Journal of Statistical Modelling: Theory and Applications
Vol. 6, No. 2, 2025, pp. 137-151
Yazd University Press 2025

JSMTA

Research Paper

Leverage and influential observations on the restricted Liu estimator in the linear mixed measurement error models

Fatemeh Ghapani*
Department of Mathematics and Statistics, Sho.C., Islamic Azad University, Shoushtar, Iran

Received: November 17, 2025/ Revised: May 14, 2026/ Accepted: Augest 12, 2026

Abstract: A fundamental step in constructing linear regression models is the identification of atypical observations, specifically high-leverage and influential points. The coexistence of atypical observations and severe multicollinearity complicates statistical analysis and reduces the effectiveness of conventional diagnostic methods. To address these challenges, this paper proposes diagnostic methods based on the restricted Liu estimator for identifying atypical observations in linear mixed measurement error models. Specifically, generalized leverage matrices are derived using the restricted Liu estimator to identify high-leverage points. Furthermore, several case-deletion measures are developed based on the restricted Liu estimator to serve as robust tools for diagnosing influential points. The performance of the proposed methods is evaluated through two simulation studies and an application to a real-life example.

Keywords: Diagnostic methods; Leverage points; Linear mixed measurement error model; Restricted Liu estimator.
Mathematics Subject Classification (2010): 62J05, 62J20.

1 Introduction

Linear mixed models generalize simple linear models by allowing for both random and fixed effects simultaneously. One of the basic assumptions in regression analysis is that all observations are correctly observed. However, independent variables are often measured with unavoidable errors in statistical models. The presence of measurement errors in the observations violates the essential properties of estimators. An important

*Corresponding author: `fatemeh.ghapani@iau.ac.ir`

issue in the area of measurement errors is finding consistent estimators of the parameters. In order to correct the effects of measurement error on parameter estimation, Nakamura (1990) considered an approach based on the correction of the score function. This approach makes it possible to perform inference as well as parameter estimation without additional assumptions.

The linear mixed measurement error (LMME) model is expressed as

$$\begin{aligned} y_i &= Z_i\beta + U_i b_i + \varepsilon_i, \quad i = 1, \ldots, l, \\ X_i &= Z_i + \delta_i, \end{aligned}$$

where y_i denotes an $n_i \times 1$ vector of responses for variables measured on subject i; β is a $p \times 1$ parameter vector of fixed effects; Z_i is an $n_i \times p$ known design matrix of full column rank for the fixed effects, which can be observed through the matrix X_i along with the measurement error δ_i, where δ_i is an $n_i \times p$ vector of uncorrelated random variables with $E(\delta_i) = 0$ and $Var(\delta_i) = \Lambda$, with Λ being a $p \times p$ matrix of known values having nonnegative diagonal elements (Fuller, 1987). U_i is an $n_i \times q_i$ known design matrix for the random effect of factor i, and b_i is a $q_i \times 1$ vector of unobservable random effects distributed as $N(0, \sigma_i^2 I_{q_i})$ and independent of ε_i, which is an $n_i \times 1$ vector of unobservable random errors distributed as $N(0, \sigma^2 I_{n_i})$. The variances σ^2 and σ_i^2 are called variance components. It is assumed that ε_i and δ_i are mutually independent. Let $y = (y_1', y_2', \ldots, y_l')'$, $n = \sum_{i=1}^{l} n_i$, $Z = (Z_1', Z_2', \ldots, Z_l')'$, $U = \oplus_{i=1}^{l} U_i$, where $\oplus$ denotes the direct sum, $q = \sum_{i=1}^{l} q_i$, $b = (b_1, b_2, \ldots, b_l')'$, $\varepsilon = (\varepsilon_1', \varepsilon_2', \ldots, \varepsilon_l')'$, and $\Delta = [\delta'_1, \delta'_2, \ldots, \delta'_n]'$. Then, the general LMME can be expressed as

$$\begin{aligned} y &= Z\beta + Ub + \varepsilon, \\ X &= Z + \Delta, \end{aligned} \tag{1}$$

where $b \sim N(0, \sigma^2\Sigma)$, and Σ is a block diagonal matrix with the i-th block being $\gamma_i I_{q_i}$ for $\gamma_i = \sigma_i^2/\sigma^2$. Consequently, y follows $MN(Z\beta, \sigma^2 V)$, where $V = I_n + \sum_{i=1}^{l} \gamma_i U_i U_i' = I_n + U\Sigma U'$. Also, Δ is an $n \times p$ random matrix distributed as $N(0, I_n \otimes \Lambda)$. The conditional distribution of $b \mid y$ is $N\left(\Sigma U' V^{-1}(y - Z\beta), \sigma^2 \Sigma T\right)$, where $T = (I_q - U'V^{-1}U\Sigma) = (I_q + U'U\Sigma)^{-1}$.

The corrected likelihood estimates (CLE) of the parameters are given by

$$\hat{\beta} = \left[X'V^{-1}X - \mathrm{tr}(V^{-1})\Lambda\right]^{-1} X'V^{-1}y = A^{-1}X'V^{-1}y,$$

where $A = \left[X'V^{-1}X - \mathrm{tr}(V^{-1})\Lambda\right]$, and

$$\hat{\sigma}^2 = n^{-1}\left[(y - X\hat{\beta})'V^{-1}(y - X\hat{\beta}) - \mathrm{tr}(V^{-1})\hat{\beta}'\Lambda\hat{\beta}\right] = n^{-1}\left(y'V^{-1}y - \hat{\beta}'X'V^{-1}y\right),$$

where $\hat{D}_i = \hat{\gamma}_i U_i' V^{-1} = \frac{\hat{\sigma}_i^2}{\hat{\sigma}^2} U_i' V^{-1}$, and $\tilde{b} = \Sigma U' V^{-1}(y - X\hat{\beta})$ is the prediction of random effects, and T_{ij} is the ij-th block of the matrix T (see Zhong et al., 2002; Zare et al., 2012).

