A DIAGNOSTIC METHOD FOR IDENTIFYING MULTIVARIATE OUTLYING OBSERVATIONS

by

YE JAIN HWANG LEE

B.S. in Agriculture Economics, National Chung-Hsing University, 1973

A MASTER'S REPORT

submitted in partial fulfillment of the
requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1983

Approved by:

Shie Shien Yang
Major Professor

Table of Contents

I. Introduction

An outlier (outlying data point) can be roughly defined as an odd point which is far away from or inconsistent with the rest of the data points. Technically, outlying data points lie near the boundary of the smallest convex set that contains all the data points. In figure 1a, set A is the smallest convex set containing all the data points. Points P1, P2, and P3 are outlying data points.

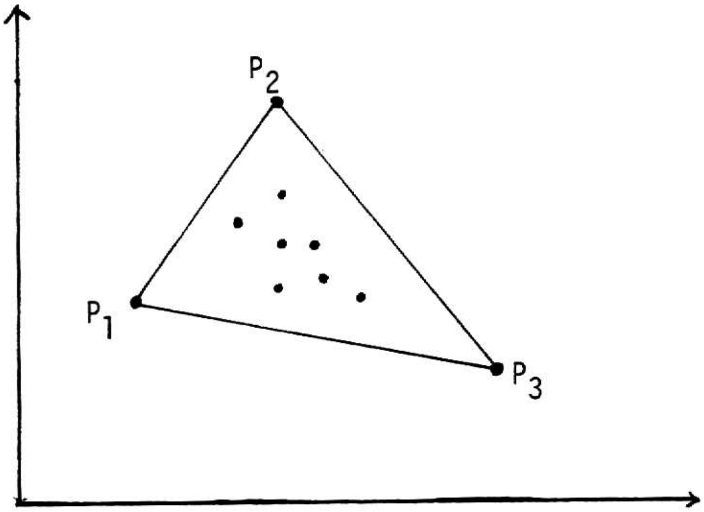

Figure 1.a

Outlying data points may contain gross error due to, for example, malfunction of the instrument taking the measurements, the presence of extreme conditions under which the measurements were taken, or the sudden presence of unexpected external factor or events altering the conditions under which the measurements were taken. These outlying data points often have strong impact on parameter estimation, on hypothesis testing and on statistical inferences in general. In this report, we propose a diagnostic procedure for identifying multivariate outlying observations.

Recently, the problems of detecting influential observations have received great attention in the statistical literature (see e.g., Hoaglin and Welsch (1978), Belsley, and et. al (1980) and Cook (1977, 1979)). An influential observation is one which has large influence on the estimated parameters. Specifically, deleting an influential observation in the calculation of the estimator will produce an estimator which is substantially different from the estimator based on all the data points. An outlier may or may not be an influential data point. For example, in figure 1b, point C is an outlier, but it is not an influential observation. Most of the diagnostic procedures for identifying influential observations considered in the literature are in the framework of regression analysis (i.e., a model is postulated and is fitted by the data). The diagnostic procedure proposed here is a model independent. It is intended to be as a diagnostic procedure for identifying outlying observations.

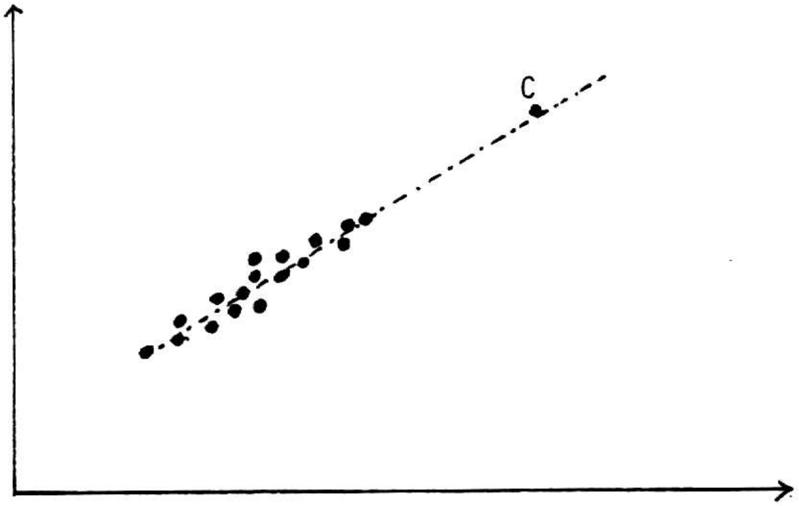

Figure 1.b

In Chapter 3, we show that the proposed diagnostic statistic is related to quantities such as studentized residuals and residual variances. Hence, we also hope that the proposed diagnostic procedure will be complementary to procedures for detecting influential observation when a model is postulated.

In Chapter 2, Cook's (1977) diagnostic method for detecting influential observations is reviewed. In Chapter 3, the proposed diagnostic procedure is developed. This procedure is applied to Longley's (1967) and Brownlee's (1965) data. A graphical method, based on multidimensional scaling technique, in representing the data points is also presented.

II. Cook's Method for Detecting Influential Observations

Cook (1977) developed a method for detecting influential data points in estimation of the parameters $\underline{\beta}$ of the following linear model,

$$\underline{Y} = X\underline{\beta} + \underline{e} \qquad (2.1)$$

where $\underline{Y}$ is an nx1 vector of observed responses, X is an nxp full rank matrix of known constants, $\underline{\beta}$ is a px1 vector of unknown parameters and $\underline{e}$ is an nx1 vector of randomly distributed error such that $E(\underline{e}) = \underline{0}$ and $E(\underline{e}\ \underline{e}') = \sigma^2 I$. Here (2.1) is written in detail

$$\underline{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} \underline{x}_1' \\ \underline{x}_2' \\ \vdots \\ \underline{x}_n' \end{bmatrix} \quad \underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \quad \underline{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

The least squares estimate of $\underline{\beta}$, is given by

$$\hat{\underline{\beta}} = X\ (X'X)^{-1}\ X'\underline{Y} \qquad (2.2)$$

The normal theory $(1-\alpha)100\%$ confidence ellipsoid for the unknown parameter vector $\underline{\beta}$ is given by the set of vector $\underline{\beta}^*$ that satisfy

$$\frac{(\underline{\beta}^* - \hat{\underline{\beta}})'(X'X)(\underline{\beta}^* - \hat{\underline{\beta}})}{p\ s^2} \leq F_{1-\alpha}(p,\ n-p) \qquad (2.3)$$

where s^2 is the mean square error for fitting model (2.1) and $F_{1-\alpha}(p,\ n-p)$ is the $(1-\alpha)$ percentage point of the F-distribution with p and n-p degrees of freedom.

Let $\hat{\underline{\beta}}_{(i)}$ denote the least squares estimate of $\underline{\beta}$ with the $\underline{ith}$ data point being deleted. In view of (2.3), Cook (1977) suggested the weighted distance between $\hat{\underline{\beta}}_{(i)}$ and $\hat{\underline{\beta}}$,

$$D_i = \frac{(\underline{\hat{\beta}}_{(i)} - \underline{\hat{\beta}})'(X'X)\ (\underline{\hat{\beta}}_{(i)} - \underline{\hat{\beta}})}{p\ s^2} \qquad (2.4)$$

as a measure of degree of influence that the ith data point has on the estimate $\underline{\hat{\beta}}$ of $\underline{\beta}$. This provides a measure of the distance between $\underline{\hat{\beta}}(i)$ and $\underline{\hat{\beta}}$ in terms of level of significance of a F-distribution. Suppose, for example that $Di \approx F_{1-\alpha}(p,n-p)$. Then it implies that the removal of the ith data point moves the least squares estimate to the edge of the $(1-\alpha)100\%$ confidence region for $\underline{\beta}$ based on $\hat{\beta}$. Cook (1977) felt that for an uncomplicated analysis, one would like each $\underline{\hat{\beta}}(i)$ to stay well within a 10% confidence region.

Let $H=X(X'X)^{-1}X'$, the projection matrix for fitting the linear model (2.1). H is also called the hat matrix. Let v_{ij} be an element of the ith row and the jth column of H, then clearly the ith diagonal element of H, $v_{ii} = \underline{x}_i'\ (X'X)^{-1}\underline{x}_i$. Let $\hat{y}_i$ and $y_i - \hat{y}_i$ be respectively the predicted value and residual corresponding to the ith data point. Then it can be shown that,

$$\mathrm{Var}\ (r_i) = \sigma^2\ (1 - v_{ii}) \qquad (2.5)$$

$$\mathrm{Trace}\ (H) = p \qquad (2.6)$$

$$0 \leq \mathrm{Max}\ v_{ii} \leq 1 \qquad (2.7)$$

If v_{ii} is close to 1, $\mathrm{Var}(r_i) \approx 0$ and thus $\hat{y}_i$ is essentially determined by y_i alone, y_i (and thus the ith data point) may have an undue influence on the determination of certain parameters of $\underline{\beta}$. Design point with $v_{ii} = \underset{j}{\mathrm{Max}}\ v_{jj}$ (large v_{ii} relative to other data points) lie on (near) the boundary of the smallest convex set containing all the design points. Cook (1979) called it independent variable hull (IVH). In figure 1.c, the convex set A is an example if IVH when the number of independent variables is two, point P has the largest v_{ii}.

