Pak. J. Statist.
2015 Vol. 31(3), 121-134

MIXED MODEL APPROACHES FOR DETECTING INFLUENTIAL OBSERVATIONS IN GENETIC DATA ANALYSIS

Yousaf Hayat[1§]**, Jian Yang**[2] and **Jun Zhu**[3]

[1] Department of Mathematics, Statistics and Computer Science
The University of Agriculture, Peshawar, Pakistan.
[2] Queensland Brain Institute, The University of Queensland, Australia
[3] Key Laboratory of Crop Germplasm Resource of Zhejiang Province,
Institute of Bioinformatics, Zhejiang University, Hangzhou,
Zhejiang 310058, China.
[§] Corresponding author: yhayat@aup.edu.pk

ABSTRACT

Influence diagnostics for detection of influential data points are widely used in statistical modeling to gauge out their impact on various aspect of the analysis. We proposed a method for detection of influential observations in mixed linear model using MINQUE (1) for estimation of variance components and linear unbiased prediction (LUP) for prediction of random effects. The method is based on the analogue of the Cook distance statistics for detection of influential observations affecting both the fixed and random components of a mixed linear model. The method is illustrated with both simulated and the real data sets based on an experimental model for genetic analysis.

KEYWORDS

Mixed Linear Model; Influence Analysis; Linear Unbiased Prediction; MINQUE (1).

1. INTRODUCTION

Genetic analysis is often conducted to assess the effect of different genotypes across diverse locations and several years. These assist growers and plant breeders in selecting suitable genotypes (Yan and Rajcan, 2003). The genotype effects can be further partitioned into various genetics effects and their interaction with environments (Zhu, 1994). In such trials, researchers are often interested to study the main effects of different genotypes as well as their interaction with specific environments. Mixed linear models are often applied to cope with a situation, which can handle factors both of the fixed and random effects involved in the experiments. Generally, it is impractical to make perfect measurements in the agricultural experiments because the complex traits with continuous phenotypic variation may depend on variations of different genotypes, environmental and genotype × environment interaction effects. Therefore, some measurement noise associated with each data point may exist and it is therefore important for the data analyst to have prior knowledge of these data points which exhibit unusually large influence on the results of the analysis and their subsequent interpretation (Öfversten, 1998). In reality, this problem is almost arises if there exist anomalous observations in a data set (Öfversten, 1998; Hayat et al., 2007). Such cases may be assessed as either being appropriate and remain in the analysis, or may indicate inappropriate data and be eliminated from the analysis, or may advocate that the current modeling scheme is inadequate, or else may show a data reading or recording error (Christensen et al., 1992). Thus, the identification of these data points is necessary before any valid inferences about the characteristics of the model can be drawn which can be overcome with the help of influence diagnostic analysis.

Influence diagnostics methods that evaluate the fit of a linear regression model have been well established (Cook, 1977; Belsley et al., 1980; Cook and Weisberg, 1982) and are available in most of the statistical software like SAS, SPSS, BMDP and Minitab. Interest in mixed model diagnostics has grown recently for the detection of outliers and influential observations (Christensen et al., 1992; Demidenko and Stukel, 2005; Cavanaugh and Shang, 2005; Nobre and Singer, 2011; Turkan and Toktamış, 2012; Mun and Lindstrom, 2013) but a related problem that has received less attention is the detection of influential observations to variety performance trials and other experiments of biological importance (like genetic models), in particular. Zewotir and Galpin (2005) developed influence statistics for mixed linear models based on basic building blocks and upgrading formula with no iterative procedure for deletion of observations. We applied the analogue of Cook's distance statistics (Zewotir and Galpin, 2005) for detecting influential data points in the experimental model using the framework of LUP via MINQUE (1) (Zhu and Weir, 1994a, b). The method is illustrated by means of simulations, and real data set was analyzed to address the effect of influential observations in agricultural trials. In addition, the results of simulation are compared with the best linear unbiased prediction (BLUP) via restricted maximum likelihood (REML).

2. METHODOLOGY

Description of Mixed Linear Model

Most of the design models in terms of mixed linear model can be expressed in a general form

$$y = Xb + Ue + e_\varepsilon = Xb + \sum_{u=1}^{r+1} U_u e_u \sim MVN\left(Xb, V = \sum_{u=1}^{r+1} \sigma_u^2 U_u U_u^T \right) \tag{1}$$

where y is the $(n \times 1)$ vector of phenotype values; b is a $(p \times 1)$ vector of fixed effect with known design matrix X of order $(n \times p)$; U_u is $(n \times q_u)$ known coefficient matrix of the u-th random vector e_u $(u = 1, 2..., r)$ and $U_{r+1} = I_n$ related to the $(n \times 1)$ vector of random error $e_{r+1} = e_\varepsilon$; each e_u is the $(q_u \times 1)$ vector of the u-th random factor and $e_u \sim N\left(0, \sigma_u^2 I_{q_u}\right)$; and $e_\varepsilon \sim N\left(0, \sigma_\varepsilon^2 I\right)$. For random effects vector e, it can also be written that $e \sim N(0, D)$. Where D is the block diagonal matrix with u^{th} block being $\psi_u I_{q_u}$, for $\psi_u = \sigma_u^2$ which indicates that the vectors of random components are independent.