In linear regression models, independent variables are often susceptible to non-negligible errors, and then it is more appropriate to consider measurement error models (Fuller, 1987). However, in measurement error models, ordinary maximum likelihood

(ML) estimates lose consistency. In order to correct the bias in parameter estimation, a method that corrects the score function itself for measurement errors is available for parameter estimation. This method is based on the corrected log-likelihood of Nakamura (1990). Zhong et al. (2002) used the corrected score method of Nakamura (1990) and obtained estimates of fixed and random effects parameters. Also, Zare et al. (2012) introduced the corrected score estimates of variance components in these models. To overcome multicollinearity, authors such as Ghapani and Babadi (2022b), Ganjealivand and Ghapani (2024) and others investigated parameter estimates in LMME models.

In LMME models, the existing work in the field of diagnostic methods is related to Fung et al. (2003) and Zare and Rasekh (2011, 2014). Borhani et al. (2023) used the Liu estimator to derive diagnostic measures for the detection of outliers and influential observations in LMME models. Also, Borhani et al. (2024) investigated diagnostic methods for identifying influential observations using the Liu corrected likelihood estimator in LMME models in the presence of multicollinearity. Recently, Ghapani (2024) investigated diagnostic measures in LMME models based on the restricted ridge estimator. Since there is no outstanding work on diagnostic methods for LMME models using the restricted Liu estimator (RLE), our primary aim in this paper is to concentrate on diagnostic methods in LMME models using the RLE based on the corrected log-likelihood function of Nakamura (1990).

The paper is organized as follows. In Section 2, we introduce the RLE in LMME models using Nakamura's approach. Using the proposed estimates, diagnostic methods are investigated in Section 3. Then, several diagnostic methods are constructed for the detection of high leverage and influential observations. In Section 4, two simulation studies are conducted to demonstrate the performance of the proposed diagnostic criteria. In Section 5, the proposed diagnostic methods are evaluated through the analysis of a real dataset. Finally, Section 6 presents the summary and concluding remarks, followed by the references.

2 Stochastic restricted Liu estimator

Consider model in (1) under the stochastic linear restrictions given by

$$r = R\beta + e, \tag{2}$$

where r is an $m \times 1$ random vector, R is a known $m \times p$ matrix of rank $m < p$, and e is an $m \times 1$ vector of unobservable random errors distributed as $N(0, \sigma^2 W)$. Also, we assume that ε is stochastically independent of e. Under models in (1) and (2), we have

$$y_r = Z_r\beta + U_r b + \varepsilon_r, \qquad \text{or equivalently} \qquad \begin{bmatrix} y \\ r \end{bmatrix} = \begin{bmatrix} Z \\ R \end{bmatrix} \beta + \begin{bmatrix} U \\ 0 \end{bmatrix} b + \begin{bmatrix} \varepsilon \\ e \end{bmatrix}, \tag{3}$$

where b and y_r are jointly distributed as

$$\begin{bmatrix} b \\ y_r \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ Z_r\beta \end{bmatrix}, \begin{bmatrix} \sigma^2\Sigma & \sigma^2\Sigma U_r' \\ \sigma^2 U_r\Sigma & \sigma^2 V_r \end{bmatrix} \right),$$

with $V_r = \begin{bmatrix} V & 0 \\ 0 & W \end{bmatrix}$. Then the conditional distribution of b given y_r is

$$b \mid y_r \sim N\left(\Sigma U_r' V_r^{-1}(y_r - Z_r\beta),\, \sigma^2 \Sigma T_r \right),$$

where $T_r = I_q - U_r' V_r^{-1} U_r \Sigma$.

The corrected log-likelihood of model (3) is written as

$$\begin{aligned} l^*(\theta; X, y, r) &= -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2}\log|V| - \frac{1}{2}\log|W| \\ &\quad -\frac{1}{2\sigma^2}\left[(y - X\beta)'V^{-1}(y - X\beta) + (r - R\beta)'W^{-1}(r - R\beta)\right. \\ &\quad \left. - \operatorname{tr}(V^{-1})\beta'\Lambda\beta\right], \\ l_b^*(\theta; X, y, r) &= -\frac{q}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\log|\Sigma T_r| \\ &\quad -\frac{1}{2\sigma^2}\left\{\left[b - \Sigma U_r' V_r^{-1}(y_r - X_r\beta)\right]'(\Sigma T_r)^{-1}\left[b - \Sigma U_r' V_r^{-1}(y_r - X_r\beta)\right]\right. \\ &\quad \left. - \operatorname{tr}(I_n - V^{-1})\beta'\Lambda\beta\right\}, \end{aligned}$$

where $\theta = (\beta, \sigma^2, \gamma)$ and $N = n+m$. If V and W are known, by solving $\partial l^*(\theta; X, y, r)/\partial\beta = 0$, $\partial l^*(\theta; X, y, r)/\partial\sigma^2 = 0$, and $\partial l_b^*(\theta; X, y, r)/\partial b = 0$, the corrected stochastic restricted estimates of β, σ^2, and the predictor of b are given by

$$\begin{aligned} \hat{\beta}_r &= A_r^{-1}\left(X'V^{-1}y + R'W^{-1}r\right), \\ \hat{\sigma}_r^2 &= N^{-1}\left[(y - X\hat{\beta}_r)'V^{-1}(y - X\hat{\beta}_r) + (r - R\hat{\beta}_r)'W^{-1}(r - R\hat{\beta}_r) - \operatorname{tr}(V^{-1})\hat{\beta}_r'\Lambda\hat{\beta}_r\right], \\ \tilde{b}_r &= \Sigma U'V^{-1}(y - X\hat{\beta}_r), \end{aligned}$$

where $A_r = X'V^{-1}X - \operatorname{tr}(V^{-1})\Lambda + R'W^{-1}R$.

