	0.817	0.181	0.001	0.011
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table B.1
continued

	r_l				c_l			
	Axis 1 (97.48%)	Axis 2 (1.89%)	Axis 3 (0.63%)	h_i	Axis 1 (97.48%)	Axis 2 (1.89%)	Axis 3 (0.63%)	DIST
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
62	0.005	0.000	0.002	0.007	0.997	0.000	0.003	0.016
63	0.004	0.004	0.015	0.023	0.954	0.020	0.026	0.011
64	0.001	0.007	0.005	0.013	0.854	0.118	0.027	0.003
65	0.003	0.006	0.009	0.018	0.937	0.042	0.021	0.008
66	0.005	0.018	0.000	0.023	0.938	0.062	0.000	0.016
67	0.003	0.000	0.002	0.005	0.995	0.000	0.005	0.009
68	0.006	0.007	0.019	0.032	0.961	0.021	0.018	0.020
69	0.004	0.011	0.001	0.016	0.944	0.054	0.002	0.011
70	0.003	0.006	0.004	0.013	0.951	0.040	0.009	0.009
71	0.002	0.002	0.002	0.006	0.974	0.019	0.007	0.005
72	0.004	0.011	0.000	0.015	0.952	0.047	0.000	0.013
73	0.004	0.017	0.008	0.029	0.920	0.069	0.010	0.014
74	0.005	0.030	0.001	0.036	0.889	0.109	0.002	0.016
75	0.005	0.008	0.035	0.048	0.929	0.029	0.042	0.016

B.2 Example 1: Hawkins, Bradu and Kass data (1984) — Chapter 4

Table B.2
Hawkins, Bradu and Kass data: r_l, c_l, h_z and DIST values

	r_l					c_l				
	Axis 1 (88.99%)	Axis 2 (9.19%)	Axis 3 (1.39%)	Axis 4 (0.43%)	h_z	Axis 1 (88.99%)	Axis 2 (9.19%)	Axis 3 (1.39%)	Axis 4 (0.43%)	DIST
1	0.056	0.022	0.002	0.001	0.081	0.960	0.039	0.000	0.000	0.209
2	0.058	0.028	0.004	0.001	0.091	0.952	0.047	0.001	0.000	0.217
3	0.066	0.025	0.003	0.003	0.097	0.962	0.037	0.001	0.000	0.245
4	0.063	0.013	0.005	0.004	0.085	0.977	0.021	0.001	0.000	0.229
5	0.065	0.020	0.000	0.001	0.086	0.970	0.030	0.000	0.000	0.239
6	0.063	0.021	0.007	0.004	0.095	0.964	0.034	0.002	0.000	0.234
7	0.066	0.034	0.001	0.010	0.111	0.948	0.051	0.000	0.001	0.248
8	0.059	0.032	0.000	0.000	0.091	0.946	0.054	0.000	0.000	0.220
9	0.060	0.017	0.004	0.006	0.087	0.970	0.029	0.001	0.000	0.220
10	0.057	0.026	0.005	0.012	0.100	0.953	0.045	0.001	0.001	0.212
11	0.041	0.162	0.000	0.048	0.251	0.707	0.289	0.000	0.004	0.207
12	0.045	0.180	0.017	0.133	0.375	0.697	0.289	0.004	0.010	0.229
13	0.050	0.146	0.003	0.000	0.199	0.767	0.232	0.001	0.000	0.232
14	0.058	0.213	0.093	0.227	0.591	0.702	0.267	0.018	0.013	0.294
15	0.001	0.002	0.029	0.014	0.046	0.639	0.105	0.223	0.032	0.007
16	0.001	0.000	0.029	0.034	0.064	0.693	0.017	0.212	0.078	0.008
17	0.005	0.000	0.021	0.000	0.026	0.937	0.008	0.056	0.000	0.021
18	0.002	0.000	0.008	0.000	0.010	0.948	0.000	0.051	0.000	0.009
19	0.003	0.000	0.013	0.003	0.019	0.932	0.004	0.060	0.004	0.012
20	0.001	0.000	0.013	0.021	0.035	0.746	0.004	0.168	0.082	0.004
21	0.001	0.001	0.009	0.006	0.017	0.817	0.074	0.093	0.017	0.006
22	0.003	0.000	0.030	0.002	0.035	0.857	0.014	0.127	0.002	0.013
23	0.003	0.002	0.003	0.009	0.017	0.930	0.046	0.012	0.012	0.012
24	0.003	0.002	0.001	0.012	0.018	0.910	0.064	0.006	0.020	0.011
25	0.005	0.000	0.000	0.003	0.008	0.991	0.005	0.001	0.003	0.019
26	0.003	0.002	0.014	0.000	0.019	0.898	0.045	0.058	0.000	0.013
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

