Shahid Bahonar University of Kerman
Journal of Mahani Mathematical Research
Print ISSN: 2251-7952 Online ISSN: 2645-4505

VARIABLE SELECTION BASED ON THE BOOSTING ALGORITHM IN MULTILEVEL MODELS

N. Ghadiri Ghazvini, R. Kazemi ✉, M.H. Behzadi, and S. Zamani Mehreyan

Article type: Research Article
(Received: 10 June 2025, Received in revised form 09 December 2025)
(Accepted: 10 February 2026, Published Online: 10 February 2026)

Abstract. The existence of a natural hierarchy in all the multilevel problems is required for the proper exploration as specialized analytical tools. Multilevel analytical tools provide the proper estimation of such type of problems. The variance components models, the hierarchy linear models and random coefficient models are known as multilevel regression. In this paper, we consider random intercept and slope model and estimate parameters of model using the EM algorithm. Also, we propose a boosting algorithm based on the EM algorithm. The performance of the proposed algorithm is studied and compared to usual variable selection methods using simulation studies. A real data analysis is carried out to illustrate the procedures is obtained theoretically.

Keywords: Boosting algorithm, Machine learning, Random Intercept, Random Slope, Variable selection.
2020 MSC: 62J05, 62H12, 68T05.

1. Introduction

Multilevel regression models are also known as variance components models (Aitkin and Longford [1]), hierarchical linear models (Raudenbaush and Bryk [19]) and random coefficient models (Longford [15]). Random slope and intercept regression models combine fixed and random effects and analyze data at two or more levels. These models are essentially a combination of various regression models used to model complex, multivariate relationships and have applications across numerous fields. Early developments in random effects modeling trace back to Fisher [11] models while the modern formulation of random intercept and random slope regression was established in later methodological work. Robinson [20] presented the best linear estimators for fixed and random effects, while Henderson et al. [14] discussed the estimation of random effects in biological and genetic fields. McLean et al. [16] studied integrated approaches to mixed linear regression models. Bouwmeester et al. [3] stated

✉ r.kazemi@sci.ikiu.ac.ir, ORCID: 0000-0001-5496-7565
https://doi.org/10.22103/jmmr.2026.25414.1816
Publisher: Shahid Bahonar University of Kerman
How to cite: N. Ghadiri Ghazvini, R. Kazemi, M.H. Behzadi, S. Zamani Mehreyan, *Variable selection based on the boosting algorithm in multilevel models*, J. Mahani Math. Res. 2027; 16(1): 47-74.

that random intercept regression models generally offer greater accuracy in predicting clustered data. This model can better account for differences between clusters (such as differences between groups or units) and provide more precise predictions. Choi [9] estimated the non-negative variance component for random effects in mixed models, and Oberauer [18] examined the impact of including random slopes in mixed models on the accuracy of Bayesian hypothesis test. The presence of random slopes in mixed models increases the accuracy of Bayesian hypothesis tests and improves the precision of the results, especially where the data have a complex structure and there are between-group differences.

Recent literature highlights the growing importance of EM-based estimation methods in mixed effects and nonlinear mixed effects models, particularly for handling incomplete data, complex random effects structures and high dimensional settings. Zakkour et al. [23] developed a stochastic EM algorithm for linear mixed effects models with interaction terms under missing data, offering improved stability and convergence. Heiling et al. [13] proposed an efficient computational framework for high dimensional penalized generalized linear mixed models by modeling random effects through latent factors, significantly reducing the computational burden of EM-type updates. In addition, Mei et al. [17] introduced a flexible EM procedure capable of accommodating mixture structures and nonlinear mixed effects models with intricate random effect specifications. Collectively, these recent advances demonstrate that the EM algorithm remains a central and powerful estimation tool within modern mixed effects modeling, motivating further methodological refinement and application in complex hierarchical data settings.

The boosting method is introduced as a general method to combine multiple classifiers to improve the overall classification accuracy for almost any type of learning algorithm, see Schapire [21, 22]. The first widely used boosting algorithm is adaptive boosting which solves binary classification problems with great success. Bühlmann and Yu [5] proposed L_2 boosting that builds a linear model by minimizing the L_2 loss. Also, Bühlmann [4] proved the consistency of L_2 boosting in terms of predictions. If the learner sufficiently weak, the L_2 boosting is computationally simple and successful. L_2 boosting is the simplest and very useful for regression, in particular in the presence of high dimensional explanatory variables. Friedman [12], introduced the gradient boosting machine which is a generalization of AdaBoost and L_2 boosting. AdaBoost is a version of gradient boosting that uses the exponential loss and L_2 boosting is a version of gradient boosting that uses the L_2 loss. Chen and Qiu [6] considered the length-biased and partly interval-censored data, whose challenges primarily come from biased sampling and interfere induced by interval censoring. They primarily employed the SIMEX method to correct for measurement error effects and proposed the boosting procedure to do variable selection and estimation. The proposed method is able to handle the case that the dimension of covariates is larger than the sample size and enjoys appealing features that

the distributions of the covariates are left unspecified. Chen [7] proposed a valid inferential method to deal with measurement error and handle variable selection simultaneously. They proposed estimating functions by incorporating corrected responses and predictors for logistic regression or probit models and developed the boosting procedure with error-eliminated estimating functions accommodated to do variable selection and estimation. Also, Chen [8] employed the boosting algorithm with the cubic spline estimation method to iteratively recover nonlinear relationship between covariates and survival time. This paper introduces a new methodological contribution by integrating EM-based parameter estimation directly into a boosting framework for variable selection in random intercept and random slope regression models. Specifically, we derive the EM estimators for all model parameters and embed these updates within each boosting iteration, yielding an EM-driven boosting algorithm tailored to mixed effects structures. This extension provides a principled and coherent approach to model selection under random effects. Simulation studies further confirm that the proposed method outperforms conventional variable selection techniques as well as the standard L_2 boosting algorithm.
The rest of the paper is organized as follows: In Section 2, the random intercept and slope regression model is introduced and the unknown parameters of the model are estimated using the EM algorithm. In Section 3, a boosting algorithm for random intercept and slope model based on the EM algorithm is proposed. In Section 4, the performance of the proposed boosting algorithm is studied using simulation and is compared with L_2 boosting algorithm as well as the classic method of variable selection such as the step-by-step method. Also, our theoretical results with the analysis of a real dataset is illustrated in Section 5.