The presence of collinearity has some destructive effects on regression analysis. In order to resolve this problem, we use the stochastic RLE. For this purpose, the stochastic linear restriction is given by $d\hat{\beta}_r = \beta + \varphi$, where $\varphi \sim N(0, \sigma^2 I_p)$ and $0 < d < 1$ is the Liu biasing parameter. This can be unified with model (3) as follows:

$$\begin{aligned} y_r &= Z_r\beta + U_r b + \varepsilon_r, \\ d\hat{\beta}_r &= \beta + \varphi. \end{aligned} \tag{4}$$

The appropriate corrected log-likelihood function for model (4) is given by

$$l^*(\theta, d; X, y, r) = l^*(\theta; X, y, r) - \frac{1}{2\sigma^2}\left[(d\hat{\beta}_r - \beta)'(d\hat{\beta}_r - \beta)\right].$$

The RLE of β and σ^2 are obtained by differentiating $l^*(\theta, d; X, y, r)$ with respect to β and σ^2. Then we have

$$\begin{aligned} \hat{\beta}_{rl} &= \left(X'V^{-1}X - \operatorname{tr}(V^{-1})\Lambda + R'W^{-1}R + I_p\right)^{-1}\left(X'V^{-1}y + R'W^{-1}r + d\hat{\beta}_r\right), \\ \hat{\sigma}_{rl}^2 &= N^{-1}\left[(y - X\hat{\beta}_{rl})'V^{-1}(y - X\hat{\beta}_{rl}) + (r - R\hat{\beta}_{rl})'W^{-1}(r - R\hat{\beta}_{rl})\right. \\ &\quad \left. +(d\hat{\beta}_r - \hat{\beta}_{rl})'(d\hat{\beta}_r - \hat{\beta}_{rl}) - \operatorname{tr}(V^{-1})\hat{\beta}_{rl}'\Lambda\hat{\beta}_{rl}\right]. \end{aligned}$$

$\hat{\beta}_{rl}$ can be expressed as $\hat{\beta}_{rl} = A_p^{-1}(A_r + dI_p)A_r^{-1}X_r'V_r^{-1}y_r$, where $A_p = A_r + I_p$. Note that $l_b^*(\theta, d; X, y, r) = l_b^*(\theta; X, y, r)$; therefore, the restricted Liu predictor of b is given by

$$\tilde{b}_{rl} = \Sigma U'V^{-1}(y - X\hat{\beta}_{rl}).$$

Similar to Ghapani and Babadi (2022a), it can be shown that $\hat{\beta}_{rl}$ has asymptotic properties as stated in Theorem 2.1.

Theorem 2.1. *$\hat{\beta}_{rl}$ has an asymptotic normal distribution with mean vector $E(\hat{\beta}_{rl}) = M_{rl}M^{-1}\beta$ and covariance matrix $\mathrm{AVar}(\hat{\beta}_{rl}) \cong M_{rl}^{-1}(B + \sigma^2 M)M_{rl}^{-1}$, where $M = Z'V^{-1}Z + R'W^{-1}R$, $B = (\sigma^2 \operatorname{tr}(V^{-1}) + \beta' Z'V^{-2}Z\beta)\Lambda$, and*

$$M_{rl} = (Z'V^{-1}Z+R'W^{-1}R+I_p)^{-1}(Z'V^{-1}Z+R'W^{-1}R+dI_p)(Z'V^{-1}Z+R'W^{-1}R)^{-1}.$$

Corollary 2.2. *$\hat{\beta}_r$ is asymptotically normally distributed with mean $E(\hat{\beta}_r) = \beta$ and asymptotic covariance matrix $\mathrm{AVar}(\hat{\beta}_r) = M^{-1}(B + \sigma^2 M)M^{-1}$.*

3 Diagnostic methods

To determine the observations that exhibit abnormal behavior compared to other observations, various methods have been used by statisticians. In the next subsections, some of these criteria are examined for LMME models with the Liu condition and under simultaneous stochastic linear restrictions.

3.1 Residuals

Residual analysis is one of the important tools in evaluating the appropriateness of any statistical model to the data. In linear models, residuals are used to examine model assumptions, including homogeneity of variance, linearity of effects, the existence of influential observations, normality of observations, and independence of error. Because a residual is considered a deviation between the observed value and the fitted value, it is a criterion or measure that is not explained by the model. Therefore, any deviation from the assumptions about errors can be observed in residuals. Residual analysis is one of the diagnostic methods used to detect various deviations in the model.

The model presented in (1) can be written as follows:

$$y = X\beta + Ub + v,$$

with the assumption that $v \sim N(0, \sigma_v^2)$ as the model errors, where $\sigma_v^2 = \sigma^2 + \beta'\Lambda\beta$. Then, the conditional fitted values based on the RLE are obtained by

$$\hat{y} = X\hat{\beta}_{rl} + U\tilde{b}_{rl} = y - V^{-1}(y - X\hat{\beta}_{rl}).$$

Using the RLE, we can write the (conditional) residuals of the model as $\hat{v}_{rl} = y - \hat{y} = V^{-1}(y - X\hat{\beta}_{rl})$. Therefore, the i-th standardized residual for model 4 is $t_i = \hat{v}_{rli}/\sqrt{\hat{v}_{rl}'\hat{v}_{rl}}$, where $\hat{v}_{rli} = y_i - x_i'\hat{\beta}_{rl} - u_i'\tilde{b}_{rl} = c_i'(y - X\hat{\beta}_{rl})$ is the i-th element of $\hat{v}_{rl}$ and c_i' is the i-th row of the matrix V^{-1}. Based on Zare and Rasekh (2014), it can be concluded that the asymptotic distribution of

$$t_i^{*2} = \frac{(n-1)t_i^2}{1 - t_i^2},$$

is Fisher's distribution with 1 and $n - 1$ degrees of freedom, and the i-th observation may be an outlier if $t_i^{*2} > F(1, n-1, \alpha)$.

3.1.1 Studentized residuals

Since the residuals have different variances, they are not comparable. Accordingly, we use the studentized residuals, which are good criteria for the detection of outliers. The i-th studentized residual of the model is defined as

$$s_i = \frac{\hat{v}_{rli}}{\hat{\sigma}_v \sqrt{h_{ii}}},$$

with $\hat{\sigma}_v^2 = \hat{\sigma}_{rl}^2 + \hat{\beta}_{rl}' \Sigma \hat{\beta}_{rl}$ and $h_{ii} = c_{ii} - c_i' X A_p^{-1} X' c_i$. Studentized residuals are more effective for detecting outlying observations than standardized residuals. If an observation has a large s_i^2, then it is considered an outlier observation.

