Jour. Ind. Soc. Ag. Statistics
56(2), 2003 : 142-154

Study of Outliers Under Variance-inflation Model in Experimental Designs

Lalmohan Bhar and V.K. Gupta
Indian Agricultural Statistics Research Institute, New Delhi–110012
(Received : January, 2001)

SUMMARY

For detecting outliers in linear regression model, Cook [2] developed a statistic under 'mean-shift' model. This statistic is appropriately developed under 'variance-inflation' model for detecting outliers in the general linear model ($\mathbf{y}, \mathbf{X}\theta, \sigma^2\mathbf{I}$) for experimental designs. The simplified form of this statistic is obtained for both one-way and two-way elimination of heterogeneity settings. The relationship between Cook-statistic under 'mean-shift' model and Cook-statistic under 'variance-inflation' model for a single outlier is established and it has been shown that an observation which is an outlier under 'mean shift' model may not be so under 'variance-inflation' model. Through an example it has been shown that heteroscedasticity in the data may not influence the estimation of parameters.

Key words : Outlier, General linear model, Variance-inflation model, Cook-statistic.

1. Introduction

A common approach to modelling outliers in the fixed effects linear model is to assume that outliers result from slippages in the expected values of contaminated observations. However, this assumption may not hold good in every occasion; instead the variances of all the observations may not remain constant which may severely affect the estimation of parameters. It may be a common occurrence in agricultural experiments where all the plots in a block may not be uniform, causing inflated variances for some of the observations. We may term these observations as outliers. In case of 'mean-shift' single outlier model, the test-statistic for the presence of an outlier is a monotonic function of the largest absolute Studentized residual (Srikantan [10]). This is not always true for 'variance-inflation' model. Cook *et al.* [4] showed that if the largest absolute residual corresponds to the largest absolute Studentized residual, then the estimated outlier is same for both the models.

For detecting outliers in linear regression model, Cook [2] developed a statistic under 'mean-shift' model. In designed experiments, the experimenter is

generally interested only in the estimation of a subset of parameters, rather than the whole set of parameters. One may, therefore, be interested to assess the influence of an outlying observation on the estimation of such function of subset of parameters. Since the design matrix in experimental designs is deficient in rank, the statistic developed by Cook [2] cannot be applied to it. Bhar and Gupta [1] modified this test-statistic for detecting outliers in experimental designs. However, no diagnostic measures seem to be developed so far for dealing with outliers under 'variance-inflation' model. Cook *et al.* [4] studied such outliers in linear regression model. Their study, however, was confined to the behaviour of the estimated outlier from 'mean-shift' model under 'variance-inflation' model. In the present paper we develop Cook-statistic under 'variance-inflation' model for detecting outliers in experimental designs. It is important to note that 'variance-inflation' model is actually a heteroscedastic model. However, our interest here is not in the estimation of parameters under a heteroscedastic model; instead we are interested to see whether inflated variances of some of the observations cause a severe effect on the estimation of a subset of parameters assuming that the model is homoscedastic.

In Section 2 this statistic has been developed under a general linear model set up of experimental designs. Section 3 considers the application of this statistic to some specific designs. The paper concludes with a Section on Discussion.

Throughout we use the following notations. All matrices and vectors are real, vectors being written as column vectors. We denote an n-component vector of all unities by $\mathbf{1}_n$, by $\mathbf{J}_n$ an $n \times n$ matrix of all unities and by $\mathbf{I}_n$ an identity matrix of order n. Matrices $\mathbf{A}'$, $\xi(\mathbf{A})$, $\mathbf{A}^-$ and $\mathbf{A}^+$ will respectively denote the transpose, column space (range), a generalized inverse (g-inverse) and the Moore-Penrose inverse of $\mathbf{A}$.

2. *Development of Cook – statistic*

2.1 Cook-statistic in Mean-shift Model

Consider the general linear model

$$\mathbf{y} = \mathbf{X}\theta + \mathbf{e};\ E(\mathbf{e}) = \mathbf{0};\ D(\mathbf{e}) = \sigma^2\mathbf{I}_n;\ \sigma^2 > 0 \tag{2.1}$$

where $\mathbf{y}$ is an $n \times 1$ vector of observations, $\mathbf{X}$ is an $n \times p$ matrix of known constants with full column rank p, θ is a $p \times 1$ vector of unknown parameters, and $\mathbf{e}$ is an $n \times 1$ vector of independent random variables each with zero mean and common variance $\sigma^2 (>0)$.