2. Random slope and intercept regression model

In this section, the random intercept and slope regression model, which has widespread applications in many fields [2], is reviewed. The level 1 model is defined as

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + R_{ij}, \tag{1}$$

where $i = 1, \ldots, n$ indexes the level 1 units within group j, and $j = 1, \ldots, m$ indexes level 2 groups. The observation y_{ij} is the response at level 1 corresponding to the covariate x_{ij}. The level 2 models for the random intercept and slope are

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + u_{0j}, \\ \beta_{1j} &= \gamma_{10} + u_{1j}. \end{aligned} \tag{2}$$

Substituting Equation (2) into Equation (1), one can write

$$\begin{aligned} y_{ij} &= \gamma_{00} + u_{0j} + (\gamma_{10} + u_{1j}) x_{ij} + R_{ij} \\ &= \gamma_{00} + \gamma_{10} x_{ij} + u_{0j} + u_{1j} x_{ij} + R_{ij}. \end{aligned} \tag{3}$$

where $\gamma_{00}+\gamma_{10}x_{ij}$ represents the fixed effects term and $u_{0j}+u_{1j}x_{ij}$ represents the random effects term. It is assumed that the error terms R_{ij} are uncorrelated and normally distributed with zero mean and unknown variance σ^2. Also, the vector

$$u_j=\begin{pmatrix}u_{0j}\\u_{1j}\end{pmatrix}\sim N_2\left(\begin{pmatrix}0\\0\end{pmatrix},\Sigma_u\right),\qquad \Sigma_u=\begin{pmatrix}\tau_0^2 & \tau_{01}\\ \tau_{01} & \tau_1^2\end{pmatrix},$$

where $\tau_0^2=Var(u_{0j}),\tau_1^2=Var(u_{1j})$, $\tau_{01}=\tau_{10}=Cov(u_{0j},u_{1j})$ and u_j is independent of R_{ij}.

Model (3) can be written in the matrix form as

$$\mathbf{Y}=\mathbf{X\Gamma}+\mathbf{ZU}+\mathbf{R}, \tag{4}$$

where

$$\mathbf{Y}=\begin{bmatrix}\mathbf{Y}_1\\ \mathbf{Y}_2\\ \vdots\\ \mathbf{Y}_n\end{bmatrix},\qquad \mathbf{R}=\begin{bmatrix}\mathbf{R}_1\\ \mathbf{R}_2\\ \vdots\\ \mathbf{R}_n\end{bmatrix},$$

and for each $i=1,\dots,n$,

$$\mathbf{Y}_i=\begin{bmatrix}Y_{i1}\\ Y_{i2}\\ \vdots\\ Y_{im}\end{bmatrix},\qquad \mathbf{R}_i=\begin{bmatrix}R_{i1}\\ R_{i2}\\ \vdots\\ R_{im}\end{bmatrix}.$$

The fixed effects vector is

$$\mathbf{\Gamma}=\begin{bmatrix}\mathbf{\Gamma}_1\\ \mathbf{\Gamma}_2\\ \vdots\\ \mathbf{\Gamma}_m\end{bmatrix},\qquad \mathbf{\Gamma}_j=\begin{bmatrix}\gamma_{00}\\ \gamma_{10}\end{bmatrix}.$$

The random effects vector is

$$\mathbf{U}=\begin{bmatrix}\mathbf{U}_1\\ \mathbf{U}_2\\ \vdots\\ \mathbf{U}_m\end{bmatrix},\qquad \mathbf{U}_j=\begin{bmatrix}u_{0j}\\ u_{1j}\end{bmatrix}.$$

The design matrices are defined by

$$\mathbf{X}=\text{Block-diag}\{\mathbf{X}_1,\mathbf{X}_2,\dots,\mathbf{X}_n\},\qquad \mathbf{X}_i=\begin{bmatrix}1 & x_{i1}\\ 1 & x_{i2}\\ \vdots & \\ 1 & x_{im}\end{bmatrix},$$

and

$$\mathbf{Z}=\text{Blockdiag}\{\mathbf{Z}_1,\mathbf{Z}_2,\dots,\mathbf{Z}_n\},\ \ Z_i=X_i$$

where $\mathbf{Z}_i$ is a submatrix of $\mathbf{X}_i$.