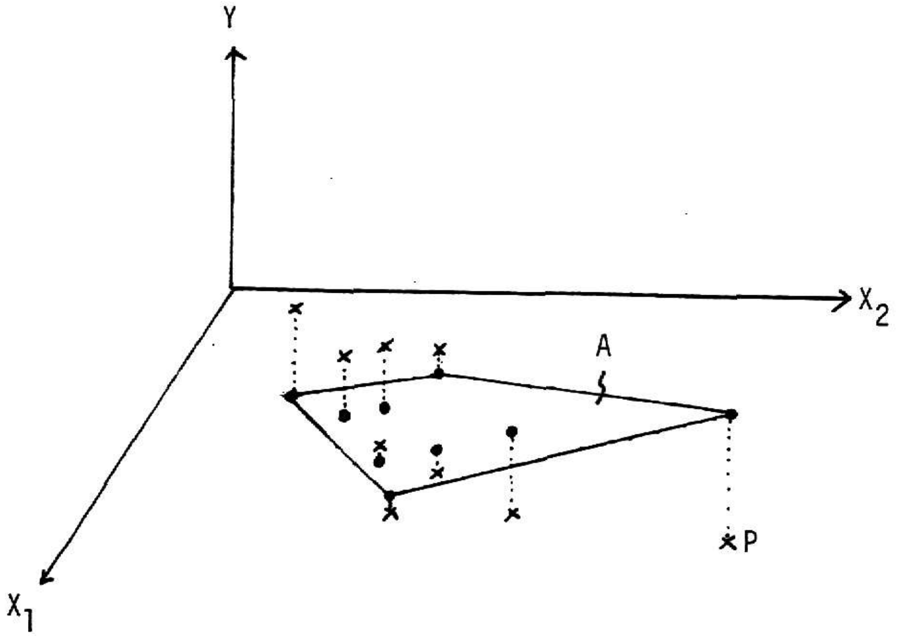

Figure 1.c

In fitting the simple linear regression model,

$$v_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \tag{2.8}$$

Hence, the design point x_i that is farthest away from $\bar{x}$ has the largest v_{ii}. In figure 1.d, we see that $v_{ii} > v_{kk}$. The ith data point (x_i, y_i) has more influence on estimating the slope of the regression line than the kth data point (x_k, y_k).

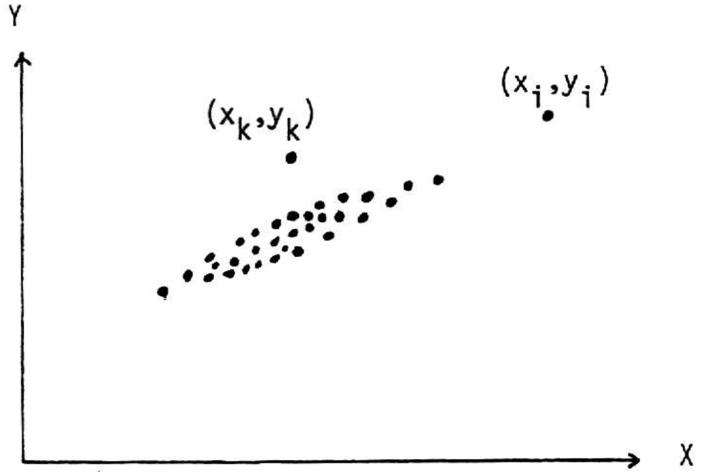

Figure 1.d

The studentized residual is defined as:

$$t_i = \frac{r_i}{\sqrt{s^2(1 - v_{ii})}} \tag{2.9}$$

Hence $|t_i|$measure how far the ith data point is away from the hyperplane defined by the fitted regression equation, and large $|t_i|$ indicates that the ith data point is potentially a critical observation. For example, in figure 1.e, point B has small v_{ii} but large $|t_i|$.

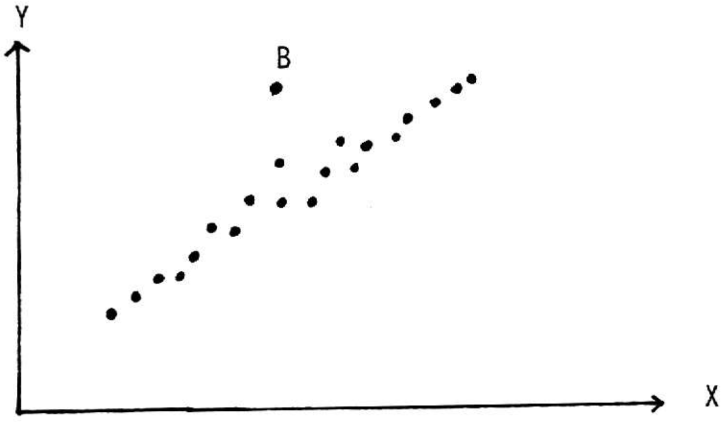

Figure 1.e

Cook showed that,

$$D_i = t_i^2 \cdot \frac{v_{ii}}{1 - v_{ii}} \cdot \frac{1}{p} \tag{2.10}$$

Hence D_i provide a combined measure which enables us to jugde simultaneously the measure $|t_i|$ and v_{ii}.

The distance measure D_i can easily be extended to accommodate the situation in which q(< = p) linear independent combinations of the elements of $\underline{\beta}$ are of interest (Cook, 1977). Let $\hat{\underline{\phi}} = C\hat{\underline{\beta}}$ where C is a qxp rank q matrix. The distance $D_i(\underline{\phi})$, between $\hat{\underline{\phi}} = C\hat{\underline{\beta}}$ and $\hat{\phi}_{(i)} = C\hat{\underline{\beta}}_{(i)}$ is defined to be:

$$D_i(\underline{\phi}) = \frac{(\hat{\underline{\phi}} - \hat{\underline{\phi}}_{(i)})'[C(X'X)^{-1}C]^{-1}(\hat{\underline{\phi}} - \hat{\underline{\phi}}_{(i)})}{q\ s^2} \tag{2.11}$$

When ϕ consists of a subset of, $\underline{\beta}_2$ of $\underline{\beta}' = (\underline{\beta}_1', \underline{\beta}_2')$, Cook found that,

$$D_i(\underline{\beta}_2) = \frac{t_i^2}{q} \cdot \frac{(v_{ii} - v_{ii}^*)}{1 - v_{ii}} \tag{2.12}$$

where $v_{ii}^* \cdot \sigma^2$ is the variance of the $\underline{ith}$ predicted value from the regression on the first p-q variables.

III. A Diagnostic Procedure Based on Prediction Region

3.1 Description of Procedure

In this chapter, we propose a diagnostic method for identifying outlying multivariate observations. This method unlike Cook's method, has no model assumption, and is based on only the relationship between a single data point and the rest of the data points. Therefore, it will be useful for data screening before any model is postulated for the data. It will be shown that the proposed procedure is related to Cook's method. Hence, we hope that the present procedure will be complementary to Cook's method and other diagnostic procedures for regression analysis.

$$\text{Let } Z = \begin{bmatrix} \underline{z}_1' \\ \underline{z}_2' \\ \vdots \\ \underline{z}_n' \end{bmatrix} \qquad (3.1)$$

Suppose $\underline{z}_1, \underline{z}_2, \ldots, \underline{z}_n$ are n random vectors from p-variate distribution with mean vector $\underline{\mu}$ and covariance matrix Σ of rank p, let $\bar{\underline{z}}$ and S described below be the usual sample mean vector and sample covariance matrix. i.e.,

$$\bar{\underline{z}} = \frac{\sum_{i=1}^{n} \underline{z}_i}{n} \qquad (3.2)$$

$$S = \frac{1}{n-1} \sum_{i=1}^{n} (\underline{z}_i - \bar{\underline{z}}) (\underline{z}_i - \bar{\underline{z}})'$$

Under the normality assumption, the Hotelling T^2-statistic

$$T^2 = n(\bar{\underline{z}} - \underline{\mu})' S^{-1} (\bar{\underline{z}} - \underline{\mu})$$

is distributed as $\frac{p(n-1)}{(n-p)}$ F, where F denoted an F-distribution with degrees of freedom p and (n-p).

Let Z(i) be a (n-1)xp matrix obtained by deleting the ith row $\underline{z}_i$ from Z. Let $\bar{\underline{z}}(i)$ and S(i) be the sample mean and covariance matrix based on the data matrix Z(i). A measure of the degree of influence the ith data point $\underline{z}_i$ has on the centroid (sample mean) of the data set may be based on the Mahalanobis distance between $\bar{\underline{z}}(i)$ and $\bar{\underline{z}}$:

$$T_i^2 = n(n-1)\ (\bar{\underline{z}}(i) - \bar{\underline{z}})'\ S(i)^{-1}\ (\bar{\underline{z}}(i) - \bar{\underline{z}}) \qquad (3.4)$$

Therefore, a data point with exceptionally large T_i^2-value relative to those of other data points is potentially an outlying observation. Since $(\bar{\underline{z}}(i) - \bar{\underline{z}}) = \frac{(\bar{\underline{z}}(i) - \underline{z}_i)}{n}$ T_i^2 can also be written as

$$T_i^2 = \frac{n-1}{n}\ (\underline{z}_i - \bar{\underline{z}}(i))'\ S(i)^{-1}\ (\bar{\underline{z}}_i - \bar{\underline{z}}(i)) \qquad (3.5)$$

A $(1-\alpha)100\%$ normal theory prediction region for a new observation $\underline{z}^*$ based on the data set Z(i) is the set of all $\underline{z}^*$ values satisfying

$$\frac{n-1}{n}\ (\underline{z}^* - \bar{\underline{z}}(i))'\ S(i)^{-1}\ (\underline{z}^* - \bar{\underline{z}}(i)) \le \frac{p(n-2)}{(n-p-1)}\ F_{1-\alpha}(p,n-p-1)$$

where $F_{1-\alpha}(p,n-p-1)$ is the $(1-\alpha)$x100 percentage point of the F-distribution with p and (n-p-1) degrees of freedom. Therefore, this also provides a measure of the distance between $\underline{z}_i$ and $\bar{\underline{z}}(i)$ in terms of descriptive level of significance. Suppose, for example, that the ith data point $\underline{z}_i$ lies outside of its 95% prediction region (i.e., $T_i^2 > \frac{p(n-2)}{(n-p-1)}\ F_{.95}(p,n-p-1)$. Such a situation may be cause for concern. For an uncomplicated situation, we would like each data point $\underline{z}_i$ to stay well within its 90% prediction region.

