Statistics and Applications
Volume 6, Nos.1 & 2, 2008 (New Series), pp. 275-292

Outliers in multi-response experiments[1]

Rajender Parsad, P. K. Nandi, L. M. Bhar and V. K. Gupta
IASRI, Library Avenue, New Delhi 110012

Abstract

The purpose of this article is to develop a test statistic for detection of a single outlying observation vector in multi-response experiments conducted in a block design set up. The test statistic developed is a multivariate extension of the Cook statistic for detection of a single outlier in the usual block design set up for uni-response experiments. The use of the proposed test statistic has been illustrated with an example.

Key words: Multi-response experiments; Block design; Outlier; Cook-statistic.

1 Introduction

An outlier in a set of data is an observation (or an observation vector) that appears to be inconsistent with the remainder of the observations in that data set. Occurrence of outlier(s) is common in every field in which data collection is involved. In many experimental situations, data on more than one response variable is recorded from the same experimental unit through application of same treatment. Such experiments are known as multi-response experiments. Outlier(s) in multi-response experiments is/are likely to appear. If an experimental plot is heavily infested with pests, disease and/or weeds then all the responses observed from that plot may be outlier(s). Outlier(s) may also occur because of heavy irrigation on some experimental plot(s) by mistake. Outlier(s) could very well be due to transcription errors.

[1]The authors of this article are three generations of students of Professor Aloke Dey. We feel privileged to write this article in this special issue being brought out to felicitate him on superannuation. His contributions to the theory of statistics particularly to Design of Experiments have been monumental.

The presence of outlier(s) in the data generated from multi-response experiments may cause departures from the assumptions of parameter estimation. The analysis of data in presence of outlier(s) may give misleading results. Therefore, it becomes pertinent to detect outlier before analyzing the data from these experiments. Barrett and Ling (1992) proposed a measure of influence for multivariate regression as an extension of measure given by Cook and Weisberg (1980) for univariate regression. Test statistics, available in the literature for detecting outlier(s) in multivariate regression cannot be applied directly to the multi-response experimental settings because

i) design matrix of multi-response experiments is not of full column rank as in multivariate regression.

ii) In multi-response experiments, interest is in a sub set of parameters (linear function of treatment effects) rather than the complete set of parameters, as in multivariate regression.

Most of the literature available for detection of outlier(s) in the experimental data and obtaining robust experimental designs in presence of outlier(s) is for single response situations [see e.g. Box and Draper (1975), Gopalan and Dey (1976), John (1978), Ghosh (1983, 1989), Singh et al. (1987), Ben and Yohai (1992), Bhar (1997), Bhar and Gupta (2001, 2003), Sarker (2002) and Sarker et al. (2003, 2005)]. John (1978) studied the problems that arise in detecting the presence of outliers in the results from factorial experiments by applying the Q_k-statistic of Gentleman and Wilk (1975). Ben and Yohai (1992) studied the asymptotic theory of M-estimates and their associated test for a single-factor experiment in a randomized complete block (RCB) design. Gopalan and Dey (1976) studied the robustness of general block designs in the presence of a single outlier by minimizing the variance of discrepancy or bias in the measurement of error variance (σ^2). Singh et al. (1987) showed that the variance balanced row-column designs satisfying the property of adjusted orthogonality are robust against the presence of a single outlier. Bhar (1997) investigated the problem of outlier(s) in the experimental data for the block designs and modified the Cook-statistic, Q_k-statistic and AP-statistic for detection of single outlier in experimental data for both mean shift and variance inflation models. Bhar and Gupta (2001) studied the robustness of block designs by minimizing the value of

Cook-statistic. Bhar and Gupta (2003) made a study of outliers under variance-inflation model in experimental designs. Sarker et al. (2003) extended these results to the experimental situations where the interest of the experimenter is only in a subset of all possible elementary treatment contrasts (test treatments-control treatment comparisons) rather than the complete set of all the possible elementary treatment contrasts. Sarker et al. (2005) formulated a test statistic for detection of a single outlier in block designs for diallel crosses. They also established a correspondence between two existing criteria of robustness i.e. minimization of average Cook-statistic and minimization of variance of discrepancy or bias in estimation of error variance. It has been shown that a proper binary balanced block design for diallel crosses is robust against the presence of a single outlier. Block designs for diallel crosses in which every line appears an equal number of times in each block are also found to be robust against the presence of a single outlier.

In multi-response experiments, for taking the advantage of correlation structure among the response variables, multivariate analysis of variance (MANOVA) of data should be performed for testing the equality of treatment effects. The inference(s) drawn from MANOVA may be misleading if outlier(s) are present in the data. Gananadesikan and Lee (1970), Gnanadesikan and Kettenring (1972) and Kang and Bates (1990) investigated the problems of outlier(s) in multi-response data mostly in regression analysis situations. Very little work seems to have been done on detection of outlier(s) in data from multi-response experiments. Therefore, in the present investigation an attempt has been made to develop a test statistic for detection of an outlier from multi-response data generated through block design. The test statistic is given in Section 3. We begin with some preliminaries in Section 2.

2 Preliminaries

Let there be v treatments laid out in a block design containing b blocks such that j^{th} block contains k_j experimental units; $j = 1, 2, \ldots, b$ and treatment i is replicated r_i times, $\sum_{j=1}^{b} k_j = \sum_{i=1}^{v} r_i = n$, the total number of experimental units. From each experimental unit p responses are

observed. Let $\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{y}_1 & \boldsymbol{y}_2 & \dots & \boldsymbol{y}_p \end{bmatrix}$ be an $n \times p$ matrix of observations, where $\boldsymbol{y}$ is an $n \times 1$ vector of observations corresponding to the s^{th} response $(s = 1, 2, \dots, p)$.

For the s^{th} response, the model is given by

$$\boldsymbol{y}_s = \boldsymbol{X}\boldsymbol{\theta}_s + \boldsymbol{\varepsilon}_s, \qquad s = 1, 2, \dots, p \tag{1}$$

where $\boldsymbol{X} = \lfloor \boldsymbol{\Delta}' \quad \mathbf{1} \quad \boldsymbol{D}' \rfloor$ is the design matrix for s^{th} response partitioned in conformity with the parameters, $\boldsymbol{\Delta}'$ is $(n \times v)$ design matrix of treatments, $\mathbf{1}$ is the n dimensional column vector of all elements unity and $\boldsymbol{D}'$ is the design matrix of blocks.