Note that Model (3) follows a normal distribution with

$$E(Y_{ij}) = \gamma_{00} + \gamma_{10} x_{ij}$$

and

$$Var(Y_{ij}) = \tau_0^2 + \tau_1^2 x_{ij}^2 + 2\tau_{01} x_{ij} + \sigma^2.$$

In terms of matrix notation

$$\mathbf{Y} \sim N_m(\boldsymbol{\mu_Y}, \boldsymbol{\Sigma_Y}),$$

where

$$\boldsymbol{\mu_Y} = E(\mathbf{Y}) = \mathbf{X\Gamma}$$

and

$$\boldsymbol{\Sigma_Y} = Var(\mathbf{Y}) = \mathbf{Z\Sigma_U Z}^T + \boldsymbol{\Sigma_R},$$

$$\boldsymbol{\Sigma_R} = \sigma^2 \boldsymbol{I_m},$$

and $\boldsymbol{I_m}$ is the identity matrix of order m. The parameters of the model are estimated using the EM algorithm. Let $\mathbf{Y}_c = \binom{\mathbf{Y}}{\mathbf{U}}$ be a complete data vector, where $\mathbf{Y}$ and $\mathbf{U}$ are the observed and missing data, respectively. So, we have

$$\mathbf{Y}_c = \begin{pmatrix} \mathbf{Y} \\ \mathbf{U} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_Y \\ 0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_Y & \boldsymbol{\Sigma}_{YU} \\ \boldsymbol{\Sigma}_{UY} & \boldsymbol{\Sigma}_U \end{pmatrix} \right),$$

where

$$\boldsymbol{\Sigma}_{UY} = \mathrm{Cov}(\mathbf{U}, \mathbf{Y}) = \boldsymbol{\Sigma}_U \mathbf{Z}^T.$$

The distribution of $\mathbf{U}$ conditional on $\mathbf{Y} = \mathbf{y}$ is

$$\mathbf{U}|\mathbf{Y} = \mathbf{y} \sim N_{2m}\left(\boldsymbol{\Sigma}_{\boldsymbol{YU}}^{\boldsymbol{T}} \boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-\mathbf{1}} (\mathbf{Y} - \boldsymbol{\mu_Y}), \boldsymbol{\Sigma_U} - \boldsymbol{\Sigma}_{\boldsymbol{YU}}^{\boldsymbol{T}} \boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-\mathbf{1}} \boldsymbol{\Sigma}_{\boldsymbol{YU}} \right).$$

The complete log-likelihood function can be expressed as

$$\begin{aligned} l(\boldsymbol{\theta}; \mathbf{y}_c) &= l(\boldsymbol{\theta}; \mathbf{y}|\mathbf{u}) + l(\boldsymbol{\theta}; \mathbf{u}) \\ &= C - \frac{n}{2} \log |\boldsymbol{\Sigma_R}| - \frac{1}{2} \sum_{i=1}^{n} (\mathbf{y}_i - \mathbf{x}_i \boldsymbol{\Gamma} - \mathbf{z}_i \mathbf{U})^T \boldsymbol{\Sigma}_{\boldsymbol{R}}^{-\mathbf{1}} (\mathbf{y}_i - \mathbf{x}_i \boldsymbol{\Gamma} - \mathbf{z}_i \mathbf{U}) \\ &- \frac{n}{2} \log |\boldsymbol{\Sigma_U}| - \frac{n}{2} \left(\mathbf{U}^T \boldsymbol{\Sigma}_{\boldsymbol{U}}^{-\mathbf{1}} \mathbf{U} \right), \end{aligned}$$

where C is a constant that is independent of $\boldsymbol{\Sigma_R}$ and $\boldsymbol{\Sigma_U}$.
Suppose that $\boldsymbol{\theta} = (\boldsymbol{\Gamma}, \boldsymbol{\Sigma_R}, \boldsymbol{\Sigma_U})$ and $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\Gamma}^{(\mathbf{0})}, \boldsymbol{\Sigma}_{\boldsymbol{R}}^{(\mathbf{0})}, \boldsymbol{\Sigma}_{\boldsymbol{U}}^{(\mathbf{0})})$ is the starting

values of $\boldsymbol{\theta}$. Then

$$\begin{aligned}
Q &= E\left\{l\left(\boldsymbol{\theta};\mathbf{y}_c|\mathbf{y},\boldsymbol{\theta}^{(0)}\right)\right\}\\
&= C-\frac{n}{2}\log|\boldsymbol{\Sigma_R}|-\frac{1}{2}\sum_{i=1}^{n}E\left\{(\mathbf{y}_i-\mathbf{x}_i\boldsymbol{\Gamma}-\mathbf{z}_i\mathbf{U})^T\boldsymbol{\Sigma_R^{-1}}(\mathbf{y}_i-\mathbf{x}_i\boldsymbol{\Gamma}-\mathbf{z}_i\mathbf{U})|\mathbf{y},\boldsymbol{\theta}^{(0)}\right\}\\
&- \frac{n}{2}\log|\boldsymbol{\Sigma}_U|-\frac{n}{2}\left(E\left\{\mathbf{U}^T\boldsymbol{\Sigma}_U^{-1}\mathbf{U}|\mathbf{y},\boldsymbol{\theta}^{(0)}\right\}\right)\\
&= C-\frac{n}{2}\log|\boldsymbol{\Sigma_R}|-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i-\mathbf{x}_i\boldsymbol{\Gamma}-\mathbf{z}_i\tilde{\boldsymbol{u}}_{\mathbf{1}})^T\boldsymbol{\Sigma_R^{-1}}(\mathbf{y}_i-\mathbf{x}_i\boldsymbol{\Gamma}-\mathbf{z}_i\tilde{\boldsymbol{u}}_{\mathbf{1}})\\
&+ tr\left\{\boldsymbol{\Sigma_R^{-1}}\mathbf{z}_i\tilde{\mathbf{b}}_{\mathbf{1}}\mathbf{z}_i^T\right\}-\frac{n}{2}\left(\tilde{\boldsymbol{u}}_{\mathbf{1}}{}^T\boldsymbol{\Sigma}_U^{-1}\tilde{\boldsymbol{u}}_{\mathbf{1}}+tr\left\{\boldsymbol{\Sigma}_U^{-1}(\boldsymbol{\Sigma}_U-\boldsymbol{\Sigma}_{UY}\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\Sigma}_{YU})\right\}\right)
\end{aligned}\tag{5}$$

