*Statistics and Applications*
Volume 11, Nos. 1&2, 2013 (New Series), pp. 147-160

# Detection of outliers in designed experiments in presence of masking

Lalmohan Bhar, V. K. Gupta and Rajender Parsad
*Indian Agricultural Statistics Research Institute, Library Avenue, PUSA, New Delhi*

## Abstract

A method of identifying subset of outliers in presence of masking has been developed for designed experiments. An influence matrix comprising of Cook-statistics in its diagonal and product of two Cook-statistics in its off-diagonal positions has been defined. On the basis of eigenvectors corresponding to large eigenvalues of this matrix, the influential subsets can be identified. The proposed procedure has been illustrated with an example.

*Key words:* Outliers, masking, eigenvalue, eigenvector

## 1 Introduction

From the very beginning when the people started exploiting and employing the information in the collected data as a tool to understand the world he lives in, there has been concern over the unrepresentative or outlying observation (or subset of observations) that appears to be inconsistent with the remaining observations in the data set. Occurrence of outlier(s) is (are) very common in all the fields where collection data is involved. Generally outlier(s) arises (arise) from heavy tailed distributions or is (are) simply bad data point(s) due to error. When outlier(s) is (are) present in the data, the conclusion drawn from the experiment may be erroneous. Therefore, it is important to detect and handle the outlier(s) efficiently. Several statistics for detection of a single outlier or an isolated influential point in linear regression analysis are now available in the literature. However, these statistics are developed under the assumption that the data are generated from a kind of linear model where the design matrix is of full rank. However, in case of design of experiments, the design matrix is not of full rank and the interest of the experimenter is in estimation of some linear functions of parameters, say treatment effects. Estimation of these functions may be severely affected in the presence of outliers. One may, therefore, be interested to study the effect of outliers on the estimation of this subset of parameters. On this line of interest, Bhar and Gupta (2001) developed some statistics for detecting outliers in designed experiments. They modified Cook statistic for its application to design of experiments, which is a follow up work of Cook (1977).

The detection of a subset of outliers, *i.e.*, multiple outliers in comparison to detection of a single outlier is more difficult, owing to masking and swamping problems. Masking occurs when one outlier is not detected because of the presence of others and swamping occurs when a non-outlier is wrongly identified owing to the effect of some hidden outliers. Several procedures have been proposed for dealing with multiple outliers in linear regression models. Marasinghe (1985) and Kianifard and Swallow (1990) have suggested a sequential testing strategy to identify a set of $k$ points, where the maximum number of outliers in the sample, $k$, is fixed in advance. Atkinson (1986), Rousseeuw and Leroy (1987) and Rousseeuw and van Zomeren (1990) have suggested the use of robust estimates with high breakdown point for the regression parameters to overcome the masking problem. These estimates are computed by using a resampling scheme. Hawkins (1980) has proposed a diagnostic procedure which is also based on a resampling scheme. Gray and Ling (1984) proposed the use of cluster analysis for identification of multiple outliers in presence of masking. Hocking (1984) has suggested that the eigenstructure of the matrix $(\mathbf{X}:\mathbf{y})'(\mathbf{X}:\mathbf{y})$ should be computed, where $\mathbf{y}$ is the vector of responses and $\mathbf{X}$ is the corresponding design matrix. Pena and Yohai (1995) proposed a method to identify influential subsets by looking at the eigenvalues of an ‘influence matrix’. This matrix is defined as the uncentred covariance of a set of vectors which represent the effect on the fit of the deletion of each data point. This matrix is normalized to have the univariate Cook (1979) statistics on the diagonal.

Pena and Yohai (1995) used difference between predicted values of observations obtained from full data and after deleting the suspected outlier to form the influence matrix. In case of designed experiment, we are generally interested in the estimation of some subset of parameters, not the whole set of parameters. Bhar and Gupta (2001) developed Cook-statistic for detecting outliers in designed experiments when our interest is in estimation of some set of treatment contrasts. In the present investigation, we developed a method to identify outliers in designed experiments in presence of masking. Following Pena and Yohai (1995), we also formed an influence matrix. But the elements of this matrix are derived from Cook-statistics, since this statistics is useful in identifying outliers in designed experiments when interest is in the estimation of a set of treatment contrasts. In the next section this method is developed and it is illustrated through an example in Section 3.