T_i^2 is related to the diagnostic statistic $\Lambda(i)$ suggested by Belsley, Kuh, and Welsch (1980, p.27). $\Lambda(i)$ is the Wilk's Λ-statistic for testing the differences in mean between two populations where one such population is represented by the ith data point and the second by the rest of data points. It can be shown that,

$$T_i^2 = (n-2)\,\frac{1-\Lambda(i)}{\Lambda(i)} \tag{3.7}$$

3.2. Computation Formulae for T_i^2

T_i^2 can be expressed in terms of $\underline{z}$ and S as defined in (3.2) and (3.3). The following two equations are useful:

$$\underline{z}(i) = \frac{(n\bar{\underline{z}} - \underline{z}_i)}{n-1} \tag{3.8}$$

$$(n-2)\ S(i) = (n-1)\ S - \frac{n}{n-1}\ (\underline{z}_i - \bar{\underline{z}})\ (\underline{z}_i - \bar{\underline{z}})' \tag{3.9}$$

Therefore,

$$\underline{z}_i - \bar{\underline{z}}(i) = \frac{n}{n-1}\ (\underline{z}_i - \bar{\underline{z}}) \tag{3.10}$$

$$S(i)^{-1} = \frac{n-2}{n-1} \left\{ S^{-1} + \frac{\frac{n}{(n-1)^2}\ S^{-1}(\underline{z}_i - \bar{\underline{z}})(\underline{z}_i - \bar{\underline{z}})'S^{-1}}{1 - \frac{n}{(n-1)^2}\ (\underline{z}_i - \bar{\underline{z}})'S^{-1}(\underline{z}_i - \bar{\underline{z}})} \right\} \tag{3.11}$$

Now using (3.10) and (3.11), we may write

$$T_i^2 = (n-2) \left\{ \frac{1}{1 - \frac{n}{(n-1)^2}\ C_{ii}} - 1 \right\} \tag{3.12}$$

$$\text{where } C_{ii} = (\underline{z}_i - \bar{\underline{z}})'\ S^{-1}\ (\underline{z}_i - \bar{\underline{z}}) \tag{3.13}$$

Hence, we only need to invert the matrix S for computing T_i^2 . (for i=1,2,...,n)

3.3 Outlying Observations

In general we would like each data point to stay well within a 90% prediction region. However, data points with exceptionally large T_i^2 relative to those of other data points should also be cause for concern.

Since $T_i^2 > = 0$ and by equation (3.12),

$$0 < = C_{ii} < = \frac{(n-1)^2}{n}$$

T_i^2 is an increasing function of C_{ii} and $T_i^2 = \infty$ when $C_{ii} = (n-1)^2/n$. Since C_{ii} is the Mahalanobis distance between $\underline{z}_i$ and $\bar{\underline{z}}$, large values of T_i^2 indicate "outlying" data points - data points which occur near the boundary of the smallest convex set containing all the data points. Furthermore,

$$\underset{\underline{x}}{\text{Max}} \quad \frac{\underline{x}'S^{-1}\underline{x}}{\underline{x}'\underline{x}} = \frac{\underline{x}_0'S^{-1}\underline{x}_0}{\underline{x}_0'\underline{x}_0} = \lambda \qquad (3.14)$$

where λ is the largest eigenvalue of S^{-1} and $\underline{x}_0$ is its corresponding eigenvector. Hence a data point lying in the direction along which the data has the least variability is more likely to yield large C_{ii} than those lying in the direction along which the data has large variability. A data point with the largest C_{ii} need not be the one whose Euclidean distance from the centroid $\bar{\underline{z}}$ of the data is the greatest. If $\underline{z}_i$ is any data point with k replicates, then

$$C_{ii} \leq \left\{ \frac{(n-1)^2}{n} \right\} \frac{1}{k}$$

This suggests that data point corresponding to large C_{ii} will tend to lie in a region near the boundary of the smallest convex set containing all the data points where the density of the data points is low.

3.4 Relation to Studentized Residuals and Residual Variances in Regression Analysis.

Let $X = \begin{bmatrix} \underline{x}_1' \\ \underline{x}_2' \\ \vdots \\ \underline{x}_n' \end{bmatrix}$

$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$

where y_i is the ith response depending on $\underline{x}_i' = (x_{i1}, x_{i2}, \ldots, x_{ip})$ which is the corresponding vector of p explanatory variables. Here we are fitting a linear model,

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i \qquad (3.15)$$

for i=1,2,..., n where the ε_i are uncorrelated random errors with zero means and common variance σ^2.

Let $Z = (X,\underline{Y})$ be nx(p+1) data matrix. It is shown that (Belsley, Kuh, and Welsch, 1980, p.27)

$$T_i^2 = (n-2)\ \frac{1 - \Lambda(i)}{\Lambda(i)} \qquad (3.16)$$

and

$$\Lambda(i) = \frac{n}{n-1}\ (1 - v_{ii})\ (1 + \frac{t_i^*}{n-p-2})^{-1} \qquad (3.17)$$

where v_{ii} is the ith diagonal element of the hat matrix of X. t_i^* is the ith residual standardized by its estimated standard deviation and is given by

$$t_i^* = \frac{r_i}{s(i)^2\ (1 - v_{ii})} \qquad (3.18)$$

where r_i is the ith residual and $s(i)^2$ is the mean square error for fitting the linear model (3.15) without the ith data point $\underline{z}_i = (\underline{x}_i', y_i)$.

Since,

$$(n-p-1)\ s(i)^2 = (n-p)\ s^2 - \frac{r_i^2}{1 - v_{ii}} \qquad (3.19)$$

equation (3.17) can be expressed as,

$$\Lambda(i) = \frac{n}{n-1}\ (1 - v_{ii})\ (1 - \frac{t_i^2}{n-p-1}) \qquad (3.20)$$

where t_i is the usual studentized residual as defined in (2.5).

We have observed in Chapter II that a potentially influential data point will tend to have large v_{ii} and/or large t_i (or t_i^*). Therefore, a small value of $\Lambda(i)$ (large value of T_i^2) indicated the critical nature of the ith data point $\underline{z}_i$. Note that if (n-p-1) is large, then the size of t_i (or t_i^*) has very little effect on the value of $\Lambda(i)$ which suggests the model independent nature of this procedure. Hence, the present approach sometimes will identify outliers which may not be influential in fitting the postulated linear model. However, when the "correctness" of the postulated model is in doubt, the present approach will identify the critical nature of a data point that may be overlooked by Cook's method or other regression diagnostic method.

To illustrate this point, two examples are considered below. In figure 3.a, point A will have a large T_i^2 but a small D_i. Point A is an outlier rather than an influential data point. In figure 3.b, point B is an outlying data point. If the data is fitted by a simple regression line, then point B will not be an influential point. However, if the data is fitted by a quadratic function of X, then point B will be influential. In particular, it will have a strong impact on estimating the regression coefficient of X^2. This point will be further examplified in Section 3.5 when we consider Brownlee's data (1965).