Considering (1), the fitted values for response y can be written as: $\hat{y} = X\hat{b} + U\hat{e}$, yielding the residuals $\hat{e}_{\varepsilon} = y - \hat{y}$; where $\hat{b}$ is the generalized least square (GLS) estimate of the fixed effects vector $b = \left(X^T V^{-1} X\right)^{-1} X^T V^{-1} y$. The symmetric matrix, $Q = V^{-1} - V^{-1} X \left(X^T V^{-1} X\right)^{-1} X^T V^{-1}$ is the projection matrix of order $(n \times n)$ which transform the observed phenotypic values into residuals (Zhu, 1997), that is: $\hat{e}_{\varepsilon} = \sigma_{\varepsilon}^2 Qy$ (using BLUP for prediction), so that, $\hat{e}_{\varepsilon} \sim N\left(0, \sigma_{\varepsilon}^4 Q\right)$, and thus $\hat{e}_{\varepsilon_i} \sim N\left(0, \sigma_{\varepsilon}^4 Q_{ii}\right)$, for $(i = 1,\ 2,\ ...,\ n)$, where Q_{ii} shows the main diagonal element of matrix Q.

The linear unbiased prediction (LUP) method uses the following equations to predict the random residuals:

$$\hat{e}_{\varepsilon(\alpha)} = \alpha_{\varepsilon} U_{\varepsilon}^T Q_{\alpha} y = \alpha_{\varepsilon} Q_{\alpha} y \text{, and } Var\left(\hat{e}_{\varepsilon(LUP)}\right) = \left(\alpha_{\varepsilon}\right)^2 P \qquad (2)$$

where, $P = Q_{\alpha} V Q_{\alpha}$. The matrices V_{α} and Q_{α} could be obtained by using the following equations:

$$V_{\alpha} = \sum_{u=1}^{r} \alpha_u U_u U_u^T + \alpha_{\varepsilon} I_n \text{ , and } Q_{\alpha} = V_{\alpha}^{-1} - V_{\alpha}^{-1} X \left(X^T V_{\alpha}^{-1} X\right)^{+} X^T V_{\alpha}^{-1}$$

From (2), it is evident that $\hat{e}_{\varepsilon(LUP)} \sim N\left(0, \alpha_{\varepsilon}^2 P\right)$ and $\hat{e}_{\varepsilon_i(LUP)} \sim N\left(0, \alpha_{\varepsilon}^2 P_{ii}\right)$, P$_{ii}$ being the main diagonal elements of matrix P. In case of MINQUE (1), V_{α} and Q_{α} can be reduced to the following expressions (Zhu and Weir, 1996);

$$V_{(1)} = \sum_{u=1}^{r} U_u U_u^T + I_n \text{, and } Q_{(1)} = V_{(1)}^{-1} - V_{(1)}^{-1} X \left(X^T V_{(1)}^{-1} X\right)^{+} X^T V_{(1)}^{-1}$$

One of the commonly used methods in plant and animal breeding experiments for prediction of random effects of a mixed linear model is known as BLUP (Henderson, 1948). Kackar and Harville (1981) determined that BLUP provide unbiased estimates when the estimates of variances are used in place of actual values (as is usually the case), although they are not guaranteed to be the best of all unbiased linear estimators (Lynch and Walsh, 1998). For known variance components, BLUP uses the following equation for predicting the random components of a mixed linear model:

$$\hat{e}_{u\left(\sigma^2\right)} = \sigma_u^2 U_u^T V^{-1} \left(y - X\hat{\beta}\right) = \sigma_u^2 U_u^T Q y$$

where σ^2 in the bracket describe the fact that BLUP uses known variances; however in case of unknown variance components, their estimates can be used instead, which can be expressed as:

$$\hat{e}_{u\left(\hat{\sigma}^2\right)} = \hat{\sigma}_u^2 U_u^T \hat{V}^{-1} \left(y - X\hat{\beta}\right) = \hat{\sigma}_u^2 U_u^T \hat{Q} y$$

So the corresponding random effect vector for residuals can be expressed as

$$\hat{e}_{\varepsilon\left(\hat{\sigma}^2\right)} = \hat{e}_{BLUP} = \hat{\sigma}^2_{\varepsilon}\hat{Q}y \text{ (as } U^T_{\varepsilon} = I_n = \text{ identity matrix of order } n\text{)}$$

where, the definition of $\hat{Q}$ and $\hat{V}$ is same like those explained above and can be obtained as:

$$\hat{Q} = \hat{V}^{-1} - \hat{V}^{-1}X\left(X^T\hat{V}^{-1}X\right)^{+} X^T\hat{V}^{-1} \text{ and } \hat{V} = \sum_{u=1}^{r+1}\hat{\sigma}^2_u U_u U^T_u = \sum_{u=1}^{r}\hat{\sigma}^2_u U_u U^T_u + \hat{\sigma}^2_{\varepsilon} I_n$$