$\boldsymbol{\theta} = [\boldsymbol{\tau}_s' \quad \boldsymbol{\mu}_s \quad \boldsymbol{\beta}_s']$ is a $(v + b + 1)$ component vector, $\boldsymbol{\tau}_s$ being v-component vector of treatment effects, μ_s the general mean and $\boldsymbol{\beta}_s$ the b-component vector of block effects for the s^{th} response. $\boldsymbol{\varepsilon}_s$ is the residual vector for s^{th} response variable distributed as $N(0, \sigma_{ss}\boldsymbol{I}_n)$.

So the model for multi-response experiments in block design set up is

$$\boldsymbol{Y} = \boldsymbol{Z}\theta + \boldsymbol{\varepsilon} \tag{2}$$

where $\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{y}_1' & \boldsymbol{y}_2' & \dots & \boldsymbol{y}_p' \end{pmatrix}'$.

Now we can roll out the matrix $\boldsymbol{Z}$ as

$$\boldsymbol{Z} = \lfloor \boldsymbol{I}_p \otimes \boldsymbol{\Delta}' \quad \boldsymbol{I}_p \otimes \mathbf{1} \quad \boldsymbol{I}_p \otimes \boldsymbol{D}' \rfloor = \boldsymbol{I}_p \otimes \boldsymbol{X} \text{ and } \boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\tau} \\ \boldsymbol{\mu} \\ \boldsymbol{\beta} \end{bmatrix}, \tag{3}$$

where $\boldsymbol{\tau} = \begin{bmatrix} \boldsymbol{\tau}_1', \boldsymbol{\tau}_2', \dots, \boldsymbol{\tau}_p' \end{bmatrix}'$, $\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \dots, \boldsymbol{\beta}_p' \end{bmatrix}$ and $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_p)'$. We also assume that $\boldsymbol{\varepsilon} \sim N_p(0, \boldsymbol{\Omega})$. where

$$\boldsymbol{\Omega} = D(\boldsymbol{\varepsilon}) \begin{bmatrix} \sigma_{11}\boldsymbol{I}_n & \sigma_{12}\boldsymbol{I}_n & \cdots & \sigma_{1p}\boldsymbol{I}_n \\ \sigma_{21}\boldsymbol{I}_n & \sigma_{22}\boldsymbol{I}_n & \cdots & \sigma_{2p}\boldsymbol{I}_n \\ \vdots & \vdots & & \vdots \\ \sigma_{p1}\boldsymbol{I}_n & \sigma_{p2}\boldsymbol{I}_n & \cdots & \sigma_{pp}\boldsymbol{I}_n \end{bmatrix} = \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n. \tag{4}$$

Here $\otimes$ denotes Kronecker product of matrices and $D(\cdot)$ denotes the dispersion matrix. Further $\boldsymbol{\Sigma} = (\sigma_{ij})$. Using the Generalized Least Square (GLS) estimation procedure, the normal equations are

$$\left(\boldsymbol{Z}'\boldsymbol{\Omega}^{-1}\boldsymbol{Z}\right)\boldsymbol{\theta} = \boldsymbol{Z}'\boldsymbol{\Omega}^{-1}\boldsymbol{Y}. \tag{5}$$

The reduced normal equations for estimating the linear functions of treatment effects are

$$\boldsymbol{C}^*\boldsymbol{\tau} = \boldsymbol{Q}^* \tag{6}$$

where

$$\boldsymbol{C}^* = \boldsymbol{\Sigma}^{-1} \otimes \left(\boldsymbol{\Delta}\boldsymbol{\Delta}' - \boldsymbol{\Delta}\boldsymbol{D}'(\boldsymbol{D}\boldsymbol{D}')^{-1}\boldsymbol{D}\boldsymbol{\Delta}'\right) = \boldsymbol{\Sigma}^{-1} \otimes (\boldsymbol{\Delta}\boldsymbol{S}\boldsymbol{\Delta}') = \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{C} \tag{7}$$

$$\boldsymbol{Q}^* = \left[\boldsymbol{\Sigma}^{-1} \otimes \left(\boldsymbol{\Delta} - \boldsymbol{\Delta}\boldsymbol{D}'(\boldsymbol{D}\boldsymbol{D}')^{-1}\right)^{-1}\boldsymbol{D}\right]\boldsymbol{Y} = \boldsymbol{\Sigma}^{-1} \otimes (\boldsymbol{\Delta}\boldsymbol{S}\boldsymbol{Y}) = \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{Q} \tag{8}$$

and $\boldsymbol{S} = \boldsymbol{I} - \boldsymbol{D}'(\boldsymbol{D}\boldsymbol{D}')^{-1}\boldsymbol{D}$.

Here $\boldsymbol{C}$ is the information matrix and $\boldsymbol{Q}$ is the vector of adjusted treatment totals in the usual setup for the univariate case. A solution of the reduced normal equations in (6) is

$$\widehat{\boldsymbol{\tau}} = \boldsymbol{C}^{*-}\boldsymbol{Q}^*. \tag{9}$$

The following theorem can be given for multi-response experiments.