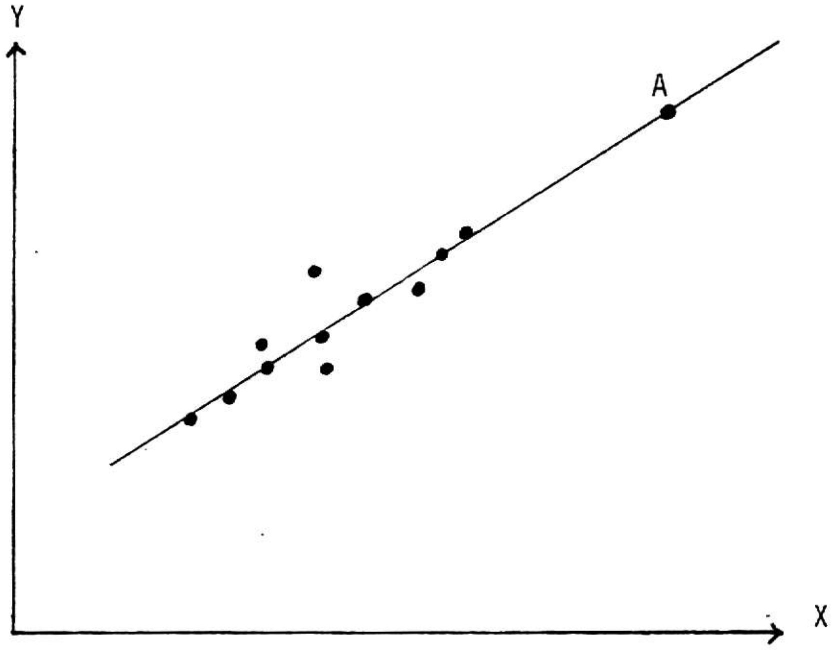

Figure 3.a

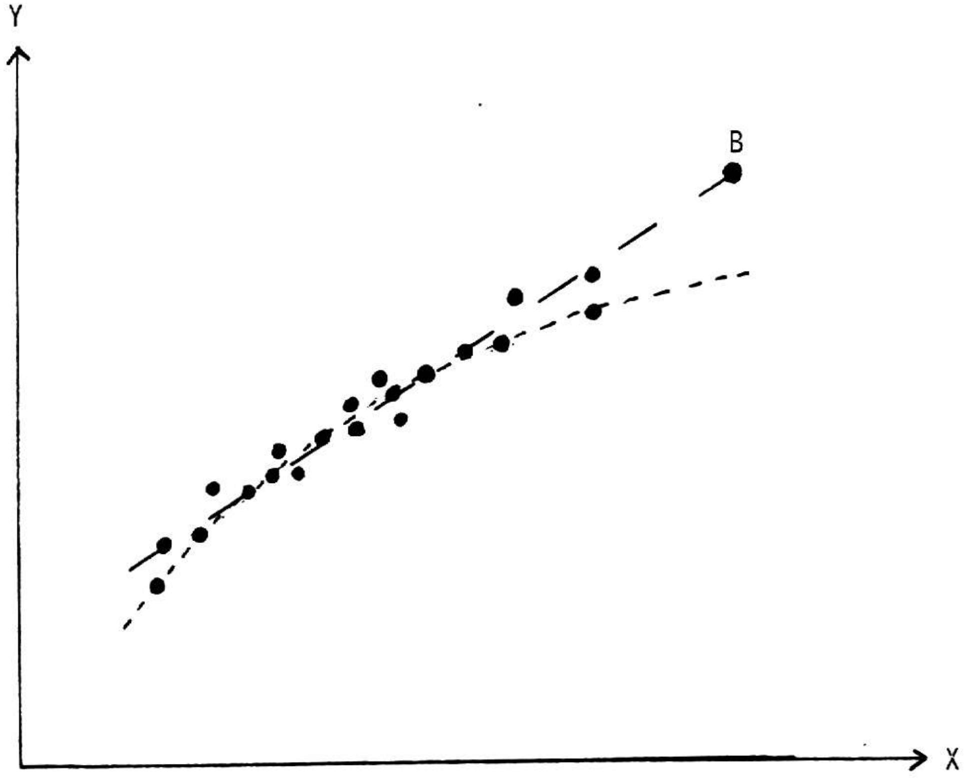

Figure 3.b

3.5 Consequence of Deleting a Data Point

In this section, we shall examine the effect on T_i^2 when an outlying observation is deleted. Suppose an outlying data point $\underline{z}_i$ is deleting from the data set, let $C_{jj}(i)$ be the C_{jj}-value computed without the outlying data point $\underline{z}_i$. As in Section 3.2, we can show that:

$$C_{jj}(i) = \frac{n-2}{n-1} \left\{ C_{jj} + \frac{\frac{n}{(n-1)^2} (C_{ij} + \frac{n-1}{n})^2}{1 - \frac{n}{(n-1)^2} C_{ii}} - \frac{1}{n} \right\} \qquad (3.21)$$

where

$$C_{ij} = (\underline{z}_i - \bar{\underline{z}})' \, S^{-1} \, (\underline{z}_j - \bar{\underline{z}})$$

Define the correlation between two data points $\underline{z}_i$ and $\underline{z}_j$ as

$$\rho_{ij} = \frac{C_{ij}}{\sqrt{C_{ii} \, C_{jj}}} \qquad (3.22)$$

Then from equation (3.21), we see that deleting an outlying data point $\underline{z}_i$ with C_{ii} being close to its upper bound $\frac{(n-1)^2}{n}$ and being highly correlated to $\underline{z}_j$ (in the sense of (3.22)), will substantially increase the C_{jj}-value (and thus the T_j^2-value) for data point $\underline{z}_j$. Hence, in general when one of two highly correlated outlying observations is deleted, the remaining observation is likely to become extremely critical as an outlier.

3.6 Consequence of Deleting a Subset of Variables.

Partition an nxm data matrix Z into

$$Z = \begin{bmatrix} \underline{z}'_{11} & \underline{z}'_{12} \\ \underline{z}'_{21} & \underline{z}'_{22} \\ \vdots & \vdots \\ \underline{z}'_{n1} & \underline{z}'_{n2} \end{bmatrix}$$

$$= \quad [Z_1, \quad Z_2] \qquad (3.23)$$

where Z_1 and Z_2 have dimensions p and (m-p) respectively. Corresponding to this partition, write

$$\underline{z}_i - \bar{\underline{z}} = \begin{bmatrix} \underline{z}_{i1} - \bar{\underline{z}}_1 \\ \underline{z}_{i2} - \bar{\underline{z}}_2 \end{bmatrix} \qquad (3.24)$$

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

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{12}' & S_{22} \end{bmatrix} \qquad (3.25)$$

The inverse of this partition matrix in terms of its submatrices is given by

$$S^{-1} = \begin{bmatrix} A_{11} & A_{12} \\ A_{12}' & A_{22} \end{bmatrix} \qquad (3.26)$$

where

$$A_{11} = S_{11}^{-1} + S_{22}^{-1} S_{12}(S_{22} - S_{12}'S_{11}^{-1}S_{12})^{-1}S_{12}'S_{11}^{-1} \qquad (3.27)$$

$$A_{12} = -S_{11}^{-1}S_{12}(S_{22} - S_{12}'S_{11}^{-1}S_{12})^{-1} \qquad (3.28)$$

$$A_{22} = (S_{22} - S_{12}'S_{11}^{-1}S_{12})^{-1} \qquad (3.29)$$

Therefore,

$$\begin{aligned} C_{ii} &= (\underline{z}_{i1} - \bar{\underline{z}})'S_{11}^{-1}(\underline{z}_{i1} - \bar{\underline{z}}_1) \\ &\quad + (S_{12}'S_{11}^{-1}(\underline{z}_{i1} - \bar{\underline{z}}_1) - (\underline{z}_{i2} - \bar{\underline{z}}_2))'A_{22} \\ &\quad (S_{12}'S_{11}^{-1}(\underline{z}_{i1} - \bar{\underline{z}}_1) - (\underline{z}_{i2} - \bar{\underline{z}}_2)) \end{aligned} \qquad (3.30)$$

Note that $(S_{11}^{-1}S_{12})$ is the estimated regression coefficients of $(\underline{z}_{i2} - \bar{\underline{z}}_2)$ on $(z_{i1} - \bar{\underline{z}}_1)$.

$(S_{12}'S_{11}^{-1}(\underline{z}_{i1} - \bar{\underline{z}}) - (\underline{z}_{i2} - \bar{\underline{z}}_2))$ is the ith residual vector obtained from regressing Z_2 on Z_1. If $p = m-1$, z_{i2} is a scalar, then (3.30) can be expressed as:

$$C_{ii} = (\underline{z}_{1i} - \bar{\underline{z}}_1)'S_{11}^{-1}(\underline{z}_{1i} - \bar{\underline{z}}_1) + \frac{n-1}{n-p-1}(1 - v_{ii})t_i^2 \qquad (3.31)$$

where v_{ii} is the ith diagonal element of the hat matrix of Z_1, t_i is the ith studentized residual of Z_2 regressing on Z_1. From equation (3.30), we see that deleting a subset of (m-p) variables in Z_2 will decrease the C_{ii}-value. Suppose that the deletion of (m-p) variables in Z_2 reduces the C_{ii}-value of an outlying data point i to such an extent that it is no longer an outlier. In this case, we may conclude that the (m-p) measurements of $\underline{z}_{2i}$ of the ith data point $\underline{z}_i$ is responsible for the ith data point $\underline{z}_i$ to be an outlier. Also if the (m-p) variables in Z_2 are included as part of the independent variables used in fitting a linear model involving the variables in the data set Z, then the ith data point $\underline{z}_i$ will likely have strong influence on the estimated regression coefficients associated with Z_2. That is deletion of $\underline{z}_i$ may substantially change the estimated regression coefficients associated with the (m-p) variables in Z_2.

3.7 Examples

Example 1

Longley (1967) considered a data set (table 3.1) relating six economic variables to total derived employment for the years 1947 to 1962. Cook (1977) applied his diagnostic method to this data set yields the conclusion that the data points corresponding to 1951 and 1962 have the largest impact on the estimation of regression coefficients $\underline{\beta}$. Removal of these points moves the least squares estimate of $\underline{\beta}$ outside of the 10% confidence region for $\underline{\beta}$ based on $\hat{\underline{\beta}}$. Cook also observed that the data point corresponding to 1956 although has the largest studentized residual, the effect of this point on estimating $\hat{\underline{\beta}}$ is not significant.

The method based on the Hotelling statistic T_i^2 shows that the data points corresponding to 1951 and 1962 fall outside of the 90% prediction region (see table 3.2). Also, their T_i^2-values are large relative to those of the rest of the data points.

Example 2

The data set in table 3.3 is from Brownlee (1965). It contains 21 successive days of operation of a plant oxidizing ammonia to nitric acid. Factor X_1 is the flow of air to the plant, factor X_2 is the temperature of the cooling water entering the countercurrent nitric oxide absorption tower, factor X_3 is the concentration of nitric acid in the absorbing liquid (it is not presented in the table) and the response value of Y is ammonia lost as unabsorbed nitric oxides.

After a detail analysis Daniel and Wood (1971) concluded a linear model which relates a linear and quadratic term for variable X_1 and a linear term for variable X_2 to the response variable Y. The fitting of the model is based on 17 "valid" observations, where 1st, 3rd, 4th and 21st observations were declared outliers.

Cook (1979) used the same model with all 21 observations to illustrate the interesting results he obtained. Cook's diagnostic statistic D_i suggests that observations 1, 2, 4 and 21 are influential observations. Observations 4 and 21 are the two most influential ones (see table 3.4). We shall use the data set (excluding the variable X_3) to illustrate the results of the previous sections.

First we shall consider Brownlee's data set containing variables Y, X_1, X_1^2 and X_2. From Table 3.4, we see that observations 1, 2, 4 and 21 lie outside of their 90% prediction regions. Their T_i^2-values are large relative to those of other observations. Observation 21 being outside of the 99% prediction region is clearly the most extreme outlying observation. Table 3.5 gives the correlation between observations 1, 2, 3, 4 and 21. The correlations between observations 1 and 2, 1 and 3, 4 and 21 are respectively .821, .798 and -.766. We should anticipate that 2 and 3 will become extremely critical when 1 is deleted, and 4 will become extremely critical when 21 is deleted. Also, since the correlation between 2 and 3 is low, we expect that the deletion of 3 will have very little effect on observation 2. The 3rd, 4th and 5th columns of Table 3.2 show the effect of deletion of observation 1, 3 and 21.