Table B.2
continued

	r_l					c_l				
	Axis 1	Axis 2	Axis 3	Axis 4		Axis 1	Axis 2	Axis 3	Axis 4	
	(88.99%)	(9.19%)	(1.39%)	(0.43%)	h_z	(88.99%)	(9.19%)	(1.39%)	(0.43%)	DIST
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮	
27	0.001	0.003	0.028	0.001	0.033	0.652	0.156	0.189	0.003	0.008
28	0.003	0.001	0.003	0.004	0.011	0.956	0.020	0.018	0.006	0.010
29	0.004	0.001	0.000	0.000	0.005	0.968	0.032	0.000	0.000	0.014
30	0.003	0.000	0.000	0.034	0.037	0.954	0.000	0.001	0.045	0.013
31	0.002	0.000	0.037	0.007	0.046	0.760	0.002	0.225	0.013	0.009
32	0.005	0.000	0.013	0.006	0.024	0.952	0.000	0.042	0.006	0.017
33	0.003	0.001	0.001	0.008	0.013	0.947	0.034	0.007	0.012	0.011
34	0.006	0.000	0.003	0.011	0.020	0.985	0.000	0.006	0.008	0.023
35	0.001	0.000	0.019	0.001	0.021	0.772	0.004	0.220	0.003	0.005
36	0.004	0.002	0.004	0.004	0.014	0.930	0.051	0.014	0.005	0.015
37	0.004	0.000	0.039	0.002	0.045	0.863	0.008	0.128	0.002	0.017
38	0.003	0.004	0.001	0.005	0.013	0.877	0.113	0.003	0.007	0.012
39	0.004	0.000	0.017	0.005	0.026	0.924	0.009	0.062	0.005	0.015
40	0.005	0.000	0.012	0.000	0.017	0.960	0.001	0.039	0.000	0.017
41	0.001	0.001	0.040	0.000	0.042	0.657	0.028	0.314	0.000	0.007
42	0.005	0.000	0.013	0.027	0.045	0.936	0.003	0.037	0.025	0.019
43	0.003	0.002	0.040	0.010	0.055	0.798	0.045	0.146	0.011	0.015
44	0.003	0.001	0.004	0.028	0.036	0.926	0.020	0.017	0.037	0.013
45	0.004	0.000	0.012	0.001	0.017	0.955	0.001	0.043	0.001	0.016
46	0.003	0.000	0.004	0.023	0.030	0.946	0.004	0.018	0.033	0.012
47	0.003	0.002	0.054	0.000	0.059	0.770	0.036	0.194	0.000	0.015
48	0.001	0.000	0.028	0.001	0.030	0.757	0.006	0.236	0.001	0.006
49	0.001	0.001	0.020	0.013	0.035	0.676	0.049	0.230	0.045	0.005
50	0.002	0.001	0.001	0.000	0.004	0.946	0.050	0.004	0.000	0.008
51	0.002	0.001	0.010	0.011	0.024	0.857	0.058	0.064	0.021	0.009
52	0.002	0.001	0.058	0.000	0.061	0.706	0.024	0.269	0.001	0.012
53	0.004	0.004	0.012	0.047	0.067	0.824	0.082	0.043	0.051	0.016
54	0.003	0.002	0.006	0.025	0.036	0.875	0.056	0.029	0.039	0.011
55	0.004	0.000	0.015	0.002	0.021	0.935	0.006	0.056	0.003	0.015
56	0.002	0.000	0.000	0.023	0.025	0.953	0.000	0.000	0.046	0.009
57	0.003	0.002	0.005	0.000	0.010	0.898	0.078	0.024	0.000	0.011
58	0.002	0.000	0.002	0.022	0.026	0.903	0.023	0.018	0.055	0.007
59	0.003	0.000	0.000	0.003	0.006	0.983	0.011	0.001	0.006	0.010
60	0.005	0.001	0.018	0.030	0.054	0.905	0.015	0.053	0.028	0.019
61	0.003	0.000	0.034	0.000	0.037	0.854	0.004	0.141	0.001	0.013
62	0.004	0.004	0.000	0.001	0.009	0.918	0.081	0.000	0.001	0.017
63	0.004	0.000	0.003	0.017	0.024	0.965	0.001	0.014	0.021	0.014
64	0.002	0.002	0.009	0.002	0.015	0.804	0.118	0.071	0.006	0.007
65	0.002	0.001	0.005	0.009	0.017	0.898	0.050	0.034	0.018	0.009
66	0.006	0.000	0.017	0.000	0.023	0.950	0.007	0.044	0.000	0.021
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮	