Theorem 2.1

(i) $E(\boldsymbol{Q}^*) = \boldsymbol{C}^*\boldsymbol{\tau}$

(ii) $D(\boldsymbol{Q}^*) = \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{C} = \boldsymbol{C}^*$ (10)

(iii) A design for multi-response experiment is connected for parameters $\boldsymbol{\tau}$ *iff* $\text{Rank}(\boldsymbol{C}^*) = p(v-1)$. *In a connected design all contrasts of* $\boldsymbol{\tau}$ *are estimable.*

Here we assume that the design is connected i.e. all $p(v-1)$ orthonormalized treatment contrasts are estimable or equivalently Rank $(\boldsymbol{C}^*) = p(v-1)$. Let the set of all $p(v-1)$ orthonormalized treatment contrasts for the parameters $\boldsymbol{\tau}$ be given by $\boldsymbol{P}\boldsymbol{\tau}$, where $\boldsymbol{P} = \boldsymbol{I}_p \otimes \boldsymbol{L}$ and $\boldsymbol{L}$ is such that $\boldsymbol{L}\boldsymbol{L}' = \boldsymbol{I}_{v-1}$ and $\boldsymbol{L}'\boldsymbol{L} = \boldsymbol{I}_v - \frac{1}{v}\boldsymbol{1}\boldsymbol{1}'$, $\boldsymbol{P}\boldsymbol{P}' = \boldsymbol{I}_p \otimes \boldsymbol{L}\boldsymbol{L}' = \boldsymbol{I}_p \otimes \boldsymbol{I}_{v-1}$ and $\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{I}'\boldsymbol{P} = \boldsymbol{I}_p \otimes \boldsymbol{L}'\boldsymbol{L} = \boldsymbol{I}_p \otimes \left(\boldsymbol{I}_v - \frac{1}{v}\boldsymbol{1}_v\boldsymbol{1}_v'\right)$. The Best Linear Unbiased Estimator (BLUE) of $\boldsymbol{P}\boldsymbol{\tau}$ is given by $\boldsymbol{P}\widehat{\boldsymbol{\tau}}$, where $\widehat{\boldsymbol{\tau}}$ is any solution of the reduced normal equations in (6).

We have the following lemma:

Lemma 2.1 *For a connected design for multi-response experiments, the dispersion matrix of* $\boldsymbol{P}\widehat{\boldsymbol{\tau}}$ *can be written as*

$$D(\boldsymbol{P}\widehat{\boldsymbol{\tau}}) = \boldsymbol{\Sigma} \otimes (\boldsymbol{L}\boldsymbol{C}\boldsymbol{L}')^{-1} = (\boldsymbol{P}\boldsymbol{C}^*\boldsymbol{P}')^{-1}. \tag{11}$$

Proof: We know that the information matrix for estimation of a linear function of treatment effects for multi-response experiments run in a block design is given by $\boldsymbol{C}^* = \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{C}$. Therefore, $\boldsymbol{P}'\boldsymbol{P}\boldsymbol{C}^* = \boldsymbol{\Sigma}^{-1} \otimes \left(\boldsymbol{I}_v - \frac{1}{v}\boldsymbol{I}_v\boldsymbol{I}_v'\right)\boldsymbol{C} = \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{C}$. Also $\boldsymbol{C}^*\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{C}$, so we can write $\boldsymbol{C}^*\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{P}'\boldsymbol{P}\boldsymbol{C}^*$.

Premultiplying $\boldsymbol{P}$ we get,

$$\begin{aligned} \boldsymbol{P}\boldsymbol{C}^*\boldsymbol{P}'\boldsymbol{P} &= \boldsymbol{P}\boldsymbol{P}'\boldsymbol{P}\boldsymbol{C}^* = \boldsymbol{P}\boldsymbol{C}^* \\ \Rightarrow \quad \boldsymbol{P} &= (\boldsymbol{P}\boldsymbol{C}^*\boldsymbol{P}')^{-1}\boldsymbol{P}\boldsymbol{C}^*. \end{aligned}$$

This follows from the fact that $\boldsymbol{P}\boldsymbol{C}^*\boldsymbol{P}' = \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{L}\boldsymbol{C}\boldsymbol{L}'$ and $\boldsymbol{L}\boldsymbol{C}\boldsymbol{L}'$ is positive definite using Lemma 2.1 of Bhar and Gupta (2001). Therefore, $\boldsymbol{P}\boldsymbol{C}^*\boldsymbol{P}'$ is positive definite.

Post multiplying $\widehat{\boldsymbol{\tau}}$ we get,

$$\begin{aligned} \boldsymbol{P}\widehat{\boldsymbol{\tau}} &= (\boldsymbol{P}\boldsymbol{C}^*\boldsymbol{P}')^{-1}\boldsymbol{P}\boldsymbol{C}^*\widehat{\boldsymbol{\tau}} \\ &= (\boldsymbol{P}\boldsymbol{C}^*\boldsymbol{P}')^{-1}\boldsymbol{P}\boldsymbol{C}^*(\boldsymbol{C}^{*-}\boldsymbol{Q}^*). \end{aligned}$$

The dispersion matrix of $\boldsymbol{P}\widehat{\boldsymbol{\tau}}$ is given by

$$\begin{aligned} D(\boldsymbol{P}\widehat{\boldsymbol{\tau}}) &= (\boldsymbol{P}\boldsymbol{C}^*\boldsymbol{P}')^{-1}\boldsymbol{P}\boldsymbol{C}^*\boldsymbol{C}^{*-}\boldsymbol{C}^*\boldsymbol{C}^{*-}\boldsymbol{C}^*\boldsymbol{P}'(\boldsymbol{P}\boldsymbol{C}^*\boldsymbol{P}')^{-1} \\ &= \left[\boldsymbol{P}\boldsymbol{C}^*\boldsymbol{P}'\right]^{-1}. \end{aligned} \tag{12}$$

3 Detection of outlier in multi-response experiments

Let us assume that a single observation vector is suspected to be an outlier in the sense that its expected value is shifted from the expected value of other observations. We consider the mean-shift model of the form,

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\theta} + \boldsymbol{U}\gamma + \boldsymbol{\varepsilon} \tag{13}$$

where $\boldsymbol{U} = (\boldsymbol{I}_p \otimes \boldsymbol{u})$, $\boldsymbol{u} = (0 \ldots 0 \quad 1(t^{\text{th}})0 \ldots 0)'$, if t^{th} observation vector is suspected as an outlier and $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are as given in (2). The dispersion matrix of $\boldsymbol{\varepsilon}$ from (4) is $\boldsymbol{\Omega}^* = D(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma} \otimes \boldsymbol{I}$.