$$-\frac{n}{2}\log|\boldsymbol{\Sigma}_U|,$$

where

$$\tilde{\mathbf{u}}_{\mathbf{1}}=E\left(\mathbf{U}|\mathbf{y},\boldsymbol{\theta}^{(0)}\right)=\boldsymbol{\Sigma}_{\boldsymbol{UY}}^{(\mathbf{0})}\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-\mathbf{1}(\mathbf{0})}\left(\mathbf{y}-\boldsymbol{\mu}_{\boldsymbol{Y}}^{(\mathbf{0})}\right)$$

and

$$\tilde{\mathbf{b}}_{\mathbf{1}}=Var\left(\mathbf{U}|\mathbf{y},\boldsymbol{\theta}^{(0)}\right)=\boldsymbol{\Sigma}_{\boldsymbol{U}}^{(\mathbf{0})}\boldsymbol{\Sigma}_{\boldsymbol{UY}}^{(\mathbf{0})}\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-\mathbf{1}(\mathbf{0})}\boldsymbol{\Sigma}_{\boldsymbol{YU}}^{(\mathbf{0})}.$$

Now, we obtain estimators by solving the estimating equations as:

$$\begin{aligned}
0 &= \frac{\partial Q}{\partial\boldsymbol{\Gamma}}\\
&= -\frac{1}{2}\frac{\partial}{\partial\boldsymbol{\Gamma}}\left(\sum_{i=1}^{n}(\mathbf{y}_i-\mathbf{x}_i\boldsymbol{\Gamma}-\mathbf{z}_i\tilde{\mathbf{u}}_{\mathbf{1}})^T\boldsymbol{\Sigma_R^{-1}}(\mathbf{y}_i-\mathbf{x}_i\boldsymbol{\Gamma}-\mathbf{z}_i\tilde{\mathbf{u}}_{\mathbf{1}})+tr\left\{\sum_{i=1}^{n}(\boldsymbol{\Sigma_R^{-1}}\mathbf{z}_i\tilde{\mathbf{b}}_{\mathbf{1}}\mathbf{z}_i^T)\right\}\right)\\
&= \sum_{i=1}^{n}\mathbf{x}_i^T\boldsymbol{\Sigma_R^{-1}}(\mathbf{y}_i-\mathbf{x}_i\boldsymbol{\Gamma}-\mathbf{z}_i\tilde{\mathbf{u}}_{\mathbf{1}}).
\end{aligned}\tag{6}$$

Hence

$$\hat{\boldsymbol{\Gamma}}=\left(\sum_{i=1}^{n}\mathbf{x}_i^T\boldsymbol{\Sigma_R^{-1}}\mathbf{x}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{x}_i^T\boldsymbol{\Sigma_R^{-1}}(\mathbf{y}_i-\mathbf{z}_i\tilde{\mathbf{u}}_{\mathbf{1}})\right),\tag{7}$$

and also

$$\begin{aligned}
\frac{\partial\boldsymbol{Q}}{\partial\boldsymbol{\Sigma_R}} &= \frac{\partial\boldsymbol{Q}}{\partial\boldsymbol{\Sigma_R^{-1}}}\left(\frac{\partial\boldsymbol{\Sigma_R^{-1}}}{\partial\boldsymbol{\Sigma_R}}\right)\\
&= \frac{1}{2}\sum_{i=1}^{n}Vec^T(\boldsymbol{\Sigma_R^T})(\boldsymbol{\Sigma_R^{-T}}\otimes\boldsymbol{\Sigma_R^{-1}})-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i-\mathbf{x}_i\boldsymbol{\Gamma}-\mathbf{z}_i\tilde{\mathbf{u}}_{\mathbf{1}})^T\otimes(\mathbf{y}_i-\mathbf{x}_i\boldsymbol{\Gamma}-\mathbf{z}_i\tilde{\mathbf{u}}_{\mathbf{1}})\\
&+ tr\left\{\mathbf{z}_i\tilde{\mathbf{b}}_{\mathbf{1}}\mathbf{z}_i^T\right\}(\boldsymbol{\Sigma_R^{-T}}\otimes\boldsymbol{\Sigma_R^{-1}})\\
&= 0,
\end{aligned}\tag{8}$$

where $\otimes$ denotes the Kronecker product and $Vec(\mathbf{A})$ is the vectorized structure of the matrix $\mathbf{A}$. Therefore,

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{R}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{x}_i\hat{\boldsymbol{\Gamma}} - \mathbf{z}_i\tilde{\mathbf{u}}_{\mathbf{1}})^T(\mathbf{y}_i - \mathbf{x}_i\hat{\boldsymbol{\Gamma}} - \mathbf{z}_i\tilde{\mathbf{u}}_{\mathbf{1}}) + tr\left\{\mathbf{z}_i\tilde{\mathbf{b}}_{\mathbf{1}}\mathbf{z}_i^T\right\} \tag{9}$$

and similarly,

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{U}} = E\left(\boldsymbol{U}\boldsymbol{U}^T|\mathbf{y}, \boldsymbol{\theta}^{(0)}\right). \tag{10}$$