Now we shall consider Brownlee's data set containing variables Y, X_1 and X_2. Table 3.7 shows that observations 1 and 21 lie outside of the 90% prediction region. Again 21 is an extreme outlying observation. Observations 2, 3 and 4 lie just outside of the 80% prediction region. From table 3.8, we see that 2 and 3 will become critical when 1 is deleted, and 4 will become extremely critical when 21 is deleted (see column 3 and 4 of table 3.9). Column 4 to column 8 of table 3.9 also shows the effects of deleting observations 21, 4, 1, 3 and 2 sequentially until all five observations are deleted. It is interesting to note that there is no data point in the remaining data set that has exceptionally large T_i^2-value.

From tables 3.4 and 3.7, we see that the T_i^2-value for observation 2 decreases from 13.79 to 6.47 and observation 4 decreases when the variable X_1^2 is deleted. Therefore, we should expect that deletion of observations 2 and 4 will have strong influence on the estimate of the regression coefficient associated with X_1^2. This can be seen by Cook's (1979) analysis based on the partial F statistics and the diagnostic statistics D_i corresponding to the regression coefficient associated with X_1^2 (see table 4 of Cook (1979) and the last paragraph of Cook's (1979) article). In fact, Cook (1979) felt that "for the final data set of Daniel and Wood, the quadratic term is needed to model a single observation".

From tables 3.7 and 3.10, we see that the T_i^2-value for observation 21 decreases from 23.7 to 0.13375. This suggests that the measurement of variable X_1 in observation 21 is responsible for its being an extreme outlying observation. From the data set in table 3.3, we see that observations 9, 10, 11 and 20 are alike. Observation 21 has the 3rd largest measurement on X_1, and its other two measurements on X_2 and Y are the same as that of observation 20. Observations 1 and 2 have the largest measurement on X_1, but their measurements on X_2 and especially Y are quite different from that of observations 9, 10, 11, 20 and 21.

This may explain why measurement on X_1 in observation 21 has such a strong impact on observation 2 being an extreme outlying observation.

Table 3.1

The Total Derived Employment and 6 Related Economic Variables

Presented by Longley (1967)

Obs.	X_1	X_2	X_3	X_4	X_5	X_6	Y
1	83.0	234,289	2,356	1,590	107,608	1947	60,323
2	88.5	259,426	2,325	1,456	108,632	1948	61,122
3	88.2	258,054	3,682	1,616	109,773	1949	60,171
4	89.5	284,599	3,351	1,650	110,929	1950	61,187
5	96.2	328,975	2,099	3,099	112,075	1951	63,221
6	98.1	346,999	1,932	3,594	113,270	1952	63,639
7	99.0	365,385	1,870	3,547	115,094	1953	64,989
8	100.0	363,112	3,578	3,350	116,219	1954	63,761
9	101.2	397,469	2,904	3,048	117,388	1955	66,019
10	104.6	419,180	2,822	2,857	118,734	1956	67,857
11	108.4	442,769	2,936	2,798	120,445	1957	68,169
12	110.8	444,546	4,681	2,637	121,950	1958	66,513
13	112.6	482,704	3,813	2,552	123,366	1959	68,655
14	114.2	502,601	3,931	2,514	125,368	1960	69,564
15	115.7	518,173	4,806	2,572	127,852	1961	69,331
16	116.9	554,894	4,007	2,827	130,081	1962	70,551

X1 = GNP implicit price deflater 1954 = 100.
X2 = Gross National Product.
X3 = Unemployment.
X4 = Size of Armed Forces.
X5 = Noninstitutional Population 14 years of Age and Over.
X6 = Time.
Y = Total Derived Employment.

Table 3.2

List of Values of T_i^2, D_i and Their Associated Confidence Coefficients Based on Longley's Data

Obs. #	Year	T_i^2	C.C. Based on T_i^2 (%)	D_i	C.C. Based on D_i (%)
1	1947	12.78	52.81	.14	.85
2	1948	16.92	67.15	.04	.02
3	1949	6.66	21.82	.00	.00
4	1950	16.76	66.69	.24	3.82
5	1951	34.65	91.57*	.61	26.64
6	1952	9.60	38.05	.09	.21
7	1953	13.56	55.98	.08	.15
8	1954	12.51	51.76	.00	.00
9	1955	10.19	41.04	.00	.00
10	1956	17.14	67.77	.24	3.48
11	1957	6.52	21.02	.00	.00
12	1958	11.48	47.26	.00	.00
13	1959	7.99	29.36	.04	.01
14	1960	3.21	4.70	.00	.00
15	1961	12.93	53.51	.17	1.47
16	1962	36.43	92.51	.47	16.37

* 1951 is the full year of the Korean war.

Table 3.3

Data from Operation of a Plant for the Oxidation of Ammonia to Nitric Acid, Brownlee (1965)

Obs #	X_1	X_2	Y
1	80	27	42
2	80	27	37
3	75	25	37
4	62	24	28
5	62	22	18
6	62	23	18
7	62	24	19
8	62	24	20
9	58	23	15
10	58	18	14
11	58	18	14
12	58	17	13
13	58	18	11
14	58	19	12
15	50	18	8
16	50	18	7
17	50	19	8
18	50	19	8
19	50	20	9
20	56	20	15
21	70	20	15

X1 = Air flow
X2 = Cooling water inlet temperature
X3 = Acid concentration (not presented here)
Y = Stack loss

Table 3.4

List of Values of T_i^2 , D_i and Their Associated Confidence Coefficients Based on Brownlee's Data Containing Variables Y, X_1, X_2 and X_1^2

Obs. #	T_i^2	C.C. Based on T_i^2 (%)	D_i	C.C. Based on D_i (%)
1	13.423	94.00	.162	4.54
2	13.790	94.45	.193	6.15
3	6.485	71.02	.125	2.88
4	13.100	93.57	.304	12.88
5	1.283	10.71	.003	.00
6	2.568	29.18	.021	.09
7	4.337	51.99	.042	.36
8	3.692	44.40	.014	.05
9	3.802	45.75	.043	.38
10	2.915	34.11	.028	.17
11	2.915	34.11	.028	.17
12	5.292	61.59	.062	.78
13	2.048	11.59	.001	.00
14	.991	7.01	.001	.00
15	3.317	39.58	.002	.00
16	3.319	39.58	.002	.00
17	3.341	39.90	.004	.00
18	3.341	39.90	.004	.00
19	4.129	49.64	.008	.02
20	.856	5.47	.008	.01
21	23.838	99.18	.699	39.70

Table 3.5

Correlations Between Observations 1, 2, 3, 4, 21

Based on Brownlee's Data Containing Variables Y, X_1, X_2, X_1^2

Obs. #	1	2	3	4	21
1	1	.821	.798	.097	-.120
2		1	.383	-.377	.322
3			1	.550	-.286
4				1	-.766
21					1

Table 3.6

List of Values of T_i^2

Based on Brownlee's Data Containing Variables Y, X_1, X_2 and X_1^2

Obs.	T_i^2 Obs. Deleted			
#	none	1	3	21
1	13.42	*	21.20	12.65
2	13.79	38.73	15.36	19.09
3	6.48	12.81	*	6.33
4	13.10	13.06	16.36	30.77
5	1.28	1.19	1.19	1.65
6	2.57	2.44	2.38	2.99
7	4.34	4.12	4.06	4.34
8	3.69	3.47	3.45	3.48
9	3.80	3.71	3.65	3.67
10	2.92	2.71	2.98	2.92
11	2.92	2.71	2.98	2.92
12	5.29	4.96	5.34	5.64
13	2.05	2.97	1.89	3.87
14	.99	.97	.89	2.00
15	3.32	3.13	3.10	3.23
16	3.32	3.10	3.16	3.09
17	3.34	3.12	3.21	3.23
18	3.34	3.12	3.21	3.23
19	4.13	3.87	3.99	4.28
20	.86	.76	.87	.94
21	23.84	22.49	22.89	*

Table 3.7

List of Values of T_i^2 , D_i and Their Associated Confidence Coefficients Based on Brownlee's Data Containing Variables Y, X_1 and X_2

Obs. #	T_i^2	C.C. Based on T_i^2 (%)	D_i	C.C. Based on D_i (%)
1	8.96	91.96	.235	12.91
2	6.47	83.70	.029	.71
3	6.41	83.40	.174	8.76
4	6.78	85.08	.172	8.60
5	.42	5.65	.006	.06
6	1.72	32.03	.027	.61
7	3.27	57.26	.060	1.98
8	2.48	45.63	.029	.70
9	3.49	59.99	.065	2.23
10	2.33	43.17	.024	.52
11	2.33	43.17	.024	.52
12	4.63	71.78	.059	1.95
13	1.85	34.54	.004	.04
14	.78	12.78	.004	.04
15	1.68	31.36	.010	.14
16	1.44	26.49	.001	.00
17	1.54	28.60	.000	.00
18	1.54	28.60	.000	.00
19	2.30	42.63	.000	.00
20	.62	9.55	.007	.09
21	23.70	99.73	.949	56.20

Table 3.8

Correlations Between Observations 1, 2, 3, 4, 21

Based on Brownlee's Data Containing Variables Y, X_1 and X_2

Obs. #	1	2	3	4	21
1	1	.763	.961	.539	-.163
2		1	.569	-.003	.374
3			1	.619	-.283
4				1	-.918
21					1