While using BLUP for prediction of random effect, restricted maximum likelihood method (REML) (Patterson and Thompson, 1971) was used for estimating the variance components of a mixed linear model. REML method uses the following equation for estimating the variance components:

$$\left[tr\left(\hat{Q}_{[h]}U_u\hat{Q}_{[h]}U_v\right)\right]\left[\hat{\sigma}^2_{u[h+1]}\right] = \left[y^T\hat{Q}_{[h]}U_u\hat{Q}_{[h]}y\right]$$

where, $\hat{Q}_{[h]} = \hat{V}^{-1}_{[h]} - \hat{V}^{-1}_{[h]}X\left(X^T\hat{V}^{-1}_{[h]}X\right)X^T\hat{V}^{-1}_{[h]}$, $\hat{V}_{[h]} = \sum_{u}^{r+1}\hat{\sigma}^2_{u[h]}U_u$, for $u, v = 1, 2, ..., r+1$

In case of balanced data, REML provides identical estimates of variance components to that of ANOVA, and is preferred for large data set (Lynch and Walsh, 1998).

Experimental Model

We considered a general genetic experiment in a randomized complete block design with 'g' genotypes $\left(G_h\right)$, 'y' years $\left(Y_i\right)$, 'l' locations $\left(L_j\right)$ and 'b' blocks $\left(B_{k(ij)}\right)$ of each genotype within each year and location. Let y_{hijk} be the phenotypic value of the h^{th} genotype, i^{th} year; j^{th} location and k^{th} block within each year and location, which can thus be expressed by the following linear model

$$\begin{aligned} &y_{hijk} = \mu + G_h + Y_i + GY_{hi} + L_j + GL_{hj} + YL_{ij} + B_{k(ij)} + \varepsilon_{hijk} \\ &h = 1, 2, ..., g;\ i = 1, 2, ..., y;\ j = 1, 2, ..., l;\ k = 1, 2, ..., b \end{aligned} \tag{3}$$

For simplicity of exposition and computations, the higher-order interaction $(G \times L \times Y)$ was ignored. In model (3) only the population mean (μ) and the genotype effect $\left(G_h\right)$ were considered fixed while all other effects were considered as random. Equation (3) can be expressed by the following matrix notation of a mixed linear model

$$\begin{aligned} y &= 1\mu + X_G b_G + U_Y e_Y + U_{GY} e_{GY} + U_L e_L \\ &\quad + U_{GL} e_{GL} + U_{YL} e_{YL} + U_{B(YL)} e_{B(YL)} + U_{\varepsilon} e_{\varepsilon} \\ &= Xb + \sum_{u=1}^{r} U_u e_u + e_{\varepsilon} \sim MVN\left(Xb, V = \sum_{u=1}^{r+1}\sigma^2_u U_u U^T_u + \sigma^2_{\varepsilon} I\right) \end{aligned} \tag{4}$$

Equation (4) has the same mathematical structure as (1) with usual assumptions of normality of all the random terms but we have considered (2) when the studentized residuals were of particular interest.

Influence Functions

In linear regression, the influence diagnostic of Cook (Cook, 1977) *i.e.* Cook's distance is based on the case deletion and measure the effect of deleting an observation on the estimated regression coefficient and the fitted values without iteration because of the availability of its closed form upgrading formula (Beckman and Trussell, 1974; Miller, 1974). However, the analogue of the Cook distance statistics (Zewotir and Galpin, 2005) were adopted to measure the influence of each deleted observations on the fixed as well as random effects of a mixed linear model (4) in the framework of LUP *via* MINQUE (1), in the present study.

The analogue of Cook's distance (Zewotir and Galpin, 2005) which measure the effect of deleted observations on the estimation of fixed effects, denoted by $CD_i(b)$, can be expressed as

$$\begin{aligned} CD_i(b) &= \left(\left(\hat{b}_{(i)}-\hat{b}\right)^T\left(X^T\hat{V}^{-1}X\right)\left(\hat{b}_{(i)}-\hat{b}\right)\right)\Big/p \\ &= \left(\left(\hat{v}^{ii}-\hat{Q}_{ii}\right)t_i^2\right)\Big/\hat{Q}_{ii}\,p,\ (i=1,2,..,n) \end{aligned} \tag{5}$$

where, $\hat{v}^{ii}$ indicates the main diagonal elements of matrix V^{-1} computed by MINQUE (1), $\hat{Q}_{ii}$ is the main diagonal element of the $Q_{(1)}$ matrix; p shows the number of parameters to be estimated; and t_i is the studentized residual computed by:

$$t_i = \hat{e}_{\varepsilon i}\Big/\alpha\sqrt{P_{ii}} \tag{6}$$

It follows a *t*-distribution with $df = n - rank\,|XU|$ (SAS, 1999). A large value of $CD_i(b)\ (i=1,2,...,n)$ will indicate that a particular observation is influential. However, in case of BLUP *via* REML the following expression was used to compute the studentized residuals:

$$t_i = \hat{e}_{\varepsilon_i}\Big/\hat{\sigma}_\varepsilon^2\sqrt{\hat{Q}_{iiii}} \sim t(n-p-1) \tag{7}$$

Similarly, analogue of the Cook's distance for the influence of predicted random effects can be defined as

$$CD_i(e) = \left(\left(\hat{e}-\hat{e}_{(i)}\right)^T D^{-1}\left(\hat{e}-\hat{e}_{(i)}\right)\right)\Big/\hat{\sigma}_\varepsilon^2 = t_i^2\left(1-\left(\hat{\sigma}_\varepsilon^2\times ssq\left(\hat{Q}_i\right)\right)\Big/\hat{Q}_{ii}\right) \tag{8}$$

where t_i is the studentized residual obtained by using (6), and $ssq\left(\hat{Q}_i\right)(i=1,2,...,n)$ is the sum of squares of the elements of the i^{th} column of $Q_{(1)}$ matrix. Any large value of

$CD_i(e)$ for the i^{th} observation will indicate that the observation is influential (Zewotir and Galpin, 2005).