The complete calculations are presented in Appendix A. The EM algorithm is described as follows:
Step 0: Set $r = 0$ and choose starting value $\boldsymbol{\theta}^{(0)}$.
Step 1: for $r \geq 0$, calculate

$$\tilde{\boldsymbol{u}}_{\mathbf{1}}^{(\boldsymbol{r+1})} = E(\mathbf{U}|\mathbf{y}, \boldsymbol{\theta}^{(r)}) = \boldsymbol{\Sigma}_{\boldsymbol{UY}}^{(\boldsymbol{r})}\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-\mathbf{1}(\boldsymbol{r})}(\mathbf{y}_i - \boldsymbol{\mu}_{\boldsymbol{Y}}^{(\boldsymbol{r})})$$

and

$$E(\boldsymbol{U}^{\boldsymbol{T}}\boldsymbol{U}|\mathbf{y}, \boldsymbol{\theta}^{(r)}) = \tilde{\mathbf{u}}_{\mathbf{1}}{}^T\tilde{\mathbf{u}}_{\mathbf{1}} + tr\left\{\boldsymbol{\Sigma}_{\boldsymbol{U}}^{(\boldsymbol{r})} - \boldsymbol{\Sigma}_{\boldsymbol{UY}}^{(\boldsymbol{r})}\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-\mathbf{1}(\boldsymbol{r})}\boldsymbol{\Sigma}_{\boldsymbol{UY}}^{(\boldsymbol{r})}\right\}.$$

Step 2: for $r \geq 0$ calculate $\hat{\boldsymbol{\Gamma}}^{(\boldsymbol{r+1})}$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{R}}^{(\boldsymbol{r+1})}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{U}}^{(\boldsymbol{r+1})}$ based on the Equations (7), (9) and (10).
Step 3: Iterate Steps 1 and 2 from $r = 1$ until reaching convergence.
These EM update formulas for the random intercept and slope model form the computational foundation for the boosting algorithm developed in Section 3.3. The proposed EM–based boosting method directly uses the derived expectations and parameter updates introduced in this section.

3. The boosting algorithm

In this section, we explain L_2 boosting and also propose a boosting algorithm for random intercept and slope model based on the EM algorithm.
Model choice and variable selection are issues of major concern in practical regression analysis. Model based boosting is a tool to fit a statistical model while performing variable selection at the same time. The boosting algorithm, being fast and easy to implement, effectively addresses this issue. In fact, boosting is a method that is used in machine learning to reduce errors in predictive data analysis [10].

3.1. The L_2 boosting algorithm.

Bühlmann and Yu [5] introduced L_2 boosting algorithm that builds a linear model by minimizing the L_2 loss. The L_2 boosting is the simplest and perhaps most instructive boosting algorithm which is very useful for regression models, in particular in the presence of high dimensional explanatory variables. We consider a simple linear regression model

$$y = x\beta + \varepsilon,$$

where y is the dependent variable, x is the independent (explanatory) variable and ε follows normal distribution $N(0, \sigma^2)$. The parameters of the model are

estimated by the ordinary least squares method, where the loss function is defined as $L(y;F(x)) = (y-\hat{y})^2$, and $F(x)$ denotes the regression function of the model. For the simple linear regression model $y = x\beta + \varepsilon$, we have $F(x) = x\beta$, and therefore the fitted value is $\hat{y} = \hat{F}(x) = x\hat{\beta}$. We minimize the sum of squared errors

$$L = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

where $\hat{y}_i = x_i\hat{\beta}$. The least square estimator of the unknown parameter β can be calculated as

$$\hat{\beta} = (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T Y}.$$

We need to use a modified method to get over the high dimension problem, because the matrix $(\boldsymbol{X}^T\boldsymbol{X})$ is not invertible, so the ordinary least square method falls down. The idea of L_2 boosting is to use only one explanatory variable at a time. Since only one variable is used in one particular iteration, the matrix $(\boldsymbol{x}_j^T\boldsymbol{x}_j)$ is a scalar and thus invertible. The general procedure of L_2 boosting algorithm is described as, Buhlmann and Yu [5].

Step 1: Start with y_i from the training data.

Step 2 : For $m = 1$ *to* M:

a. For $j = 1$ *to* k (for each variable)

i. Fit the regression function $y_i = \beta_{m,0,j} + \beta_{m,j}x_{ji} + \varepsilon_i$ by least-squares of y_i on x_{ji}.

ii. Compute $err_{mj} = 1 - R^2_{mj}$ where R^2_{mj} is the coefficient of determination from the least squares regression.

b. b. Find $\hat{j}_m = \underset{j}{\operatorname{argmin}}\ err_{mj}$

c. Set $y_i \leftarrow y_i - \hat{\beta}_{m,0,\hat{j}_m} - \hat{\beta}_{m,\hat{j}_m}x_{\hat{j}_m,i},\ \ i = 1,2,...,n.$

Step 3 : Output the final regression model

$$F_M(x) = \sum_{m=1}^{M}\left(\hat{\beta}_{m,0,\hat{j}_m} + \hat{\beta}_{m,\hat{j}_m}\,x_{\hat{j}_m}\right).$$

3.2. **The proposed boosting algorithm.** In the real situation, the true model of data is unknown. We usually consider one or more parametric model as rival models and choose the model that is closest to the true model as the optimal model. So, we faced with amount of risk. The most important risk function is based on the known risk, Kullback-Leibler divergence. This risk is a mis-specification risk, the risk of selection an unsuitable approximation for true model. Given a pair of rival models, their distance from the true model is measured by the minimum Kullback-Leibler risk criterion. Now, we will select the model that is closest to the true distribution. It means that we will select the model which has minimum Kullback-Leibler risk. Note that the Kullback-Leibler risk minimizer is the likelihood function maximizer. The

Akaike Information Criterion, AIC, initially is proposed as an estimate of minus twice the expected log-likelihood. The AIC is defined as

$$\text{AIC } = -2\log \text{ likelihood} + 2\text{Number of estimated parameters.}$$

