Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 69(1) 2015 57-63

Detection of Outliers in Designed Experiments with Correlated Errors

Sankalpa Ojha and Lalmohan Bhar
ICAR-Indian Agricultural Statistics Research Institute, New Delhi

Received 15 March 2014; Revised 10 March 2015; Accepted 13 March 2015

SUMMARY

Two statistics for detecting outliers in designed experiments with correlated errors have been developed. These statistics are Cook-statistic and AP-statistic. General expressions of these statistics for detecting any *t* outliers have been obtained. Equal correlation structure has been considered for general variance-covariance matrix. Developed Cook-statistic has been illustrated with an example. However, case of occurrence of a single outlier has been considered in the example.

Keywords: Cook-statistic, AP-statistic, Outlier, Block design, Correlated error, Autocorrelation.

1. INTRODUCTION

The presence of outlier in the data is the most serious illness in any data set. The analysis of data from designed experiments is valid only when the assumptions like normality and homogeneity of error variances hold. The departures from these assumptions may take place in presence of outlier(s). Therefore, it is important to detect and handle the outlier(s) efficiently. However, most of the studies thus far conducted in design of experiment to detect outliers have been restricted to models having uncorrelated disturbances with constant variances. Bhar and Gupta (2001) investigated the problem of outliers in block designs and modified the Cook-statistic (Cook 1977, 1979), Q_k-statistic (Gentleman and Wilk 1975) and *AP*-statistic (Andrews and Pregibon 1978) for detection of outliers in experimental data. Some more reference on outliers in block designs are due to Bhar and Gupta (2003), Sarker *et al*. (2003), Sarker *et al.* (2005), Parsad *et al*. (2008) and Bhar and Ojha (2014). A little bit different kind of study on outlier in designed experiments is found in Ghosh (1983). He considered the measures for detecting the influential observations *w.r.t*. one or several parameters of interest at the design stage. He considered Cook-statistic for detecting the influential observations at the inference stage. All these literature dealing with the problem of outliers in block designs are for spherical error structure. But it has been frequently observed, the dispersion matrix may not always be spherical. In designed experiments, it is generally assumed that the observations are independently and identically distributed. However, there are many experimental situations in which the assumption of independence of observations gets violated. In field experiments, the observations are mutually correlated through some systematic pattern of environmental variations. For example, plots occurring close together within a field are well known to be more similar than plots occurring far away from each other. Thus in field experiments, blocks are often formed using adjacent plots within a field. Whenever spatial contiguity is used as a criterion for blocking, it is often the case that the experimental units occurring close together within spatial blocks created are correlated. We have a vast literature on design and analysis of experiments in the presence of correlated errors in general. For an excellent review references may be made to Williams (1952), Atkinson (1969), Berenblut and Webb (1974), Bartlett (1978), Herzberg (1982),

Corresponding author: Lamohan Bhar
E-mail address: lmbhar@gmail.com

Wilkinson *et al.* (1983). Different types of correlation structures that may exist among the observations within a block are nearest neighbour, autoregressive and equi-correlated observations, etc. In such cases it is necessary to develop methodology to detect outliers. Some work on detection of outliers in linear regression with correlated errors are due to Schall and Dunne (1991), Martin (1992), Kim and Huggins (1998) and Sen Roy and Guria (2004, 2009). However, it seems that no work for detection of outliers in designed experiments with correlated errors is available in the literature. Test statistics as available in literature for regression analysis cannot be applied directly to designed experiments, because of rank deficiency problem of its design matrix. In the present investigation, Cook-statistic and AP-statistic has been developed for detecting any t outliers from design of experiments conducted using a block design when errors are correlated.

In Section 2, these statistics are developed for detecting any t outliers for designed experiments with correlated errors. In Section 3, a particular type of correlation structure has been considered and its estimation procedure is given. In Section 4, an example is given to illustrate the procedure so developed under Section 2. Throughout we use $\mathbf{1}_n$ to denote an n-component vector of ones and $\mathbf{1}_n$ an identity matrix of order n. Further $\mathbf{A}'$, $\mathbf{A}^-$ and $\mathbf{A}^+$ respectively denote the transpose, a generalized inverse (g-inverse) and the Moore-Penrose inverse of a matrix $\mathbf{A}$.