## 2 Development of influence matrix

Consider the linear regression model

$$\mathbf{y} = \mathbf{X\theta} + \boldsymbol{\varepsilon}\,;\ \mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}\,,\ \mathrm{D}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_\mathrm{n}\,,\ \sigma^2 > 0 \tag{2.1}$$

where $\mathbf{y}$ is an $n\times 1$ vector of observations, $\mathbf{X}$ is $n\times p$ design matrix for explanatory variables with rank $p$. According to Pena and Yohai (1995), $\mathbf{t}_i = \hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)}$ summarizes the effect on the fit of deleting the $i^{th}$ observation, where $\hat{\mathbf{y}}$ and $\hat{\mathbf{y}}_{(i)}$ are the estimated values of $\mathbf{y}$ with full data and after deleting the $i^{th}$ data point respectively. They defined

$$\mathbf{N} = c\mathbf{T}'\mathbf{T}\,,$$

where $\mathbf{T} = (\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_n)$ and $c$ is a constant.

According to them one of the most important types of masking situations occurs when several observations have similar effects on the least squares fit. Two observations $i$ and $j$ have similar effects on the set of treatment contrasts when $\mathbf{t}_i \approx \lambda\mathbf{t}_j$ for some scalar $\lambda > 0$ and have opposite effects when $\lambda < 0$ .

Let $r_{ij} = \dfrac{n_{ij}}{n_{ii}^{1/2}\, n_{jj}^{1/2}}$ . If two outlying observations $i$ and $j$ are masked, then it is expected that $r_{ij} = 1$ or $-1$ according to they have similar effects or opposite effects. Final determination is done after assessing the effects on deletion of these observations. Hence Pena and Yohai (1995) proposed a procedure of detecting outliers in presence of masking on the basis of the coordinates of eigenvectors corresponding to large eigenvalues of $\mathbf{N}$.

As mentioned earlier, in case of design of experiments, our interest is the estimation of some treatment contrasts. Estimation of these contrasts may be severely affected by masked outliers, if any. The procedure based on $\mathbf{t}_i$ may not be able to reveal the fact whether masked outliers has affected the estimation of treatment contrasts or not, because $\mathbf{t}_i$ is based on whole design matrix $\mathbf{X}$, therefore, the effect of one outlier may be compensated by the effect of another outlier when we estimate some subset of parameters. We, therefore, develop similar procedure keeping in mind that our main objective is the estimation of some treatment contrasts.

We now consider the same linear model (2.1) for an experimental design $d$ (say). The rank of the design matrix $\mathbf{X}$ is now $m(< p)$. Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1' \quad \boldsymbol{\theta}_2')'$ , where $\boldsymbol{\theta}_1$ is a $v$-component vector containing all parameters of interest to the experimenter (say treatment effects) and $\boldsymbol{\theta}_2$ is (*p*-*v*) component vector containing the set of nuisance parameters in the model which are not of much interest to the experimenter.

Thus $$\mathbf{y} = (\mathbf{X}_1 \quad \mathbf{X}_2)\begin{pmatrix}\boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2\end{pmatrix} + \boldsymbol{\varepsilon}, \tag{2.2}$$

where $\mathbf{X}$ is partitioned in conformity with the parameters, $\mathbf{X}_1$ is an $n\times v$ matrix of rank *v* and $\mathbf{X}_2$ is an $n\times$ $(p-v)$ matrix such that $\mathbf{X} = [\mathbf{X}_1 \quad \mathbf{X}_2]$. The normal equations obtained by least squares method for estimating the parameters are given by

$$\mathbf{X}'\mathbf{X}\boldsymbol{\theta} = \mathbf{X}\mathbf{y}$$

From these equations on eliminating $\boldsymbol{\theta}_2$, we obtain the reduced normal equations involving only $\boldsymbol{\theta}_1$ as $\mathbf{C}_{\boldsymbol{\theta}_1}\boldsymbol{\theta}_1 = \mathbf{Q}_{\boldsymbol{\theta}_1}$, (2.3)