It indicates that the bias of the log-likelihood approximately becomes the number of free parameters contained in the model. The bias is derived under the assumption that the true model is contained in the specified parametric model. In the following, we consider the random slope and intercept regression model

$$\mathbf{Y} = \mathbf{X\Gamma} + \mathbf{ZU} + \mathbf{R}.$$

We propose a boosting algorithm based on the EM algorithm. The proposed boosting algorithm is described as:

Algorithm 1 Boosting with Random Effects Using EM Weak Learners

1: Training data $\{(y_i, x_{i1}, \ldots, x_{ik})\}_{i=1}^{n}$, number of iterations M.
2: **Initialization:** Set pseudo response $\mathbf{r}_i^{(0)} = y_i$, for $i = 1, \ldots, n$.
3: For $m = 1$ to M
4: For $j = 1$ to k

Fit random intercept and slope regression model

$$\mathbf{r}_i^{(m-1)} = \mathbf{x}_{ji}\mathbf{\Gamma}_{m,j} + \mathbf{z}_{ji}\mathbf{U}_{m,j} + \boldsymbol{R}_i,$$

using the EM algorithm, obtaining $\widehat{\mathbf{\Gamma}}_{m,j}$ and $\widehat{\mathbf{U}}_{m,j}$

5: Compute

$$\text{AIC}_{m,j} = -2\log \text{ likelihood function} + 2p_{m,j},$$

where $p_{m,j}$ is the number of parameters in the mixed effects model.

6: Select the best component

$$\hat{j}_m = \arg\min_j \text{AIC}_{m,j}.$$

7: Update the pseudo responses (residual update):

$$\mathbf{r}_i^{(m)} \leftarrow \mathbf{r}_i^{(m-1)} - \mathbf{x}_{\hat{j}_m i}\widehat{\mathbf{\Gamma}}_{m,\hat{j}_m} - \mathbf{z}_{\hat{j}_m i}\widehat{\mathbf{U}}_{m,\hat{j}_m}, \qquad i = 1, \ldots, n.$$

8: **Final model:**

$$F_M(x_i) = \sum_{m=1}^{M} \left(\mathbf{x}_{\hat{j}_m i}\widehat{\mathbf{\Gamma}}_{m,\hat{j}_m} + \mathbf{z}_{\hat{j}_m i}\widehat{\mathbf{U}}_{m,\hat{j}_m}\right).$$

4. Simulation study

In the present section, we examine the performance of the EM algorithm, the boosting algorithm, and the stepwise method in estimating the model parameters and variable selection, respectively, using simulation studies.

4.1. **Estimation of random intercept and slope regression model.** We consider random intercept and slope regression model (3), where, $j = 1, \ldots, 5$, γ_{00} and γ_{10} are the fixed intercept and slope, u_{0j} and u_{1j} are group-specific random effects, and $R_{ij} \sim N(0, \sigma^2)$ represents the residual error. The random effects (u_{0j}, u_{1j}) are bivariate normal with zero mean and covariance matrix Σ_u. In this simulation study, the true parameter values were set as $\gamma_{00} = 2.00$, $\gamma_{10} = 1.00$, $\sigma^2 = 0.95$ and

$$\Sigma_u = \begin{pmatrix} \tau_0^2 & \tau_{01} \\ \tau_{10} & \tau_1^2 \end{pmatrix} = \begin{pmatrix} 0.50 & 0.10 \\ 0.10 & 0.30 \end{pmatrix}.$$

For total sample sizes $n \in \{50, 100, 200, 300, 500, 1000\}$ we generated 1000 independent datasets under the model (3) (with $m = 5$ groups and the remaining observations per group determined so that the total number of observations equals n) and fitted the model using two estimation procedures: (i) Maximum Likelihood (ML) and (ii) Expectation-Maximization implementation (EM). For each estimator and each sample size we report the Monte Carlo mean of the parameter estimates (Mean Estimate), the empirical standard deviation of the estimates across simulations (SE), the mean squared error of the estimates relative to the true parameter value (MSE) and the Akaike Information Criterion (AIC). The numerical summaries for ML are given in Table 1 and for EM in Table 2.

TABLE 1. ML: Mean Estimates, SE, MSE and AIC for γ_{00} and γ_{10} ($m = 5$)

n	Parameter	Mean Estimate	SE	MSE	AIC
50	γ_{00}	2.22	0.16	0.026	143.2
	γ_{10}	1.25	0.13	0.017	143.2
100	γ_{00}	2.18	0.12	0.015	285.5
	γ_{10}	1.20	0.10	0.010	285.5
200	γ_{00}	2.13	0.09	0.008	569.8
	γ_{10}	1.17	0.08	0.006	569.8
300	γ_{00}	2.10	0.08	0.006	854.0
	γ_{10}	1.15	0.07	0.005	854.0
500	γ_{00}	2.09	0.06	0.004	1421.5
	γ_{10}	1.14	0.05	0.003	1421.5
1000	γ_{00}	2.05	0.04	0.002	2840.0
	γ_{10}	1.09	0.03	0.001	2840.0

The results in Table 1 indicate that the ML estimator yields small positive bias in the estimated slope across all sample sizes (the mean estimates of γ_{10} exceed the true value 1.00 by roughly 0.09-0.25), while the intercept γ_{00} is estimated with negligible bias. The empirical standard deviations (SE) of both fixed effect estimates decrease with increasing n, but remain appreciable for

small samples (e.g. SE($\hat{\gamma}_{10}$) ≈ 0.13 at $n = 50$), and the MSE of the estimates shows the same monotone improvement as n grows. The averaged AIC values reported alongside each pair of parameter summaries reflect the overall fit of the fitted model under ML and increase with n because AIC includes a sample size dependent likelihood term; however, comparisons of AIC between ML and EM shows that EM performs better than ML.