2. TEST STATISTICS FOR DETECTION OF OUTLIERS

Consider the general linear model for an experimental design d (say),

$$\mathbf{y} = \mu\mathbf{1}_n + \Delta'\tau + \mathbf{D}'\beta + \varepsilon \qquad (1)$$

where $\mathbf{y}$ is an $n \times 1$ vector of observations, $\mathbf{1}_n$ is the n dimensional column vector of all elements unity, Δ' is an $n \times v$ design matrix for treatment effects, $\mathbf{D}'$ is an $n \times b$ design matrix of block effects, τ is a $v \times 1$ vector of treatment effects, and β is a $b \times 1$ vector of block effects, μ is general mean and ε is the vector of random errors. We have $\Delta'\mathbf{1}_v = \mathbf{1}_n = \mathbf{D}'\mathbf{1}_b$, $\Delta\mathbf{1}_n = \mathbf{r}$, $\mathbf{D}\mathbf{1}_n = \mathbf{k}$, where $\mathbf{r} = (r_1, r_2, ..., r_v)'$ and $\mathbf{k} = (k_1, k_2, ... k_b)'$ are the vectors of replications and block sizes respectively. We also assume that $E(\varepsilon) = \mathbf{0}$ and $D(\varepsilon) = \Omega$. The dispersion matrix Ω is positive definite and symmetric. We also assume that the variance-covariance matrix of ε for the j^{th} block is Σ_j for $j = 1, 2,\ldots, b,$ where Σ_j is a $k_j \times k_j$ positive definite matrix. Thus $\Omega = diag(\Sigma_1, \Sigma_2,..., \Sigma_b)$. That is, it is assumed that the observations belonging to the same block are correlated and they are uncorrelated when they belong to different blocks.

Now by applying Aitkin's transformation, we rewrite the model (1) as

$$\Omega^{-1/2}\mathbf{y} = \mu\Omega^{-1/2}\mathbf{1} + \Omega^{-1/2}\Delta'\tau + \Omega^{-1/2}\mathbf{D}'\beta + \Omega^{-1/2}\varepsilon. \qquad (2)$$

From (2) we obtain on eliminating β and μ, the equations involving only τ as

$$\mathbf{C}_\tau\tau = \mathbf{Q}_\tau, \qquad (3)$$

where

$$\mathbf{C}_\tau = (\Delta\Omega^{-1}\Delta' - \Delta\Omega^{-1}\mathbf{D}'(\mathbf{D}\Omega^{-1}\mathbf{D}')^{-1}\mathbf{D}\Omega^{-1}\Delta') = \Delta\Phi\Delta', \qquad (4)$$

$$\mathbf{Q}_\tau = (\Delta\Omega^{-1}\mathbf{y} - \Delta\Omega^{-1}\mathbf{D}'(\mathbf{D}\Omega^{-1}\mathbf{D}')^{-1}\mathbf{D}\Omega^{-1}\mathbf{y}) = \Delta\Phi\mathbf{y}, \qquad (5)$$

and

$$\Phi = \Omega^{-1} - \Omega^{-1}\mathbf{D}'(\mathbf{D}\Omega^{-1}\mathbf{D}')^{-1}\mathbf{D}\Omega^{-1}. \qquad (6)$$

It is easy to verify that $\mathbf{C}_\tau$ is symmetric with row sums and column sums equal to zero.

Now writing $\Delta' = (\mathbf{1}_n\ \Delta'\ \mathbf{D}')$, we denote.

$$\mathbf{V} = \mathbf{1}_n - \mathbf{X}\Omega^{-1}(\mathbf{X}'\Omega^{-1}\mathbf{X})^-\Omega^{-1}\mathbf{X}' = \Phi - \Phi\Delta'\mathbf{C}_\tau^+\Delta\Phi. \qquad (7)$$

Then we define a set of residuals under model (1) as

$$\mathbf{r}^* = \mathbf{V}\mathbf{y} \qquad (8)$$

Now the following result can easily be proved (Bhar and Gupta 2001).

Theorem 1

(i) $E(\mathbf{Q}_\tau) = \mathbf{C}_\tau\,\tau$

(ii) $D(\mathbf{Q}_\tau) = \mathbf{C}_\tau\,\tau$

We assume that the design d considered here is connected, *i.e*., all $(v - 1)$ orthonormalized contrasts of parameters of τ are estimable or equivalently Rank $(\mathbf{C}_\tau) = v - 1$. Let the set of all $(v - 1)$ orthonormalized contrasts of parameters of τ be given by $\mathbf{P}\tau$. Where the matrix $\mathbf{P}$ is of dimension $(v - 1) \times v$ and such that

$$\mathbf{P}\mathbf{P}' = \mathbf{I}_{(v-1)} \text{ and } \mathbf{P}'\mathbf{P} = \mathbf{I}_v - (1/v)\,\mathbf{J}. \qquad (9)$$

The best linear estimator of $\mathbf{P}\tau$ is given by $\mathbf{P}\hat{\tau}$ where $\hat{\tau}$ is any solution of the normal equations (3). Hence $\mathbf{P}\hat{\tau} = \mathbf{P}\mathbf{C}_\tau^+\mathbf{Q}_\tau$.