where $$\mathbf{C}_{\boldsymbol{\theta}_1} = \mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{+}\mathbf{X}_2'\mathbf{X}_1 = \mathbf{X}_1'\mathbf{B}\mathbf{X}_1,$$

$$\mathbf{Q}_{\boldsymbol{\theta}_1} = \mathbf{X}_1'\mathbf{y} - \mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{+}\mathbf{X}_2'\mathbf{y} = \mathbf{X}_1'\mathbf{B}\mathbf{y},$$

$$\mathbf{B} = \mathbf{I}_n - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{+}\mathbf{X}_2' \quad , \tag{2.4}$$

and $\mathbf{A}^{+}$ is the Moore-Penrose inverse of **A**. The matrix **B** is symmetric and idempotent.

We assume that the design *d* considered here is connected, *i.e.*, all (*v*–1) orthonomalized contrasts for the parameters $\boldsymbol{\theta}_1$ are estimable or equivalently Rank ($\mathbf{C}_{\boldsymbol{\theta}_1}$) = *v*–1, and let the set of all (*v*–1) orthonormalized contrasts for the parameters $\boldsymbol{\theta}_1$ be given by $\mathbf{P}\boldsymbol{\theta}_1$. The (*v*–1)×*v* matrix **P** is such that $\mathbf{P}\mathbf{P}' = \mathbf{I}_{v-1}$, $\mathbf{P}'\mathbf{P} = \mathbf{I}_v - \frac{1}{v}\mathbf{J}_v$ and $\mathbf{I}_n$ denotes an identity matrix of order *n* and $\mathbf{J}_n$ denotes an $n\times n$ matrix whose all elements are ones.

Let *k* observations be suspected of being outliers in the sense that their expected values are shifted from the expected values of other observations. Assuming that the design *d* remains connected after deletion of any *k* observations, Bhar and Gupta (2001) have shown that the difference between the estimators of contrasts of $\mathbf{P}\boldsymbol{\theta}_1$ and $\mathbf{P}\boldsymbol{\theta}_{1(k)}$ can be expressed as

$$\mathbf{P}(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_{1(k)}) = \mathbf{P}\mathbf{C}^{+}_{\theta_1}\mathbf{X}_1'\mathbf{V}\mathbf{U}(\mathbf{U}'\mathbf{V}\mathbf{U})^{-1}\mathbf{U}'\mathbf{V}\mathbf{y}, \tag{2.5}$$

where $\mathbf{P}\hat{\boldsymbol{\theta}}_1$ is the least squares estimator of $\mathbf{P}\boldsymbol{\theta}_1$, $\hat{\boldsymbol{\theta}}_1$ is any solution of the normal equations (2.3), $\mathbf{P}\hat{\boldsymbol{\theta}}_{1(k)}$ is the least squares estimator of $\mathbf{P}\boldsymbol{\theta}_{1(k)}$ obtained after deleting the suspected $k$ outlying observations, $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k)$, $\mathbf{u}_i = (0, 0, \ldots, 1(i^{th}), \ldots, 0, 0)$ and $\mathbf{V} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{+}\mathbf{X}'$.

Thus the effect of deleting a single data point (say $i^{\text{th}}$ ) on the set of treatment contrasts can be obtained from (2.5) as

$$\mathbf{f}_i = \mathbf{P}(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_{1(i)}) = \mathbf{P}\mathbf{C}_{\theta_1}^{+}\mathbf{X}_1'\mathbf{B}\mathbf{u}_i(\mathbf{u}_i'\mathbf{V}\mathbf{u}_i)^{-1}\mathbf{u}_i'\mathbf{V}\mathbf{y}\,. \tag{2.6}$$

The ($v$–1) component vector $\mathbf{f}_i$ summarizes the effect on the set of treatment contrasts of deleting the observation $i$. The individual deletion statistics identify influential points as those with large values of $\mathbf{f}_i$ in some suitable norm. For instance, Cook statistic for a set of treatment contrasts for the $i^{th}$ observation is given by $\frac{1}{(v-1)\hat{\sigma}^2}\mathbf{f}_i'(\mathbf{P}\mathbf{C}_{\boldsymbol{\theta}_1}\mathbf{P}')\mathbf{f}_i$. However, when masking is present, the $\mathbf{f}_i$ values corresponding to outliers tend to be small, and therefore they are not detected. Now applying the same logic of Pena and Yohai (1995) for detecting possible sets of influential observations having similar or opposite effects on the fit, we look at the uncentred covariance matrix of $\mathbf{f}_i$. Let us call $\mathbf{F}$ the ($v$–1)×$n$ matrix $\mathbf{F} = (\mathbf{f}_1 \ \ \ldots \ \ \mathbf{f}_n)$ whose columns are the vectors $\mathbf{f}_i$. Then we define the $n$×$n$ influence matrix $\mathbf{M}$ as