Cook [2] defined a statistic which measures the effect of an outlying observation on the estimation of parameters of the model as

$$D_i = \frac{\left(\hat{\theta} - \hat{\theta}_{(i)}\right)' \left[D(\hat{\theta})\right]^{-1} \left(\hat{\theta} - \hat{\theta}_{(i)}\right)}{\text{Rank}\left[D(\hat{\theta})\right]} \tag{2.2}$$

where $\hat{\theta}$ and $\hat{\theta}_{(i)}$ are the least-square estimates of θ with and without the i-th data point respectively and D(**x**) denotes the dispersion matrix of **x**. The statistic provides a measure of distance between $\hat{\theta}$ and $\hat{\theta}_{(i)}$ in terms of descriptive levels of significance, because D_i is actually $(1-\alpha)\times 100\%$ confidence ellipsoid for the vector θ under normal theory, which satisfies $D_i \le F(p, n-p, (1-\alpha))$. Extension of D_i to more than one outlier is straight forward. For usual interpretation of Cook-statistic, see Cook ([2], [3]).

2.2 Cook-statistic in Variance-inflation Model

Now consider the general linear model (2.1), for an experimental design d (say) with the design matrix **X** deficient in rank, i.e., Rank(**X**) = m(< p). Let $\theta = (\theta_1' \quad \theta_2')'$, where θ_1 be a v-component vector containing all parameters of interest to the experimenter and θ_2 be a (p – v) component vector containing the set of nuisance parameters in the model that are not of much interest to the experimenter. Thus

$$\mathbf{y} = (\mathbf{X}_1 \quad \mathbf{X}_2)\begin{pmatrix}\theta_1 \\ \theta_2\end{pmatrix} + \mathbf{e} \tag{2.3}$$

where **X** is partitioned in conformity with the partition of the parameter vector. We also assume that the ranks of $\mathbf{X}_1$ and $\mathbf{X}_2$ are m_1 and m_2 respectively. From the normal equations $\mathbf{X'X}\theta = \mathbf{X'y}$ we obtain, on eliminating θ_2, the equations involving only θ_1 as

$$\mathbf{C}_{\theta_1}\theta_1 = \mathbf{Q}_{\theta_1} \tag{2.4}$$

where $\mathbf{C}_{\theta_1} = \mathbf{X}_1'\mathbf{B}\mathbf{X}_1$, $\mathbf{Q}_{\theta_1} = \mathbf{X}_1'\mathbf{B}\mathbf{y}$ and $\mathbf{B} = \mathbf{I}_n - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-}\mathbf{X}_2'$

The matrix **B** is symmetric and idempotent and the matrix $\mathbf{C}_{\theta_1}$ is symmetric. Since the linear model considered here is for experimental designs, it is not unrealistic to assume that the column space of $\mathbf{X}_1$ and $\mathbf{X}_2$ contain the vector **1**, thus $\mathbf{C}_{\theta_1}\mathbf{1} = \mathbf{0}$.

Suppose that $t\{t \le \min(m_1, m_2)\}$ of the n observations are suspected to be outliers in the sense that the variance of each of the t observations is shifted from the error variance of others. Then we have the variance-inflation model as

$$\mathbf{y} = (\mathbf{X}_1 \quad \mathbf{X}_2)\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} + \mathbf{z} \tag{2.5}$$

where $E(\mathbf{z}) = \mathbf{0}$, $D(\mathbf{z}) = \sigma^2 \begin{bmatrix}\mathbf{W} & \mathbf{0}'\\ \mathbf{0} & \mathbf{I}_{n-t}\end{bmatrix}$ and $\mathbf{W} = \text{diag}(w_1, ..., w_t)$ with $\sigma^2 > 0$, and $w_i > 1$; $\forall$ i = 1, ..., t are unknown. We also assume that the distribution of $\mathbf{z}$ is normal with mean vector $\mathbf{0}$ and dispersion matrix as given above. The model (2.5) can alternatively be written as

$$\mathbf{y} = (\mathbf{X}_1 \quad \mathbf{X}_2)\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} + \mathbf{U}\delta + \mathbf{e} \tag{2.6}$$

where $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_t)$; $h_Y^2 = (0, ..., 1(i-th), 0, ...0)'$; i = 1, ..., t

and $\delta = (\delta_1, ..., \delta_t)'$

Now δ_i's are distributed normally with mean 0 and common variance $\sigma^2(w_i - 1)$; $\forall$ i = 1, ..., t; and the components of $\mathbf{e}$ are as usual distributed normally with mean 0 and variance σ^2. Thus from (2.5) it follows that

$$D(\mathbf{y}) = {}^{x}{}_{3}\begin{bmatrix}\mathbf{W} & \mathbf{0}'\\ \mathbf{0} & \mathbf{I}_{n-t}\end{bmatrix} = \sigma^2 [\mathbf{I}_n - \mathbf{U}(\mathbf{I}_t - \mathbf{W})\mathbf{U}'] = \Sigma \text{ (say)} \tag{2.7}$$

Clearly Σ is non-singular and , δ and $\mathbf{e}$ are independent by assumption.

$$\text{Now}, \Sigma^{-1} = \frac{1}{\sigma^2}\mathbf{I}_n + \frac{1}{\sigma^2}\mathbf{U}\mathbf{W}_*\mathbf{U}' \tag{2.8}$$

where $$\mathbf{W}_* = \text{diag}\left(\frac{1-w_1}{w_1}, ..., \frac{1-w_t}{w_t}\right) \tag{2.9}$$