Since row sums and column sums of $\mathbf{C}_\tau$ is zero, *i.e.*, $\mathbf{C}_\tau \mathbf{1}_v = \mathbf{1}'_v \mathbf{C}_\tau = \mathbf{0}$, the dispersion matrix of $\mathbf{P}\hat{\boldsymbol{\tau}}$ can be written as (Gupta and Mukerjee 1989).

$$D(\mathbf{P}\hat{\boldsymbol{\tau}}) = (\mathbf{P}\mathbf{C}_\tau\mathbf{P}')^{-1}. \quad (10)$$

2.1 Cook-statistic

Cook (1977, 1979) developed a statistic to detect outliers in linear regression model which is actually a measure of distance between the parameter estimates obtained from the full model and parameter estimates obtained after deleting the suspected outliers. On this line Bhar and Gupta (2001) modified this statistic for detecting outliers in designed experiments and when the experimenter's interest is in estimation of some functions of parameters of interest. They developed this statistic when the dispersion matrix $\Omega = \mathbf{I}_n\sigma^2$. Consider the linear model (1) in which $\Omega = \mathbf{I}_n\sigma^2$. If any t observations are suspected to be outliers, then the Cook-statistic for the set of contrasts $\mathbf{P}\tau$ is given by (Bhar and Gupta 2001),

$$D_t = \frac{(\mathbf{P}\hat{\boldsymbol{\tau}} - \mathbf{P}\hat{\boldsymbol{\tau}}_{(t)})'[D(\mathbf{P}\hat{\boldsymbol{\tau}})]^{-1}(\mathbf{P}\hat{\boldsymbol{\tau}} - \mathbf{P}\hat{\boldsymbol{\tau}}_{(t)})}{Rank[D(\mathbf{P}\hat{\boldsymbol{\tau}})]} \quad (11)$$

where $\mathbf{P}\hat{\boldsymbol{\tau}}$ is the least squares estimator of $\mathbf{P}\tau$, $\mathbf{P}\hat{\boldsymbol{\tau}}_{(t)}$ is the least squares estimator of $\mathbf{P}\tau$ obtained after deleting the suspected t outlying observations.

Thus to obtain Cook-statistic, we have to obtain the estimate of $\mathbf{P}\tau$ after deleting the suspected outlying observations. To obtain this estimate we need to estimate $\mathbf{P}\tau$ after deleting the suspected t outliers in the model with correlated errors. Now we write the model after deleting these t outlying observations as

$$\mathbf{y}_{(t)} = \mu_{(t)}\mathbf{1}_{(t)} + \Delta'_{(t)}\tau_{(t)} + \mathbf{D}'_{(t)}\beta_{(t)} + \varepsilon_{(t)}, \quad (12)$$

where $\mathbf{y}_{(t)}$ has $(n\text{-}t)$ observations, $\mathbf{y}_{(t)}$, $\mathbf{1}_{(t)}$, $\Delta'_{(t)}$, $\mathbf{D}'_{(t)}$ and $\varepsilon_{(t)}$ have $(n\text{-}t)$ rows. The parameters $\mu_{(t)}$, $\tau_{(t)}$ and $\beta_{(t)}$ denote that they are obtained after deleting t observations. The variance-covariance matrix of $\varepsilon_{(t)}$ under this model is denoted by $\Omega_{(tt)}$ indicating that this is obtained after deleting t rows and t columns. We assume that any t observations from the n observations are outliers in the sense that expected values of these observations are shifted from the expected value of other observations. Without loss of generality, we assume that the first t observations are outliers. Further it is also assumed that the design remains connected after deletion of t observations.

We now define a matrix $\mathbf{A} = \mathbf{1}_n - \mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'$, where $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_t)$ and $\mathbf{u}_j = (0, 0, ..., 1(j^{th}), ..., 0, 0)$ an n-component vector with 1 in the j^{th} position if the j^{th} observation is an outlier and all other elements as zero. The matrix $\mathbf{A}$ is symmetric and idempotent. Then the model (12) can alternatively be written as

$$\mathbf{A}\Omega^{-1/2}\mathbf{y} = \mu\mathbf{A}\Omega^{-1/2}\mathbf{1} + \mathbf{A}\Omega^{-1/2}\Delta'\tau + \mathbf{A}\Omega^{-1/2}\mathbf{D}'\beta + \mathbf{A}\Omega^{-1/2}\varepsilon \quad (13)$$

If we denote the treatment effects under this model by $\tau_{(t)}$, then,

$$\mathbf{C}_{\tau(t)}\,\tau_{(t)} = \mathbf{Q}_{\tau(t)}, \quad (14)$$

where

$$\mathbf{C}_{\tau(t)} = \Delta\Omega^{-1/2}\mathbf{A}\Omega^{-1/2}\Delta' - \Delta\Omega^{-1/2}\mathbf{A}\Omega^{-1/2}\mathbf{D}'(\mathbf{D}\Omega^{-1/2}\mathbf{A}\Omega^{-1/2}\mathbf{D}')^{-1}\mathbf{D}\Omega^{-1/2}\mathbf{A}\Omega^{-1/2}\Delta'.$$