$$\mathbf{M} = \frac{1}{(v-1)\hat{\sigma}^2}\mathbf{F}'(\mathbf{P}\mathbf{C}_{\boldsymbol{\theta}_1}\mathbf{P}')\mathbf{F} \tag{2.7}$$

After doing some algebra it can easily be shown that the $ij^{\text{th}}$ element of $\mathbf{M}$ is

$$m_{ij} = \frac{r_i r_j l_{ij}}{(1-h_{ii})(1-h_{jj})(v-1)\hat{\sigma}^2}\,, \tag{2.8}$$

where $r_t$ is the $t^{\text{th}}$ residual, $h_{ij}$ is the $ij^{\text{th}}$ element of the matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{+}\mathbf{X}'$ and $l_{ij}$ is the $ij^{\text{th}}$ element of the matrix $\mathbf{B}\mathbf{X}_1\mathbf{C}_{\theta_1}^{+}\mathbf{X}_1'\mathbf{B}$.

Let $e_{ij}$ be the uncentred correlation coefficient between $\mathbf{f}_i$ and $\mathbf{f}_j$. This actually measures the effects on the least square fit of the $i^{\text{th}}$ and $j^{\text{th}}$ points. Then

$$e_{ij} = \frac{m_{ij}}{m_{ii}^{1/2} m_{jj}^{1/2}}\,. \tag{2.9}$$

Following Pena and Yohai(1995), if there are groups of outliers, then $|e_{ij}|$ would be near one, if the observations $i$ and $j$ belong to some group. This correlation coefficient would be positive if the two observations have positive effects and would be negative if they have opposite effects. In other situations this correlation coefficient would be near zero. In the ideal situation, *i.e.*, when these correlation coefficients are either exactly one or zero, the eigenvectors of **M** would be orthogonal and the elements of these vector would be either $m_{ij}^{1/2}$ , $-m_{ij}^{1/2}$ or 0, depending upon whether two observations belong to the same group or not and whether they have similar or opposite effects. The eigenvalues can then be calculated as

$$\beta_h = \sum_{i \in h} m_{ii} ,$$

where *h* denotes a group in which outliers are present and summation is taken over all values belonging to this group. In the extreme case the situation is like this as described above. For real data sets, this may not happen exactly. However, the masking effect is typically due to the presence in the sample of blocks of influential observations having similar or opposite effects. These blocks are likely to produce a matrix **M** with a structure close to that described above. Two influential observations *i* and *j* producing similar effects should have $e_{ij}$ close to 1, and close to −1 when they have opposite effects. Influential observations with non-correlated effects have $\left| e_{ij} \right|$ close to 0. The same will happen with non-influential observations. Therefore, the eigenvectors will have approximately the structure described above. This suggests the following procedure to identify influential sets:

(a) Find the eigenvectors corresponding to the non-null eigenvalues of the matrix **M.**

(b) Consider the eigenvectors corresponding to large eigenvalues, and define two sets according to the components with large positive and negative values of the eigenvectors respectively.

***Procedure for detecting influential sets***

To identify influential sets, we need to look at the eigenvectors corresponding to the large non-zero eigenvalues of the influence matrix **M**. Different influential subsets may have different eigenvectors. To find influential subsets we look at all eigenvalues corresponding to non-zero eigenvalues of **M**. In each eigenvector we search for the sets of coordinates with relatively large value and same sign. To compare the relative value of an element, we adopt the strategy adopted by Pena and Yohai (1995). Pena and Yohai (1995) suggested in case of regression analysis to look at the ratio between the components in decreasing order, searching for a clear cut-off point, to form a set of candidate outliers, and then to test the points in this set to identify the outliers.

