Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/050311

O-Blue for Outlier Tests

Arvind Pandey[1,*] *and Nibha Srivastava*[2]

[1] Department of Statistics, Pachhunga University College, MZU, Aizawl-796001, India.
[2] Department of Statistics and Planing Implementation, UP, Allahabad, India.

Received: 18 Jul. 2016, Revised: 29 Sep. 2016, Accepted: 1 Oct. 2016
Published online: 1 Nov. 2016

Abstract: The O-BLUEs defined by Moussa-Hamouda and Leone(1974) are considered and the effect of an outlying observation in these estimates are studied for a regression model. Then these estimates are used in developing two outlier test procedures. The results are highlighted with an example. The power of these procedures are studied and the power values for the same example are also tabulated.

Keywords: Linear model, likelihood ratio test, O-BLUEs, Outlier

1 Introduction

Consider a general linear model $\mathbf{y} = \mathbf{X}'\beta + \mathbf{e}$, where $\mathbf{y}$ is a column vector of observations of order k, $\mathbf{X}$ is a known matrix of order $p \times k$, is the column vector of effects which is of order p and e is the error vector of order k, which is assumed to be distributed as multivariate normal with mean vector $\mathbf{0}$ and dispersion matrix $\mathbf{I}$ i.e. $\mathbf{N}(\mathbf{0}, \mathbf{I})$.

Most technique towards the identification of outliers in experimental design can categorized in three groups- residual based, nonparametric and robust. Use of the residuals to detect multiple outliers in a two-way design may lead to unsatisfactory results. Most of the early work in identification of outlier has been developed on the examination of residuals e.g. Danial (1960)[1], Stenfansky (1971)[2], Goldsmith & boddy (1973)[3], John and Prescott (1975)[4] and usually they are successful to identify at most one or two outliers, But residual based methods are not much reliable as the residuals have some serious short comings. Barnett and Lewis (1994)[5] point out that outliers not only affect their own residuals but have a carry over effect on others. Therefore the existing residual based methods have become unreliable for detecting more than one outlier e.g. Gentleman and wilk(1975a)[6] first propose a formal test for thepresence of multiple outliers. some of the work on testing of outliers in linear model based on residuals are Gentleman and Wilk (1975b)[7], John and Draper (1978)[8], Draper and John (1980)[9], Joshi (1972)[10], Joshi and Lalitha (1986)[11],Ellenberg (1976)[12] etc. Hamounda and Leone (1974)[13] have considered a regression model and for estimating the parameters of the model,they have used O-BLUE (Ordered - Best Linear Unbaised estimators), which they have indicated to be better estimators than the one based on residuals. Here we study the effect of outliers in the estimation procedure, i.e. O-BLUE defined in this paper , so that this information can be used for developing an outlier detection test.

2 O-BLUE and the effect of outliers in them

A simple regression model is a form of a Gauss-Markov linear model. The explicit form of the model is

$$Y_{ij} = \alpha + \beta(x_i - \bar{x}) + e_{ij} \tag{1}$$

where $i = 1, 2, ..., k$ and $j = 1, 2, ..., n_i$, x_i is an independent or input variable assumed to be fixed. The error terms e_{ij} are independent and identically distributed random variables from a continuous symmetric distribution with mean 0 and variance σ^2.

* Corresponding author e-mail: arvindmzu@gmail.com

Now the O-BLUE defined by Hamouda and Leone (1974), with ordered observations in a distribution, depending only on location and scale parameter and equal values of n_i, i.e. $n_1 = n_2 = ... = n_k = n$, is as follows:
Arrange the observations Y_{ij}'s an ascending order of magnitude, i.e. $Y_{i(1)} \leq Y_{i(2)} \leq ... \leq Y_{i(n)}$ for each i. Further let

$$Z_{i(j)} = Y_{i(j)} - \alpha - \beta(x_i - \bar{x})/\sigma \tag{2}$$

i.e. the $Z_{i(j)}$ is a standardized residual about the regression line. Then $Z_{i(j)}$ is the j^{th} order statistic of a random sample of size n from a standardized symmetric distribution F with mean zero and variance one.
Defining the first and second moments of the order statistic as

$$E(Z_{i(j)}) = c_{ij} \text{ and } Cov(Z_{i(j)}, Z_{i(j')}) = \omega_{jj'}$$

the O-BLUE of α, β and σ are

$$\hat{\alpha}_0 = \frac{\mathbf{1}'\Omega^{-1}}{k(\mathbf{1}'\Omega^{-1}1)} \sum_{i=1}^{k} Y_i, \tag{3}$$

$$\hat{\beta}_0 = \frac{\mathbf{1}'\Omega^{-1}}{(\mathbf{1}'\Omega^{-1}\mathbf{1}) \sum_{i=1}^{k} (x_i - \bar{x})} \sum_{i=1}^{k} (x_i - \bar{x}) Y_i \tag{4}$$