Table B.2
continued

	r_l					c_l				
	Axis 1 (88.99%)	Axis 2 (9.19%)	Axis 3 (1.39%)	Axis 4 (0.43%)	h_z	Axis 1 (88.99%)	Axis 2 (9.19%)	Axis 3 (1.39%)	Axis 4 (0.43%)	DIST
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮	
67	0.004	0.001	0.000	0.001	0.006	0.974	0.024	0.001	0.002	0.013
68	0.005	0.005	0.009	0.014	0.033	0.878	0.084	0.026	0.012	0.020
69	0.003	0.001	0.011	0.001	0.016	0.934	0.017	0.047	0.002	0.012
70	0.002	0.002	0.008	0.003	0.015	0.892	0.059	0.044	0.005	0.009
71	0.002	0.000	0.002	0.002	0.006	0.979	0.000	0.017	0.005	0.007
72	0.004	0.000	0.011	0.000	0.015	0.959	0.001	0.040	0.000	0.015
73	0.004	0.001	0.019	0.005	0.029	0.885	0.037	0.072	0.006	0.015
74	0.005	0.001	0.030	0.002	0.038	0.911	0.009	0.078	0.001	0.021
75	0.004	0.001	0.010	0.033	0.048	0.905	0.030	0.032	0.033	0.017

B.3 Example 1: Hawkins, Bradu and Kass data (1984) — Chapter 5

Table B.3
Hawkins, Bradu and Kass data – Influential Observations (B_j)

	X1	X2	X3
1	3.134*	1.987	1.025
2	1.028	0.178	2.151*
3	2.559*	6.852*	4.977*
4	2.917*	2.859*	5.412*
5	0.025	3.296*	4.052*
6	5.481*	1.608	0.501
7	4.257*	0.716	1.201
8	1.533	3.260*	2.985*
9	2.82*5	4.034*	6.122*
10	3.790*	5.993*	8.029*
11	0.434	0.377	0.869
12	0.038	2.172*	1.880
13	0.240	0.471	0.327
14	0.513	5.063	2.845
15	1.722	0.416	1.378
16	0.784	0.314	0.718
⋮	⋮	⋮	⋮

* Regression outlier, and B_j exceeds 2.

Table B.4
Hawkins, Bradu and Kass data – Influential observations (DFBETAS)

	X1	X2	X3
1	0.115	−0.061	0.039
2	−0.043	0.006	0.092
3	0.079	−0.179	0.159
4	−0.084	−0.069	0.160
5	−0.001	−0.089	0.134
6	0.200	−0.049	−0.019
7	0.188	0.027	−0.054
8	0.060	−0.107	0.120
9	−0.085	−0.102	0.189
10	−0.125	−0.166	0.271
11	0.236	0.172	−0.486
12	−0.024	1.178	−1.247
13	−0.255	−0.420	0.356
14	0.329	−2.727	1.873
15	−0.061	−0.012	0.050
16	0.091	0.031	−0.086
⋮	⋮	⋮	⋮