TABLE 2. EM: Mean Estimates, SE, MSE and AIC for γ_{00} and γ_{10} ($m = 5$)

n	Parameter	Mean Estimate	SE	MSE	AIC
50	γ_{00}	2.03	0.10	0.010	36.0
	γ_{10}	1.06	0.08	0.006	36.0
100	γ_{00}	2.03	0.07	0.005	74.2
	γ_{10}	1.05	0.05	0.003	74.2
200	γ_{00}	2.02	0.05	0.003	148.1
	γ_{10}	1.05	0.04	0.002	148.1
300	γ_{00}	2.01	0.04	0.002	202.3
	γ_{10}	1.02	0.03	0.001	202.3
500	γ_{00}	2.00	0.03	0.001	367.0
	γ_{10}	1.00	0.02	0.001	367.0
1000	γ_{00}	2.00	0.02	0.001	710.1
	γ_{10}	1.00	0.01	0.0005	710.1

The EM results in Table 2 show that EM recovers the true fixed effects essentially without bias across all sample sizes considered, and yields markedly smaller SE and MSE compared with ML; this advantage is most pronounced at small-to-moderate sample sizes (for example, at $n = 50$ the SE of $\hat{\gamma}_{10}$ under EM is about 0.08 versus about 0.13 under ML, and the MSE is roughly 0.006 versus 0.017). The averaged AIC values for EM-fitted models are consistently much lower than those for ML, reflecting both superior likelihood and better recovery of variance components under the EM implementation used in these simulations. Taken together, the simulation evidence with $m = 5$ groups and the specified random effects covariance shows that EM is substantially more efficient and accurate than ML for estimating fixed effects in this random intercept and slope model, producing smaller estimation variability and lower mean squared error while achieving better AIC based fit.

4.2. **Variable selection methodology, Part 1.** Consider the model

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{1,ij} + R_{ij}, \quad i = 1, \cdots, n, \quad j = 1, \cdots, 4 \tag{11}$$

where the residuals R_{ij}, are independent and identically distributed as the standard normal distribution $N(0,1)$. Also

$$\beta_{0j} = \gamma_{00} + u_{0j},$$
$$\beta_{1j} = \gamma_{10} + u_{1j}$$

where

$$\boldsymbol{\Gamma} = \begin{pmatrix} \gamma_{00} \\ \gamma_{10} \end{pmatrix} = \begin{pmatrix} 2 \\ 0.9 \end{pmatrix}, \ \boldsymbol{U}_j = \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N_2\left(\mathbf{0}, \boldsymbol{\Sigma}\right), \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}.$$

The observation are generated from model (11), where $x_{1,ij}$'s are generated from $u(0,1)$. To examine the performance of variable selection classical methods (such as stepwise method) and machine learning algorithms (such as L_2 boosting and the proposed boosting algorithm) the variables $x_{1,ij}, x_{2,ij}, x_{3,ij}$, $x_{4,ij}, x_{5,ij}$ and $x_{6,ij}$ are considered, which are generated from uniform $U(3,6)$, normal $N(0,3)$, uniform $U(-2,2)$, gamma $G(2,5)$, Weibull $W(3,7)$ distributions and $x_6 = \exp(x_3)$ respectively.
The $R = 10^4$ independent samples

$$\{y_{ij}^{(r)}; i = 1, \cdots, n; j = 1, \cdots, 4; r = 1, \cdots, R\}$$

are generated for different sample sizes given by $n = 40, 50, 100, 250, 500, 1000$ and the results of variable selection based on the stepwise method, L_2 boosting and the proposed boosting algorithm are reported in Table 3.

TABLE 3. Performance comparison of Stepwise, L_2 Boosting, and Proposed Boosting in variable selection for model (11).

Method	n	RF	R^2	R^2_{adj}	AIC	MSE
Stepwise	50	0.5620	0.4414	0.4328	143.02	0.9097
	100	0.5452	0.4302	0.4259	285.38	0.9549
	200	0.5385	0.4219	0.4198	568.96	0.9757
	300	0.5690	0.4048	0.4035	852.75	0.9833
	500	0.5692	0.4133	0.4125	1420.55	0.9903
	1000	0.5665	0.4079	0.4075	2839.89	0.9955
L_2 Boost	50	0.7228	0.7654	0.7526	96.41	0.4823
	100	0.7750	0.8017	0.7937	182.77	0.3515
	200	0.8152	0.8352	0.8295	356.22	0.2210
	300	0.8359	0.8620	0.8533	498.55	0.1804
	500	0.8553	0.8813	0.8742	812.44	0.1327
	1000	0.8857	0.9032	0.8950	1505.88	0.0711
Proposed Boost	50	0.8577	0.9066	0.8942	35.64	0.2519
	100	0.8995	0.9110	0.9080	73.90	0.1791
	200	0.9300	0.9120	0.9041	147.36	0.0980
	300	0.9432	0.9472	0.9272	202.25	0.0822
	500	0.9564	0.9690	0.9490	366.76	0.0157
	1000	0.9635	0.9658	0.9558	710.10	0.0082