and

$$\hat{\sigma} = \frac{\mathbf{c}'\Omega^{-1}}{k(\mathbf{c}'\Omega^{-\mathbf{1}}\mathbf{c})} \sum_{i=1}^{k} Y_i; \tag{5}$$

where $\mathbf{1}' = (1, 1, ..., 1)$, $\mathbf{Y}_\mathbf{i}' = (Y_{i(1)}, Y_{i(2)}, ..., Y_{i(n)})$, $\Omega = Cov(\mathbf{Z_i})$, $\mathbf{Z}_\mathbf{i}' = (Z_{i(1)}, Z_{i(2)}, , ..., Z_{i(n)})$ and $\mathbf{c}' = E(\mathbf{Z}_\mathbf{i}')$. Also let $\mathbf{Y} = (\mathbf{Y_1}, \mathbf{Y_2}, ..., \mathbf{Y_k})$ be an $n \times k$ matrix of all $\mathbf{Y}$ vectors. We have

$$\sum_{i=1}^{k} \mathbf{Y_i} = \begin{pmatrix} \sum_{i=1}^{k} Y_{i(1)} \\ \sum_{i=1}^{k} Y_{i(2)} \\ . \\ . \\ . \\ \sum_{i=1}^{k} Y_{i(n)} \end{pmatrix}$$

The variances of these estimators are

$$Var(\hat{\alpha}_0) = \frac{\sigma^2}{k(\mathbf{1}'\Omega^{-\mathbf{1}}\mathbf{1})}$$

$$Var(\hat{\beta}_0) = \frac{\sigma^2}{(\mathbf{1}'\Omega^{-\mathbf{1}}\mathbf{1}) \sum_{\mathbf{i=1}}^{\mathbf{k}} (\mathbf{x_i} - \bar{\mathbf{x}})^2}$$

$$Var(\hat{\sigma}_0) = \frac{\sigma^2}{(\mathbf{c}'\Omega^{-\mathbf{1}}\mathbf{c})} \tag{6}$$

and all covariances are zero.

Now to study the effect of outliers on these estimators, we introduce a shift (either positive or negative) in the location parameter of any component of one of the observation. If a negative shift in smallest component and a positive shift in largest component is introduced, this may not disturb the order of the observations. But if the shift is in any other component, or a positive shift in the smallest component, or a negative shift in the largest component, then this shift will affect the order of the observation and calculation of all O-BLUEs of the parameters.

Suppose we introduce a shift a in the b^{th} observation $b=(1,2,...,k)$, i.e. if a positive shift $a(>0)$ is introduced in the location parameter of r^{th} component of b^{th} observation, then the changed order of the observation, in the matrix Y is

$$\mathbf{Y}=\begin{pmatrix} y_{11} & y_{21} & \cdot\cdot & y_{b1} & \cdot\cdot & y_{k1} \\ y_{12} & y_{22} & \cdot\cdot & y_{b2} & \cdot\cdot & y_{k2} \\ \cdot & \cdot & \cdot\cdot & \cdot & \cdot\cdot & \cdot \\ \cdot & \cdot & \cdot\cdot & \cdot & \cdot\cdot & \cdot \\ \cdot & \cdot & \cdot\cdot & \cdot & \cdot\cdot & \cdot \\ y_{1(r-1)} & y_{2(r-1)} & \cdot\cdot & y_{b(r-1)} & \cdot\cdot & y_{k(r-1)} \\ y_{1r} & y_{2r} & \cdot\cdot & y_{br+a}=y_{br'} & \cdot\cdot & y_{kr} \\ \cdot & \cdot & \cdot\cdot & \cdot & \cdot\cdot & \cdot \\ \cdot & \cdot & \cdot\cdot & \cdot & \cdot\cdot & \cdot \\ \cdot & \cdot & \cdot\cdot & \cdot & \cdot\cdot & \cdot \\ y_{1(s-1)} & y_{2(s-1)} & \cdot\cdot & y_{b(s-1)'} & \cdot\cdot & y_{k(s-1)} \\ y_{1s} & y_{2s} & \cdot\cdot & y_{bs} & \cdot\cdot & y_{sr} \\ \cdot & \cdot & \cdot\cdot & \cdot & \cdot\cdot & \cdot \\ \cdot & \cdot & \cdot\cdot & \cdot & \cdot\cdot & \cdot \\ \cdot & \cdot & \cdot\cdot & \cdot & \cdot\cdot & \cdot \\ y_{1n} & y_{2n} & \cdot\cdot & y_{bn} & \cdot\cdot & y_{kn} \end{pmatrix}$$

and

$$\mathbf{a}=\begin{pmatrix} 0 \\ \cdot \\ \cdot \\ a \\ \cdot \\ \cdot \\ 0 \end{pmatrix} \tag{7}$$