Table 3.9

List of Values of T_i^2

Based on Brownlee's Data Containing Variables Y, X_1 and X_2

			T_i^2				
Obs.			Obs. Deleted			21,	21,4,
#	none	1	21	21,4	21,4,1	4,1,3	1,3,2
1	8.96	*	8.45	11.73	*	*	*
2	6.47	10.04	9.03	8.92	12.09	37.18	*
3	6.41	12.23	6.24	10.01	34.20	*	*
4	6.78	8.63	15.64	*	*	*	*
5	.42	.39	.87	.90	.80	.71	1.56
6	1.72	1.59	2.21	2.24	2.06	2.20	2.53
7	3.27	3.06	3.37	3.15	2.91	2.75	3.59
8	2.48	2.35	2.35	2.23	2.15	2.17	4.79
9	3.49	3.28	3.39	3.15	2.94	3.27	2.99
10	2.33	2.22	2.39	2.22	2.07	2.43	3.17
11	2.33	2.22	2.39	2.22	2.07	2.43	3.17
12	4.63	4.40	5.05	4.92	4.57	4.47	4.47
13	1.85	1.71	3.71	5.33	5.50	8.54	9.12
14	.78	.70	1.83	2.51	2.55	3.85	3.87
15	1.68	1.62	1.72	2.12	1.94	1.96	2.18
16	1.44	1.46	1.33	1.33	1.26	1.20	2.52
17	1.54	1.55	1.56	1.87	1.72	1.55	2.31
18	1.54	1.55	1.56	1.87	1.72	1.55	2.31
19	2.30	2.27	2.57	3.46	3.20	3.04	3.34
20	.62	.58	.70	1.40	1.58	4.27	6.53
21	23.70	22.40	*	*	*	*	*

Table 3.10

List of Values of T_i^2

Based on Brownlee's Data Containing Variables Y and X_2

Obs. #	T_i^2
1	8.81
2	4.77
3	5.88
4	1.12
5	.26
6	1.45
7	3.15
8	2.47
9	3.47
10	2.33
11	2.33
12	4.60
13	1.23
14	.45
15	1.04
16	1.16
17	1.03
18	1.03
19	1.44
20	.13
21	.13*

IV. Graphical Presentation

Here we apply the classical solution of the multidimensional scaling technique (see e.g. Chatfield and Collins, 1980) to construct a configuration of the n data points in Euclidean space. This technique is essentially based on the method of principle components. We shall use C_{ij} as a measure of dissimilarities (or similarities) of two data points.

We apply this graphical technique to Brownlee's data based on variables Y, X_1, X_2 and on variables Y, X_1, X_2, X_1^2. For these two sets of data, no reduction of dimension is realized. Hence the configuration of the n data points is plotted on selected pairs of axes (planes). Figures 3.d - 3.j show the plots. Observations 1, 2, 3, 4 and 21 are identified in each of these plots. The results agree with the analysis obtained in the previous sections.

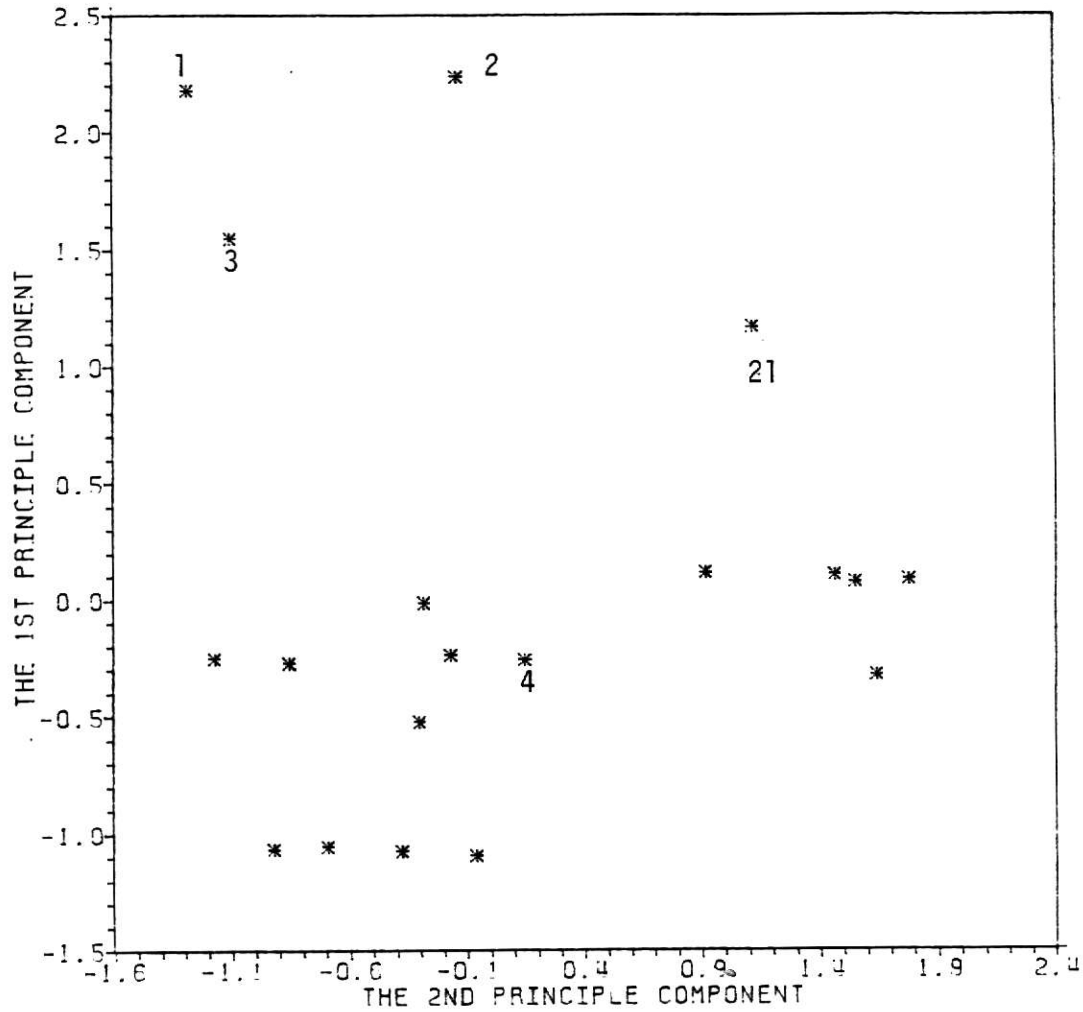

Figure 3.d : Plotting data points by multidimensional scaling method, data from Brownlee (1965) contains variables Y, X_1 , X_2 and X_1^2 .

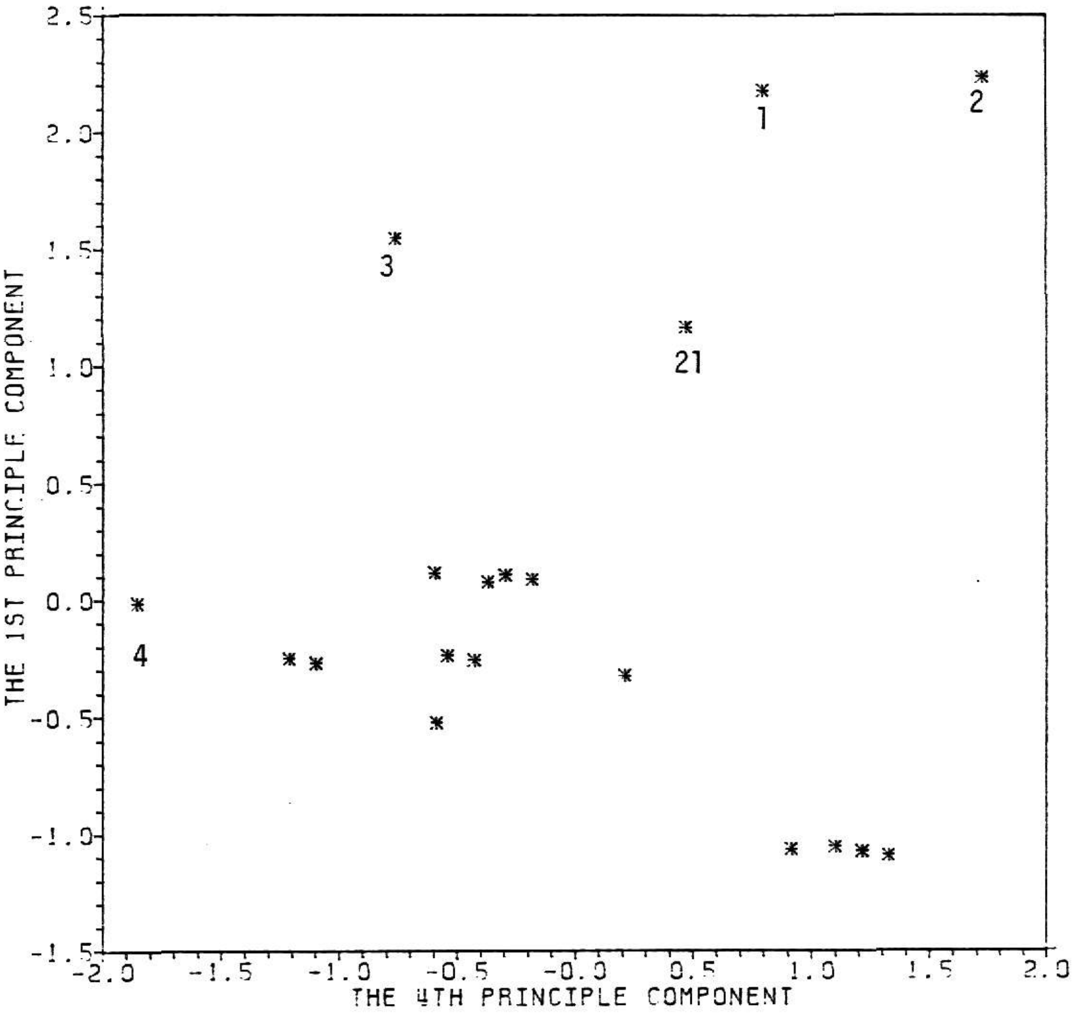

Figure 3.e : Plotting data points by multidimensional scaling method, data from Brownlee (1965) contains variables Y, X_1, X_2 and X_1^2 .