3.2 High-leverage points

One of the main topics in the sensitivity analysis of regression models is influential points. Using penetration points, the amount of the i-th observation's effect on the predicted value can be measured. According to the generalized definition of penetration stated by Wei et al. (1998), the generalized penetration matrix for the fixed effects is defined as the partial derivative of the marginal fitted values with respect to the response values, that is,

$$GL(\hat{\beta}_{rl}) = \frac{\partial X \hat{\beta}_{rl}}{\partial y'} = X A_p^{-1} (A_r + dI_p) A_r^{-1} X' V^{-1}.$$

According to Fung et al. (2002) and Demidenko and Stukel (2005), the above statement provides information on the effect of observations on fixed effects. Therefore, we can define the influence of the i-th observation. According to Hoaglin and Welsch (1978), the i-th observation is an influential point on the fixed effects.

Given that in linear mixed models, an observation can affect both the estimates of the fixed effects and the predicted values of the random effects, it is convenient to evaluate the joint influence of each observation on the fixed and random effects. The generalized leverage matrix in LMME models with estimation of fixed and random effects is defined as

$$GL(\hat{\beta}_{rl}, \tilde{b}_{rl}) = \frac{\partial \hat{y}_{rl}}{\partial y'} = P,$$

where $P = I_n - V^{-1} + V^{-1} X A_p^{-1} (A_r + dI_p) A_r^{-1} X' V^{-1}$. The i-th diagonal element of P, $p_{ii} = 1 - c_{ii} + c_i' X A_p^{-1} (A_r + dI_p) A_r^{-1} X' c_i$, where c_{ii} denotes the i-th diagonal element of V^{-1}, can be thought of as the amount of leverage of the response value on the corresponding fitted value $\hat{y}_i$. The i-th observation is said to have high leverage on the fixed effects and random effects if $GL(\hat{\beta}_{rl}, \tilde{b}_{rl}) > \frac{3\,\mathrm{tr}(P)}{n}$. Moreover, the generalized leverage matrix for the random effects is $GL(\tilde{b}_{rl}) = I_n - V^{-1}$. Then, the i-th observation is said to have high leverage on the random effects if $GL_{ii}(\tilde{b}_{rl}) > \frac{3\,\mathrm{tr}[GL(\tilde{b}_{rl})]}{n}$.

3.3 Case deletion diagnostic

To quantify the effects of deleting the i-th observation on the estimation of parameters, a fundamental approach is the case deletion model which, for the RLE with the i-th observation deleted in LMME models, is defined as

$$\begin{aligned} y_{(i)} &= Z_{(i)}\beta + U_{(i)}b + \varepsilon_{(i)}, \qquad X_{(i)} = Z_{(i)} + \Delta_{(i)}, \quad i = 1, 2, \ldots, n \quad \text{subject to} \\ r &= R\beta + e \qquad \text{and} \qquad d\hat{\beta}_r = \beta + \varphi. \end{aligned} \tag{5}$$

Theorem 3.1. *For model* (5)*, we have*

$$\begin{aligned} \hat{\beta}_{rl(i)} &\simeq \hat{\beta}_{rl} - A_p^{-1}\frac{X'c_i\hat{v}_{rli}}{h_{ii}}, \\ \hat{\sigma}^2_{rl(i)} &\simeq \frac{\hat{\sigma}^2_{rl}}{N-1}\left(N - \frac{\hat{v}^2_{rli}}{\hat{\sigma}^2_{rl}h_{ii}}\right) + O_p(N^{-1}), \end{aligned}$$

where $\hat{\beta}_{rl(i)}$, $\hat{\sigma}^2_{rl(i)}$, *and* $\tilde{b}_{rl(i)}$ *denote the estimates of* β, σ^2, *and* b *when the* i*-th case is deleted, respectively.*

Proof. When the i-th row is deleted from the data set, we have

$$\hat{\beta}_{rl(i)} = \left(X'_{(i)}V^{-1}_{[i]}X_{(i)} - \mathrm{tr}(V^{-1}_{[i]})\Lambda + I_p + R'W^{-1}R\right)^{-1}\left(X'_{(i)}V^{-1}_{[i]}y_{(i)} + d\hat{\beta}_r + R'W^{-1}r\right).$$

It can also be written as

$$\left(X'_{(i)}V^{-1}_{[i]}X_{(i)} - \mathrm{tr}(V^{-1}_{[i]})\Lambda + I_p + R'W^{-1}R\right)^{-1} = \left(A_p - \frac{X'c_ic_i'X}{c_{ii}} + c_{ii}\Lambda\right)^{-1},$$

therefore,

$$\begin{aligned} \hat{\beta}_{rl(i)} &= \left(A_p - \frac{X'c_ic_i'X}{c_{ii}} + c_{ii}\Lambda\right)^{-1}\left(X'V^{-1}y - \frac{X'c_ic_i'y}{c_{ii}} + d\hat{\beta}_r + R'W^{-1}r\right) \\ &= \left(A_p - \frac{X'c_ic_i'X}{c_{ii}}\right)^{-1}\left(X'V^{-1}y - \frac{X'c_ic_i'y}{c_{ii}} + d\hat{\beta}_r + R'W^{-1}r\right) + O_p(n^{-1}), \end{aligned}$$

Using $(A+BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1}+DA^{-1}B)^{-1}DA^{-1}$ (Rao et al., 2008, Theorem A.18), we have

$$\left(A_p - \frac{X'c_ic_i'X}{c_{ii}}\right)^{-1} = A_p^{-1} + A_p^{-1}\frac{X'c_i}{c_{ii}}\left(I_n - \frac{c_i'XA_p^{-1}X'c_i}{c_{ii}}\right)^{-1}c_i'XA_p^{-1}.$$

Then, $\hat{\beta}_{rl(i)}$ can be written as

$$\hat{\beta}_{rl(i)} \simeq \hat{\beta}_{rl} - A_p^{-1}\frac{X'c_i\hat{v}_{rli}}{h_{ii}} + O_p(n^{-1}).$$