Now making use of $\boldsymbol{Z}$ as given in (2) reduced normal equations for estimating the linear function of treatment effects under model (13) are obtained as

$$\boldsymbol{C}^*_{(t)}\boldsymbol{\tau}_{(t)} = \boldsymbol{Q}^*_{(t)} \tag{14}$$

where

$$\begin{aligned} \boldsymbol{C}^*_{(t)} &= \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Delta S} \left[\boldsymbol{I} - \boldsymbol{u}(\boldsymbol{u}'\boldsymbol{S}\boldsymbol{u})^{-1}\boldsymbol{u}'\right] \boldsymbol{S}\boldsymbol{\Delta}', \\ &= \boldsymbol{\Sigma}^{-1} \otimes [\boldsymbol{C} - \boldsymbol{f}\boldsymbol{f}'] . \qquad (15) \\ \boldsymbol{Q}^*_{(t)} &= \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Delta S} \left[\boldsymbol{I} - \boldsymbol{u}(\boldsymbol{u}'\boldsymbol{S}\boldsymbol{u})^{-1}\boldsymbol{u}'\right] \boldsymbol{S}\boldsymbol{Y}, \\ &= \boldsymbol{\Sigma}^{-1} \otimes \lfloor \boldsymbol{Q} - w^{1/2}\boldsymbol{f}\boldsymbol{u}'\boldsymbol{S}\boldsymbol{Y} \rfloor . \qquad (16) \end{aligned}$$

where $w = (\boldsymbol{u}'\boldsymbol{S}\boldsymbol{u})^{-1}$ and $\boldsymbol{f} = w^{1/2}\boldsymbol{\Delta S}\boldsymbol{u}$.

Following the definition of Cook-statistic for univariate case (Bhar 1997) we give the Cook-statistic for the set of contrasts $\boldsymbol{P\tau}$ of $\boldsymbol{\tau}$ in multi-response experiment as:

$$(\boldsymbol{CD})_t = \frac{\left[\boldsymbol{P}(\widehat{\boldsymbol{\tau}} - \widehat{\boldsymbol{\tau}}_{(t)})\right] \left[D(\boldsymbol{P}\widehat{\boldsymbol{\tau}})\right]^{-1} \left[\boldsymbol{P}(\widehat{\boldsymbol{\tau}} - \widehat{\boldsymbol{\tau}}_{(t)})\right]}{\text{Rank}\left[D(\boldsymbol{P}\widehat{\boldsymbol{\tau}})\right]} \qquad \text{for} \quad t = 1, 2, \ldots, n. \quad (17)$$

Lemma 3.2 *The difference between the estimators of the contrasts of* $\boldsymbol{\tau}$ *under the model (2) and (13) can be expressed as*

$$\boldsymbol{P}(\widehat{\boldsymbol{\tau}} - \widehat{\boldsymbol{\tau}}_{(t)}) = (\boldsymbol{I} \otimes \boldsymbol{L}\boldsymbol{C}^{-}\boldsymbol{M})\boldsymbol{Y}, \qquad (18)$$

where $\boldsymbol{M} = \boldsymbol{E}\boldsymbol{C}^{-}\boldsymbol{F} + \boldsymbol{F} - \boldsymbol{E}\boldsymbol{C}^{-}\boldsymbol{\Delta S}$, $\boldsymbol{E} = \dfrac{\boldsymbol{f}\boldsymbol{f}'}{1 - \boldsymbol{f}'\boldsymbol{C}^{-}\boldsymbol{f}}$, $\boldsymbol{F} = w^{1/2}\boldsymbol{f}\boldsymbol{u}'\boldsymbol{S}$.

Proof: From (15) we have

$$\boldsymbol{C}^*_{(t)} = \boldsymbol{\Sigma}^{-1} \otimes [\boldsymbol{C} - \boldsymbol{f}\boldsymbol{f}']$$

and a g-inverse of $\boldsymbol{C}^*_{(t)}$ is obtained as [Pringle and Rayner (1971, p.32) and Dey (1993, Theorem 2)]

$$\boldsymbol{C}^{*-}_{(t)} = \boldsymbol{\Sigma} \otimes \left[\boldsymbol{C}^{-} + \frac{\boldsymbol{C}^{-}\boldsymbol{f}\boldsymbol{f}'\boldsymbol{C}^{-}}{1 - \boldsymbol{f}'\boldsymbol{C}^{-}\boldsymbol{f}}\right].$$

Thus

$$\begin{aligned} \boldsymbol{C}^{*-}_{(t)}\boldsymbol{Q}^*_{(t)} &= \left(\boldsymbol{\Sigma} \otimes \boldsymbol{C}^{-}\right)\left(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{Q}\right) - \left(\boldsymbol{\Sigma} \otimes \boldsymbol{C}^{-}\right)\left(\boldsymbol{\Sigma}^{-1} \otimes w^{1/2}\boldsymbol{f}\boldsymbol{u}'\boldsymbol{S}\boldsymbol{Y}\right) \\ &+ \left(\boldsymbol{\Sigma} \otimes \frac{\boldsymbol{C}^{-}\boldsymbol{f}\boldsymbol{f}'\boldsymbol{C}^{-}}{1 - \boldsymbol{f}'\boldsymbol{C}^{-}\boldsymbol{f}}\right)\left(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{Q}\right) - \left(\boldsymbol{\Sigma} \otimes \frac{\boldsymbol{C}^{-}\boldsymbol{f}\boldsymbol{f}'\boldsymbol{C}^{-}}{1 - \boldsymbol{f}'\boldsymbol{C}^{-}\boldsymbol{f}}\right)\left(\boldsymbol{\Sigma}^{-1} \otimes w^{1/2}\boldsymbol{f}\boldsymbol{u}'\boldsymbol{S}\boldsymbol{Y}\right). \end{aligned}$$

Then

$$\boldsymbol{C}^{*-}\boldsymbol{Q}^* - \boldsymbol{C}^{*-}_{(t)}\boldsymbol{Q}^*_{(t)} = \boldsymbol{I}_p \otimes w^{1/2}\boldsymbol{C}^{-}\boldsymbol{f}\boldsymbol{u}'\boldsymbol{S}\boldsymbol{Y} - \boldsymbol{I}_p \otimes \frac{\boldsymbol{C}^{-}\boldsymbol{f}\boldsymbol{f}'\boldsymbol{C}^{-}}{1 - \boldsymbol{f}'\boldsymbol{C}^{-}\boldsymbol{f}}\boldsymbol{Q} + \boldsymbol{I}_p \otimes w^{1/2}\frac{\boldsymbol{C}^{-}\boldsymbol{f}\boldsymbol{f}'\boldsymbol{C}^{-}}{1 - \boldsymbol{f}'\boldsymbol{C}^{-}\boldsymbol{f}}\boldsymbol{f}\boldsymbol{u}'\boldsymbol{S}\boldsymbol{Y}.$$