Now from (2.5) the normal equations for estimating the linear functions of the parameters θ_1 are

$$\mathbf{C}_{\theta_1}^{H}\theta_1 = \mathbf{Q}_{\theta_1}^{H} \tag{2.10}$$

where $$\mathbf{C}_{\theta_1}^{H} = \left[\mathbf{X}_1'\Sigma^{-1}\mathbf{X}_1\right] - \left[\mathbf{X}_1'\Sigma^{-1}\mathbf{X}_2\right]\left[\mathbf{X}_2'\Sigma^{-1}\mathbf{X}_2\right]^{-}\left[\mathbf{X}_2'\Sigma^{-1}\mathbf{X}_1\right] \tag{2.11}$$

and $$\mathbf{Q}_{\theta_1}^{H} = \left[\mathbf{X}_1'\Sigma^{-1}\mathbf{y}\right] - \left[\mathbf{X}_1'\Sigma^{-1}\mathbf{X}_2\right]\left[\mathbf{X}_2'\Sigma^{-1}\mathbf{X}_2\right]^{-}\left[\mathbf{X}_2'\Sigma^{-1}\mathbf{y}\right] \tag{2.12}$$

Now $\mathbf{C}_{\theta_1}^{H}$ and $\mathbf{Q}_{\theta_1}^{H}$ can alternatively be written as

$$\mathbf{C}_{\theta_1}^{H} = (\mathbf{C}_{\theta_1} + \mathbf{L}\mathbf{F}^{-1}\mathbf{L}')/\sigma^2 \tag{2.13}$$

$$\mathbf{Q}_{\theta_1}^{H} = (\mathbf{Q}_{\theta_1} + \mathbf{L}\ \mathbf{F}^{-1}\ \mathbf{U}_*'\mathbf{B}\mathbf{y})/\sigma^2 \tag{2.14}$$

where $\mathbf{L} = \mathbf{X}_1'\mathbf{B}\mathbf{U}_*$, $\mathbf{F} = \mathbf{I} + \mathbf{U}_*'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-}\mathbf{X}_2'\mathbf{U}_*$, $\mathbf{U}_* = \mathbf{U}\mathbf{W}_*^{1/2}$ and $\mathbf{C}_{\theta_1}$ and $\mathbf{Q}_{\theta_1}$ are the corresponding matrices under the homoscedastic model. Clearly the matrix $\mathbf{F}$ is non- singular.

Now we have the following theorem

Theorem 2.1. (i) $\mathrm{E}\left(\mathbf{Q}_{\theta_1}^{H}\right) = \mathbf{C}_{\theta_1}^{H}\ \theta_1$

(ii) $\mathrm{D}\left(\mathbf{Q}_{\theta_1}^{H}\right) = \mathbf{C}_{\theta_1}^{H}$

We assume that the design d is connected and the rank of $\mathbf{C}_{\theta_1}^{H}$ is $(v-1)$. This can be checked by the connectedness property of the design under homoscedastic model, since a design that is connected under homoscedastic model is also connected under heteroscedastic model (see Gupta [7]). Let $\mathbf{P}\theta_1$ be the set of all $(v-1)$ orthonormalized contrasts of the parameters θ_1. The $(v-1)\times v$ matrix $\mathbf{P}$ 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 the least square estimator of $\mathbf{P}\theta_1$ is given by $\mathbf{P}\hat{\theta}_1$, where $\hat{\theta}_1$ is any solution of the normal equation (2.10). Since the least square estimator is BLUE, the set is also BLUE. Then we have

$$\mathbf{P}\hat{\theta}_1 = \mathbf{P}(\mathbf{C}_{\theta_1}^{H})^{-}\ \mathrm{r}_{P_3} = \mathbf{P}(\mathbf{C}_{\theta_1} + \mathbf{L}\mathbf{F}^{-1}\mathbf{L}')^{+}(\mathbf{Q}_{\theta_1} + \mathbf{L}\ \mathbf{F}^{-1}\ \mathbf{U}_*'\mathbf{B}\mathbf{y})$$

Now taking the Moore-Penrose inverse of $\mathbf{C}_{\theta_1} + \mathbf{L}\mathbf{F}^{-1}\mathbf{L}'$, we get

$$\mathbf{P}\hat{\theta}_1 = \mathbf{P}(\mathbf{C}_{\theta_1}^{+} - \mathbf{C}_{\theta_1}^{+}\mathbf{L}\mathbf{E}^{-1}\mathbf{L}'\mathbf{C}_{\theta_1}^{+})(\mathbf{Q}_{\theta_1} + \mathbf{L}\mathbf{F}^{-1}\mathbf{U}_*'\mathbf{B}\mathbf{y}) \tag{2.15}$$

where $\mathbf{E} = \mathbf{F} + \mathbf{L}'\mathbf{C}_{\theta_1}^{+}\mathbf{L}$ and clearly it is non- singular.