Table B.3
continued

	X1	X2	X3
⋮	⋮	⋮	⋮
17	1.029	0.625	0.098
18	0.528	0.273	0.187
19	0.559	0.641	0.160
20	0.714	0.398	0.731
21	0.225	0.020	0.164
22	0.758	0.601	0.002
23	0.836	0.904	1.223
24	0.015	0.324	0.262
25	0.113	0.596	0.376
26	1.186	0.959	0.030
27	1.679	0.272	0.887
28	0.130	0.514	0.226
29	0.041	0.113	0.046
30	0.718	1.447	1.351
31	1.363	0.092	0.816
32	0.725	1.186	0.473
33	0.073	1.154	0.831
34	0.959	0.838	1.023
35	0.698	0.191	0.321
36	1.021	0.174	0.634
37	1.999	0.665	0.661
38	0.086	0.061	0.115
39	0.927	1.315	0.236
40	1.005	0.512	0.157
41	1.291	0.728	0.326
42	1.777	1.070	1.700
43	0.542	0.645	0.152
44	0.102	1.946	1.307
45	0.783	0.791	0.033
46	0.032	1.559	1.011
47	2.346	1.235	0.679
48	0.870	0.633	0.131
49	0.336	0.057	0.256
50	0.184	0.108	0.245
51	0.390	0.107	0.338
52	2.098	0.856	0.760
53	0.565	0.557	0.692
⋮	⋮	⋮	⋮

Table B.4
continued

	X1	X2	X3
⋮	⋮	⋮	⋮
17	−0.039	0.020	0.004
18	−0.014	0.006	0.005
19	−0.027	0.026	−0.008
20	0.024	0.011	−0.026
21	0.035	0.003	−0.026
22	−0.064	0.043	−0.000
23	−0.015	−0.014	0.022
24	0.003	0.055	−0.054
25	0.002	0.008	−0.006
26	0.024	−0.016	0.001
27	−0.103	0.014	0.056
28	−0.003	0.009	−0.005
29	−0.001	−0.003	0.002
30	0.051	0.086	−0.099
31	−0.026	0.001	0.016
32	−0.021	0.029	−0.014
33	0.000	−0.005	0.004
34	0.029	0.021	−0.031
35	−0.025	0.006	0.012
36	0.045	0.006	−0.029
37	0.022	−0.006	−0.007
38	0.025	0.015	−0.034
39	−0.052	0.062	−0.014
40	0.001	−0.000	−0.000
41	−0.092	0.043	0.024
42	0.087	0.044	−0.085
43	−0.124	0.124	−0.036
44	0.008	0.131	−0.108
45	−0.029	0.025	0.001
46	0.002	0.097	−0.077
47	−0.146	0.065	0.043
48	−0.056	0.034	0.009
49	0.051	0.007	−0.040
50	−0.006	−0.003	0.009
51	0.052	0.012	−0.046
52	−0.124	0.043	0.046
53	0.010	0.009	−0.013
⋮	⋮	⋮	⋮

Table B.3
continued

	X1	X2	X3
⋮	⋮	⋮	⋮
54	0.049	0.536	0.368
55	0.676	0.692	0.137
56	0.280	0.948	0.872
57	0.200	0.121	0.069
58	0.689	0.948	1.118
59	0.299	0.367	0.367
60	2.200	1.072	1.956
61	1.513	0.626	0.416
62	0.081	0.204	0.145
63	0.095	1.304	0.921
64	0.941	0.190	0.754
65	0.098	0.524	0.269
66	1.475	0.721	0.266
67	0.205	0.437	0.521
68	0.496	0.350	0.495
69	0.576	0.166	0.294
70	0.313	0.073	0.213
71	0.298	0.102	0.292
72	0.817	0.294	0.212
73	0.737	0.091	0.460
74	2.047	0.640	0.675
75	0.932	0.830	1.071