Then

$$\sum_{i=1}^{k}\mathbf{Y_i}=\begin{pmatrix} \sum_{i=1}^{k} y_{i1} \\ \sum_{i=1}^{k} y_{i2} \\ \cdot \\ \cdot \\ \cdot \\ \sum_{i=1}^{k} y_{i(r-1)} \\ \sum_{i=1,i\neq b}^{k} y_{i(r)}+y_{br'} \\ \cdot \\ \cdot \\ \cdot \\ \sum_{i=1,i\neq b}^{k} y_{i(s-1)}+y_{b(s-1)'} \\ \sum_{i=1}^{k} y_{is} \\ \cdot \\ \cdot \\ \cdot \\ \sum_{i=1}^{k} y_{in} \end{pmatrix} \mathbf{or} \sum_{\mathbf{i=1}}^{\mathbf{k}}\mathbf{Y_i}+\mathbf{a}=\begin{pmatrix} \sum_{i=1}^{k} y_{i1} \\ \sum_{i=1}^{k} y_{i2} \\ \cdot \\ \cdot \\ \cdot \\ \sum_{i=1}^{k} y_{i(r-1)} \\ \sum_{i=1,i\neq b}^{k} y_{i(r)}+y_{br'} \\ \sum_{i=1,i\neq b}^{k} y_{i(r+1)}+y_{b(r+1)'} \\ \cdot \\ \cdot \\ \cdot \\ \sum_{i=1,i\neq b}^{k} y_{i(s-1)}+y_{b(s-1)'} \\ \sum_{i=1}^{k} y_{is} \\ \cdot \\ \cdot \\ \cdot \\ \sum_{i=1}^{k} y_{in} \end{pmatrix} \tag{8}$$

Then $\sum_{i=1}^{k} \mathbf{Y_i} + \mathbf{a}$ can be written as

$$\begin{pmatrix} \sum_{i=1}^{k} y_{i1} \\ \sum_{i=1}^{k} y_{i2} \\ . \\ . \\ . \\ \sum_{i=1,i\neq b}^{k} y_{ir} \\ \sum_{i=1,i\neq b}^{k} y_{i(r+1)} \\ . \\ . \\ . \\ \sum_{i=1,i\neq b}^{k} y_{i(s-1)} \\ \sum_{i=1,i\neq b}^{k} y_{is} \\ . \\ . \\ . \\ \sum_{i=1}^{k} y_{in} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ . \\ . \\ . \\ y_{br'} \\ y_{b(r+1)'} \\ . \\ . \\ . \\ y_{b(s-1)'} \\ 0 \\ . \\ . \\ 0 \end{pmatrix} \tag{9}$$

Hence the O-BLUE of α_0' is

$$\hat{\alpha_0}' = \frac{1'\Omega^{-1}}{k(1'\Omega^{-1}1)}(\sum_{i=1}^{k} Y_i + a), \tag{10}$$

where $(\sum_{i=1}^{k} Y_i + a)$ is as in (9).

Thus (10) gives the estimate of α, when a shift was introduced in one of the components of the observation vector Y_b. Hence, to make an assessment about the deviation caused by this shift in the estimate of α, we obtain the difference between the estimator obtained in (10) and that obtained in (3) as follows:

$$\hat{\alpha_0}' - \hat{\alpha_0} = \frac{1'\Omega^{-1}}{k(1'\Omega^{-1}1)} \begin{pmatrix} 0 \\ 0 \\ . \\ . \\ . \\ y_{br'} - y_{br} \\ y_{b(r+1)'} - y_{b(r+1)} \\ . \\ . \\ . \\ y_{b(s-1)'} - y_{b(s-1)} \\ 0 \\ . \\ . \\ 0 \end{pmatrix}$$

$$\hat{\alpha_0}' - \hat{\alpha_0} = \frac{1'\Omega^{-1}}{k(1'\Omega^{-1}1)} \underline{d}$$

where $\underline{d}$ is the vector of differences between the shifted observation and the original observation.

Thus while calculating $\hat{\alpha_0}$, if we wish an accuracy of ε, then we should have the deviation of $\hat{\alpha_0}'(\hat{\alpha_0}$ in the presence of an

outlier)from $\hat{\alpha}_0$ less than ε.

$$|\hat{\alpha}_0{}' - \hat{\alpha}_0| \leq \varepsilon \Rightarrow |d| \leq |\frac{1'\Omega^{-1}}{k(1'\Omega^{-1}1)}|\varepsilon \tag{11}$$

Similarly, in the estimation of $\hat{\beta}_0$, when there is a shift,

$$\hat{\beta}_0{}' = \frac{1'\Omega^{-1}}{(1'\Omega^{-1}1)\sum_{i=1}^{k}(x_i-\bar{x})^2}\sum_{i=1}^{k}(x_i-\bar{x})(Y_i+a),$$

$$\hat{\beta}_0{}' = \hat{\beta}_0 + \frac{1'\Omega^{-1}}{(1'\Omega^{-1}1)\sum_{i=1}^{k}(x_i-\bar{x})^2}\sum_{i=1}^{k}(x_i-\bar{x})a,$$

Hence, the shift introduced does not affect the estimate of β i.e. the estimator $\hat{\beta}_0$ is unaffected by the presence of any outlier.