Now we give the following lemma.

Lemma 2.1. $\mathrm{D}(\mathbf{P}\hat{\theta}_1) = [\mathbf{P}(\mathbf{C}_{\theta_1} + \mathbf{L}\mathbf{F}^{-1}\mathbf{L}')\mathbf{P}']^{-1}\sigma^2$

Now we consider another model in which the outliers are actually omitted

$$\mathbf{y}_{(t)} = \begin{pmatrix} \mathbf{X}_{1(t)} & \mathbf{X}_{2(t)} \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} + \mathbf{z}_{(t)} \tag{2.16}$$

where (t) denotes the removal of t rows from the corresponding vector and matrices. Note that the dispersion matrix $\Sigma_{(t)}$ now becomes $\sigma^2 \mathbf{I}_{n-t}$. The normal equations for estimating the contrasts of the parameters θ_1 under the model (2.16) are given by

$$\mathbf{C}^{H}_{\theta_{1t}} \theta_{1t} = \mathbf{Q}^{H}_{\theta_{1t}} \tag{2.17}$$

where $$\mathbf{C}^{H}_{\theta_{1t}} = \mathbf{X}'_{1(t)}\mathbf{X}_{1(t)} - \mathbf{X}'_{1(t)}\mathbf{X}_{2(t)}(\mathbf{X}'_{2(t)}\mathbf{X}_{2(t)})^{-}\mathbf{X}'_{2(t)}\mathbf{X}_{1(t)} \tag{2.18}$$

and $$\mathbf{Q}^{H}_{\theta_{1t}} = \mathbf{X}'_{1(t)}\mathbf{y}_{1(t)} - \mathbf{X}'_{1(t)}\mathbf{X}_{2(t)}(\mathbf{X}'_{2(t)}\mathbf{X}_{2(t)})^{-}\mathbf{X}'_{2(t)}\mathbf{y}_{1(t)} \tag{2.19}$$

The matrices $\mathbf{C}^{H}_{\theta_{1t}}$ and $\mathbf{Q}^{H}_{\theta_{1t}}$ can alternatively be written as

$$\mathbf{C}^{H}_{\theta_{1t}} = \mathbf{C}_{\theta_1} - \mathbf{L}_*\mathbf{F}_*^{-1}\mathbf{L}'_* \tag{2.20}$$

and $$\mathbf{Q}^{H}_{\theta_{1t}} = \mathbf{Q}_{\theta_1} - \mathbf{L}_*\mathbf{F}_*^{-1}\mathbf{U}'\mathbf{B}\mathbf{y} \tag{2.21}$$

where $\mathbf{L}_* = \mathbf{X}'_1\mathbf{B}\mathbf{U}$ and $\mathbf{F}_* = \mathbf{U}'\mathbf{B}\mathbf{U}$ (2.22)

We also assume that the residual design obtained after deleting t outlying observations remains connected. This fact can be checked by the connectedness property of the residual design under homoscedasticity, since it is known that if a design obtained after deleting t observations is connected under homoscedastic model it is also connected under heteroscedastic model (see Lal *et al.* [8]). The set of all orthonormalized contrasts for the parameters θ_1 is given by $\mathbf{P}\theta_{1t}$. If $\hat{\theta}_{1t}$ is any solution of the normal equations (2.17), then

$$\mathbf{P}\hat{\theta}_{1t} = \mathbf{P}(\mathbf{C}^{H}_{\theta_{1t}})^{-}\mathbf{Q}^{H}_{\theta_{1t}} = \mathbf{P}(\mathbf{C}_{\theta_1} - \mathbf{L}_*\mathbf{F}_*^{-1}\mathbf{L}'_*)^{+}(\mathbf{Q}_{\theta_1} - \mathbf{L}_*\mathbf{F}_*^{-1}\mathbf{U}'\mathbf{B}\mathbf{y}) \tag{2.23}$$

Now taking the Moore-Penrose inverse of $\mathbf{C}_{\theta_1} - \mathbf{L}_*\mathbf{F}_*^{-1}\mathbf{L}'_*$ we get

$$\mathbf{P}\hat{\theta}_{1t} = \mathbf{P}(\mathbf{C}^{+}_{\theta_1} + \mathbf{C}^{+}_{\theta_1}\mathbf{L}_*\mathbf{E}_*^{-1}\mathbf{L}'_*\mathbf{C}^{+}_{\theta_1})(\mathbf{Q}_{\theta_1} - \mathbf{L}_*\mathbf{F}_*^{-1}\mathbf{U}'\mathbf{B}\mathbf{y}) \tag{2.24}$$

where $$\mathbf{E}_* = \mathbf{F}_* - \mathbf{L}'_*\mathbf{C}^{+}_{\theta_1}\mathbf{L}_* \tag{2.25}$$