Then it follows

$$\begin{aligned}
\boldsymbol{P}\left(\widehat{\boldsymbol{\tau}}-\widehat{\boldsymbol{\tau}}_{(t)}\right) &= (\boldsymbol{I}_p\otimes\boldsymbol{L})\left(\boldsymbol{I}_p\otimes w^{1/2}\boldsymbol{C}^{-}\boldsymbol{f}\boldsymbol{u}'\boldsymbol{S}\boldsymbol{Y}-\boldsymbol{I}_p\otimes\frac{\boldsymbol{C}^{-}\boldsymbol{f}\boldsymbol{f}'\boldsymbol{C}^{-}}{1-\boldsymbol{f}'\boldsymbol{C}^{-}\boldsymbol{f}}\boldsymbol{Q}\right.\\
&\quad\left.+\boldsymbol{I}_p\otimes w^{1/2}\frac{\boldsymbol{C}^{-}\boldsymbol{f}\boldsymbol{f}'\boldsymbol{C}^{-}}{1-\boldsymbol{f}'\boldsymbol{C}^{-}\boldsymbol{f}}\boldsymbol{f}\boldsymbol{u}'\boldsymbol{S}\boldsymbol{Y}\right)\\
&= \boldsymbol{I}_p\otimes w^{1/2}\boldsymbol{L}\boldsymbol{C}^{-}\boldsymbol{f}\boldsymbol{u}'\boldsymbol{S}\boldsymbol{Y}-\boldsymbol{I}_p\otimes\frac{\boldsymbol{L}\boldsymbol{C}^{-}\boldsymbol{f}\boldsymbol{f}'\boldsymbol{C}^{-}}{1-\boldsymbol{f}'\boldsymbol{C}^{-}\boldsymbol{f}}\boldsymbol{Q}\\
&\quad+\boldsymbol{I}_p\otimes w^{1/2}\,\frac{\boldsymbol{L}\boldsymbol{C}^{-}\boldsymbol{f}\boldsymbol{f}'\boldsymbol{C}^{-}}{1-\boldsymbol{f}'\boldsymbol{C}^{-}\boldsymbol{f}}\boldsymbol{f}\boldsymbol{u}'\boldsymbol{S}\boldsymbol{Y}\\
&= \boldsymbol{I}_p\otimes\boldsymbol{L}\boldsymbol{C}^{-}\boldsymbol{F}\boldsymbol{Y}-\boldsymbol{I}_p\otimes\boldsymbol{L}\boldsymbol{C}^{-}\boldsymbol{E}\boldsymbol{C}^{-}\boldsymbol{Q}+\boldsymbol{I}_p\otimes\boldsymbol{L}\boldsymbol{C}^{-}\boldsymbol{E}\boldsymbol{C}^{-}\boldsymbol{F}\boldsymbol{Y}\\
&= \left(\boldsymbol{I}_p\otimes\boldsymbol{L}\boldsymbol{C}^{-}\boldsymbol{M}\right)\boldsymbol{Y}.
\end{aligned}$$

Now from (17) and (18) Cook-statistic for multi-response experiments can be written as

$$\begin{aligned}
(\boldsymbol{C}\boldsymbol{D})_t &= \frac{1}{p(v-1)}\boldsymbol{Y}'\left(\boldsymbol{I}\otimes\boldsymbol{M}'\boldsymbol{C}^{-}\boldsymbol{L}'\right)\left[\boldsymbol{P}\left(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{C}\right)\boldsymbol{P}'\right]\left(\boldsymbol{I}\otimes\boldsymbol{L}\boldsymbol{C}^{-}\boldsymbol{M}\right)\boldsymbol{Y}\\
&= \frac{1}{p(v-1)}\left[\boldsymbol{Y}'\left(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{M}'\boldsymbol{C}^{-}\boldsymbol{M}\right)\boldsymbol{Y}\right]. \qquad (19)
\end{aligned}$$

Remark 3.1 *For a Randomized Complete Block (RCB) design the matrix* $\boldsymbol{S}$ *can be written as* $\boldsymbol{S} = \operatorname{diag}\left[\left(\boldsymbol{I}_v-\frac{1}{v}\boldsymbol{I}_v\boldsymbol{I}_v'\right),\left(\boldsymbol{I}_v-\frac{1}{v}\boldsymbol{I}_v\boldsymbol{I}_v'\right),\ldots,\left(\boldsymbol{I}_v-\frac{1}{v}\boldsymbol{I}_v\boldsymbol{I}_v'\right)\right]$. *Thus the matrices* $\boldsymbol{E}$ *and* $\boldsymbol{F}$ *simplified as* $\boldsymbol{E} = \frac{r}{r-1}\boldsymbol{f}\boldsymbol{f}'$ *and* $\boldsymbol{F} = \frac{v-1}{v}\boldsymbol{f}\boldsymbol{u}'\boldsymbol{S}$, *where* $\boldsymbol{f} = \left(1-\frac{1}{v}-\frac{1}{v}\cdots-\frac{1}{v}\right)$. *Using these simplifications, one can obtain a* $(CD)_t$ *for* t^{th} *observation in a RCB design*

Belsley et al. (2004) have given a cut off point for $(CD)_t$ in case of a multiple linear regression as $4/n$. For any observation vector if calculated value of $(CD)_t$ $(t=1,2,\ldots,n)$ is more than $4/n$, then we may conclude that the observation vector from the t^{th} experimental unit is an outlier. Approximate distribution of $(CD)_t(t=1,2,\ldots,n)$ is unknown and is an open problem. A SAS code has been written for obtaining the test statistic for detection of outlier observation vector and is given in the Appendix (Table 2).