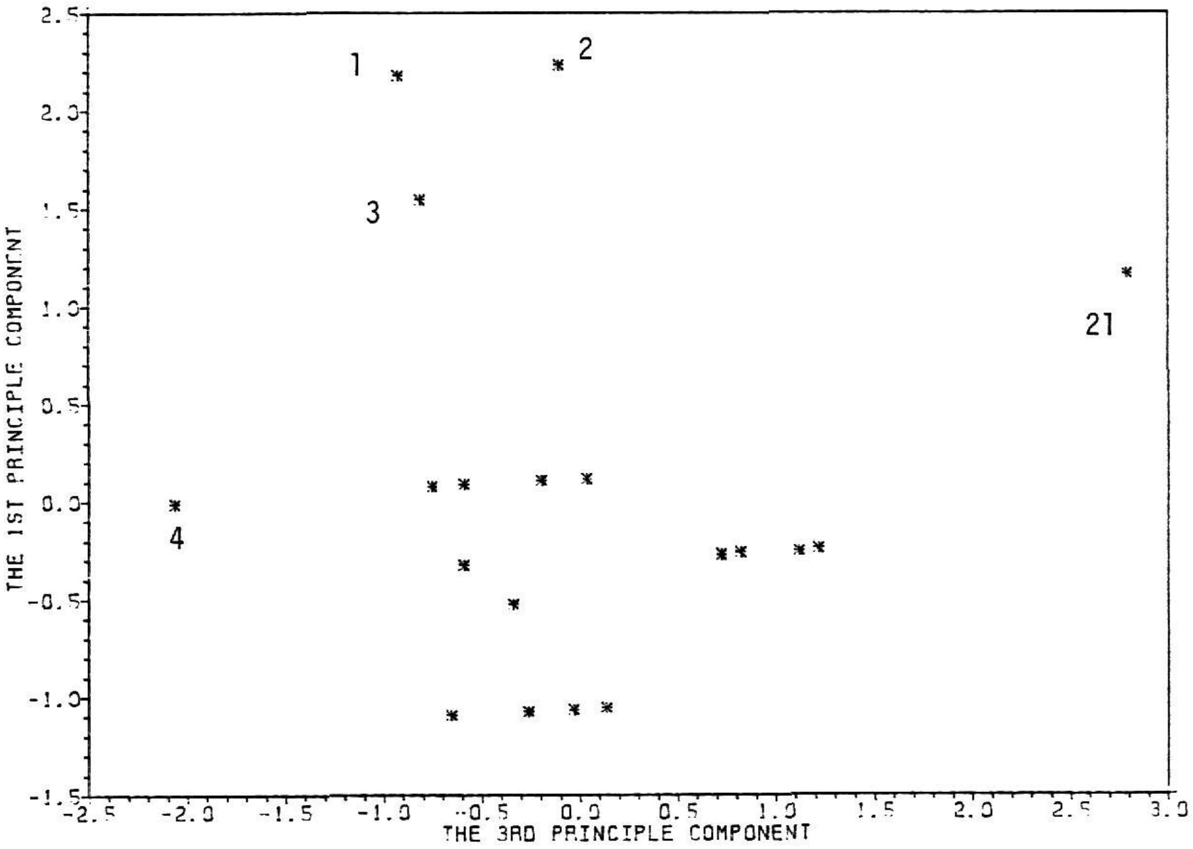

Figure 3.f : Plotting data points by multidimensional scaling method, data from Brownlee (1965) contains variables Y, X_1, X_2 and X_1^2 .

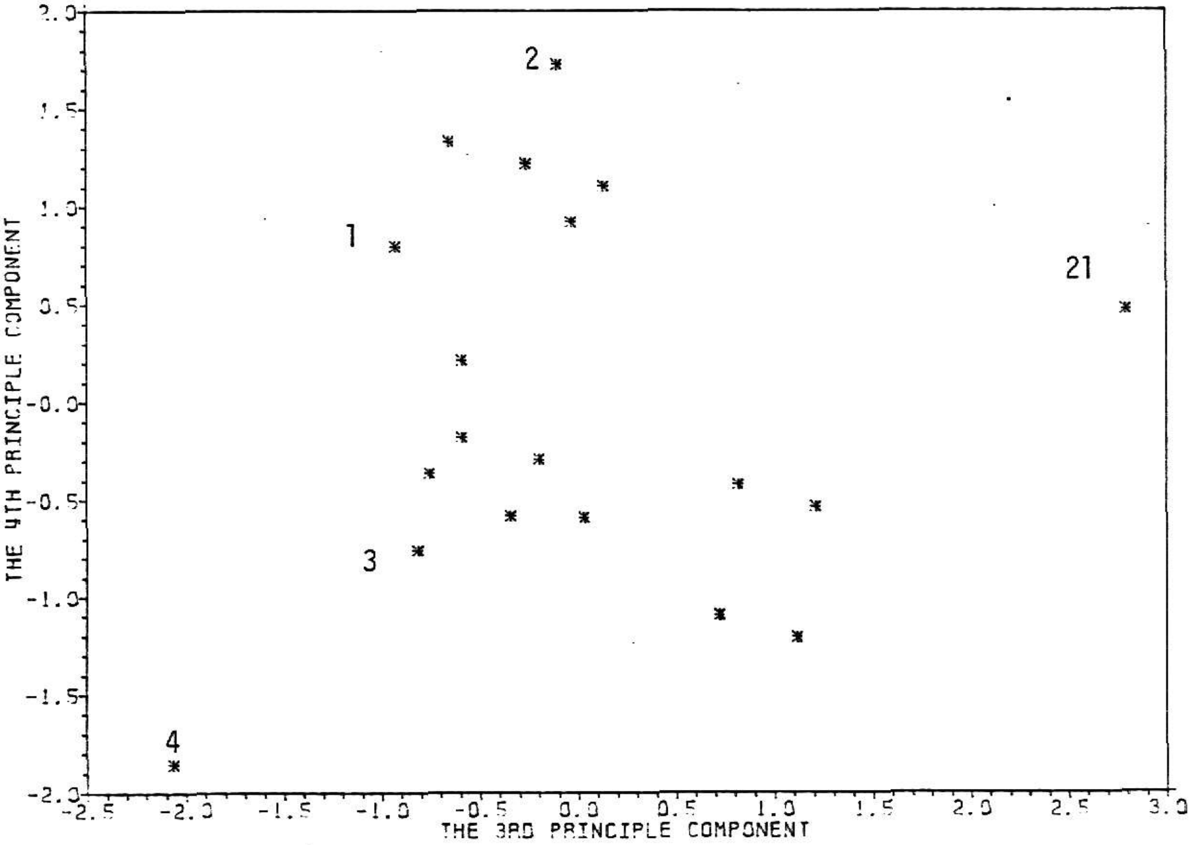

Figure 3.g : Plotting data points by multidimensional scaling method, data from Brownlee (1965) contains variables Y, X_1, X_2 and ${X_1}^2$.

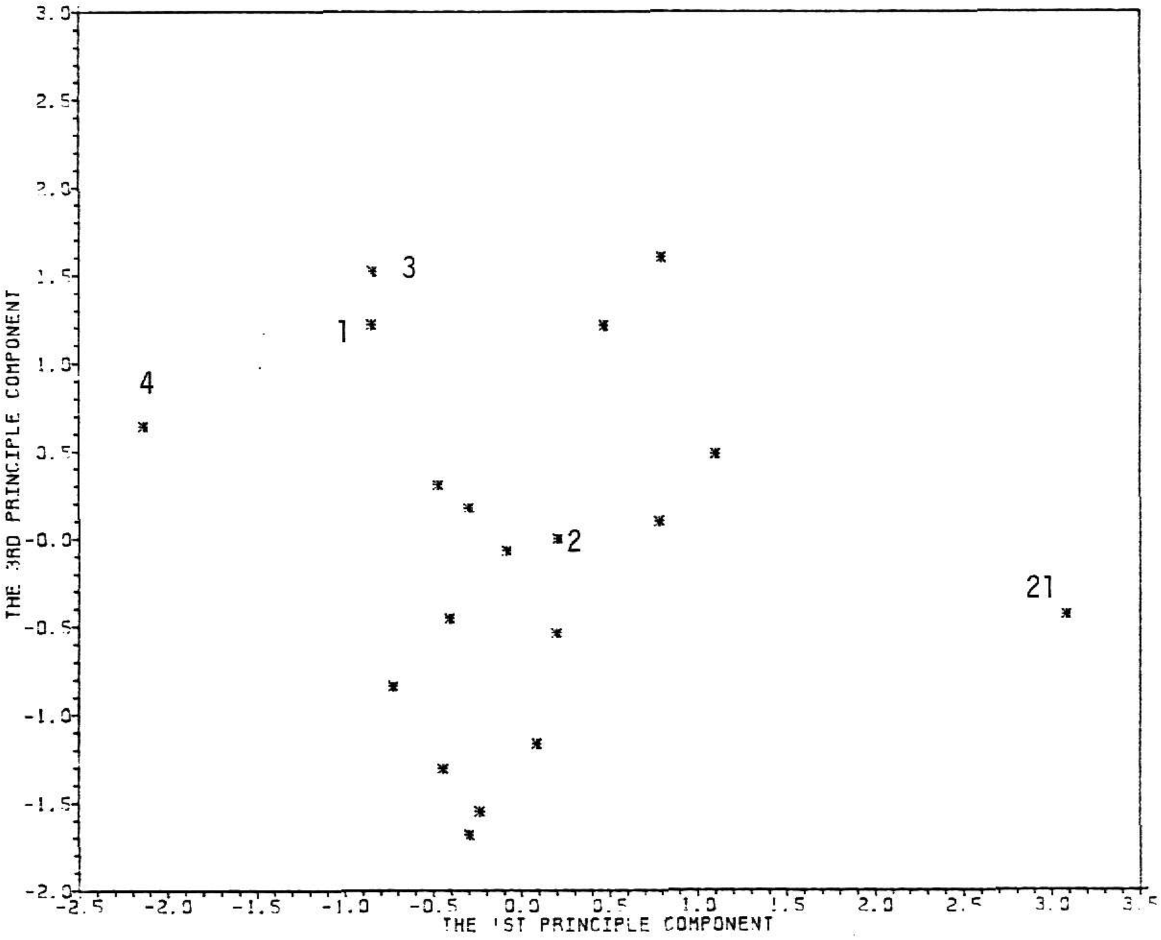

Figure 3.h : Plotting data points by multidimensional scaling method, data from Brownlee (1965) contains variables Y, X_1 and X_2 .