Now the expression for the estimator $\hat{\sigma}_0$, when a shift a is introduced is

$$\hat{\sigma}_0{}' = \frac{c^{*'}\Omega^{-1}}{k(c^{*'}\Omega^{-1}c^*)}\sum_{i=1}^{k}(\underline{Y_i}+\boldsymbol{a}) \tag{12}$$

$$\hat{\sigma}_0{}' = \frac{c^{*'}\Omega^{-1}}{k(c^{*'}\Omega^{-1}c^*)}\sum_{i=1}^{k}\underline{Y_i} + \frac{c^{*'}\Omega^{-1}}{k(c^{*'}\Omega^{-1}c^*)}\boldsymbol{a}$$

where $\mathbf{a}^*$ and Ω^{-1} are as defined in (7) and (9) respectively and

$$c^* = E(Z_i + a^*) = c + a^*, \text{ where } a^* = \tfrac{a}{\sigma}$$

Hence to make an assessment about the deviation, we obtain the difference between the estimator obtained in (12) and that obtained in (5) as follows:

$$\hat{\sigma}_0{}' - \hat{\sigma}_0 = \begin{pmatrix} 0 \\ 0 \\ . \\ . \\ . \\ y_{br'} - y_{br} \\ y_{b(r+1)'} - y_{b(r+1)} \\ . \\ . \\ . \\ y_{b(s-1)'} - y_{b(s-1)} \\ 0 \\ . \\ . \\ 0 \end{pmatrix} \text{ or } \hat{\sigma}_0{}' - \hat{\sigma}_0 = \frac{c^{*'}\Omega^{-1}}{k(c^{*'}\Omega^{-1}c^*)}\underline{d}$$

Thus while calculating $\hat{\sigma}_0{}'$, if we wish an accuracy of ε, then we should have the deviation of $\hat{\sigma}_0{}'$ from $\hat{\sigma}_0$less than ε,

$$i.e. |\hat{\sigma}_0{}' - \hat{\sigma}_0| \leq \varepsilon.$$

$$\Rightarrow |\frac{c^{*'}\Omega^{-1}}{k(c^{*'}\Omega^{-1}c^*)}\underline{d}| \leq \varepsilon$$

$$\Rightarrow |\underline{d}| \leq |\frac{c^{*'}\Omega^{-1}}{k(c^{*'}\Omega^{-1}c^*)}|^{-1}\varepsilon$$

Note that the above arguments will hold if any component of one of the observation is shifted and the shifted amount does not change the order of the observation $Y_i(j)$. Due to this, it is difficult to define an outlier detection procedure, if some of the observations in the middle are shifted . We can only discuss that how much variation is created by shifted amount from the original value. However, for end observations, an outlier detection procedure is discussed below next section.

3 Outliers tests for end observations

We propose to define outlier detection procedure based on (1) the desired accuracy of the estimate, and (2) the usual significance level, i.e. using the likelihood ratio test. For this, we consider the O-BLUE defined for a simple regression model.
In practice, while dealing with real data, neither the value of a nor the component which is deviated from the original value may be known. In such case the j^{th} component (which is supposed to have deviated) of i^{th} vector Y_i can be identified as the one for which the amount of deviation is given by

$$a = max(a_1, a_n) \tag{13}$$

$$\text{where } a_i = max(|Y_{il} - \bar{Y_l}|, |Y_{in} - \bar{Y_n}|), i = 1,2,...,k.$$

$$\text{and } \bar{Y_j} = \frac{1}{n-1} \sum_{l(\neq j)=1}^{n} Y_{il}, j = 1 \text{ or } n$$

3.1 Test based on the desired accuracy of the estimate

A test procedure based on the shifted value a can be defined as follows:
If an accuracy of ε is sought in the estimation of the parameters α and σ, then declare the j^{th} component of the i^{th} vector to be an outlier if either

$$a > max \left\{ |\frac{k \sum_{i=1}^{n} \sum_{j=1}^{n} \omega^{ij}}{\sum_{i=1}^{n} \omega^{ij}}| \varepsilon, |\frac{k \sum_{i=1}^{n} \sum_{j=1}^{n} c_i^* c_j^* \omega^{ij}}{\sum_{i=1}^{n}} c_i^* \omega^{ij}| \varepsilon \right\}, \tag{14}$$

where a is as defined in (13).

3.2 Data Analysis

We consider the example used on observations by Hamouda and Leone (1974), which is as follows: A research program to investigate the relationship between reaction condition and yield in the low pressure polymerization of ethylene included 200 laboratory polymerizations. Here we consider only the concentration of ethylene (Et) in moles/cc as the input and the polymer yield (G) in grams per batch as output. The reaction time is fixed at 30 minutes. The regression equation used to estimate the theoretical kinetic model is

$$log(G) = \alpha' + \beta' log(Et)$$

Let $Y_{ij} = log(G)$ and $x_i = \frac{[log(Et)+2.2040]}{0.301}$. In this example there are 4 observation vectors each with 5 components i.e $k = 4$ and $n = 5$. For this example the O-BLUE obtained by Hamouda and Leone (1974) for the parameters are $\hat{\alpha_0} = 1.9227$, $\hat{\beta}_0 = 0.20866$ and $\hat{\sigma_0} = 0.14908$.
Here,