A set of candidate outliers is obtained by analyzing the eigenvectors corresponding to the non-null eigenvalues of the influence matrix **M**, and by searching in each eigenvector for a set of co-ordinates with relatively large value and the same sign. The search is done in the following way.

(a) Suppose corresponding to a large eigenvalue, $\mathbf{v}$ denotes the eigenvector. Order the co-ordinates of the eigenvector $\mathbf{v}$, obtaining $v_{(1)} \leq v_{(2)} \leq ... \leq v_{(n)}$.

(b) Compute the ratios $a_j = \frac{v_{(j)}}{v_{(j-1)}}$ for $j = n, . . .,\ n - c_1$ and $b_j = \frac{v_{(j)}}{v_{(j+1)}}$ for $j = 1, . . . ,\ c_2$. The constants $c_1$ and $c_2$ are smaller than $n/2$ and are determined on the basis of the coordinates.

(c) Look for the first $j_0$ such that $\left|a_j\right| > \delta$ and first $i_0$ such that $\left|b_j\right| > \delta$

(d) If $i_0$>1 and/or $j_0$>1, consider the set $\mathbf{j}_0 = (i_{(n)}, i_{(n-1)}, ..., i_{(n-i_0+1)})$ and/or $\mathbf{i}_0 = (i_{(1)}, i_{(2)}, ..., i_{(j_0-1)})$ as candidate outlier.

In regression analysis choice of $c_1$ and $c_2$ is generally taken on the basis of break down point. However, in designed experiments, since, experiments are controlled there could be very less number of outliers. Therefore, choice of $c_1$ and $c_2$ could be made up to 4 or 5 depending on the values of coordinates. Choice of $\delta$ is again arbitrary. In case of regression analysis Pena and Yohai (1995) suggested this value to be 2.1 through a simulation study. In case of designed experiments we took this value in the neighborhood of 1.5.

For checking the outlyingness, we remove the candidate outliers from the data and reanalyze the data to see whether any drastic change has occurred in the inference. Since we are only interested in the eigenvectors corresponding to the non-null eigenvalues, the direct computation of the eigenvalues and eigenvectors of **M** can be obtained by using spectral decomposition of the matrix **M**. However we may directly calculate the eigenvalues and eigenvectors using SAS/IML software.

## 3 Illustration

To illustrate the procedure, we applied this method to experimental data obtained from Agricultural Field Experiments Information System (AFIES), Indian Agricultural Statistics Research Institute, New Delhi. It is observed that in some experiments some observations are not influential individually, but jointly with some other observations, they are influential. These observations are actually masked by some other outlying observations and, therefore, could not be detected when diagnostic test procedure for detecting a single outlier is applied. To make the exposition clear consider the following example.

An experiment with 10 treatments was conducted in a randomized complete block (RCB) design with 4 replications at Sugarcane Research Institute, Shahjahanpur, Uttar Pradesh, India to find out the suitable herbicide to control weeds in Sugarcane (net plot size: 8.00m × 5.40m.). The treatment details are

$T_0$ = Control weeded check

$T_1$ = Local conventional method

$T_2$ = Trash mulching

$T_3$ = 1.0 kg active ingredient (a.i.)/hectare of 2,4-D sodium salt and 0.50 kg a.i./ hectare of gramoxone at 3 weeks of planting followed by application of the same at 6-8 weeks of planting.

$T_4$ = 2.0 kg a.i./ hectare of Atrazine as Pre-emergence spray

$T_5$ = 1.00 kg a.i./ hectare of 2,4-D Sodium Salt at 8-10 weeks after planting

$T_6$ = 2.0 kg a.i./ hectare of 2,4-D (Amine) as Pre-emergence spray followed by spray of the same at 8-10 weeks after planting.

$T_7$ = 2.0 kg a.i./ hectare of Atrazine as Pre-emergence spray followed by spray of Glyphosate at 1.0 kg a.i./ha at 6-8 weeks after planting.