Lemma 2.2. The difference between the estimators of the contrasts under the model (2.6) and (2.16) can be expressed as

$$\mathbf{P}\hat{\theta}_1 - \mathbf{P}\hat{\theta}_{1t} = -\mathbf{P}\mathbf{C}^{+}_{\theta_1}\mathbf{G}\mathbf{B}\mathbf{y}$$

where $$\mathbf{G} = (\mathbf{L}\mathbf{E}\mathbf{L}' + \mathbf{L}_*\mathbf{E}_*\mathbf{L}'_*)\mathbf{C}^{+}_{\theta_1}\mathbf{X}'_1 - \mathbf{L}\mathbf{F}^{-1}\mathbf{U}'_* - \mathbf{L}_*\mathbf{F}_*^{-1}\mathbf{U}' + \mathbf{L}\mathbf{E}\mathbf{L}'\mathbf{C}^{+}_{\theta_1}\mathbf{L}\mathbf{F}^{-1}\mathbf{U}'_* - \mathbf{L}_*\mathbf{E}_*\mathbf{L}'_*\mathbf{C}^{+}_{\theta_1}\mathbf{L}_*\mathbf{F}_*^{-1}\mathbf{U}' \tag{2.26}$$

Cook-statistic: From (2.2) and using Lemma 2.2 we get Cook-statistic under variance-inflation model as

$$D_t^H = \mathbf{y'BG'}\, C_{\theta_t}^{+}\, \mathbf{P'}\left[D(\mathbf{P}\hat{\theta}_t)\right]^{-1} \mathbf{PC}_{\theta_t}^{+}\mathbf{GBy}/(v-1)$$

$$= \mathbf{y'BG'}\, C_{\theta_t}^{+}\, [C_{\theta_t} + \mathbf{LF}^{-1}\mathbf{L'}]\, C_{\theta_t}^{+}\mathbf{GBy}/(v-1)\hat{\sigma}^2 \qquad (2.27)$$

where $\hat{\sigma}^2$ is substituted for σ^2.

Remark 2.1. Suppose that only a single outlier is present. Without loss of generality we assume that the 1st observation is an outlier. Thus $\mathbf{U} = \mathbf{u}_1$. Then Cook- statistic takes the following form

$$D_1^H = d\left(\mathbf{L}_*'C_{\theta_t}^{+}\mathbf{Q}_{\theta_t} - \mathbf{u}_1'\mathbf{By}\right)^2 \left(\mathbf{L}_*'C_{\theta_t}^{+}\mathbf{L}_*\right) \qquad (2.28)$$

where $$d = d_0\left\{1 + \left(\frac{1}{1-w} - \mathbf{u}_1'\mathbf{Bu}_1\right)^{-1}\mathbf{L}_*'C_{\theta_t}^{+}\mathbf{L}_*\right\}$$

$$d_0 = \left\{\left(\frac{1}{1-w} - \alpha\right)^{-1} + \alpha^{-1}\right\}^2$$

$$\alpha = (\mathbf{u}_1'\mathbf{Bu}_1 - \mathbf{L}_*'C_{\theta_t}^{+}\mathbf{L}_*) \text{ and } \mathbf{L}_* = \mathbf{X}_1'\mathbf{Bu}_1 \qquad (2.29)$$

Remark 2.2. If all the w_i's are equal to 1, then $w_* = 0$ and the Cook-statistic can be written as

$$D_1^H = \frac{\mathbf{y'VU(U'VU)^{-1}U'BX}_1 C_{\theta_t}^{+}\mathbf{X}_1'\mathbf{BU(U'VU)^{-1}U'Vy}}{(v-1)\hat{\sigma}^2} = D_1 \qquad (2.30)$$

Thus variance-inflation model is transformed into mean-shift model (Bhar and Gupta [1]).

Remark 2.3. Cook-statistic under 'mean-shift' model, i.e., D_t actually represents the volume of confidence ellipsoid (see Cook and Weisberg [5]) and thus $D_t \le F(1-\alpha, p, n_e)$, where p is the rank of the corresponding dispersion matrix and n_e is the degrees of freedom (df) for estimating σ^2. Cook and Weisberg [5] suggest to compare D_t with critical value of F. Now D_t^H also represents confidence ellipsoid under heteroscedastic model. Thus the critical value of F can be taken as good approximation for D_t^H. In the literature it is found that σ^2 under heteroscedastic model is estimated with the same df as it is estimated under homoscedastic model (see, e.g. Sen and Srivastava [9]). Moreover, here w_i's are estimated either assuming that they are equal or

proportional to some character. Thus they may be treated as adhoc estimates and df may be taken as it is taken under homoscedastic model. We, therefore, suggest that D_i^H as given in (2.27) may be compared with the critical value of F with v-1 and n_e df.