Furthermore, we calculate the P-value of each observation from its corresponding t-value and define an observation as outlier if it is significant at 5% level of significance. In addition, P_{cutoff} is calculated by the FDR method (Benjamini and Hochberg, 1995) to control the false positives.

3. RESULTS

3.1 Simulations

In this section, two examples are considered which are based on 200 simulations with different perturbation schemes to illustrate the detection ability of these influence functions.

An experimental model (4) for balanced data was considered to meet the required objectives of the study. Given the known values of variance components for random effects $\sigma_u^2 (u = 1, 2, ..., r+1)$ and the true values of fixed effects (genotypes), 200 data sets were simulated using a program written in C++ programming language. To obtain the values of X and U matrices, a real data set was considered to generate the random values from stochastic residuals e_ε and the random components e_u, and thus the phenotypic values (y) were obtained. In generating the data set, it was assumed that all the random factors are independent and each of the random factors follows a normal distribution *i.e.* $e_u \sim N\left(0, \sigma_u^2\right)$. While simulating the data, two cases were considered for understanding the detection of influential data points. In the first case (case 1), the following assumptions were taken about the distribution of the random factors *i.e.* $e_Y \sim N(0,3.0)$; $e_L \sim N(0,2.0)$; $e_{GY} \sim N(0,3.5)$; $e_{GL} \sim N(0,2.5)$; $e_{YL} \sim N(0,3.75)$; $e_{B_{(YL)}} \sim N(0,3.0)$ and $e_\varepsilon \sim N(0,1.0)$.

At each replicate of 200 simulated data sets, the response vector at case number 100 was incremented by a number 5, to make it aberrant. In the second case (case 2), we considered the same sets of fixed and random effects with the same parameter values but two aberrant cases were introduced at case number 100 and 300, respectively with identical magnitude as defined for case 1. In these simulations, each data set contained a total of 450 observations, consisting of 10 genotypes and 5 locations with in each of the three years, in addition to 3 blocks for each of the genotype within locations and years. We did not change the factors level of any of the fixed genotypic effects and that of the random factors involved in model (4) to assume that there is no outlier (leverage) in the fixed and random spaces (factor space), during all the simulations.

Fig. 1 indicates the index plot of $CD(b)$ and $CD(\mathrm{e})$ for the 200 simulated data sets (case 1). It turned out that the case number 100 is highly influential in affecting both the fixed genotypic and random predicted effects. To check the effectiveness of our approach for detection of influential observations we compared it with those of Zewotir and Galpin (2005). Compared to Fig. 2b, our approach (Fig. 2a) exhibits a good agreement with both the influence statistics used for detection of influential observation for the influence of fixed as well as the random effects, respectively. In approximation, both the methods have the same detection ability and almost the same trends for detecting influential data points.

Fig. 2 illustrates the index plot of Cook's distance statistics for 200 simulated data sets, for case 2. Both the analogue of Cook's distance statistics can effectively detect that the data points at case number 100 and 300 are influential in affecting both the fixed genotypic effects and the predicted random components of model (4). Our approach (Fig. 2a) has a nice resemblance and in agreement with Zewotir and Galpin approach (Fig. 2b).

3.2 Worked Example

In this section, the experimental data of real experiments of rice yield available in the software QGA Station (Chen and Zhu, 2003) is considered for illustration. The data set has a balanced structure with 5 genotypes tested within four locations over two years in a randomized complete block design (RCBD) with three replications. It consists of a total of 120 observations with a minimum phenotype value of 9.6 and a maximum of 75.6. The genotypic effect was considered fixed while all other factors were taken as random. To demonstrate the procedure, we considered model (4) and the data were analyzed in the framework of a mixed linear model for obtaining the analogue of Cook's distance statistics and other required statistical quantities.

The diagnostics plots for screening of influential data points affecting the fixed as well as random effects by using the analogues of Cook's distance statistics for experimental data of rice yield are shown in Fig. 3, for both the methods. Fig. 3a demonstrates that there exists only one data point at case number 100 which is highly influential in affecting both of the fixed and random effects, and the same was detected by the other method (Fig. 3b). Except the case number 100, some other data points at case numbers 102 and 113 were also detected as influential but showed lower peaks as compared to the case number 100.