$T_8$ = 1.00 kg a.i./ hectare of Arochlor and 1.00 kg a.i./ha of Atrazine as pre-emergence spray

$T_9$ = 2.00 kg a.i./ hectare of Arochlor as pre-emergence spray

The table below shows the data on yield per plot in quintal(q) for different treatments:

Table 1: Yield of sugar cane in q/plot

| Replication | Treatment | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 2.52 | 2.82 | 2.42 | 2.67 | 2.50 | 3.01 | 2.65 | 2.62 | 2.18 | 2.57 |
| 2 | 2.77 | 2.77 | 2.52 | 3.69 | 3.21 | 3.05 | 2.64 | 2.53 | 2.47 | 2.82 |
| 3 | 2.32 | 2.38 | 2.44 | 2.30 | 1.90 | 2.46 | 2.35 | 2.47 | 2.15 | 2.26 |
| 4 | 2.31 | 2.14 | 2.38 | 2.13 | 2.51 | 2.79 | 2.21 | 2.52 | 2.66 | 2.35 |

Analysis of this data is presented in Table 2. The treatment effects are not significantly different at 5% level of significance. Cook-statistic (Bhar and Gupta, 2001) for each observation is computed and values are given in Table 3. It is observed from Table 3 that the observation number 14 stands out. We tested it with *F* value (Probability value of 0.3823402 with 9 and 27 degrees of freedom is 0.066) and found that this observation is statistically influential. No other observation is found to be influential.

This observation is deleted and analysis is carried out again. The result is presented in Table 4. From the table, it is observed that not much change has occurred in the inference. Though the significance level of treatment effects has been lowered, yet it remains not significant at 5% level of significance. Thus the observation number 14, in spite of being an outlier, does not have much influence in the analysis. However, there might be groups of outliers that cannot be detected by using single outlier detection technique of Cook-statistic.

We then applied our method to identify the group of observations that are influential. The largest positive eigenvalue of **M** is 0.03345, the other eigenvalues are very small (<<0.005). Since the eigenvalue of M is likely to be large if there are some groups of influential observations, we ignore the other eigenvalues. The coefficients of the eigenvector of this eigenvalue are given in Table 5.

From Table 5 we find that the first two eigenvectors give high positive weights to observations in the set {14, 39}, especially to observation number 14. The coefficients of this eigenvector are arranged in descending order in Table 5. This Table also summarizes the results when we apply the procedure for detecting outliers. We have chosen $c_1 = c_2 = 10$ and $\delta$ =1.5. The first $a_j$ value is 0.50361 corresponding to observation number 14 and the next highest value of $a_j$ is 0.27874 corresponding to observation number 39. Thus the first set of observations for which $a_j$ exceeds 1.50 {14, 39} as relevant candidate outliers. The other values of $a_j$ and $b_j$ are smaller than 1.5. Thus these two observations are likely to be influential. The data was reanalyzed after deleting these two observations. The result is presented in Table 6. The dramatic effect to note here is that the level of significance of treatment effects has been reduced to 0.0519. Thus treatment effects are now significant at almost 5% level of significance. Thus these two observations are influential. The interesting point to note here is that though the observation number 14 was detected as outlier when we applied Cook-statistic for detecting a single outlier, yet observation number 39 was not. Its effect was masked by the observation number 14. Removal of any other pair of observations does not have any effect on the analysis.

Another point to note that the choice of $\delta$ is arbitrary. Taking this value larger means that we more stringent in identifying candidate outliers. If we lower this value, some more observations are likely to be candidates. However, their influence should be assessed after deleting these observations. For example, if we choose $\delta$ =1.2, then the set {21, 25, 34} is likely to be influential on the basis of value of $b_j$. However, when we test for their influence, we find that removal of these three points or two at a time does not alter the main results of the analysis.

Table 2: ANOVA (With original data)

| Source | Degrees of freedom | Sum of Squares | Mean square | F Value | Significan ce Level |
|---|---|---|---|---|---|
| Replication | 3 | 1.731 | 0.577 | 8.64 | 0.0003 |
| Treatment | 9 | 0.637 | 0.070 | 1.06 | 0.4206 |
| Error | 27 | 1.802 | 0.066 | | |
| Total | 39 | 4.171 | | | |