The above test statistic helps in detection of a single outlier vector. Once the outlier vector is detected, the next question arises as to what to do with this observation vector? First check whether there is any transcription error. If there are transcription errors, correct them and perform the analysis. If one finds that outlying observation vector is not due to transcription errors, then one simple way is to delete the observation vector that is identified as an outlier or perform

multivariate analysis of covariance by defining a covariate for each outlier.

The above procedure is illustrated with the help of an example in Section 4.

4 Illustration

Consider an experiment conducted during winter season of 2003-04 in terai region of West Bengal to study the effect of integrated nutrient management on growth and yield of late-sown Wheat. The experiment was laid out in Randomized complete block (RCB) design with 14 treatments in 3 replications. The data on following 9 characters were observed: plant height at harvest (cm), dry matter (DM) accumulation at 90 days after sowing (DAS), leaf area index (LAI) at 75 DAS, number of spikes/sq cm, number of grains per spike, test weight (g), grain yield (q/ha), straw yield (q/ha) and harvest index (%).

The data were analyzed for detection of single outlier observation vector using the test statistic developed in Section 3. The results obtained are given in Table 1 (Appendix). From Table 1, it can be observed that the observation corresponding to treatment number 1 and replication 3 has value of $(CD)_t$-statistic (0.1043) which is more than the cut off value of $(4/n = 0.09524)$. Therefore, we can say that the observation vector pertaining to treatment number 1 in replication 3 is an outlier.

Multivariate analysis of variance for testing the equality of treatment effect vectors was performed on original data and after deleting the outlying observation vector. The significance of treatment and replication effects was tested using Wilks Lamda criterion. Multivariate analysis of covariance was also performed by defining a covariate for the outlying observation vector as defined in (13). The results obtained are given in Table 2 (Appendix). From the analysis of original data given in Table 2, it is seen that replication effects are not significantly different at 5% level of significance whereas from the analysis after deleting the outlying observation (observation number 3) it is seen that replication effects are significantly different at 5% level of significance. There is no change in the results pertaining to treatment effects, though. It has been observed that deleting any other observation does not change the result of original data.

One can also observe that the results with analysis of covariance

and by deleting the outlier observation vector are same. Therefore, these approaches may be able to take care of presence of outlier(s) in the experimental data and any one of the two options can be used in practice. However, it is necessary that the outlier(s) is (are) detected at the first instance. The statistic developed for the detection of outlier(s) in the experimental data may be very helpful.

5 Discussion

In the present investigation, a test statistic has been developed for detection of a single outlier observation vector in multi-response experiments conducted in block designs. It may happen that all the components of the observation vector obtained from an experimental unit may not be outlier. Therefore, further efforts need to be made for developing a test statistic for detection of any p_1-component sub-vector of a p-component observation vector as outlier. Further, outlier(s) may exist in more than one observation vector. Therefore, a test statistic for detection of outlier(s) in more than one observation vector needs to be developed. Once an outlier is detected, one may think of either deleting the observation(s) identified as outlier(s) or carrying out the analysis of covariance. This procedure may be subjected to criticism. Therefore, one way to deal with such a situation is to develop robust procedure of estimation of treatment contrasts. Therefore, research efforts need to be made for developing a procedure of robust estimation in presence of outlier(s) in multi-response experiments.

A lot of literature is available on designs that are robust in presence of a single outlier in single response situations see e.g. Gopalan and Dey (1976), Singh et al. (1987), Ben and Yohai (1992), Bhar (1997), Bhar and Gupta (2001, 2003), Sarker (2002) and Sarker et al. (2003, 2005). A criterion of robustness of multi-response designs in presence of a single outlying observation vector needs to be developed.

Acknowledgements: Authors are grateful to the annonymous referee for useful suggestions that led to considerable improvement in the presentation of results.

References

Barrett, B. E. and Ling, R. F. (1992). General classes of influence measure for multivariate regression. *Journal of American Statistical Association* **87(417)**, 184-191.

Belsley, D.A., Kuh, E. and Welsch, R.E. (2004). *Regression Diagnostics -Identifying Influential Data and Sources of Collinearity.* John Wiley and Sons, New York.

Ben, M.G. and Yohai, V.J. (1992). Robust analysis of variance for a randomized block design. *Commun. Statist.-Theory & Meth.* **21(7)**, 1779-1798.

Bhar, L. (1997). *Outliers in Experimental Designs.* Unpublished Ph.D. Thesis. I.A.R.I. New Delhi.

Bhar, L. and Gupta, V. K. (2001). A useful statistic for studying outliers in experimental designs. *Sankhya,* **B63(4)**, 338-350.

Bhar, L. and Gupta, V. K. (2003). Study of outliers under variance-inflation model in experimental designs. *J. Indian. Soc. Agril. Statist.* **56(2)**, 142-154.

Box, G.E.P. and Draper, N.R. (1975). Robust designs. *Biometrika* **62(2)**, 347-352.

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

Cook, R. D. and Weisberg, S. (1980). Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics* **22**, 495-508.

Dey, A. (1993). Robustness of block designs against missing data. *Statistica Sinica* **3**, 219-231.

Gnanadesikan, R. and Kettenting, J.R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics* **28**, 81-124.

Gnanadesikan, R. and Lee, E.T. (1970). Graphical techniques for internal comparisons amongst equal degree of freedom groupings in multiresponse experiments. *Biometrika* **57**, 229-237.

Gentleman, J.E. and Wilk, M.B. (1975). Detecting outliers in two-way table: 1. Statistical behavior of residuals. *Technometrics* **17**, 1-14.

Ghosh, S. (1983). Influential observations in view of design and inference. *Commu. Stat.-Theory and Methods* **12(14)**, 1675-1683.

Ghosh, S. (1989). On two methods of identifying influential sets of observations. *Statist. Prob. Letters* **7**, 241-245.

Gopalan, R. and Dey, A. (1976). On robust experimental designs, *Sankhya* **B38**, 297-299.

John, J.A. (1978). Outliers in factorial experiments. *Appl. Statist.* **27**, 111-119.