$$Y_1 = \begin{pmatrix} 1.5774 \\ 1.6232 \\ 1.6263 \\ 1.7450 \\ 1.7902 \end{pmatrix}, Y_2 = \begin{pmatrix} 1.4698 \\ 1.6901 \\ 1.9647 \\ 1.9956 \\ 1.9986 \end{pmatrix}, Y_3 = \begin{pmatrix} 1.8881 \\ 1.9020 \\ 2.0549 \\ 2.1058 \\ 2.2329 \end{pmatrix} \& Y_4 = \begin{pmatrix} 2.0310 \\ 2.1283 \\ 2.1405 \\ 2.1983 \\ 2.2913 \end{pmatrix}$$

For these vectors,c' and Ω^{-1} were found to be

$$c = \begin{pmatrix} -1.16296 & -0.49502 & 0 & 0.49502 & 1.16296 \end{pmatrix}$$

and

$$\Omega^{-1} = \begin{pmatrix} 0.848720 & 0.519080 & -0.052162 & -0.515950 & -0.415780 \\ 0.519081 & 0.111022 & -0.150125 & -0.150162 & -0.515950 \\ -0.052162 & -0.150125 & -0.132590 & -0.150125 & -0.052162 \\ -0.515950 & -0.150162 & -0.150125 & 0.111022 & 0.051908 \\ -0.415780 & -0.515950 & -0.052162 & 0.0519081 & 0.848720 \end{pmatrix}$$

Now in one of the components of one of the vectors a shift is introduced. Suppose the first component of the vector Y_1 is shifted to the left by a magnitude of 11.8903 i.e. when the vector Y_1 has a shifted value, we have

$$c = \begin{pmatrix} -10.3129 & 1.6232 & 1.6263 & 1.7450 & 1.7902 \end{pmatrix}$$

With this change in the observation, the O-BLUE of α_0 in the presence of an outlying observation is given by $\hat{\alpha}_0' = 10.1404$. The O-BLUE of β_0 is unaffected by the presence of any outlier i.e. $\hat{\beta}_0^1 = \hat{\beta}_0 = 0.20866$ and finally the O-BLUE of σ_0 in the presence of an outlier is found to be $\hat{\sigma_0'}$. From this it can be noticed that O-BLUE are highly sensitive for outlying observations, as with just one component of one vector has affected the estimates of α and σ to a great extent. Now for the outlier test procedure, suppose we wish an accuracy of $\varepsilon = 0.01$ Then we wish $\hat{\alpha_0}' - \hat{\alpha_0}$ and similarly $\hat{\sigma_0}' - \hat{\sigma_0}$. If this is violated, then we expect some outlying observations. Hence to estimate the deviation of the outlying observation, we use (13). Thus it was found

$$a = max(|-12.00911|, |3.11981|) = 12.00911.$$

This corresponds to the first component of the first vector i.e. $i = 1$ and $j = 1$. Now, to test whether this shift is significant or not, we may have to calculate $|\frac{k\sum_{i=1}^{n}\sum_{j=1}^{n}\omega^{ij}}{\sum_{i=1}^{n}\omega^{ij}}|\varepsilon = \alpha_1$ and $|\frac{k\sum_{i=1}^{n}\sum_{j=1}^{n}c_i^* c_j^* \omega^{ij}}{\sum_{i=1}^{n}}c_i^*\omega^{ij}|\varepsilon = \sigma_1$, as per (14) for comparison with the estimated value of a. Thus the calculated values of α_1 and α_2 are 0.01476 and 0.14334 respectively. Hence it can be seen that a is greater than $max\{\alpha_1, \sigma_1\}$; therefore the first component of the first vector is declared as an outlying observation.

3.3 Likelihood ratio procedure for outlier detection

In this procedure we state the null hypothesis H_0 as there is no outlier in the data and the alternative hypothesis H_1 as one of the components i.e. j^{th} component of one of the observation vector i.e. i^{th} vector is shifted by an amount a. For this the likelihood functions under H_0 and under H_1 have to be determined. Here again $i = 1, 2, .., k$ and $j = 1, 2, ..., n$. Under H_0,

$$E(Y_h) = \begin{pmatrix} Y_{h1} \\ . \\ . \\ . \\ Y_{hn} \end{pmatrix} = \begin{pmatrix} \mu_{h1} \\ . \\ . \\ . \\ \mu_{hn} \end{pmatrix} = \mu_{\mathbf{h}} = \begin{pmatrix} \alpha+\beta(x_h-\overline{x}) \\ \alpha+\beta(x_h-\overline{x}) \\ \alpha+\beta(x_h-\overline{x}) \\ \alpha+\beta(x_h-\overline{x}) \\ \alpha+\beta(x_h-\overline{x}) \end{pmatrix}, \mathbf{h} \neq \mathbf{i}$$