Table 3: Cook-statistics

| Serial No. | Replication | Treatment | Cook-Statistics | Serial No. | Replication | Treatment | Cook-Statistics |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0.000312 | 21 | 3 | 1 | 0.004440 |
| 2 | 1 | 2 | 0.044626 | 22 | 3 | 2 | 0.0060796 |
| 3 | 1 | 3 | 0.0051954 | 23 | 3 | 3 | 0.0448183 |
| 4 | 1 | 4 | 0.0062219 | 24 | 3 | 4 | 0.022109 |
| 5 | 1 | 5 | 0.0065846 | 25 | 3 | 5 | 0.1292313 |
| 6 | 1 | 6 | 0.0124363 | 26 | 3 | 6 | 0.0147602 |
| 7 | 1 | 7 | 0.0134679 | 27 | 3 | 7 | 0.0120352 |
| 8 | 1 | 8 | 0.0005345 | 28 | 3 | 8 | 0.0233389 |
| 9 | 1 | 9 | 0.0491404 | 29 | 3 | 9 | 0.0002813 |
| 10 | 1 | 10 | 0.0000906 | 30 | 3 | 10 | 0.0000347 |
| 11 | 2 | 1 | 0.0003455 | 31 | 4 | 1 | 0.0009225 |
| 12 | 2 | 2 | 0.003801 | 32 | 4 | 2 | 0.0517879 |
| 13 | 2 | 3 | 0.043674 | 33 | 4 | 3 | 0.0048107 |
| 14 | 2 | 4 | **0.3823402** | 34 | 4 | 4 | 0.1526988 |
| 15 | 2 | 5 | 0.1122303 | 35 | 4 | 5 | 0.0111566 |
| 16 | 2 | 6 | 0.0063657 | 36 | 4 | 6 | 0.0080566 |
| 17 | 2 | 7 | 0.0145407 | 37 | 4 | 7 | 0.0110611 |
| 18 | 2 | 8 | 0.0818239 | 38 | 4 | 8 | 0.0121348 |
| 19 | 2 | 9 | 0.034714 | 39 | 4 | 9 | 0.1530533 |
| 20 | 2 | 10 | 0.0000742 | 40 | 4 | 10 | 0.0001498 |

Table 4: ANOVA (After deleting observation No. 14)

| Source | Degrees of freedom | Sum of Squares | Mean square | F Value | Significanc e Level |
|---|---|---|---|---|---|
| Replication | 3 | 1.106 | 0.368 | 8.62 | 0.0004 |
| Treatment | 9 | 0.537 | 0.059 | 1.39 | 0.2411 |
| Error | 26 | 1.113 | 0.042 | | |
| Total | 38 | 2.806 | | | |

Table 5: Eigenvalue Coefficients

| Observati on N0. | Coefficients of eigenvalue | Observati on N0. | Ordered Coefficients of eigenvalue | $a_j$ | $b_j$ |
|---|---|---|---|---|---|
| 1 | 0.106274 | **14** | **0.50361** | **1.806736** | |
| 2 | 0.136534 | **39** | **0.27874** | **1.615097** | |
| 3 | -0.04642 | **15** | **0.172584** | **0.945285** | |
| 4 | -0.06424 | 31 | 0.182574 | 1.337201 | |
| 5 | -0.05634 | 2 | 0.136534 | 1.001373 | |
| 6 | 0.07003 | 23 | 0.136347 | 1.220389 | |
| 7 | 0.073182 | 11 | 0.111724 | 1.051282 | |
| 8 | 0.015021 | 1 | 0.106274 | 1.070714 | |
| 9 | -0.15228 | 28 | 0.099256 | 1.353514 | |
| 10 | 0.005872 | 35 | 0.073332 | 1.00205 | |
| 11 | 0.111724 | 7 | 0.073182 | | |
| 12 | -0.03985 | 38 | 0.07157 | | |
| 13 | -0.1346 | 6 | 0.07003 | | |
| 14 | 0.50361 | 27 | 0.06918 | | |
| 15 | 0.172584 | 36 | 0.056366 | | |
| 16 | -0.0501 | 22 | 0.050395 | | |
| 17 | -0.07604 | 33 | 0.044671 | | |
| 18 | -0.18585 | 8 | 0.015021 | | |
| 19 | -0.12799 | 29 | 0.011522 | | |
| 20 | 0.005313 | 10 | 0.005872 | | |
| 21 | -0.40057 | 20 | 0.005313 | | |
| 22 | 0.050395 | 30 | -0.00364 | | |
| 23 | 0.136347 | 40 | -0.00755 | | |
| 24 | -0.1211 | 12 | -0.03985 | | |