Kang, G. and Bates, D.M. (1990). Approximate inferences in multiresponse regression analysis. *Biometrika* **77**, 321-332.

Pringle, R.M. and Rayner, A.A. (1971). *Generalized Inverse Matrices with Applications in Statistics.* Griffin's Statistical Monographs and Courses, No. 28. *Hafner Publishing Co., New York.*

Singh, G., Gupta, S. and Singh, M. (1987). Robustness of row column designs. *Statist. Prob. Letters* **5**, 421-424

Sarker, S. (2002). *Studies on Outlier(s) in Designed Experiments.* Unpublished Ph. D. Thesis, I.A.R.I., New Delhi.

Sarker, S., Gupta, V. K. and Parsad, R. (2003). Robust block designs for making test treatment-control treatment comparisons against the presence of an outlier. *J. Indian. Soc. Agril. Statist.* **56(1)**, 7-18.

Sarker, S., Parsad, R. and Gupta, V.K. (2005). Outliers in block designs for diallel crosses. *Metron-International Journal of Statistics* **63(2)**, 177-191.

Rajender Parsad
IASRI, Library Avenue, Pusa
New Delhi 110 012, India
Email : rajender@iasri.res.in

P. K. Nandi
IASRI, Library Avenue, Pusa
New Delhi 110 012, India
Email : nandi_stat@yahoo.co.in

L. M. Bhar
IASRI, Library Avenue, Pusa
New Delhi 110 012, India
Email : lmbhar@iasri.res.in

V. K. Gupta
IASRI, Library Avenue, Pusa
New Delhi 110 012, India
Email : vkgupta@iasri.res.in

Appendix

Table 1 : Detection of outlier observation vector using the test statistic

Observation Number	Treatment	Replication	Plant Height at harvest (cm)	DM accumulation at 90 DAS	LAI at 75 DAS	No. of spikes/sq m	No. of grains/ spike	Test weight (g)	Grain yield (q/ha)	Straw yield (q/ha)	Harvest Index (%)	(CD)
1	1	1	112.0	723.1	3.3	343.2	34.1	40.7	27.3	44.3	38.1	0.076812
2	1	2	133.0	729.0	4.2	325.0	37.0	49.0	32.0	36.0	37.0	0.078462
3	1	3	124.0	745.0	3.2	356.0	36.0	78.0	25.0	37.0	26.0	**0.104318**
4	2	1	111.1	784.6	3.7	372.2	38.2	41.4	29.0	47.0	40.3	0.040911
5	2	2	123.0	765.0	3.8	354.0	35.0	47.0	27.0	45.0	41.0	0.034252
6	2	3	112.0	734.0	3.2	345.0	32.0	43.0	29.0	48.0	43.0	0.063704
7	3	1	105.1	722.5	3.1	330.3	33.0	40.4	26.3	43.2	37.9	0.015557
8	3	2	110.0	734.0	3.4	323.0	32.0	46.0	27.0	46.0	38.0	0.0346000
9	3	3	109.0	720.0	3.2	354.0	36.0	42.0	26.0	43.0	39.0	0.049034
10	4	1	104.4	715.3	3.1	325.3	33.2	40.4	25.9	54.0	37.8	0.014674
11	4	2	109.0	726.0	3.5	342.0	34.0	46.0	25.0	52.0	36.0	0.026203
12	4	3	107.0	745.0	3.4	325.0	37.0	43.0	26.0	51.0	39.0	0.032702
13	5	1	106.8	729.2	3.2	337.2	45.0	40.5	26.9	43.9	38.0	0.014317
14	5	2	110.0	765.0	3.5	335.0	46.0	39.0	27.0	41.0	36.0	0.028023
15	5	3	107.0	754.0	3.2	342.0	41.0	42.0	25.0	42.0	41.0	0.035353
16	6	1	103.1	704.2	3.0	319.8	32.0	40.0	25.5	42.4	37.5	0.016434
17	6	2	109.0	765.0	3.1	323.0	29.0	43.0	25.0	43.0	38.0	0.060864
18	6	3	111.0	702.0	3.4	312.0	32.0	47.0	26.0	41.0	36.0	0.045580
19	7	1	102.6	696.9	2.9	315.3	32.1	40.0	34.0	41.8	37.4	0.035370
20	7	2	103.0	692.0	3.2	312.0	33.0	46.0	32.0	45.0	35.0	0.032348
21	7	3	109.0	723.0	3.0	321.0	35.0	42.0	29.0	43.0	36.0	0.046583

Observation Number	Treatment	Replication	Plant Height at harvest (cm)	DM accumulation at 90 DAS	LAI at 75 DAS	No. of spikes/sq m	No. of grains/ spike	Test weight (g)	Grain yield (q/ha)	Straw yield (q/ha)	Harvest Index (%)	(CD)
22	8	1	105.8	718.8	3.1	331.2	33.1	40.4	26.0	43.1	37.6	0.037268
23	8	2	105.0	726.0	3.2	335.0	31.0	46.0	27.0	42.0	39.0	0.021785
24	8	3	106.0	765.0	3.1	345.0	36.0	48.0	26.0	45.0	41.0	0.042323
25	9	1	109.0	761.9	3.5	362.5	37.0	41.0	28.7	46.4	38.2	0.030287
26	9	2	101.0	765.0	3.6	365.0	36.0	41.0	28.0	43.0	37.0	0.032326
27	9	3	109.0	786.0	3.4	356.0	38.0	46.0	29.0	42.0	38.0	0.018721
28	10	1	107.2	757.5	3.4	357.8	36.0	40.9	27.2	45.7	38.2	0.019468
29	10	2	110.0	725.0	3.5	357.0	40.0	42.0	25.0	48.0	39.0	0.062378
30	10	3	105.0	754.0	3.4	376.0	35.0	41.0	25.0	45.0	34.0	0.057845
31	11	1	110.7	769.5	3.6	363.3	37.5	41.2	29.6	46.9	38.7	0.049584
32	11	2	105.0	754.0	3.4	387.0	38.0	43.0	25.0	43.0	36.0	0.040707
33	11	3	113.0	767.0	3.4	367.0	36.0	41.0	28.0	39.0	37.0	0.029235
34	12	1	106.0	744.3	3.4	353.8	36.0	40.6	28.0	45.3	38.2	0.007566
35	12	2	106.0	765.0	3.3	356.0	35.0	42.0	29.0	45.0	35.0	0.036392
36	12	3	109.0	723.0	3.2	354.0	32.0	43.0	24.0	47.0	41.0	0.040477
37	13	1	105.0	738.9	3.3	350.5	34.9	40.6	27.7	44.9	38.2	0.047424
38	13	2	109.0	734.0	3.2	356.0	33.0	43.0	26.0	51.0	34.0	0.028278
39	13	3	110.0	743.0	3.2	354.0	35.0	40.0	28.0	48.0	37.0	0.012904
40	14	1	107.8	755.2	3.5	358.2	36.9	41.0	29.0	46.0	38.7	0.033924
41	14	2	112.0	765.0	3.4	343.0	36.0	40.0	28.0	46.0	39.0	0.024054
42	14	3	113.0	734.0	3.5	323.0	31.0	41.0	25.0	42.0	41.0	0.056339