To further investigate the data point at case number 100, we referred to the original data set. This case number belongs to the data point (5, 1, 2, 1) of model (4) *i.e.* genotype 5, year 1, location 2 and replication 1. The yield for genotype 5 in all the three replicates of the same year 1 and location 2 were 71.5, 47.6 and 55.4. This indicates that the difference between genotype mean $\left(\bar{y}_{5...}\right)$ and the grand mean $\left(\bar{y}_{....}\right)$ could be more substantial and thus having a more profound impact on the estimates of variance components as well as the fixed genotypic effects, in addition to the predicted random effects. In a similar way, the case numbers 102 and 113 correspond to the data points (5, 1, 2, 3) and (5, 2, 2, 2), respectively, also showed high peaks to both the influence functions demonstrating to be influential data points. The data point at case number 102 corresponds to the 3rd replication of the same genotype 5, year 1 and location 2, whereas the data point (5, 2, 2, 2) corresponds to case number 113 indicating the same 5th genotype, in the 2nd location and year 2 (Table 1).

How could these data points affect the results of the analysis? We performed the case deletion diagnostics and the variance components of residuals (only) for the case deleted data sets were obtained (Fig. 4). It revealed that the deletion of case number 100 is associated with a drastic decrease in the residuals variance; whilst the effects of the other two cases was small as compared to case number 100, in the reduction of residual variance. The estimates of variance components for deleting the data of case number(s) 100; 102 and 103; 100, 102 and 103; all influential data points and outliers (Table 1); and that of the full data are listed in Table 2. The results illustrate that much improved estimates of variance components and particularly that of residuals can be obtained in the absence of influential observations and outliers. However, the net reduction in residual variance due to case number 102 and 113 was not too high as compared to only the case number 100, and the same was further clarified by the studentized residuals that the case number 100 is the most influential observation and a clear outlier $\left(p < 0.01 \right)$ (results not shown).

4. DISCUSSION, RECOMMENDATION AND CONCLUSIONS

In linear regression, the influence diagnostics for influence on the estimates of model parameters and their underlying assumptions are based on modification of the response and explanatory variables (Thomas, 1990). We considered small perturbations in response by using MINQUE (1) and LUP as tools for estimation of variance components and prediction of random effects, respectively. Monte Carlo simulations showed that the influence functions in the above defined framework works well in identifying the influential data points and has the same detection ability when BLUP was used for prediction of random effects *via* REML (Figs. 1-2). Zewotir and Galpin (2006) applied a series of simulations, and in each iterated simulated data they considered the double of the maximum response value to make it anomalous, which was not adopted in the present study. We considered only a fixed data point for a specified case number (without regard to their being a minimum or maximum) of a response and a very small positive quantity was added to make it aberrant.

The main feature of the present study was to check the performance of the proposed method and to make plant breeders aware of the fact that the estimates of parameters in the mixed linear model could be affected due to small change in the phenotype data. Monte Carlo simulations reveal that our approach works well in identifying unusual data points, if present, in the phenotype data. The simplicity of our approach is that it uses the method of MINQUE (1) for estimation of variance components and LUP for prediction of random factors (Zhu and Weir, 1994a, b). The advantageous feature of MINQUE (1) method is that of unbiased and efficient estimation of variance components of the random factors with less computational time as compared to the REML method. MINQUE (1) is a non-iterative method but the REML needs iteration for estimation of variance components, and in some cases it is not possible to converge (Zhu, 1992). It was observed that any small change in the phenotype data can made it aberrant and hence was detected by both the methods as influential.

In plant breeding, evaluation of different genotypes in multi-location trials is one of the crucial steps to select suitable cultivars for improved yield production to find for each environment the genotype that is best adapted (Yan and Rajcan, 2003). Extensive experimentation may sometimes lead to erroneous or otherwise abnormal data (Öfversten, 1998) where mixed models and the likelihood methods are not robust in the presence of these data points (Christenson et al., 1992; Haslett, 1999). In the worked example of rice yield data, we observed three influential observations (Fig. 3) with considerably larger influence on the estimates of various parameters (Fig. 4, Table 2). The imperative feature of these atypical observations is that all these observations (influential data points as well as outliers) were associated with the 5th genotype and more frequently in the 2nd location. We are not well aware of other basic steps in collecting these observations; however, one possible reason might be that this particular genotype would be more sensitive (or stable) at location 2 and thus the methods show it as the most influential or else its measurement could be a result of some sort of random error or data entry error. Another possible reason could be the complexity of the response of a genotype to a particular location, which must depend upon the effects of the total seasonal pattern to which a genotype has been subjected or a year may have a general period of some sort of stress, stresses at different sites with regard to that particular genotype at different periods in the life cycle of the plants (Hanson, 1964). In general, examining the data may help us to identify the steps in the field work that are particularly exposed to inaccuracies and errors and, consequently, to improve the entire testing process (Öfversten, 1998).

Mixed linear models are commonly used in determining sampling designs, quality control procedures and statistical genetics (Christensen et al., 1992). Statistical genetics as an active area of research for the inheritance of complex traits, most of the data sets restrain outliers and influential observations that can seriously affect the estimates of genetic variances of variable effects (additive, dominance and epistasis), prediction of random effects (breeding values) and various tests involved. In the present study, we have considered a general genetic model but the method can be easily extended to more complex genetic models like additive dominance (AD) model, additive dominance maternal (ADM) model, diploid plant seeds model, triploid endosperm model (Zhu, 1992, 1997; Zhu and Weir, 1994a, b) and that of QTL mapping model (Yang et al., 2007). It is important to mention that the current method have been applied for the influence of outliers on QTL mapping for complex traits, providing the evidence of additional QTLs and epistatic loci effecting the 1stBrain and the endbrains-OB in a cross of BAD mouse population, and reveal a remarkable increase in estimating heritability of QTL in the absence of influential observations and outliers (Hayat et al., 2008).