On simplification, we get,

$$\mathbf{C}_{\tau_{(i)}} = \mathbf{C}_\tau - \Delta\Phi\mathbf{U}'(\mathbf{U}\Phi\mathbf{U}')^{-1}\mathbf{U}\Phi\Delta' \quad (15)$$

Similarly

$$\mathbf{Q}_{\tau(t)} = \mathbf{Q}_\tau - \Delta\Phi\mathbf{U}'(\mathbf{U}\Phi\mathbf{U}')^{-1}\mathbf{U}\Phi\mathbf{y}. \quad (16)$$

Theorem 2

The difference between the estimators of contrasts in τ under the model (1) and (12), *i.e.*, $\mathbf{P}\hat{\boldsymbol{\tau}} - \mathbf{P}\hat{\boldsymbol{\tau}}_{(t)}$ can be expressed as

$$\mathbf{P}\hat{\boldsymbol{\tau}} - \mathbf{P}\hat{\boldsymbol{\tau}}_{(t)} = \mathbf{P}\mathbf{C}_\tau^{+}\Delta\Phi\mathbf{U}'(\mathbf{U}\mathbf{V}\mathbf{U}')^{-1}\mathbf{U}\mathbf{V}\mathbf{y} \quad (17)$$

Proof

A Moore-Penrose inverse of $\mathbf{C}_{\tau(t)} = \mathbf{C}_\tau - \Delta\Phi\mathbf{U}(\mathbf{U}'\Phi\mathbf{U})^{-1}\mathbf{U}'\Phi\Delta'$ is given by

$$\mathbf{C}^{+}_{\tau(t)} = \mathbf{C}^{+}_\tau + \mathbf{C}^{+}_\tau\Delta\Phi\mathbf{U}(\mathbf{U}'\Phi\mathbf{U} - \mathbf{U}'\Phi\mathbf{D}'\mathbf{C}^{+}_\tau\Delta\Phi\mathbf{U}')^{-1}\mathbf{U}'\Phi\Delta'\mathbf{C}^{+}_\tau.$$

Now $\mathbf{U}'\Phi\mathbf{U} - \mathbf{U}'\Phi\mathbf{D}'\mathbf{C}^{+}_\tau\mathbf{D}\Phi\mathbf{U} = \mathbf{U}'(\Phi - \Phi\mathbf{D}'\mathbf{C}^{+}_\tau\Delta\Phi)\mathbf{U} = \mathbf{U}'\mathbf{V}\mathbf{U}\mathbf{C}^{+}_\tau.$

Thus $\mathbf{C}^{+}_{\tau(l)} = \mathbf{C}^{+}_\tau + \mathbf{C}^{+}_\tau\Delta\Phi\mathbf{U}'(\mathbf{U}'\Phi\mathbf{U})^{-1}\mathbf{U}'\Phi\mathbf{D}'\mathbf{C}^{+}_\tau.$

Hence,

$$\mathbf{P}\hat{\tau}_{(t)} = \mathbf{P}\mathbf{C}^{+}_{\tau(t)}\mathbf{Q}_{t(t)} = \mathbf{P}(\mathbf{C}^{+}_\tau + \mathbf{C}^{+}_\tau\mathbf{D}\Phi\mathbf{U}(\mathbf{U}'\mathbf{V}\mathbf{U})^{-1}\mathbf{U}'\Phi\mathbf{D}'\mathbf{C}^{+}_\tau)\mathbf{Q}_{t(t)} \quad (18)$$