Estimation of w

In practice w_i's are unknown. These are to be estimated from the data. Here we obtain the estimates of w_i's by assuming that all w_i's are equal. The extra sum of squares due to the parameters δ under the model (2.6) is given by s_d^2, where $s_d^2 = \mathbf{y'VU(U'VU)^{-1}U'Vy}$ and $\mathbf{V = I - X(X'X)^{-}X'}$. Under the variance-inflation model s_d^2 has the expectation

$$E(s_d^2) = \sigma^2 + tr(\mathbf{U'VU})\sigma_d^2, \text{ where } \sigma_d^2 = \sigma^2(w-1) \tag{2.31}$$

where tr(.) denotes the trace of a square matrix. Since, the variance-covariance matrix of **y** is now

$D(\mathbf{y}) = [\mathbf{I}_n + (w-1)\mathbf{UU'}]\,\sigma^2 = \sigma^2\,\mathbf{I}_n + \mathbf{UU'}\,\sigma_d^2$, an estimate of σ_d^2 is obtained as

$$\hat{\sigma}_d^2 = \frac{(s_d^2 - \hat{\sigma}^2)}{tr(\mathbf{U'VU})}, \text{ i.e., } \hat{\sigma}^2(w-1) = \frac{(s_d^2 - \hat{\sigma}^2)}{tr(\mathbf{U'VU})} \tag{2.32}$$

$$\text{Finally, } \hat{w} = 1 + \frac{(s_d^2 - \hat{\sigma}^2)}{\hat{\sigma}^2 tr(\mathbf{U'VU})} \tag{2.33}$$

It is indeed possible that $\hat{\sigma}_d^2$ is negative and $\hat{w}$ is negative. However, we take the absolute value of $\hat{w}$ as an estimate of w.

3 Outliers in Some Specific Designs

In the last section, we have derived Cook-statistic for outlier detection in general linear models. In the present section, we apply this test statistic to designs for both one-way and two-way elimination of heterogeneity settings.

3.1 Outliers in Designs for One-Way Elimination of Heterogeneity

3.1.1 Some Preliminaries

Consider the usual intra-block model of n-observations

$$\mathbf{y} = \mu\mathbf{1} + \Delta'\tau + \mathbf{D}'\beta + \mathbf{e} \tag{3.1}$$

with $E(\mathbf{e}) = \mathbf{0}$ and $D(\mathbf{e}) = \sigma^2\,\mathbf{I}_n$

Here Δ' is an $n \times v$ $(0 - 1)$ design matrix for treatments, $\mathbf{D}'$ is an $n \times b\,(0-1)$ design matrix for blocks, μ is the general mean, τ is a

v-component vector of treatment effects and β is a b-component vector of block effects. Also, $\Delta'\mathbf{1} = \mathbf{1}$, $\mathbf{D}'\mathbf{1} = \mathbf{1}$, $\Delta\mathbf{1} = \mathbf{r}$, $\mathbf{D}\mathbf{1} = \mathbf{k}$, where $\mathbf{r} = (r_1, \ldots, r_v)'$ and $\mathbf{k} = (k_1, \ldots, k_b)'$ are the vectors of replications and block sizes, respectively. We also define $\mathbf{N} = \Delta\mathbf{D}' = ((n_{ij}))$, where the non-negative integers n_{ij} denote the number of times i-th treatment appears in the j-th block; $\sum_j n_{ij} = r_i$, $\forall i = 1, \ldots, v$; $\sum_i n_{ij} = k_j, \forall\, j = 1, \ldots, b$. Here the parameter vector of interest is τ, and the normal equations for τ is obtained from (2.4) as

$$\mathbf{C}_\tau \tau = \mathbf{Q}_\tau \tag{3.2}$$

where $\mathbf{C}_\tau = \Delta\mathbf{B}\Delta'$, $\mathbf{Q}_\tau = \Delta\mathbf{B}\mathbf{y}$ and $\mathbf{B} = \mathbf{I} - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'$ (3.3)

3.1.2 Single Outlier in Binary Block Designs

Cook-statistic: Without any loss of generality we assume that the observation pertaining to the first treatment in the first block is an outlier. Then the incidence matrix $\mathbf{N}$ can be written as

$$\mathbf{N} = \begin{bmatrix} 1 & \in' \\ \mathbf{f} & \mathbf{N}_0 \end{bmatrix} \tag{3.4}$$

where, $\mathbf{f}$ is a $(v-1)$ component $(0-1)$ vector of incidence of remaining $(v-1)$ treatments in the first block, $\in$ is a $(b-1)$ component $(0-1)$ vector of incidence of the first treatment in the remaining $(b-1)$ blocks and $\mathbf{N}_0$ is the incidence matrix of the remaining $(v-1)$ treatments in the remaining $(b-1)$ blocks.