Under H_1, $E(Y_{ij}) = \mu_i + \mathbf{a}$ where $\mathbf{a} = \begin{pmatrix} 0 \\ . \\ . \\ . \\ a \\ . \\ . \\ . \\ 0 \end{pmatrix} \rightarrow \mathbf{j^{th}}$ component, i.e. $E(Y_{ij}) = \mu_{ij} + a = \alpha + \beta(x_i - \overline{x}) + a$. In this, we make use of $\hat{\alpha}_0'$ and $\hat{\beta}_0'$, the O-BLUE of α and β when an outlier is present, for the values of α and β. If $Y_i, i = 1, ,k$are distributed each as normally with mean μ_i and covariance matrix Σ then under H_0,

$$L_0 = \prod_{h=1}^{k} \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(y_h-\mu_h)'\Sigma^{-1}(y_h-\mu_h)}$$

$$= \frac{1}{(2\pi)^{\frac{nk}{2}}|\Sigma|^{\frac{k}{2}}} e^{-\frac{1}{2}\sum_{h=1}^{k}(y_h-\mu_h)'\Sigma^{-1}(y_h-\mu_h)} \tag{15}$$

$$L_1 = \prod_{h=1,(h\neq i)}^{k} \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(y_h-\mu_h)'\Sigma^{-1}(y_h-\mu_h)} e^{-\frac{1}{2}(y_i-\mu_i-a)'\Sigma^{-1}(y_i-\mu_i-a)}$$

$$= \frac{1}{(2\pi)^{\frac{nk}{2}}|\Sigma|^{\frac{k}{2}}} e^{-\frac{1}{2}\left[\sum_{h=1,(h\neq i)}^{k}(y_h-\mu_h)'\Sigma^{-1}(y_h-\mu_h)+(Y_i-\mu_i-a)'\Sigma^{-1}(Y_i-\mu_i-a)\right]}$$

Now,

$$(Y_i-\mu_i-a)'\Sigma^{-1}(Y_i-\mu_i-a)$$
$$= (Y_i-\mu_i)'\Sigma^{-1}(Y_i-\mu_i) - a'\Sigma^{-1}(Y_i-\mu_i) - (Y_i-\mu_i)'\Sigma^{-1}a + a'\Sigma^{-1}a.$$

Let $C = \frac{1}{(2\pi)^{\frac{nk}{2}}|\Sigma|^{\frac{k}{2}}}$, then

$$L_1 = C.e^{-\frac{1}{2}\left[\sum_{h=1,h\neq 1}^{k}(Y_h-\mu_h)'\Sigma^{-1}(Y_h-\mu_h)+(Y_i-\mu_i)'\Sigma^{-1}(Y_i-\mu_i)-a'\Sigma^{-1}(Y_i-\mu_i)-(Y_i-\mu_i)'\Sigma^{-1}a+a'\Sigma^{-1}a\right]}$$

$$= C.e^{-\frac{1}{2}\left[\sum_{h=1}^{k}(Y_h-\mu_h)'\Sigma^{-1}(Y_h-\mu_h)-a'\Sigma^{-1}(Y_i-\mu_i)-(Y_i-\mu_i)'\Sigma^{-1}a+a'\Sigma^{-1}a\right]} \quad (16)$$

From (15) and (16), we have

$$L_1 = L_0.e^{-\frac{1}{2}\left[a'\Sigma^{-1}a-2a'\Sigma^{-1}(Y_i-\mu_i)\right]}$$
$$\frac{L_1}{L_0} = e^{-\frac{1}{2}\left[a'\Sigma^{-1}a-2a'\Sigma^{-1}(Y_i-\mu_i)\right]}$$

Now in the likelihood ratio test the critical region is given by

$$\frac{L_1}{L_0} > k \Rightarrow e^{-\frac{1}{2}\left[a'\Sigma^{-1}a-2a'\Sigma^{-1}(Y_i-\mu_i)\right]} > K. \quad (17)$$

Taking logarithm on both sides of (17)

$$-\frac{1}{2}\left[a'\Sigma^{-1}a - 2a'\Sigma^{-1}(Y_i-\mu_i)\right] > log(K)$$

$$\Rightarrow (Y_i-\mu_i) > \frac{1}{2}(a'\Sigma^{-1})^{-1}(2log(K)+a'\Sigma^{-1}a)$$

$$\Rightarrow (Y_i-\mu_i) > K_1 \text{ where } K_1 = \tfrac{1}{2}(a'\Sigma^{-1})^{-1}(2log(K)+a'\Sigma^{-1}a)$$

$$\Rightarrow (Y_i-\mu_i)'\Sigma^{-1}(Y_i-\mu_i) > K_1'\Sigma^{-1}K_1 \quad (18)$$

where $(Y_i-\mu_i)'\Sigma^{-1}(Y_i-\mu_i)$ is distributed as a χ^2 variate with n degrees of freedom under H_0. Hence, the α level critical region is given by