In the same way,

$$\begin{aligned} (N-1)\hat{\sigma}^2_{rl(i)} &= (y_{(i)} - X_{(i)}\hat{\beta}_{rl(i)})'V^{-1}_{[i]}(y_{(i)} - X_{(i)}\hat{\beta}_{rl(i)}) - \mathrm{tr}(V^{-1}_{[i]})\hat{\beta}'_{rl(i)}\Lambda\hat{\beta}_{rl(i)} \\ &\quad -(d\hat{\beta}_r - \hat{\beta}_{rl(i)})'(d\hat{\beta}_r - \hat{\beta}_{rl(i)}), \end{aligned}$$

then,

$$\hat{\sigma}^2_{rl(i)} \simeq \frac{\hat{\sigma}^2_{rl}}{N-1}\left(N - \frac{\hat{v}^2_{rli}}{\hat{\sigma}^2_{rl}h_{ii}}\right) + O_p(N^{-1}).$$

Also, for the predictor of b in model (5), we have

$$\begin{aligned}
\tilde{b}_{rl(i)} &= \Sigma U'_{(i)} V^{-1}_{[i]}(y_{(i)} - X_{(i)}\hat{\beta}_{rl(i)}) \\
&= \Sigma U'V^{-1}y - \Sigma U'\frac{c_i c'_i}{c_{ii}}y - \Sigma\left(U'V^{-1}X - \Sigma U'\frac{c_i c'_i}{c_{ii}}X\right)\hat{\beta}_{rl(i)} \\
&\simeq \tilde{b}_{rl} - \Sigma U'\left(c_i - V^{-1}XA_p^{-1}X'c_i\right)\frac{\hat{v}_{rli}}{h_{ii}} = \tilde{b}_{rl} - \Sigma U' h_i \frac{\hat{v}_{rli}}{h_{ii}}.
\end{aligned}$$

□

3.4 Generalized Cook's distance

Cook's distance is a convenient measure of influence that, based on Cook (1977) and by the deletion of the i-th observation in LMME models under RLE, is defined as

$$CD_i(\beta) = (\hat{\beta}_{rl} - \hat{\beta}_{rl(i)})' M (\hat{\beta}_{rl} - \hat{\beta}_{rl(i)}),$$

where $M = A_p \hat{\sigma}^{-2}_{rl}$. Then, we have

$$CD_i(\beta) = \hat{\sigma}^{-2}_{rl} c'_i X_i A_p^{-1} X' c_i \frac{\hat{v}^2_{rli}}{h^2_{ii}} + O_p(n^{-1}),$$

and approximately we have

$$CD_i(\beta) = \frac{(c_{ii} - h_{ii})\hat{v}^2_{rli}}{\hat{\sigma}^2_{rl} h^2_{ii}} + O_p(n^{-1}).$$

3.5 Cook's distance for random effects

Cook's distance for random effects is defined as

$$CD_i(b) = \frac{(\tilde{b}_{rl} - \tilde{b}_{rl(i)})'(U'U + \Sigma^{-1})(\tilde{b}_{rl} - \tilde{b}_{rl(i)})}{\hat{\sigma}^2_{rl}}.$$

Since $\tilde{b}_{rl(i)} \simeq \tilde{b}_{rl} - \Sigma U' h_i \frac{\hat{v}_{rli}}{h_{ii}}$, we obtain

$$CD_i(b) = h'_i (V - I_n) V h_i \frac{\hat{v}^2_{rli}}{\hat{\sigma}^2_{rl} h^2_{ii}}.$$

3.6 Conditional Cook's distance

To investigate the influence of observations on the predicted values, we define the conditional Cook's distance similarly to Tan et al. (2001) in LMME models with RLE as follows:

$$CD_{cond_i} = \hat{\sigma}^{-2}_{rl}(\hat{y}_{rl} - \hat{y}_{rl(i)})'(\hat{y}_{rl} - \hat{y}_{rl(i)}),$$

which can be decomposed into the three components

$$CD_{cond_i} = CD^1_{cond_i} + CD^2_{cond_i} + CD^3_{cond_i},$$

with

$$\begin{aligned} CD^1_{cond_i} &= \hat{\sigma}_{rl}^{-2}(\hat{\beta}_{rl} - \hat{\beta}_{rl(i)})'X'X(\hat{\beta}_{rl} - \hat{\beta}_{rl(i)}), \\ CD^2_{cond_i} &= \hat{\sigma}_{rl}^{-2}(\tilde{b}_{rl} - \tilde{b}_{rl(i)})'U'U(\tilde{b}_{rl} - \tilde{b}_{rl(i)}), \\ CD^3_{cond_i} &= 2\hat{\sigma}_{rl}^{-2}(\hat{\beta}_{rl} - \hat{\beta}_{rl(i)})'X'U(\tilde{b}_{rl} - \tilde{b}_{rl(i)}). \end{aligned}$$

Here, $CD^1_{cond_i}$ is related to the fixed effects; the observations that may influence the predictors of the random effects are determined by $CD^2_{cond_i}$; and finally, $CD^3_{cond_i}$ is related to the covariance of $\hat{\beta}$ and $\tilde{b}$, which is expected to be close to zero.

4 Simulation study

We use two simulation studies to evaluate the credibility of the proposed diagnostic measures and to provide more evidence of the good performance of the proposed methods using RLE in LMME models.