The matrix $\mathbf{L}_*$ given in (2.22) becomes in the present case as

$$\mathbf{L}_* = \Delta\mathbf{B}\mathbf{u}_1 = \frac{1}{k_1}\begin{pmatrix} k_1 - 1 \\ -\mathbf{f} \end{pmatrix} = \begin{pmatrix} k_1 - 1 \\ k_1 \end{pmatrix}^{\frac{1}{2}} \mathbf{U}_0 \tag{3.5}$$

where $$\mathbf{U}_0 = \{k_1(k_1-1)\}^{-1/2}\begin{pmatrix} k_1 - 1 \\ -\mathbf{f} \end{pmatrix} \tag{3.6}$$

Thus $$\mathbf{L}_*'\mathbf{C}_{\theta_1}^{+}\mathbf{L}_* = \begin{pmatrix} k_1 - 1 \\ k_1 \end{pmatrix}\mathbf{U}_0'\mathbf{C}_\tau^{+}\mathbf{U}_0$$

$$\mathbf{u}_1'\mathbf{B}\mathbf{u}_1 = \frac{k_1 - 1}{k_1} \text{ and } \left(\mathbf{L}_*'\mathbf{C}_{\theta_1}^{+}\mathbf{Q}_{\theta_1} - \mathbf{u}_1'\mathbf{B}\mathbf{y}\right) = \mathbf{u}_1'\mathbf{V}\mathbf{y} = -r_1^*$$

where r_1^* is the first ordinary residual. Thus from (2.28) we get

$$D_1^H = d_0 \frac{\mathbf{U}_0'\mathbf{C}_\tau^+\mathbf{U}_0}{1-\mathbf{U}_0'\mathbf{C}_\tau^+\mathbf{U}_0}\frac{t_1^2}{v-1} = d_0 D_1 \tag{3.7}$$

where t_1^2 is the Studentized residual for the first observation

$$d_0 = \left(\frac{k_1}{wk_1 - w + 1}\right)^2 \left(\frac{1}{1+\frac{(1-w)(k_1-1)}{wk_1-w+1}\mathbf{U}_0'\mathbf{C}_\tau^+\mathbf{U}_0}\right) \tag{3.8}$$

and D_1 is the corresponding Cook-statistic under the homoscedastic model.

Lemma 3.1. $D_1^H \le D_1$

Proof: d_0 given in (3.8) can alternatively be written as

$$d_0 = \left(\frac{1}{1+(w-1)\left(1-\frac{1}{k_1}\right)}\right)\left(\frac{1}{\left\{1+(w-1)\left(1-\frac{1}{k_1}\right)\right\}(1-\mathbf{U}_0'\mathbf{C}_\tau^+\mathbf{U}_0)}\right)$$

Since the residual design after deleting the outlying observation remains connected, $(1-\mathbf{U}_0'\mathbf{C}_\tau^+\mathbf{U}_0) > 0$ (cf. Dey [6]). Again $w > 1$, therefore $d_0 \le 1$. Hence the proof follows from (3.7).

Remark 3.1. From Lemma 3.1 it is clear that an observation which is an outlier under mean-shift model may not be so under variance-inflation model. This was also observed by Cook *et al.* [4] in case of regression model.

3.1.3 Multiple Outliers in Variance-inflation Model

If the magnitude of the variance of each outlying observation, *i.e.*, w_i is known then the expression for D_t^H given in (2.28) can be directly used for testing the influence of these outliers. But if the values of w_i's are unknown, then these are to be estimated from the data, which is extremely difficult. However, a convenient formula for estimating the magnitude of inflated variance can be obtained once we assume the equality of these magnitudes, and for that the estimating procedure is given in Section 2. Thus replacing the matrices $\mathbf{C}_{\theta_i}$ and $\mathbf{B}$ in the expression (2.28) by appropriate matrices for designs for one-way elimination of heterogeneity setting we can easily test for outliers.

Remark 3.2. Proceeding on the same lines as for designs for one-way elimination of heterogeneity, we can easily show that the Cook-statistic for designs for two-way elimination of heterogeneity can be written as

$$D_i^H = d_0^* \frac{U_i^{*'} C_{\theta_i}^+ U_i^*}{\left(g_2 / g_1 - U_i^{*'} C_{\theta_i}^+ U_i^* \right)} \frac{t_i^2}{v-1} = d_0^* D_i \qquad (3.9)$$

where $U_i^* = \frac{1}{\sqrt{g_1}} X_1' B u_i$

$$d_0^* = \left(\frac{1}{\{1-(1-w)g_2\}^2} \right) \left(\frac{1}{1 + \frac{(1-w)g_1}{1-(1-w)g_2} U_i^{*'} C_{\theta_i}^+ U_i^*} \right)$$

g_1 is the norm of the vector $X_1'Bu_i$, g_2 is the i-th diagonal element of B, D_i is the corresponding Cook-statistic under 'mean-shift' model. Here the matrices $C_{\theta_i}^+$, X_1 and B are to be changed appropriately with the corresponding matrices of designs for two-way elimination of heterogeneity.