*$4/n = 0.095238$

Table 2: Multivariate analysis of variance/ covariance for simultaneous comparison of treatment effects from original data, after removing the outlier observation vector and by defining a covariate corresponding to outlier observation vector

Source	Original Data		After removing the outlying observation vector		Defining a covariate corresponding to outlying observation vector	
	Wilk's Lambda	Prob > F	Wilk's Lambda	Prob > F	Wilk's Lambda	Prob > F
Treatment	0.0001	< 0.0001	0.0001	< 0.0001	0.0001	< 0.0001
Replication	0.2684	0.0556	0.2406	0.0443	0.2406	0.0443
Covariate	-	-	-	-	0.2058	0.0002

SAS code for detecting outlier observation vector from multi-response experiments:

```
options ps=2000 ls=100;
data outlier;
input trt blk y1-y9;
cards;
1   1   112.0  723.1  3.3  343.2  34.1  40.7  27.3  44.3  38.1
1   2   133.0  729.0  4.2  325.0  37.0  49.0  32.0  36.0  37.0
1   3   124.0  745.0  3.2  356.0  36.0  52.0  25.0  37.0  26.0
⋮   ⋮     ⋮      ⋮    ⋮    ⋮     ⋮     ⋮     ⋮     ⋮     ⋮
14  1   107.8  755.2  3.5  358.2  36.9  41.0  29.0  46.0  38.7
14  2   112.0  765.0  3.4  343.0  36.0  40.0  28.0  46.0  39.0
14  3   113.0  734.0  3.5  323.0  31.0  41.0  25.0  42.0  41.0
;
run;
proc iml;
use outlier;
read all into d;
run;
n = nrow(d); *number of observations;
v = max(d[,1]); *number of treatments;
b = max(d[,2]); *number of blocks;
x1 = J(n,v,0); *x1 is del prime;
x2 = j(n,b,0); *x2 is d prime;
y = d[,3:ncol(d)];
p = ncol(y); *p is number of response variables;
do i = 1 to n;
        do j = 1 to v;
        if d[i,1] = j then x1[i,j] = 1;
        end;
end;
```

```
do i = 1 to n;
        do j = 1 to b;
        if d[i,2] = j then x2[i,j] = 1;
        end;
do i = 1 to n;
        do j = 1 to b;
        if d[i,2] = j then x2[i,j] = 1;
        end;
end;

x21 = j(nrow(y),1,1);
x = x1||x2||x21;
beta = ginv(x'*x)*x'*y;
yv0 = j(1,1,0);
do i = 1 to ncol(y);
        yv0 = yv0//y[,i];
end;
print yv0;
yv = yv0[2:nrow(yv0),];
c0 = x1'*x1-x1'*x2*ginv(x2'*x2)*x2'*x1;
print c0;
q0 = (x1'-x1'*x2*ginv(x2'*x2)*x2')*y;
print q0;
run;

b0 = x2'*y;
b01 = b0[,1];
b02 = b0[,2];

tau0 = ginv(c0)*q0;

c01 = ginv(c0);

trssp = q0'*c01*q0;

tssp = j(ncol(y),ncol(y),0);

do i = 1 to ncol(y);
        do j = 1 to ncol(y);
        tssp[i,j] = y[,i]'*y[,j]-(y[+,i]*y[+,j])/(nrow(y));
        end;
end;

Repssp=j(ncol(y),ncol(y),0);
do i=1 to ncol(y);
        do j=1 to ncol(y);
        Repssp[i,j]=b0[,i]'*inv(x2'*x2)*b0[,j]-(y[+,i]*y[+,j])/(nrow(y));
        end;
end;
```

```
ressp = tssp - repssp - trssp;
wl_trt = det(ressp)/det(trssp + ressp);
wl_blk = det(ressp)/det(repssp + ressp);

print trssp;
print repssp;
print ressp;
sig_est = ressp/(nrow(y) - v - b + 1);
print sig_est;
c = inv(sig_est) @ c0;
q = (inv(sig_est) @ (x1‘-x1‘*x2*ginv(x2‘*x2)*x2‘))*yv;
tau = ginv(c)*q;
/*Finding out Cook's Distance for outlier detection */

S = i(nrow(y))-x2*inv(x2‘*x2)*x2‘;
u = i(nrow(y));
c_d = j(1,1,0);

dd = j(1,2,0);
d = j(1,1,0);


do i = 1 to nrow(y);
w= inv(u[,i]‘*S*u[,i]);
f1= sqrt(w)*x1‘*S*u[,i];
F= sqrt(w)*f1*u[,i]‘*S;
E= f1*f1‘*inv(1-f1‘*ginv(c0)*f1);
M= E*ginv(c0)*F+F-E*ginv(c0)*x1‘*S;

C_ Dt=(yv‘*(inv(sig_ est)@(M‘*ginv(c0)*M))*yv)/(p*(v-1));
          c_d=c_ d/c_ d1;
dd=dd//(i||c_d1);
end;
dd1= dd[2:nrow(dd),];
print dd1;
cut = 4/n;
print the cut off point is= cut;
run;
```