Table B.4
continued

	X1	X2	X3
⋮	⋮	⋮	⋮
54	−0.013	0.117	−0.098
55	−0.033	0.029	−0.007
56	0.016	0.046	−0.052
57	0.019	−0.009	−0.007
58	0.012	0.014	−0.020
59	0.012	0.012	−0.015
60	0.122	0.050	−0.112
61	0.002	−0.001	−0.000
62	−0.007	−0.015	0.013
63	−0.003	0.039	−0.034
64	−0.041	−0.007	0.034
65	−0.004	0.016	−0.010
66	0.016	−0.006	−0.003
67	−0.004	−0.008	0.011
68	−0.035	−0.021	0.036
69	0.009	−0.002	−0.005
70	−0.022	−0.004	0.015
71	0.000	0.000	−0.000
72	−0.007	0.002	0.002
73	−0.033	−0.003	0.021
74	0.032	−0.009	−0.011
75	0.027	0.020	−0.032

B.4 Example 2: Stack Loss data (1965)– Chapter 3

Table B.5
H matrix – Stack Loss data. Entries are rounded values of $100 \times h_{ij}$.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	25	**26**	**17**	4	2	3	−1	−1	−2	−1	−7	−7	−2	−9	−16	−14	−4	−9	−8	−3	7
2	**26**	27	**17**	4	2	3	−2	−2	−2	0	−8	−8	−1	−11	−18	−15	−1	−8	−8	−2	6
3	**17**	**17**	13	1	1	1	0	0	−2	−1	−2	−2	−2	−3	−10	−9	−8	−9	−9	−3	8
4	4	4	1	8	2	5	8	8	8	−8	−8	−11	−8	−4	−2	−2	1	1	4	−0	−11
5	2	2	1	2	0	1	2	2	1	−2	−2	−2	−2	−1	−1	−1	−1	−1	−0	−1	−1
6	3	3	1	5	1	3	5	5	5	−5	−5	−7	−5	−3	−1	−1	−0	0	2	−0	−6
7	−1	−2	0	8	2	5	17	17	10	−16	−3	−8	−13	6	6	2	−15	−5	−0	−4	−8
8	−1	−2	0	8	2	5	17	17	10	−16	−3	−8	−13	6	6	2	−15	−5	−0	−4	−8
9	−2	−2	−2	8	1	5	10	10	9	−10	−7	−10	−9	−2	2	1	0	2	6	−0	−13
10	−1	0	−1	−8	−2	−5	−16	−16	−10	15	4	8	13	−5	−5	−1	14	5	0	4	8
11	−7	−8	−2	−8	−2	−5	−3	−3	−7	4	11	13	5	11	7	5	−10	−4	−6	−2	12
12	−7	−8	−2	−11	−2	−7	−8	−8	−10	8	13	17	9	11	6	4	−8	−4	−8	−1	16
13	−2	−1	−2	−8	−2	−5	−13	−13	−9	13	5	9	11	−1	−2	0	9	3	−1	3	9
14	−9	−11	−3	−4	−1	−3	6	6	−2	−5	11	11	−1	16	12	7	−19	−7	−7	−5	9
15	−16	−18	−10	−2	−1	−1	6	6	2	−5	7	6	−2	12	14	10	−9	1	2	−1	−2
16	−14	−15	−9	−2	−1	−1	2	2	1	−1	5	4	0	7	10	8	−1	3	4	1	−3
17	−4	−1	−8	1	−1	−0	−15	−15	0	14	−10	−8	9	−19	−9	−1	36	18	16	11	−14
18	−9	−8	−9	1	−1	0	−5	−5	2	5	−4	−4	3	−7	1	3	18	11	11	6	−11
19	−8	−8	−9	4	−0	2	−0	−0	6	0	−6	−8	−1	−7	2	4	16	11	13	5	−15
20	−3	−2	−3	−0	−1	−0	−4	−4	−0	4	−2	−1	3	−5	−1	1	11	6	5	3	−4
21	7	6	8	−11	−1	−6	−8	−8	−13	8	12	16	9	9	−2	−3	−14	−11	−15	−4	24

Appendix C

R Functions

C.1 R Functions for Chapter 3

In the following codes, the word "*matrix*" should be replaced by the appropriate name of the data used, and the response variable is always located in the last column.