$$v_i = (Y_i-\mu_i)'\Sigma^{-1}(Y_i-\mu_i) > \chi_n^2(\alpha) \quad (19)$$

$\chi_n^2(\alpha)$ being the 100α percent tail of a χ^2 distribution. Hence if for a given set of observations (19) holds then the i^{th} vector is an outlier and to identify the j^{th} component of such a vector, we make use of (13). Here, in our case we have $\Sigma = \sigma^I$. Hence (19) becomes

$$v_i = \frac{1}{\sigma^2}(Y_i-\mu_i)'(Y_i-\mu_i) > \chi_n^2(\alpha)$$

or

$$v_i = \frac{1}{\sigma^2}\sum_{j=1}^{n}(Y_{ij}-\mu_{ij})^2 > \chi_n^2(\alpha), i = 1,2,...,k. \quad (20)$$

Hence to perform this test, if we make use of the O-BLUE of $\hat{\alpha}'$ and $\hat{\sigma}'$ for the value of σ and for the estimation of $\mu_i, i = 1,2,...,k$, then the determination of the critical region will be erroneous, as the outlying observation affects both $\hat{\alpha}'$ and $\hat{\sigma}'$ in the calculation of the estimate of $\mu_i, i = 1,2,...,k$. Hence, we take a censored sample for these estimations, as it is shown by Hamouda and Leone (1974) that the estimates obtained from censored samples are very close to the uncensored estimates. Now to decide as to which component has to be deleted to obtain a censored sample , we make use of (14) and delete that component, which corresponds to the calculation of maximum of a. The same component has to be deleted from all the observations and we estimate all the parameters, which are further used for estimating $\mu_i, i = 1,2,...,k$

4 Data Analysis

We shall consider the same example as was discussed in previous procedure. On the basis of maximum of a, the first component of all the observations are deleted and a censored sample of size 4 each with 4 components are formed. From this the estimates of α, β and σ were found to be $\hat{\alpha}'' = 1.9286$, $\hat{\beta}'' = 0.20078$ and $\hat{\sigma}'' = 0.17703$. Now from these the estimates of $\mu_i, i = 1,2,...,k$ are $\mu_{1j} = 1.6274$, $\mu_{2j} = 1.8282$, $\mu_{3j} = 2.0290$ and $\mu_{4j} = 2.2298$, $j = 1,..,5$. Using these estimates the values of $v_i, i = 1,2,...,k$ are $v_1 = 36.4667, v_2 = 7.1222, v_3 = 2.6844$, and $v_4 = 1.9962$. Also the tabulated value of χ^2_5 is 11.070 at 5% level of significance. Hence, except for the first value all are less than the tabulated value, thereby indicating that the first observation to be the one which is contaminated. Again, since the maximum of a corresponds to the first component of the first vector, that component can be declared as the outlying component.

5 Performance of the test

For the performance calculation, we have to obtain the probability

$$\eta = Pr\left\{(Y_i - \mu_i - a)'\Sigma^{-1}(Y_i - \mu_i - a) \geq \chi^2_n\right\}$$

here we have $\Sigma = \sigma^2\mathbf{I}$. Hence

$$\frac{1}{\sigma^2}(Y_i - \mu_i - a)'(Y_i - \mu_i - a) =$$

$$(Y_i - \mu_i - a)'\Sigma^{-1}(Y_i - \mu_i - a) = \frac{1}{\sigma^2}\left\{(Y_i - \mu_i)'(Y_i - \mu_i) - a'(Y_i - \mu_i) - (Y_i - \mu_i)'a + a'a\right\}$$

$$= \frac{1}{\sigma^2}\left\{\sum_{k=1}^{n}(Y_{ik} - \mu_{ik})^2 - 2a(Y_{ij} - \mu_{ij}) + a^2\right\}$$

Now, if $a = 2a(Y_{ij} - \mu_{ij})$, then

$$\frac{1}{\sigma^2}(Y_i - \mu_i - a)'(Y_i - \mu_i - a) = v_i - \frac{1}{\sigma^2}\left\{a^2 - a^a\right\} = v_i.$$

Hence $\eta = P\left\{v_i \geq \chi^2_n(\alpha)\right\} = \alpha$. Let $b = \frac{1}{\sigma^2}\left\{a^2 - 2(Y_{ij} - \mu_{ij})a\right\}$. Then,
$\eta = Pr\left\{v_i + b \geq \chi^2_n(\alpha)\right\} = Pr\left\{v_i \geq \chi^2_n(\alpha) - b\right\}$
Now if $b > 0$, then $\eta > \alpha$; otherwise it will be less than α. Hence, for the test to be effective, we should have a positive value of b i.e.

$$a^2 > 2(Y_{ij} - \mu_{ij})a \Rightarrow a > 2(Y_{ij} - \mu_{ij}).$$

The performance of the test for the above example was done with different values of a .The performance of the test for positive shift is given in table 1 and performance of the test for negative shiftis given in table 2.