4.1 Simulation study one

To investigate the behavior of $GL(\hat{\beta}_{rl})$, $GL(\tilde{b}_{rl})$, and $GL(\hat{\beta}_{rl}, \tilde{b}_{rl})$ in LMME models, we generated the i-th set of simulated data as

$$\begin{aligned} y_i &= \beta_0 + Z\beta + Ub_i + \varepsilon_i, \quad i = 1, \dots, 1000, \\ X &= Z + \Delta \quad \text{and} \quad r_i = R\beta + e_i, \end{aligned} \tag{6}$$

where $y_i = (y_{11i}, \dots, y_{1n_1i}, y_{21i}, \dots, y_{2n_2i}, \dots, y_{k1i}, \dots, y_{kn_ki})$, $Z = (z^{(1)}, \dots, z^{(p)})$, in which $z^{(t)} = (z^{(t)}_{11}, \dots, z^{(t)}_{1n_1}, z^{(t)}_{21}, \dots, z^{(t)}_{2n_2}, \dots, z^{(t)}_{k1}, \dots, z^{(t)}_{kn_k})'$, $t = 1, \dots, p$, $b_i = (b_{1i}, b_{2i}, \dots, b_{ki})'$, and ε_i is rewritten in accordance with y_i. In addition, k can be interpreted as the number of independent groups; n_i is the group size, so $n = \sum_{i=1}^k n_i$ is the total sample size of the data set. $U = 1_{n_i} \oplus 1_{n_i} \oplus \cdots \oplus 1_{n_i}$ is an $n \times q$ matrix, where 1_{n_i} is an $n_i \times 1$ vector with all elements equal to 1. Furthermore, $r_i = (r_{1i}, \dots, r_{mi})'$, $R = (R^{(1)}, \dots, R^{(p)})$, $R^{(t)} = (R_{1j}, \dots, R_{mj})'$, and $e_i \sim N(0, \sigma^2 I_m)$. To achieve different degrees of collinearity, following McDonald and Galarneau (1975), the fixed effects variables are computed as $z_{it} = (1 - \rho^2)^{1/2} w_{it} + \rho w_{i,p+1}$, where w_{it} are independent standard normal pseudo-random numbers and ρ represents the correlation between any two fixed effects. In this simulation study, the value of ρ was chosen as 0.90. For each set of explanatory variables, the parameter vector is chosen as the eigenvector corresponding to the largest eigenvalues of $Z'V^{-1}Z$. To create observations with high leverage, we considered $w_{it} \sim N(6, 2)$ for $i = 5$ and $i = 9$. The random errors are $\varepsilon_i \sim N(0, \sigma^2)$ for $\sigma^2 = 0.5$, except for the high leverage observations with $\varepsilon_i = 5$ for $i = 5$. This structure was intended to generate two high-influence observations (observations 5 and 9). The simulation study was carried out using the R software (The R codes are available from the author upon request). The following combinations were used for the simulation: $q = 10$, $n = 50$, $p = 3$, $\Lambda = \text{diag}(0, 0.05, 0.05, 0.05)$, $m = 2$, $R_{ij}^{(t)} \sim N(0,1)$, and $\sigma_1^2 = 0.1$.

To visualize the performance of $GL(\hat{\beta}_{rl})$, $GL(\tilde{b}_{rl})$, and $GL(\hat{\beta}_{rl}, \tilde{b}_{rl})$ in the simulation study, we calculate their values for the 1000 simulated data sets by generating new error terms and then take the average value over the simulation runs. Figures 1–2 show plots of the generalized leverage measures from the simulated data. All the dotted lines correspond to three times the mean leverages. Figures 1–2 show that observations 5 and 9 have high generalized leverage on fixed effects and random effects as well as high joint generalized leverage on both fixed and random effects.

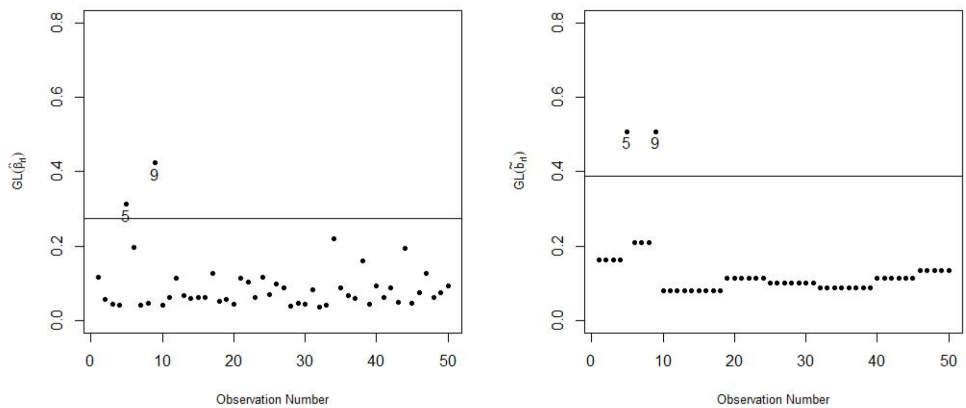

Figure 1: Generalized leverage plot on fixed effects and random effects.

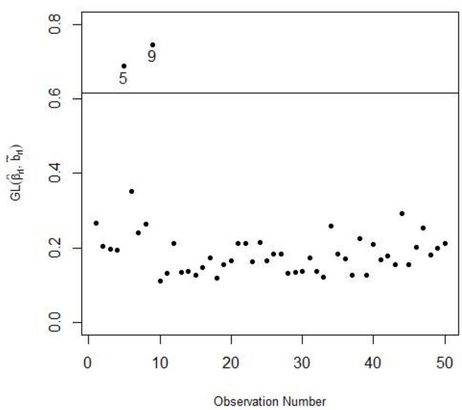

Figure 2: Generalized leverage plot on fixed and random effects.

4.2 Simulation study two

Under Model (6), we change the fixed effects β to 3β when generating the response of the first observation. The Cook's distances for fixed and random effects based on CDM are computed for each simulated dataset; then the experiment was repeated 1000 times by generating new error terms. Figures 3–4 show plots of the average values of $CD_i(\beta)$, $CD_i(b)$, and CD_{cond_i}, respectively.