Apart from the analogue of Cook's distance statistic $\left(CD_i(b)\right)$, some other influence diagnostic statistics like variance ratio (VR), Andrew-Region (AP) statistic and Cook-Weisberg statistic (COW) are also used for identifying outliers and influential cases affecting the fixed effects of a mixed linear model (results not reported). Analogue of these influence statistics (Zewotir and Galpin, 2005) can be easily applied in mixed linear model for identification of aberrant cases. The analogue of VR measures the change in the determinant of the variance-covariance matrix of the fixed effect parameters estimates $\hat{\beta}$ when the *i*-th case $(i = 1, 2, ..., n)$ is deleted. The analogue of AP statistic measures the influence of *i*-th observation $(i = 1, 2, ..., n)$ on the fit of the data. Similarly, the analogue of the Cook-Weisberg statistic measure the change of the confidence ellipsoid volume of fixed effect parameters estimates (Zewotir and Galpin, 2005). All these influence statistics also detected the same points as influential which were detected by using the analogue of Cook's distance statistic $\left(CD_i(b)\right)$.

Therefore, it is suggested that when performing any sort of agricultural trials, it is necessary to perform a thorough data quality check by searching the data set for possible outliers and influential data points so that a correct interpretation of the genetic phenomena can be carried out. It is worth mentioning that the methods can be easily affected by the well-known masking and swamping effects. However, in real data analysis it is suggested to perform influence diagnostic analysis as the prerequisite requirements so that the effects of unusual observations on the estimates of various parameters of a genetic model can be minimized, to ensure efficient and unbiased estimates of the model parameters.

REFERENCES

1. Balsley, D.A. Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Source of Collinearity*. John Willey and Sons.
2. Beckman, R.J. and Trussell, H.J. (1974). The distribution of an arbitrary studentized residuals and the effects of updating in multiple regression. *J. Amer. Statist. Assoc.*, 69(345), 199-201.
3. Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Series B,* 57, 289-300.
4. Cavanaugh, J.E. and Shang, J. (2005). A diagnostic for assessing the influence of cases on the prediction of random effects in a mixed model. *Journal of Data Science,* 3, 137-151.
5. Chen, G.B. and Zhu, J. (2003). Software for the classical quantitative genetics. Copy Right, Institute of Bioinformatics, Zhejiang University, Hangzhou, China: http://ibi.zju.edu.cn/software/qga/index.htm.
6. Christensen, R., Pearson, L.M. and Johnson, W. (1992). Case-deletion diagnostics for mixed models. *Technometrics*, 34(1), 38-45.
7. Cook, R.D. (1977). Detection of influential observations in linear regression. *Technometrics*, 19, 15-18.
8. Cook, R.D. and Weisberg, S. (1982). *Residual and Influence in Regression*. Chapman and Hall.
9. Demidenko, E. and Stukel, T.A. (2005). Influence analysis for linear mixed-effects models. *Statist. Med.,* 24, 893-909.
10. Hanson, W.D. (1964). Genotype-environment interaction concepts for field experimentation. *Biometrics*, 20(3), 540-552.
11. Haslett, J. (1999). A simple derivation of deletion diagnostics for the general linear model with correlated errors. *J. Roy. Statist. Soc., B*, (61), 603-609.
12. Hayat, Y., Salahuddin, Mahmood, Q., Islam, E. and Yang, J. (2007). Comparative study of outliers based on statistical methods to evaluate and select the optimum regression model for fertilizers utilization. *Sci. Res. Monthly,* 3, 81-84.
13. Hayat, Y., Yang, J., Xu, H.M. and Zhu, J. (2008). Influence of outliers on QTL mapping for complex traits. *Journal of Zhejiang Univ. Sci. B.*, 9(12), 931-937.