Now using the value of $\mathbf{Q}_{\tau(t)}$ as given in (16) in (18), we get

$$\begin{aligned}\mathbf{P}\hat{\boldsymbol{\tau}}_{(t)} &= \mathbf{P}\mathbf{C}_\tau^{+}\mathbf{Q}_\tau\Delta\Phi\mathbf{U}(\mathbf{U}'\mathbf{V}\mathbf{U})^{-1}\mathbf{U}'\Phi\Delta'\mathbf{C}_\tau^{+}\mathbf{Q}_\tau-\mathbf{P}\mathbf{C}_\tau^{+}\\ &\quad\Delta\Phi\mathbf{U}(\mathbf{U}'\Phi\mathbf{U})^{-1}\mathbf{U}'\Phi\mathbf{y}-\mathbf{P}\mathbf{C}_\tau^{+}\Delta\Phi\mathbf{U}(\mathbf{U}'\mathbf{V}\mathbf{U})^{-1}\\ &\quad\mathbf{U}'\Phi\Delta\,\mathbf{C}_\tau^{+}\Delta\Phi\mathbf{U}(\mathbf{U}'\Phi\mathbf{U})^{-1}\,\mathbf{U}'\Phi\mathbf{y}\\ &= \mathbf{P}\mathbf{C}_\tau^{+}\Delta\Phi\mathbf{U}(\mathbf{U}'\mathbf{V}\mathbf{U})^{-1}\mathbf{U}'\Phi\Delta'\mathbf{C}_\tau^{+}\Delta\Phi\mathbf{y}-\mathbf{P}\mathbf{C}_\tau^{+}\\ &\quad\Delta\Phi\mathbf{U}(\mathbf{U}'\Phi\mathbf{U})^{-1}\mathbf{U}'\Phi\mathbf{y}-\mathbf{P}\mathbf{C}_\tau^{+}\Delta\Phi\mathbf{U}(\mathbf{U}'\mathbf{V}\mathbf{U})^{-1}\\ &\quad\mathbf{U}'\Phi\Delta'\mathbf{C}_\tau^{+}\Delta\Phi\mathbf{U}(\mathbf{U}'\Phi\mathbf{U})^{-1}\mathbf{U}'\Phi\mathbf{y}\\ &= \mathbf{P}\mathbf{C}_\tau^{+}\mathbf{Q}_\tau+\mathbf{P}\mathbf{C}_\tau^{+}\Delta\Phi\mathbf{U}(\mathbf{U}'\mathbf{V}\mathbf{U})^{-1}\mathbf{U}'\Phi\Delta'\mathbf{C}_\tau^{+}\\ &\quad\Delta(\Phi-\Phi\mathbf{U}(\mathbf{U}'\Phi\mathbf{U})^{-1}\mathbf{U}'\Phi)\mathbf{y}-\mathbf{P}\mathbf{C}_\tau^{+}\Delta\Phi\mathbf{U}\\ &\quad(\mathbf{U}'\Phi\mathbf{U})^{-1}\mathbf{U}'\Phi\mathbf{y}\\ &= \mathbf{P}\mathbf{C}_\tau^{+}\mathbf{Q}_l+\mathbf{P}\mathbf{C}_\tau^{+}\Delta\Phi\mathbf{U}(\mathbf{U}'\mathbf{V}\mathbf{U})^{-1}\mathbf{U}(\Phi-\mathbf{V})\mathbf{y}-\\ &\quad\mathbf{P}\mathbf{C}_\tau^{+}\Delta\Phi\mathbf{U}(\mathbf{U}'\mathbf{V}\mathbf{U})^{-1}\mathbf{U}'(\Phi-\mathbf{V})\mathbf{U}(\mathbf{U}'\Phi\mathbf{U})^{-1}\\ &\quad\mathbf{U}'\Phi\mathbf{y}-\mathbf{P}\mathbf{C}_\tau^{+}\Delta\Phi\mathbf{U}(\mathbf{U}'\Phi\mathbf{U})^{-1}\mathbf{U}'\Phi\mathbf{y}\\ &= \mathbf{P}\mathbf{C}_\tau^{+}\mathbf{Q}_l+\mathbf{P}\mathbf{C}_\tau^{+}\Delta\Phi\mathbf{U}(\mathbf{U}'\mathbf{V}\mathbf{U})^{-1}\mathbf{U}'\Phi\mathbf{y}-\mathbf{P}\mathbf{C}_\tau^{+}\\ &\quad\Delta\Phi\mathbf{U}(\mathbf{U}'\mathbf{V}\mathbf{U})^{-1}\mathbf{U}'\mathbf{V}\mathbf{y}-\mathbf{P}\mathbf{C}_\tau^{+}\Delta\Phi\mathbf{U}(\mathbf{U}'\mathbf{V}\mathbf{U})^{-1}\\ &\quad\mathbf{U}'\Phi\mathbf{y}\,\mathbf{P}\mathbf{C}_\tau^{+}\Delta\Phi\mathbf{U}(\mathbf{U}'\Phi\mathbf{U})^{-1}\mathbf{U}'\Phi\mathbf{y}-\mathbf{P}\mathbf{C}_\tau^{+}\Delta\Phi\\ &\quad\mathbf{U}(\mathbf{U}'\Phi\mathbf{U})^{-1}\mathbf{U}'\Phi\mathbf{y}\\ &= \mathbf{P}\mathbf{C}_\tau^{+}\mathbf{Q}_\tau-\mathbf{P}\mathbf{C}_\tau^{+}\Delta\Phi\mathbf{U}(\mathbf{U}'\mathbf{V}\mathbf{U})^{-1}\mathbf{U}'\mathbf{V}\mathbf{y}\\ &= \mathbf{P}\hat{\boldsymbol{\tau}}-\mathbf{P}\mathbf{C}_\tau^{+}\Delta\Phi\mathbf{U}(\mathbf{U}'\mathbf{V}\mathbf{U})^{-1}\mathbf{U}'\mathbf{V}\mathbf{y}\end{aligned}$$

Hence the proof.