Table 1: Performance of the test for positive shift

Parameter	Estimate	Standard Error	Lower Credible Limit	Upper Credible Limit
0.001	-0.03231	0.00295	11.06753	0.05005
0.025	-0.03231	0.10081	10.96967	0.05198
0.05	-0.03231	0.25786	10.81262	0.05522
0.075	-0.03231	0.471162	10.59932	0.059929
0.1	-0.03231	0.740703	10.32978	0.066412
0.2	-0.03231	2.381299	8.689184	0.122122
0.3	-0.03231	4.921789	6.148694	0.292018
0.4	-0.03231	8.362173	2.70831	0.744847
0.5	-0.03231	12.70245	0	1
1	-0.03231	47.90224	0	1
1.25	-0.03231	73.93864	0	1
1.5	-0.03231	105.5994	0	1
2	-0.03231	185.7939	0	1
2.5	-0.03231	288.4857	0	1
3	-0.03231	413.6748	0	1
3.5	-0.03231	561.3614	0	1

Table 2: Performance of the test for negative shift

Parameter	Estimate	Standard Error	Lower Credible Limit	Upper Credible Limit
-3.5	-0.03231	541.0085	0	1
-3.0	-0.03231	396.2295	0	1
-2.5	-0.03231	273.9479	0	1
-2.0	-0.03231	174.1636	0	1
-1.5	-0.03231	96.87671	0	1
-1.25	-0.03231	66.66975	0	1
-1.0	-0.03231	42.08713	0	1
-0.5	-0.03231	9.794893	1.275589	0.937423
-0.4	-0.03231	6.036127	5.034355	0.411702
-0.3	-0.03231	3.177255	7.893228	0.162219
-0.2	-0.03231	1.218276	9.852206	0.079533
-0.1	-0.03231	0.159191	10.91129	0.053167
-0.075	-0.03231	0.035028	11.03545	0.050682
-0.05	-0.03231	-0.03289	11.10337	0.049368
-0.025	-0.03231	-0.04457	11.11505	0.049146
-0.001	-0.03231	-0.00286	11.07335	0.049945

From the above tables, it can be seen that the test is effective for all the shifts a as long as $a > 2(Y_{ij} - \mu_{ij})$, and the power value goes less than the significance level α for $a = -0.001, -0.025$ and -0.05 i.e. when this inequality does not hold. In fact, in these cases the observation goes closer and closer to the ideal value i.e. μ_{ij}, because of the shift and hence it is no longer an outlying observation.

Acknowledgement

The authors are grateful to the anonymous referee for a careful checking of the details and for helpful comments that improved this paper.

References

[1] Daniel, C. (1960). Locating outliers in factorial experiments'. Technometrics, 2, 149-156. (235, 238, 240)
[2] Stefansky, W. (1971). 'Rejecting outliers by maximum normal residual'. Ann. Math. Statist., 42, 35-45. (94, 242, 254, 255)
[3] Goldsmith, P. L., and Boddy, R. (1973). 'Critical analysis of factorial experiments and orthogonal fractions'. Applied Statistics, 22, 141-160. (241, 244, 264)
[4] John, J. A., and Prescott, P. (1975). 'Critieal values of a test to detect outliers in factorial experiments'. Appl. Statistics, 24, 56-59. (241, 244)
[5] Barnett, V. and Lewis,T. 1994, 3rd edition, John Wiley & Sons, Chichester.
[6] Gentleman, J.F. and Wilk M.B. (1975a). Detecting outliers in a two-way table:I. Statistical behaviour of residuals, Technometrics 17, 1-14.
[7] Gentleman, J.F. and Wilk M.B. (1975b). Detecting outliers in a two-way table:II. Supplementing the direct analysis of residuals, Biometrics 3, 387-410.
[8] John, J.A. and Draper, N.R. (1978). On testing for two outliers or one outlier in two-way tables, Technometrics 20, 69-78.
[9] Draper, N.R. and John J.A. (1980). Testing for three or fewer outliers in two-way tables, Technometrics 22, 9-15.
[10] Joshi, P.C. (1972). Some slippage tests of mean for a single outlier in a linear regression, Biometrika 59, 109-120.
[11] Joshi, P.C. and S. lalitha, (1986). Tests for two outliers in a linear model, Biometrika 73 , no.1, 236-239.
[12] Ellenberg, J. H. (1976). Testing for a single outliers from a general linear regression, Biometrics 32, 637-64.
[13] Hamouda and Leone(1974). HAMOUDA , . M., and LEONE , F. C. (1974). The O-BLUE estimators for complete and censored samples in linear regression, Technometrics 16, 441-446.

Arvind Pandey received the PhD degree in Statistics at Savitribai Phule Pune University, Pune. His research interests are in the areas of frailty models and applied statistics and Bayesian Survival Analysis. He has published research articles in reputed international journals of Statistics. He is referee of statistical journals.

Nibha Srivastava is an officer in Department of Statistics and Planing Implementation, UP,Allahabad, India. She received the PhD degree in Statistics at department of statistics, University of Allahabad. Her area of interest is Outliers, Linear regression and Sample Survey.