14. Henderson, C.R. (1948). *Estimation of general, specific and maternal combining abilities in crosses among inbred lines of swine*. Ph.D. Thesis, Iowa State.
15. Kackar, R.N. and Harville, D.A. (1981). Unbiasedness of two-stage estimation and prediction for mixed linear models. *Comm. Stat. Theor. Meth.* A10, 1249-1261.
16. Lynch, M. and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits.* Sinauer Associates, Sunderland, Massachusetts, USA.
17. Miller, R.G. (1974). The Jackknife: A review. *Biometrika*, 61, 1-15.
18. Mun, J. and Lindstrom, M.J. (2013). Diagnostics for repeated measurements in linear mixed effects models. *Journal of Applied Statistics*, 32 1361-1375.
19. Nobre, J.S. and Singer, J.M. (2011). Leverage analysis for linear mixed models. *Journal of Applied Statistics*, 38(5), 1063-1072.
20. Öfversten, J. (1998). Assessing sensitivity of agricultural crop varieties. Journal of Agricultural, Biological and Environmental Statistics, 3(1), 37-47.
21. Patterson, A.H., and Thompson, R., 1971. Recovery of inter-block information when block sizes are unequal. Biometrika, 58: 545–554.
22. SAS (1999). SAS STAT User's Guide Version 7 and 8, SAS Institute Inc, Cary, NC USA. pp. 2118.
23. Thomas, W. (1990). Influence on the confidence regions for regression coefficients in generalized linear models. *American Statistical Association*, 85(410), 393-397.
24. Turkan, S. and Toktamış, Ö. (2012). Influence analysis in the mixed model. Pak. J. Statist., 28(3), 341-349.
25. Yan, W. and Rajcan, I. (2003). Prediction of cultivar performance based on single- versus multiple-year tests in soybean. *Crop Science*, 43, 549-555.
26. Yang, J., Zhu, J. and Williams, R.W. (2007). Mapping genetic architecture of complex trait in experimental populations. *Bioinformatics*, 23, 1527-1536.
27. Zewotir, T. and Galpin, J.S. (2005). Influence diagnostics for linear mixed models. *Journal of Data Science*, 3, 153-177.
28. Zewotir, T. and Galpin, J.S. (2006). Evaluation of linear mixed model case deletion diagnostic tools by Monte Carlo simulation. *Commun. in Statist. – Simul. and Compu.*, 35, 645-682.
29. Zhu, J. (1992). Mixed model approaches for estimating genetic variances and covariances. *Journal of Biomathematics*, 7(1), 1-11.
30. Zhu, J. (1994). General genetic models and new analysis methods for quantitative traits (in Chinese). *Journal of Zhejiang Agricultural University*, 20, 551-559.
31. Zhu, J. and Weir, B.S. (1994a). Analysis of cytoplasmic and maternal effects-I. A genetic model for diploid plant seeds and animals. *Theoretical and Applied Genetics*, 89, 153-159.
32. Zhu, J. and Weir, B.S. (1994b). Analysis of cytoplasmic and maternal effects-II. Genetic model for triploid endosperms. *Theoretical and Applied Genetics*, 89, 160-166.
33. Zhu, J. and Weir, B.S. (1996). Diallel analysis for sex-linked and maternal effects. *Theoretical and Applied Genetics*, 92, 1-9.
34. Zhu, J. (1997). *Analysis Methods for Genetic Models*. Beijing Agricultural Publication House of China (in Chinese).

Table 1
Summary Statistics of Outliers and Influential Data Points for Rice Yield Data

Case No.	Data point[+]	t_i[‡]	*P*-value
100	(5, 1, 2, 1)	3.574	0.0006**
102	(5, 1, 2, 3)	2.615	0.0108**
110	(5,2, 1, 2)	1.995	0.0496*
113	(5, 2, 2, 2)	-2.678	0.0091**
114	(5, 2, 2, 3)	-2.051	0.0437*

[+] the numbers in brackets (*h, i, j, k*) correspondingly shows the *h*-th genotype, *i*-th year, *j*-th location and *k*-th replication, respectively; [‡] Studentized residual; ** significant at $P_{cuttof} = 0$ .011304; * significant at $P < 0.05$.

Table 2
MINQUE (1) Estimates of Variance Components for Random Effects of the Rice Yield

Parameter	Full Data	a	b	c	d
Year	87.62	93.90	95.82	103.38	102.73
Loc	85.60	77.50	83.87	70.19	74.18
Gen*Year	7.25	5.74	5.41	5.53	5.67
Gen*Loc	10.95	12.88	12.04	16.17	16.68
Year*Loc	24.63	28.48	26.46	32.81	32.32
B(Year*Loc)	12.75	9.83	18.95	15.82	16.35
Residual	42.63	36.88	35.95	28.06	25.96

“a”: datum of case number 100 was deleted; “b”: data points at case number 102 and 113 were deleted; “c”: deletion of all the data points *i.e.* at case number 100, 102 and 113; “d”: deletion of all data points with case numbers listed in Table 1.

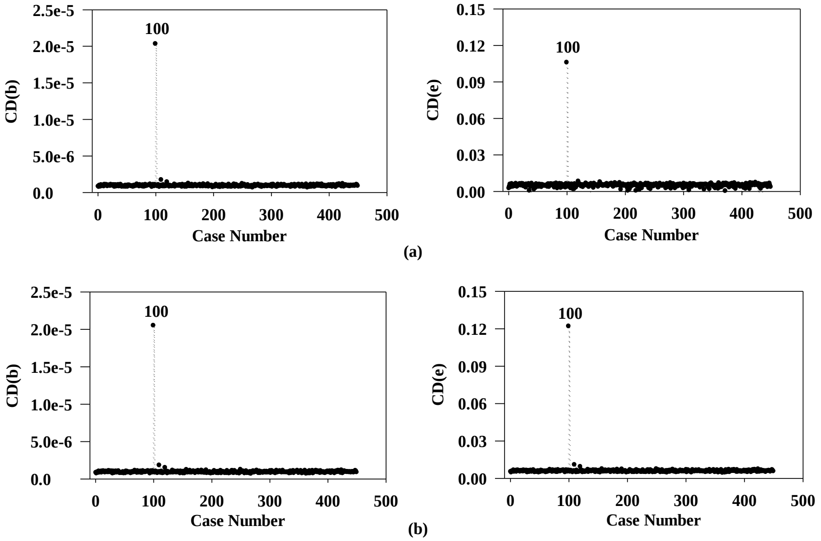

Fig. 1: Index plot of influence diagnostics functions *i.e.* the analogues of Cook's distance (CD(b) and CD(**e**)) for influence on the fixed effects and prediction of random effects, respectively for the simulated data (Case 1) using (**a)** MINQUE (1) for estimation of variance components and LUP for prediction of random effects **(b)** REML for estimation of variance components and BLUP for prediction of random effects.