Now following the definition of Cook-statistic for uncorrelated error (Bhar and Gupta 2001) as given in (11) we give the Cook-statistic for the set of contrasts $\mathbf{P}\boldsymbol{\tau}$ of $\boldsymbol{\tau}$ in designed experiments with correlated error for t outliers as

$$D_l = \frac{(\mathbf{P}\hat{\boldsymbol{\tau}}-\mathbf{P}\hat{\boldsymbol{\tau}}_{(t)})'[D(\mathbf{P}\hat{\boldsymbol{\tau}})]^{-1}(\mathbf{P}\hat{\boldsymbol{\tau}}-\mathbf{P}\hat{\boldsymbol{\tau}}_{(t)})}{Rank[D(\mathbf{P}\hat{\boldsymbol{\tau}})]}$$

$$= \frac{\mathbf{y}'\mathbf{V}\mathbf{U}(\mathbf{U}'\mathbf{V}\mathbf{U})^{-1}\mathbf{U}'\Phi\Delta'\mathbf{C}_\tau^{+}\mathbf{P}'[D(\mathbf{P}\hat{\boldsymbol{\tau}})]\mathbf{P}\mathbf{C}_\tau^{+}\Delta\Phi\mathbf{U}(\mathbf{U}'\mathbf{V}\mathbf{U})^{-1}\mathbf{U}'\mathbf{V}\mathbf{y}}{(v-1)} \quad (19)$$

Now since $\mathbf{C}_\tau\,\mathbf{1}_y = \mathbf{1}'_v\,\mathbf{C}_\tau = \mathbf{0}$, using (10), we get

$$D_l = \frac{\mathbf{y}'\mathbf{V}\mathbf{U}(\mathbf{U}'\mathbf{V}\mathbf{U})^{-1}\mathbf{U}'\Phi\Delta'\mathbf{C}_\tau^{+}\Delta\Phi\mathbf{U}(\mathbf{U}'\mathbf{V}\mathbf{U})^{-1}\mathbf{U}'\mathbf{V}\mathbf{y}}{(v-1)} \quad (20)$$

$$D_l = \frac{\mathbf{r}_t^{*'}(\mathbf{U}'\mathbf{V}\mathbf{U})^{-1}\mathbf{U}'\Phi\Delta'\mathbf{C}_\tau^{+}\Delta\Phi\mathbf{U}(\mathbf{U}'\mathbf{V}\mathbf{U})^{-1}\mathbf{r}_t^{*}}{(v-1)}, \quad (21)$$

where $\mathbf{r}_t^{*'} = \mathbf{U}'\mathbf{V}\mathbf{y}$, vector of residuals corresponding to outlying observations. (22)

This follows approximately an F-distribution with $v-1$ and $n-v-b+1$ degrees of freedom (Bhar and Gupta 2001).

Now we consider a special case of occurrence of a single outlier.

2.1.1 Single outlier

Without loss of generality, we assume that the first observation in the first block is an outlier, then from (21) we get the Cook-statistic for the present case as

$$D_t = \frac{h_{11}}{(v-1)}\left(\frac{r_1^{*}}{v_{11}}\right)^2, \quad (23)$$

where v_{11} is the first diagonal element of the matrix $\mathbf{V}$ as defined in (7), h_{11} is the first diagonal element of matrix $\mathbf{H} = \Phi\Delta'\mathbf{C}_\tau^{+}\Delta\Phi$ and r_1^{*} is the first element of $\mathbf{r}_t^{*}$ as defined in (22).

2.2 AP-statistic

Bhar and Gupta (2001) also defined another test statistic which is very useful in detecting outlier(s) in experimental data. This statistic is also useful in determining the degree of influence of outlier(s) on parameter estimation. Consider again the model (1) in which $\mathbf{X}$ has full column rank and define a matrix $\mathbf{Z}$ as $\mathbf{Z} = (\mathbf{X}\ \mathbf{U})$, where $\mathbf{X}$ and $\mathbf{U}$ are as defined before. Then AP-statistic given by Andrews and Pregibon (1978) is defined as

$$AP_t = \frac{\left|\mathbf{Z}^{*'}\mathbf{Z}^{*}\right|}{\left|\mathbf{X}^{*'}\mathbf{X}\right|}, \quad (24)$$

where $\mathbf{X}^{*} = (\mathbf{X}\ \mathbf{y})$ and $\mathbf{Z}^{*} = (\mathbf{X}\ \mathbf{U}\ \mathbf{y})$ and $|\mathbf{A}|$ denotes determinant value of $\mathbf{A}$. The quantity $(1\text{-}AP_t)$ corresponds to the proportion of volume generated by $\mathbf{X}^{*}$ attributed to the t outlying observations. Small values of AP_t statistic are associated with deviant or influential observations. Bhar and Gupta (2001) modified this statistic for application into designed experiments with uncorrelated errors. We also modified this statistic for application in designed experiments with correlated errors as

$$AP_t = |\mathbf{U}'\mathbf{V}\mathbf{U}|\left(1-\frac{\mathbf{r}_t^{*'}(\mathbf{U}'\mathbf{V}\mathbf{U})^{-1}\mathbf{r}_t^{*}}{\mathbf{y}'\mathbf{V}\mathbf{y}}\right) \quad (25)$$

where $\mathbf{r}_t^{*}$ and $\mathbf{V}$ are as defined earlier.