C.1.1 Functions for r_l

Contributions of Observations to Axes

```
rows<-function(x1)
{
x2<-x1[,-ncol(x1)]
A<- scale(x2)
n1<-dim(A)[1]; n2<-dim(A)[2]
X<-svd(A)
D<-(diag(X$d)^2)/(n1-1)
F<-matrix(X$u%*%sqrt(D),nrow=n1)
FF<-F^2
Prop_F<-FF%*%solve(diag(apply(FF,2,sum)))
# appending h_i values to the last column
ReC_F<-cbind(round(Prop_F,4),round(apply(Prop_F,1,sum),3))
ReC_F
}
rows(matrix)
```

Alternatively

```
rows<-function(x1)
{
x2<-x1[,-ncol(x1)]
A<- scale(x2)
n1<-dim(A)[1]; n2<-dim(A)[2]
X<-svd(A)
Usq<-(X$u)^2
# appending h_i values to the last column
ReC_F<-cbind(round(Usq,4),round(apply(Usq,1,sum),3))
ReC_F
}
rows(matrix)
```

C.1.2 Functions for c_l

```
cols<-function(x1)
{
x2<-x1[,-ncol(x1)]
A<-scale(x2)
n1<-dim(A)[1]; n2<-dim(A)[2]
X<-svd(A)
D<-(diag(X$d)^2)/(n1-1)
F<-matrix(X$u%*%sqrt(D),nrow=n1)
FF<-F^2
AbC_F<- round(solve(diag(apply(FF,1,sum)))%*%FF,3)
AbC_F
}
cols(matrix)
```

C.2 R Functions for Chapter 4

```
residuals<-function(x1)
{
x2<-x1[,-ncol(x1)]
A<- scale(x2)
n1<-dim(A)[1]; n2<-dim(A)[2]; n3<-dim(x1)[2]
X<-svd(A)
y_vec<-matrix(scale(x1[,n3]),ncol=1)
X2<-matrix(NA,nrow=n1,ncol=n1)
for (i in 1:nrow(X$u))X2[i, ]<- apply(X$u, 1, function(x) x %*% X$u[i,])*y_vec[i,]
diag(X2)<-0
m_X2<-matrix(X2,nrow=n1,byrow=F)
resi<-apply(m_X2,1,sum)
Hi<-matrix(abs(resi),nrow=n1)
Hj<-apply(Hi,2,sum)
Prop_Hj<-round(matrix(Hi/Hj,nrow=n1),3)
Rj<-Prop_Hj*n1
Rj
}
```

residuals(*matrix*)

C.3 R Functions for Chapter 5

```
coefficients<-function(x)
{
x2<-x[,-ncol(x)]
n1<- dim(x)[1];n2<- dim(x2)[2];n3<- dim(x)[2]
A<- scale(x2); X<-svd(A)
Y<- matrix(scale(x[,n3]),ncol=1)
D<-(X$d)/sqrt(n1)
Aij<-matrix((X$v)%*% solve(diag(D))%*%t(X$u),ncol=n2, byrow= T)
YAij<-matrix(NA,nrow=n1,ncol=n2)
for (i in 1:n2) YAij [,i ]<- Y*Aij[,i]
AbC_B<- round(abs(YAij)%*%solve(diag(apply(abs(YAij),2,sum)))*n1,3)
AbC_B
}
```

coefficients(*matrix*)

C.4 R Functions for Chapter 6

```
K_j<-function(x)
{
x1<-scale(x) x2<-x[,-ncol(x)]
n1<-nrow(x);n2<-ncol(x2);n3<-ncol(x1)
A<-scale(x2)
X<-svd(A)
Y<- x1[,n3]
V<-(X$v)%*% solve(diag(X$d))
UY<-t(X$u)%*%Y
Bhat_kj<-V%*%diag(apply(UY,1,sum))
AbC_Bhat_kj <- round(abs(Bhat_kj)%*%solve(diag(apply(abs(Bhat_kj),2,sum))),3)*n2
AbC_Bhat_kj
}
K_j(matrix)
```