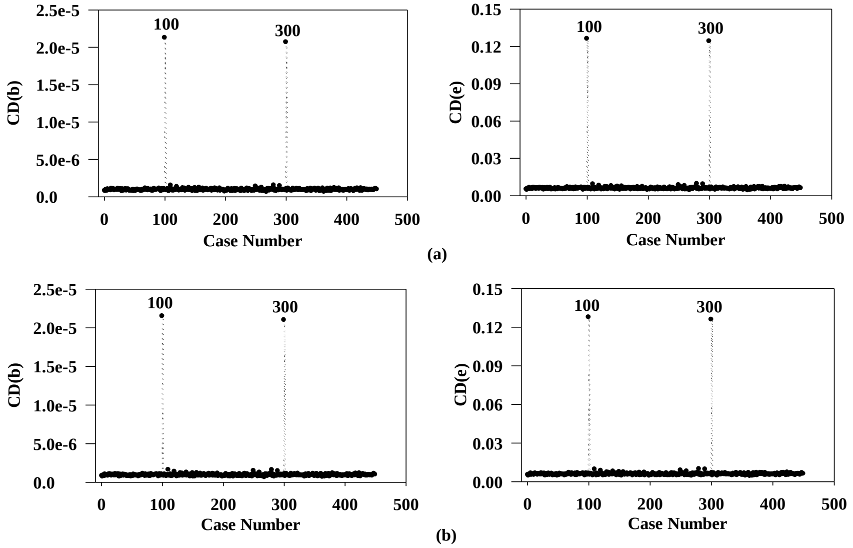

Fig. 2: Index plot of influence diagnostics functions *i.e.* the analogues of Cook's distance (CD(b) and CD(**e**)) for influence on the fixed effects and prediction of random effects, respectively for the simulated data (Case 2) using (**a)** LUP *via* MINQUE (1) **(b)** BLUP *via* REML.

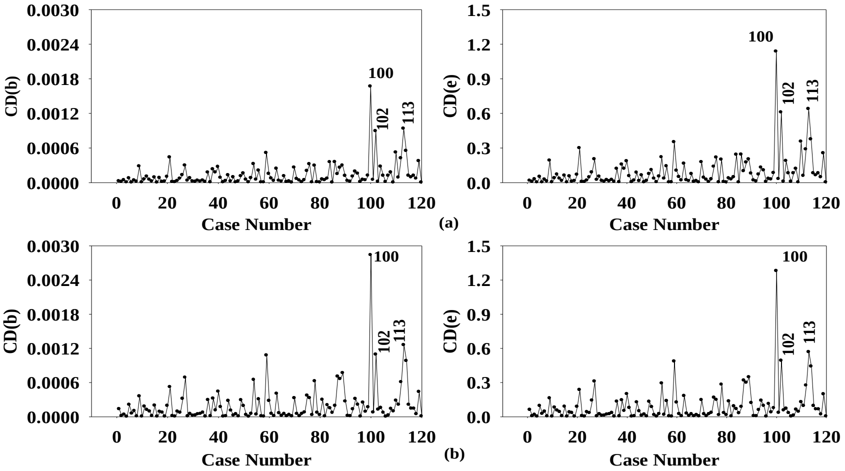

Fig. 3: Index plot of influence diagnostics functions *i.e.* the analogues of Cook's distance (CD(b) and CD(**e**)) for influence on the fixed effects and prediction of random effects, respectively for the rice yield data using **(a)** LUP *via* MINQUE (1) **(b)** BLUP *via* REML.

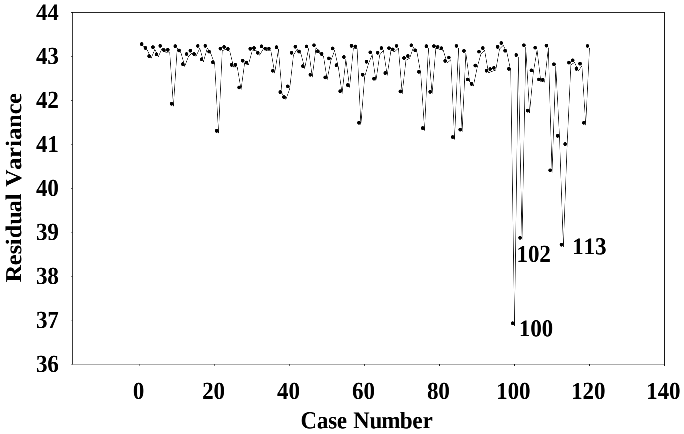

Fig. 4: The index plot for the estimates of residuals variance for stepwise deleted observations of the rice yield data set by using MINQUE (1)