4. Discussion

In this section, we discuss various aspects of heteroscedasticity in data generated from designed experiments. We begin with the following data set. Table 1 contains the grain yield (in Kg.) of rice variety IR8 with six different rates of seeding, from a Randomized Complete Block Design experiment with four replications. Figure within the parentheses indicates the observation number.

Table 1. Grain yield of rice variety IR8 (in Kg.)

Treatment No.	Replication No.			
	1	2	3	4
1	5113(1)	5398(7)	5307(13)	4678(19)
2	5346(2)	5952(8)	4719(14)	4264(20)
3	5272(3)	5713(9)	5483(15)	4749(21)
4	5164(4)	4831(10)	4986(16)	4410(22)
5	4804(5)	4848(11)	4432(17)	4748(23)
6	5254(6)	4542(12)	4919(18)	4098(24)

We compute the Cook-statistic given in (3.7) for all the 24 observations. The observations are numbered block wise in the ascending order of treatments. The largest value of Cook-statistic among the 24 observations is 0.018 and it pertains to observation number 8 corresponding to the treatment number 2 in the second block. The table value of F for 5 and 16 degrees of freedom and at 5% level of significance is 0.21. It, therefore, follows that none of the 24 observations is influential due to inflated variance.

We now consider the possibility of the presence of two outliers. We calculate Cook-statistic for two observations obtained from (2.27) for all

possible pair of observations. It is found that the maximum value of Cook-statistic is 0.201 corresponding to the pair of observations (2, 8). Again we find that two observations at a time are not influential. Naturally, a question may arise as to whether the data really exhibits heteroscedasticity? To ascertain this we construct a simple F-test taking mean square error (mse) for 2 outlying observations and mse for remaining 22 observations and perform all possible ${}^{n}c_{2}$ F-tests. Following pairs of observations are found to be significant: (2,13), (3, 6), (3, 13) (5, 10), (7, 24), (8, 12), (8, 17), (8, 19), (8, 20), (8,22), (8, 24), (9, 17), (9, 20), (9, 22), (9, 24), (14, 19), (14, 21), (14, 23), (17, 22) and (21, 23). This result reveals one point clearly that the heteroscedasticity may not always affect the estimation of parameters. Even if heteroscedasticity is confirmed, we should not analyze the data assuming the model as heteroscedastic. Instead, we should check whether this heteroscedasticity really affects the parameter estimation or not. The present study provides a statistical tool to assess the influence by the presence of inflated variance of a few observations. If the heteroscedasticity of some observations does not affect the parameter estimation, we do not bother about their presence. We can as usual carry out our analysis assuming the model as homoscedastic.

Generally, in a particular situation we expect the presence of only a few outlying observations. For detecting t outliers through this statistic, one has to consider Cook-statistic for all ${}^{n}c_{t}$ sets. But before analysing the data one can test for heterogeneity of error variance. Some simple tests can be constructed for this purpose. For example, if w_i's are all equal, then one can construct an F-test as described earlier. This can also be generalized to the case when all w_i's are not equal. Thus one can identify the subsets of data, in which heteroscedasticity is present, and Cook-statistic can be applied to that particular subset and not to all the ${}^{n}c_{t}$ subsets.

In general if heteroscedasticity is confirmed, we take some appropriate measures such as variance stabilizing transformations and so on. But before applying transformation, one should check whether this heteroscedasticity is really influential in terms of parameter estimation. If this heteroscedasticity is not influential, we need not transform the data.

Another point worth noting is that sometimes heteroscedasticity arises due to the fact that some important covariates are not included in the model. For example inflated variance may occur due to heterogeneous plots within a block as explained in the introduction. But variance may not be inflated for all the plots. Moreover, it may not happen in all the blocks. Thus, one should identify the outlying observations in a particular block and then test whether they are influential or not. If they are influential, then one can include size or some other character of plot to ensure homoscedasticity.

ACKNOWLEDGEMENT

The authors are grateful to the referee for making valuable suggestions that helped to improve the results considerably.

REFERENCES

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

[2] Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics*, **21**, 15–18.

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

[4] Cook, R. D., Holschuh, N. and Weisberg, S. (1982). A note on an alternative outlier model. *J. Roy. Statist. Soc.*, **B44**, 370–376.

[5] Cook, R. D. and Weisberg, S. (1982). *Residual and Influence in Regression*. Chapman and Hall, New York.

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

[7] Gupta, V. K. (1995). Universally optimal designs under heteroscedastic model. *Sankhya*, **B57**, 420–427.

[8] Lal, K., Gupta, V. K. and Bhar, L. (2001). Robustness of designed experiments against missing data. *J. Appl. Statist.*, **28**, 63–79.

[9] Sen, A. and Srivastava, M. (1990). *Regression Analysis: Theory, Methods, and Applications*. Springer-Verlag, New York, 120.

[10] Srikantan, K. S. (1961). Testing for a single outlier in a regression model. *Sankhya*, **A23**, 251–260.