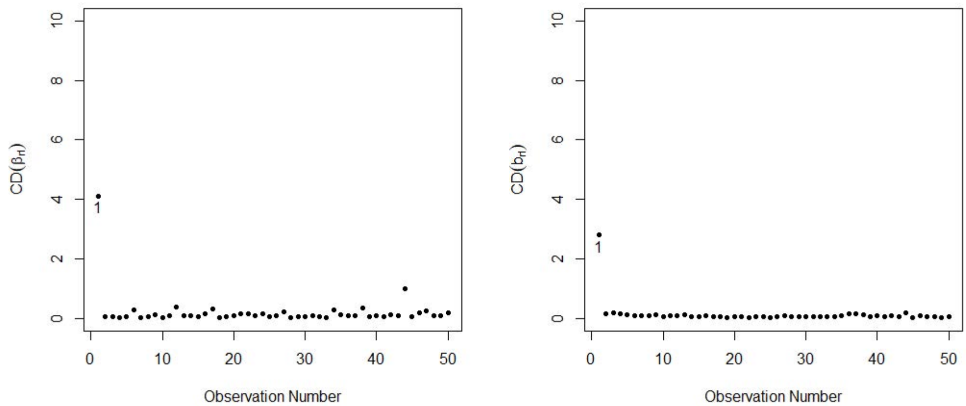

Figure 3: Plot of Cook's distance for fixed effects and random effects.

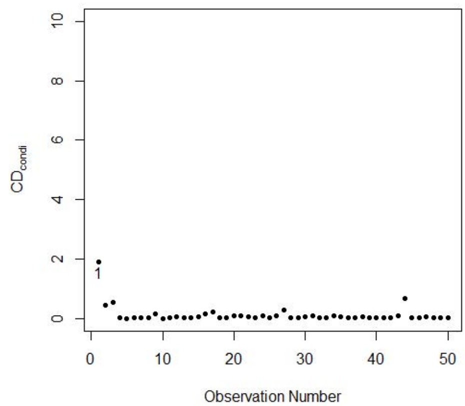

Figure 4: Plot of conditional Cook's distance.

A glance at Figures 3–4 shows that observation 1 has greater influence on fixed effects, random effects, and predicted values. Also, from Figures 3–4, we conclude that observation 1 is an influential observation and affects the model more than usual. Also, for each replicate, the RMSE (Root Mean Squared Error) values were computed for the full data and with deletion of observation 1. In this study, we consider $n = 50$ or $n = 100$ and three values of $\rho = 0.85, 0.90$, and 0.95. The mean RMSE values are presented in Table 1.

Table 1: RMSE values for full data and removing of observations 1.

ρ	Full data		Observation 1 deleted	
	$n = 50$	$n = 100$	$n = 50$	$n = 100$
0.85	0.829	0.605	0.592	0.558
0.90	0.897	0.639	0.645	0.590
0.95	1.385	0.685	1.144	0.634

Based on Table 1, the following conclusions can be drawn:
- The RMSE values increase as the collinearity level increases.
- The RMSE values decrease as n increases.
- After deleting observation 1, the RMSE values decrease.

5 Real data analysis

In this section, the market historical data set of real estate valuation from Sindian Dist., New Taipei City, Taiwan is used to evaluate the performance of the proposed diagnostic criteria. The dataset is publicly available online at the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets.html) and is labeled as the "Real estate valuation" data set. The inputs are as follows:

Dependent Variable: House price of unit area

Independent (Explanatory) Variables:
- House age,
- Distance to the nearest MRT station,
- Number of convenient stores in the living circle on foot,
- Latitude(geographic coordinate, latitude),
- Longitude(geographic coordinate, longitude).

Among these variables, X_2, X_4, and X_5 are subject to measurement error. The data set consists of 414 observations, two of which were used as random restrictions. Therefore, the data set was fitted with a LMME model of the form $y = X\beta + Ub + \varepsilon$, where y is a 412×1 vector of response variables, and X and U are regression matrices with dimensions 412×5 and 412×2, respectively. The transaction date (2012 and 2013) is considered as a random effect. The initial values for estimating the variance components are $\sigma^2 = 0.1$ and $\sigma^2 = 0.5$. The condition number of $\hat{Z}'V^{-1}\hat{Z}$ is equal to 126456.9, which indicates that there is severe multicollinearity among the fixed effects variables. The 114-th and 220-th elements of the dependent variable and the 114-th and 220-th rows of the independent variables were considered as r and R, respectively.

Table 2 gives the parameter estimates, percentage change, and RMSE values for the LMME model with the full data and with deletion of 269 observations. As can be seen, after deleting observation 269, the RMSE values and estimates of some coefficients decrease. Also, the various Cook's distances introduced in the previous sections are calculated for each observation. Figures 5–6 show the plots of Cook's distance for fixed effects, random effects, and the conditional Cook's distance, respectively.

A glance at Figures 5–6 shows that observation 269 has greater influence on fixed effects, random effects, and predicted values. According to Figures 5–6, it is obvious that observation 269 stands out as an influential observation with a relatively large value of Cook's distance.

Table 2: Parameter estimates for the real estate valuation data set.

Parameter	Full data	Observation 269 deleted	% change
β_1	-0.2561	0.2483	3.05
β_2	-0.0054	-0.0051	5.56
β_3	1.3088	1.4511	10.87
β_4	-15.9094	-16.5989	-4.33
β_5	3.6187	3.7505	4.64
b_1	-1.1976	-1.078	-9.99
b_2	1.2550	1.1434	-8.89
σ^2	83.7473	69.3794	-17.16
σ_1^2	1.8523	1.5199	-17.95
RMSE	8.8233	7.9744	-9.62

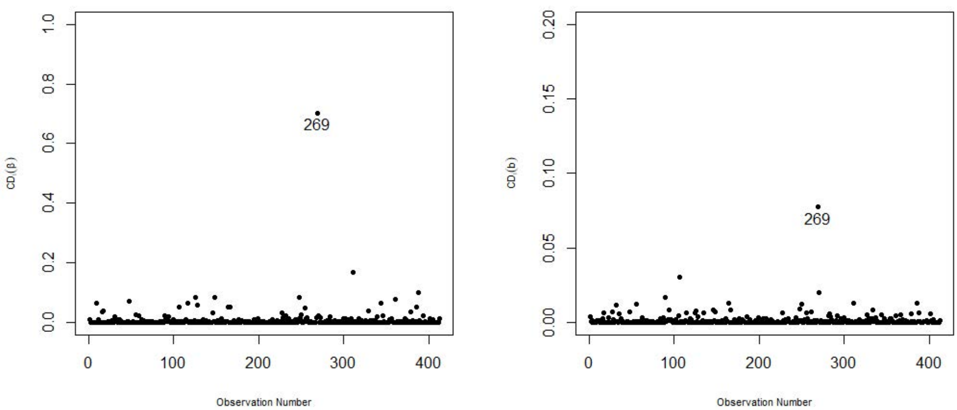

Figure 5: Plot of Cook's distance for fixed effects and random effects.