2.2.1 Single outlier

Without loss of generality, we assume that the first observation in the first block is an outlier, then from (25) we get the AP_t-statistic for a single outlier as

$$AP_t = v_{11}\left(1 - \frac{v_{11} r_t^{*2}}{\mathbf{y}'\mathbf{V}\mathbf{y}}\right) \tag{26}$$

where v_{11} and r_t^* are as defined earlier.

3. CORRELATION STRUCTURE AND ITS ESTIMATION

As mentioned earlier that in designed experiments, there are many experimental situations in which the assumption of independence of observations gets violated. In field experiments, the observations are mutually correlated through some systematic pattern of environmental variations. We now discuss various types of such correlation structures. The correlation structures that may exist among the observations within a block are nearest neighbour, autoregressive and equi-correlated etc. However for illustration purpose, we consider only equi-correlation structure for the present study. In case of the equi-correlation structure, it is assumed that the same amount of correlation (ρ) exists between the observations within a block. The amount of correlation is constant for all pair of observations taken from a block. The correlation between $(y_{ij}, y_{j't'})$ in the same block is same.

$$Corr(y_{jt}, y_{jt'}) = \begin{cases} 1, & if\ t = t' \\ \rho_j, & otherwise \end{cases} \tag{27}$$

where t, $t' = 1, 2, ..., k_j$, $j = 1, 2, b$ and ρ_j is the correlation coefficient in the j^{th} block.

That is,

$$\Sigma_{k_j} = \sigma^2 \begin{bmatrix} 1 & \rho_j & \rho_j & ... & \rho_j & \rho_j \\ \rho_j & 1 & \rho_j & ... & \rho_j & \rho_j \\ \rho_j & \rho_j & 1 & ... & \rho_j & \rho_j \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_j & \rho_j & \rho_j & ... & 1 & \rho_j \\ \rho_j & \rho_j & \rho_j & ... & \rho_j & 1 \end{bmatrix} \tag{28}$$

To calculate Cook-statistic, we need to estimate Ω. This involved a particular structure of correlation of blocks. Therefore an estimate of the correlation coefficient ρ is required. One way to estimate this coefficient is to apply auto-correlation method. Thus ρ_j for j^{th} block can be estimated by the following formulae.

$$\hat{\rho}_j = \frac{\sum_{t=2}^{k_j} r_t r_{t-1}}{\sum_{t=2}^{k_j} r_{t-1}^2} \tag{29}$$

and estimate of σ^2 can be obtained as $\hat{\sigma}^2 = \mathbf{r'r}/(n - p)$, where $\mathbf{r}$ is the vector of residuals as obtained from the model (1) without considering the correlation structure, *i.e.*, $\mathbf{r} = (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}')\mathbf{y}$ and r_t is the t^{th} component of these residuals in the j^{th} block.

4. ILLUSTRATION

In this Section we illustrate this statistic through an example. An experiment with 5 manurial treatments was conducted in the randomized complete block (RCB) design with 4 replications in 2001 to evaluate the N, P and K status in the soil on paddy (net plot size 20.00 *m* × 5.00 *m*). The treatment details are as follows:

Treatment 1(T_1) = Control

Treatment 2(T_2) = 125 kg/ha of Nitrogen (N)

Treatment 3(T_3) = 125 kg/ha of N + 50 kg/ha of Phosphorous (P_2O_5)

Treatment 4(T_4) = 125 kg/ha of N + 50 kg/ha of Potash (K_2O)

Treatment 5(T_5) = 125 kg/ha of N + 50 kg/ha of P_2O_5 + 50 kg/ha of K_2O

The data is given in Table 1.

Table 1. Grain yield

	Block 1	Block 2	Block 3	Block 4
Treatment 1	27.80	28.40	28.80	28.60
Treatment 2	29.80	31.20	31.60	29.80
Treatment 3	31.20	30.60	33.20	32.40
Treatment 4	32.60	32.60	33.60	34.40
Treatment 5	32.80	33.80	33.00	33.60

The usual analysis of this data was carried out by considering that there is no correlation among the observations. The ANOVA is presented in Table 2. From the table, it is seen that the treatment effects are highly significant and block effects are significant at around 9% level of significance.

Table 2. ANOVA with original data

Source	Degrees of Freedom	Sum of Squares	Mean Square	F Value	Prob. > F
Treatment	4	68.088	17.022	33.289	<0.000001
Block	3	4.134	1.378	2.6949	0.0929276
Error	12	6.136	0.511		
Corrected Total	19	78.358			

We then calculated block wise equal correlation coefficients for observations. These correlation coefficients are –0.23 for the first block, –0.19 for the second block, 0.18 for the third block and 0.01 for the fourth block. Mean square error estimate under usual analysis has been used as the estimate of σ^2. This value is 0.511. Using these values the variance-covariance matrix has been computed. This variance-covariance matrix has been used to re-analyze the data under correlated error structure. The result of this analysis is presented in Table 3. From this table it is seen that treatment effects remain highly significant, whereas the significance level of block effects has been reduced to 6%.