| Observation N0. | Coefficients of eigenvalue | Observation N0. | Ordered Coefficients of eigenvalue | $a_j$ | $b_j$ |
|---|---|---|---|---|---|
| 25 | -0.24958 | 3 | -0.04642 | | |
| 26 | -0.07629 | 16 | -0.0501 | | |
| 27 | 0.06918 | 5 | -0.05634 | | |
| 28 | 0.099256 | 4 | -0.06424 | | |
| 29 | 0.011522 | 37 | -0.06632 | | |
| 30 | -0.00364 | 17 | -0.07604 | | |
| 31 | 0.182574 | 26 | -0.07629 | | 1.146545 |
| 32 | -0.14708 | 24 | -0.1211 | | 1.003327 |
| 33 | 0.044671 | 19 | -0.12799 | | 1.587341 |
| 34 | -0.31826 | 13 | -0.1346 | | 1.056836 |
| 35 | 0.073332 | 32 | -0.14708 | | 1.051638 |
| 36 | 0.056366 | 9 | -0.15228 | | 1.092775 |
| 37 | -0.06632 | 18 | -0.18585 | | 1.035314 |
| 38 | 0.07157 | 25 | -0.24958 | | 1.220455 |
| 39 | 0.27874 | 34 | -0.31826 | | 1.34294 |
| 40 | -0.00755 | 21 | -0.40057 | | 1.275198 |

Table 6: ANOVA (With 2 data points deleted)

| Source | Degrees of freedom | Sum of Squares | Mean sum of squares | F Value | Significance Level |
|---|---|---|---|---|---|
| Replication | 3 | 1.207 | 0.402 | 11.58 | <.0001 |
| Treatment | 9 | 0.706 | 0.078 | 2.26 | 0.0519 |
| Error | 25 | 0.868 | 0.034 | | |
| Total | 37 | 2.782 | | | |

Conclusions drawn from an experiment may be misleading due to presence of some outliers as we have demonstrated through this example. Therefore, detection of outliers in the experimental data is very important. Detection of outliers is very difficult if masking is present in the data. Several methods have been developed for dealing with outliers in presence of masking in linear regression analysis. However, no work in designed experiments seems to be available. We have attempted to apply one such method to designed experiments. But there is a lot of scope to develop other methods as well. In the present study, we assumed that masking occurs due to proportional effects of two outliers. These proportional effects may be in positive direction or in negative direction. However, there may be many other reasons, due to which outliers may occur. These need to be investigated thoroughly.

## References

Atkinson, A. C. (1986). Masking unmasked. *Biometrika*, **73**, 533-541.

Bhar, L. and Gupta, V.K. (2001). A useful statistic for studying outliers in experimental designs. *Sankhyā* , **B63**, 338-350.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics,* **19(1)**, 15-18.

Cook, R. D. (1979). Influential observations in linear regression. *J. Amer. Statist. Assoc*., **74**, 169 – 174.

Gray, J. B. and Ling, R. F. (1984). K-clustering as a detection tool for influential subsets in regression. *Technometrics*, **26**, 197-208.

Hawkins, D. M.(1980). *Identification of outliers*. Chapman and Hall, London.

Hocking, R. R. (1984). Duiscussion of Gray and Ling Paper. *Technometrics*, **26**, 321-323.

Kianifard, F. and Swallow, W. (1990). A Monte Carlo comparison of five procedures for identifying outliers in linear regression. *Comm. Statist. Theory Methods*, **19**, 1913-1938.

Dr. L. M. Bhar
Indian Agricultural Statistics Research Institute
Library Avenue, PUSA
New Delhi-110012
E-mail: lmbhar@iasri.res.in

Dr. V. K. Gupta
Indian Agricultural Statistics Research Institute
Library Avenue, PUSA
New Delhi-110012
E-mail: vkgupta@iasri.res.in

Dr. Rajender Parsad
Indian Agricultural Statistics Research Institute
Library Avenue, PUSA
New Delhi-110012
E-mail: rajender@iasri.res.in