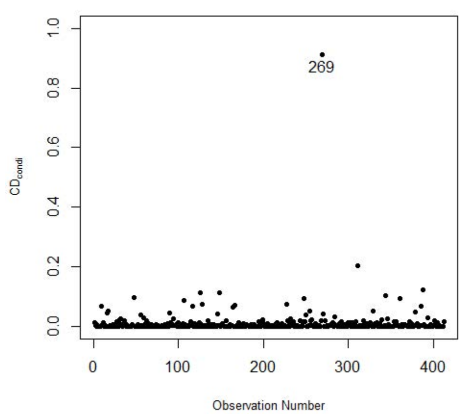

Figure 6: Plot of Conditional Cook's distance.

6 Summary and concluding remarks

This paper presents diagnostic methods based on the RLE for identifying high-leverage points and influential observations in LMME models subject to stochastic linear restrictions. To detect influential observations, a case-deletion approach is applied to the proposed model. Various forms of Cook's distance are employed to identify observations that exert significant influence on the estimation of fixed effects, the prediction of random effects, or both. Simulation studies and real-data analysis demonstrate that the proposed diagnostic measures, utilizing the RLE within LMME models, perform exceptionally well in detecting influential observations.

Acknowledgement

The author would like to sincerely thank the Editor, Associate Editor, and anonymous reviewers for their constructive and valuable comments, which greatly improved the manuscript.

References

Borhani, Sh., Ghapani, F. and Jafaraghaie, R. (2023). Detection of outliers and influential observations in linear mixed measurement error models with Liu estimation. *Journal of Statistical Modelling: Theory and Applications*, **4**(2):1–16.

Borhani, Sh., Ghapani, F. and Jafaraghaie, R. (2024). Detecting Influential Observations Using Liu Estimator in Linear Mixed Measurement Error Models. *Journal of Statistical Theory and Practice*, **18**(2):18.

Cook, R.D. (1977). Detection of influential observations in linear regression. *Technometrics*, **19**(1):15–18.

Demidenko, E. and Stukel, T.A. (2005). Influence analysis for linear mixed-effects models. *Statistics in medicine,* **24**(6):893–909.

Fuller, W.A. (1987). *Measurement Error Models.* New York: Wiley.

Fung, W.K., Zhong, X.P. and Wei, B.C. (2003). On estimation and influence diagnostics in linear mixed measurement error models. *American Journal of Mathematical and Management Sciences*, **23**(1-2):37–59.

Fung, W.K., Zhu, Z.Y., Wei, B.C. and He, X. (2002). Influence diagnostics and outliers tests for semiparametric mixed models. *Journal of the Royal Statistical Society. Series B*, **64**(3):565–579.

Ganjealivand, N. and Ghapani, F. (2024). The weighted two-parameter estimation for linear mixed models with measurement error under stochastic linear mixed restrictions. *Journal of Statistical Modelling: Theory and Applications*, **5**(2):137–152.

Ghapani, F. (2024). Diagnostic measures based on restricted ridge estimator in linear mixed measurement error models. *Journal of Data Science and Modeling*, **3**(1):95–122.

Ghapani, F. and Babadi, B. (2022a). Diagnostic measures for the restricted Liu estimator in linear measurement error models. *Journal of Statistical Computation and Simulation,* **92**(11):2257–2272.

Ghapani, F. and Babadi, B. (2022b). Two parameter weighted mixed estimator in linear easurement error models. *Communications in Statistics-Simulation and Computation,* **51**(12):6936–6946.

Hoaglin, D.C. and Welsch, R.E. (1978). The hat matrix in regression and ANOVA. *The American Statistician*, **32**(1):17–22.

Liu, K. (1993). A new class of biased estimate in linear regression. *Communications in Statistics-Theory and Methods*, **22**(2):393–402.

McDonald, G.C. and Galarneau, D.I. (1975). A Monte Carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association*, **70**(350):407–416.

Nakamura, T. (1990). Corrected score function for errors-in-variables models: Methodology and application to generalized linear models. *Biometrika,* **77**(1):127–137.

Rao, C.R., Shalabh, Toutenburg, H. and Heumann, C. (2008). *Linear Models and Generslizations.* Berlin, Heidelberg: Springer Berlin Heidelberg.

Tan, F.E.S., Ouwens, M.J.N. and Berger, M.P.F. (2001). Detection of influential observations in longitudinal mixed effects regression models. *Journal of the Royal Statistical Society Series D*, **50**(3):271–284.

Wei, B.C., Hu, Y.Q. and Fung, W.K. (1998). Generalized leverage and its applications. *Scandinavian Journal of Statistics*, **25**(1):25–37.

Welsch, R.E. (1982). Influence functions and regression diagnostics. In: Launer, R.L. and Siegel, A.F. eds., *Modern Data Analysis.* pp. 149-169, New York: Academic Press.

Zare, K. and Rasekh, A. (2011). Diagnostic measures for linear mixed measurement error models. *Sort*, **35**(2):125–144.

Zare, K. and Rasekh, A. (2014). Residuals and leverages in the linear mixed measurement error models. *Journal of Statistical Computation and Simulation*, **84**(7):1427–1443.

Zare, K., Rasekh, A. and Rasekhi, A.A. (2012). Estimation of variance components in linear mixed measurement error models. *Statistical Papers,* **53**(4):849–863.

Zhong, X.P., Fung, W.K. and Wei, B.C. (2002). Estimation in linear models with random effects and errors-invariables. *Annals of the Institute of Statistical Mathematics*, **54**(3):595–606.