Table 3. ANOVA with correlated observations

Source	Degrees of Freedom	Sum of Squares	Mean Square	F Value	Prob. > F
Treatment	4	128.806	32.201	16.735	0.00007
Block	3	18.640	6.213	3.229	0.060
Error	12	23.089	1.924		
Corrected Total	19	170.536			

We now applied Cook-statistic to detect outliers, if any. These Cook-statistics are presented in Table 4. Cook-statistics are worked out by writing programme in SAS IML. The distribution of Cook-statistic is approximately F-distribution with 3 and 12 degrees of freedom. The highest value of Cook-statistic is 0.33476 corresponding to observation number 17. This observation corresponds to the treatment number 2 in the fourth block. Comparing with F-distribution, we find that this observation is highly significant (Probability value of F(0.33476, 3, 12) = 0.15).

We then remove this observation and re-analyze the data under correlated error structure. Note that design remains connected after deletion of any single observation. Once a data point is deleted, the design becomes non-orthogonal. We analyzed this non-orthogonal data by GLM. Appropriate programme is written in IML of SAS software. The result is presented in Table 5.

The treatment effects remain highly significant, whereas block effects now become significant at about 4% level of significance.

Table 4. Cook-statistic

Observation	Cook-Statistic No.	Observation No.	Cook-Statistic
1	0.00077	11	0.00866
2	0.00366	12	0.10154
3	0.00136	13	0.25053
4	0.00064	14	0.03421
5	0.00902	15	0.29287
6	0.0064	16	0.00079
7	0.10782	17	**0.33476**
8	0.24795	18	0.01055
9	0.05322	19	0.19619
10	0.10257	20	0.00866

Table 5. ANOVA with correlated observations after deletion of observation. No. 17

Source	Degrees of Freedom	Sum of Squares	Mean Square	F Value	Prob. > F
Treatment	4	163.809	40.952	16.494	0.0001
Block	3	29.375	9.791	3.943	0.039
Error	12	27.310	2.482		
Corrected Total	19	220.495			

REFERENCES

Andrews, D.F. and Pregibon, D. (1978). Finding the outliers that matter. *J. Roy. Statist. Soc.,* **B40**, 87-93.

Atkinson, A.C. (1969). The use of residuals as concomitant variables. *Biometrika*, **56**, 33-41.

Bartlett, M.S. (1978). Nearest neighbour models in the analysis of field experiments (with discussion). *J. Roy. Statist. Soc.,* **B40**, 147-174.

Berenblut, I.I. and Webb, G.I. (1974). Experimental design in the presence of autocorrelated errors. *Biometrika*, **61**, 427-437.

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

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

Bhar, L. and Ojha, S. (2014). Outliers in multi-response experiments. *Comm. Statist.-Theory Methods*, **43(13)**, 2782-2798

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

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

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. *Comm. Statist.-Theory Methods*, **12(14)**, 1675-1683.

Gupta, S. and Mukerjee, R. (1989). *A Calculus for Factorial Arrangements*. Springer-Verlag, New York.

Herzberg, A.M. (1982). The design of experiments for correlated error structures: Layout and robustness. *Cand. J. Statist.,* **10(2)**, 133-138.

Kim, S.W. and Huggins, R. (1998). Diagnostics for autocorrelated regression models. *Austral. & New Zealand J. Statist.*, **40**, 65-71.

Martin, R.J. (1992). Leverage, influence and residuals in regression models when observations are correlated. *Comm. Statist.-Theory Methods.,* **21(5)**, 1183-1212.

Parsad, R., Nandi, P.K., Bhar, L. and Gupta, V.K. (2008). Outliers in multi-response experiments. *Stat. Appl.,* **6**, 275-292.

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*, **63(2)**, 177-191.

Schall, R. and Dunne, T.T. (1991). Diagnostics for regression-ARMA time series. In: *Directions in Robust Statistics and Diagnostics*. Ed. W. Stahel and S. Weisberg, Part **2,** 205-221.

Sen Roy, S. and Guria, S. (2004). Regression diagnostics in an autocorrelated model. *Braz. J. Prob. Stat.*, **18**, 103-112.

Sen Roy, S. and Guria, S. (2009). Estimation of regression parameters in the presence of outliers in the response. *Statistics*, **43(6),** 531-539.

Wilkinson, G.N., Eckert, S.R., Hancock, T.W. and Mayo, O. (1983). Nearest Neighbour (NN) analysis of field experiments. *J. Roy. Statist. Soc.,* **B45**, 151-211.

Williams, R.M. (1952). Experimental designs for serially correlated observations. *Biometrika*, **39**, 151-167.