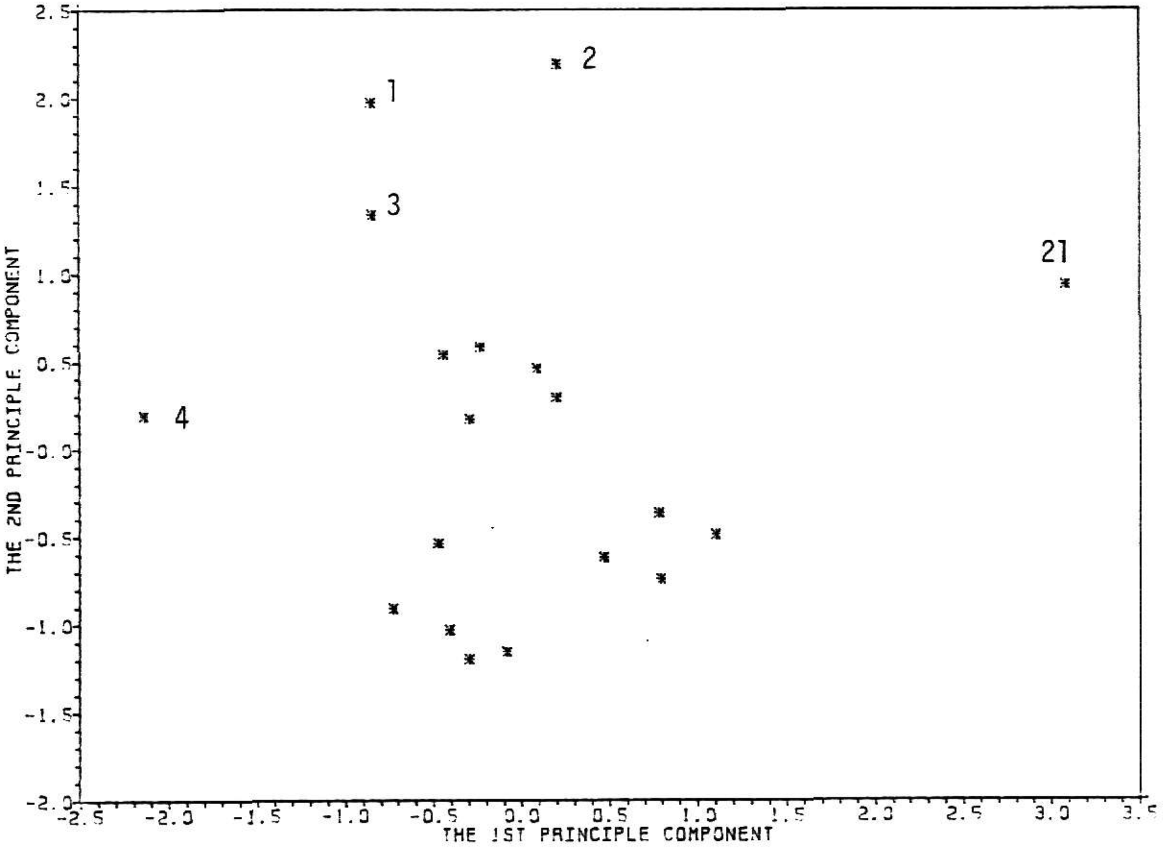

Figure 3.i : Plotting data points by multidimensional scaling method, data from brownlee (1965) contains variables Y, X_1 and X_2.

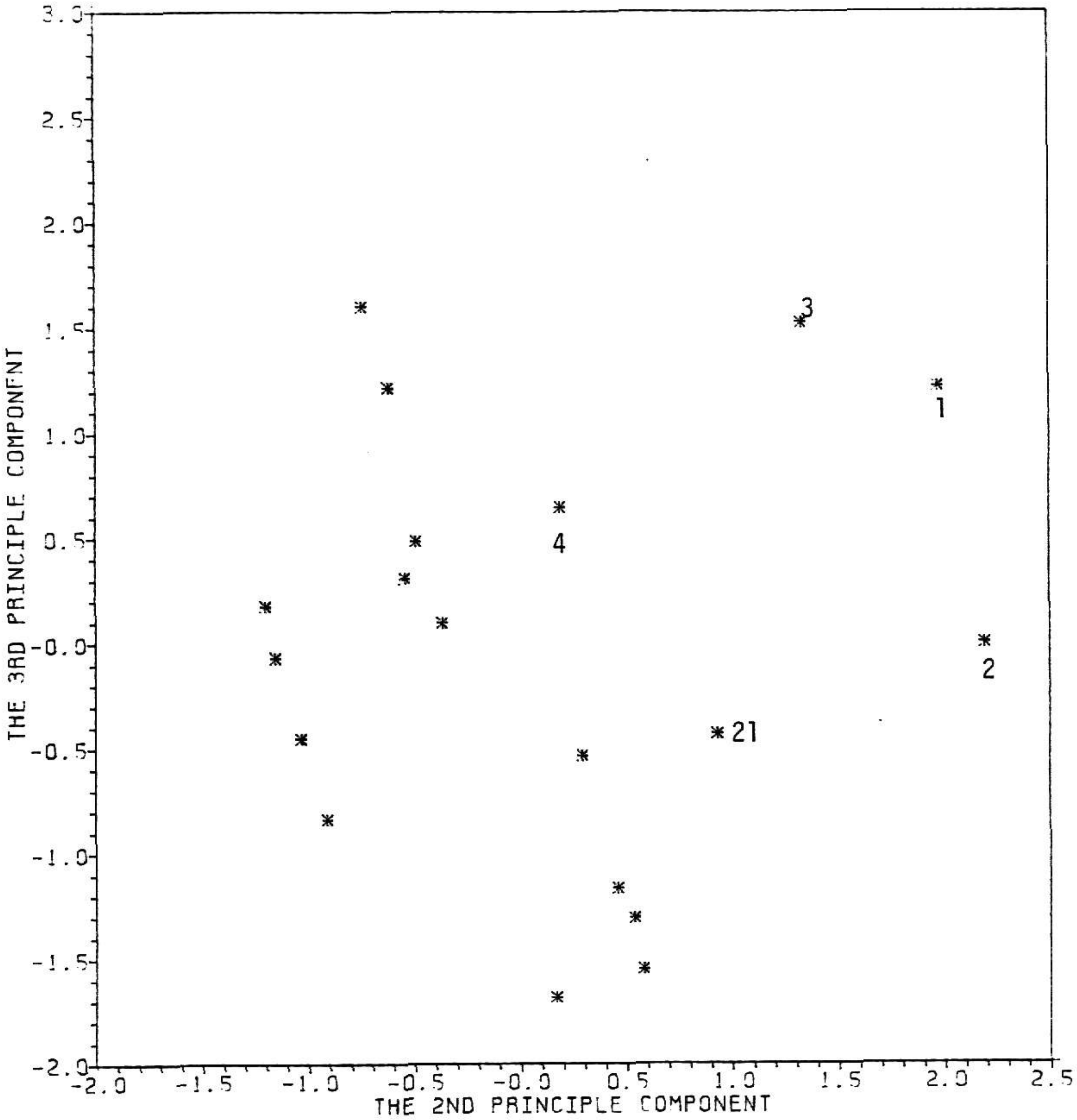

Figure 3.j : Plotting data points by multidimensional scaling method, data from Brownlee (1965) contains variables Y, X_1 and X_2 .

REFERENCES

Belsley, D.A., E. Kuh and R.E. Welsch, "Regression Diagnostics", (1980), Hohn Wiley & Sons, N.Y.

Brownlee, K.A., "Statistical Theory and Methodology in Science and Engineering", (1965), 2nd ed., John Wiley & Sons, N.Y.

Chatfield, C. and A.J. Collins, "Introduction to Multivariate Analysis", (1980), Chapman and Hall, N.Y.

Cook, R.D., "Detection of Influential Observation in Linear Regression", (1977), Technometrics, 19, 15-18.

Cook, R.D., "Influential Observations in Linear Regression", (1979), Journal of the American Statistical Association, 74, 169-174.

Daniel, C. and F.S. Wood, "Fitting Equations to Data", (1971), John Wiley & Sons, N.Y.

Draper, N.R. and H. Smith, "Applied Regression Analysis", (1980), John Wiley & Sons, N.Y.

Draper, N.R. and D.M. Stoneman, "Testing for the Inclusion of Variables in Linear Regression by a Randomisation Technique", (1966), Techbometrics, 8, 695-699.

Hoaglin, D.C. and R.E. Welsch, "The Hat Matrix in Regression and ANOVA",(1978), The American Statistician, 32, 17-22.

Longley, J.W., "An Appraisal of Least Squares Programs for the Electronic Computer from the Point of View of the User", (1967), 62, 819-841.

Morrison, D.F., "Multivariate Statistical Methods", (1976), 2nd ed., McGraw - Hill.

A DIAGNOSTIC METHOD FOR IDENTIFYING MULTIVATIATE OUTLYING OBSERVATIONS

by

YE JAIN HWANG LEE

B.S. in Agriculture Economics, National Chung-Hsing University,1973

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY

1983

ABSTRACT

When the sample size and dimension of the data points are large, outlying multivariate data points are difficult to spot. A simple method for identifying outlying multivariate observations is developed. The proposed method is related to the method of detecting influential observations in regression analysis when a linear model is postulated. Two examples are considered to illustrate the results obtained in this article.