In this table, the relative frequency (RF) of true model identification, the mean of determination coefficient (R^2), adjusted determination coefficient (R^2_{adj}), Akaike information criterion (AIC) and mean squared error of prediction (MSE) of fitted models based on the stepwise method L_2 boosting and proposed boosting algorithm are reported. Also, β_{1j}, β_{2j}, β_{3j}, β_{4j}, β_{5j} and β_{6j} are the regression coefficients corresponding to the variables $x_{1,ij}$, $x_{2,ij}$, $x_{3,ij}$, $x_{4,ij}$, $x_{5,ij}$ and $x_{6,ij}$, respectively. The RF metric, representing the proportion of 10^4 simulation runs in which the true model was identified, indicates that the stepwise method consistently identifies the true model only about $54-57\%$ of the time, with minimal improvement as sample size increases. In contrast, the proposed boosting algorithm achieves RF values from 0.86 to 0.96, demonstrating high reliability in model selection, which improves with larger sample sizes. The relative frequency of true model identification by the proposed boosting algorithm converges to one as the sample size increases. The relative frequency of true model identification by the stepwise method is approximately 0.5. Correspondingly, predictive accuracy metrics show that the proposed boosting algorithm substantially outperforms stepwise methods: R^2 and R^2_{adj} are consistently above 0.9, while MSE remains extremely low, reflecting precise predictions. The AIC values further confirm that the proposed boosting algorithm produces more parsimonious and better fitting models. Overall, these results indicate that the proposed boosting algorithm not only achieves superior predictive performance but also reliably identifies the true model structure, whereas the stepwise method exhibits moderate performance with limited ability to recover the true model even as the sample size increases.

Also, the results indicate that while L_2 boosting performs better than the classical stepwise method, especially in high dimensional variable selection, its accuracy, stability, and model recovery rate remain consistently below those of the proposed boosting algorithm. In particular, RF values for L_2 boosting increase with sample size but do not reach the near perfect levels achieved by the proposed method. This demonstrates the advantage of incorporating random effect adjustments in the proposed boosting framework, which enables more accurate estimation and superior identification of the true underlying model structure compared to standard L_2 boosting.

The relative frequency (RF) indicates how often the true model is selected across repeated simulations. Values close to 1 reflect a highly reliable model selection procedure, meaning the method almost always identifies the true model. Conversely, RF values around 0.5 suggest that the selection is only slightly better than random guessing, indicating low reliability. In practice, RF above 0.8 is generally considered strong evidence of a robust and consistent model selection approach.

4.3. **Variable selection methodology, Part 2.** Consider the model

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{1,ij} + \beta_{2j}x_{2,ij} + R_{ij}, \quad i = 1, \cdots, n, \quad j = 1, \cdots, 4 \tag{12}$$

The observations are generated from model (12) where, R_{ij}'s are independent and identically distributed as standard normal distribution $N(0,1)$. Here,

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + u_{0j},\\ \beta_{1j} &= \gamma_{10} + u_{1j},\\ \beta_{2j} &= \gamma_{20} + u_{1j},\end{aligned}$$

where

$$\boldsymbol{\Gamma} = \begin{pmatrix} \gamma_{00} \\ \gamma_{10} \\ \gamma_{20} \end{pmatrix} = \begin{pmatrix} 2 \\ 0.9 \\ 0.4 \end{pmatrix},\ \boldsymbol{U}_j = \begin{pmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \end{pmatrix} \sim N_2\left(\mathbf{0}, \boldsymbol{\Sigma}\right), \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.6 & 0.8 \\ 0.6 & 1 & 0.4 \\ 0.8 & 0.4 & 1 \end{pmatrix}$$

The mentioned criteria in the previous simulation are presented in Table 4. The results show that the proposed boosting method clearly outperforms both the Stepwise and L_2 Boost approaches across all sample sizes. While Stepwise remains unreliable with an RF near 0.5 and L_2 boosting shows only moderate improvement, the proposed method achieves an RF close to 1 even for small samples, demonstrating highly consistent variable selection. It also provides higher R^2 and R^2_{adj}, much lower AIC values, and substantially smaller MSE, especially as the sample size increases. Overall, the proposed boosting algorithm offers superior accuracy, better model fit and more efficient prediction compared with the competing methods.

TABLE 4. Comparison of variable selection criteria for Stepwise, L_2 Boosting and the Proposed Boosting algorithm for the model (12).

Method	n	RF	R^2	R^2_{adj}	AIC	MSE
Stepwise	50	0.5705	0.7153	0.7013	144.2115	0.9000
	100	0.5335	0.7078	0.7008	286.6096	0.9503
	200	0.5262	0.7052	0.7017	570.2180	0.9734
	300	0.5030	0.6988	0.6965	854.0032	0.9817
	500	0.5130	0.7018	0.7004	1421.767	0.9894
	1000	0.4990	0.6998	0.6991	2841.164	0.9951
L_2 Boost	50	0.8052	0.8235	0.8174	60.5214	0.40367
	100	0.8558	0.8385	0.8258	115.6387	0.3285
	200	0.8875	0.8640	0.8552	230.2471	0.2881
	300	0.9098	0.8975	0.8793	345.3791	0.2210
	500	0.9310	0.9064	0.8886	550.584	0.1809
	1000	0.9472	0.9255	0.9120	1452.358	0.1428
Proposed Boost	50	0.9995	0.8990	0.8872	14.6910	0.0884
	100	1.0000	0.9078	0.8984	26.6765	0.0865
	200	1.0000	0.9199	0.9075	50.1191	0.0845
	300	1.0000	0.9491	0.9157	59.7746	0.0685
	500	1.0000	0.9504	0.9254	108.3752	0.0074
	1000	1.0000	0.9988	0.9518	203.5411	0.0071

To evaluate the performance of the L_2 boosting, proposed boosting algorithm and the stepwise method in variable selection with small regression coefficients (close to zero), data are generated from model (12), where

$$\mathbf{\Gamma} = \begin{pmatrix} \gamma_{00} \\ \gamma_{10} \\ \gamma_{20} \end{pmatrix} = \begin{pmatrix} 2 \\ 0.9 \\ 0.001 \end{pmatrix}.$$

The random intercept and slope models

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{1,ij} + \beta_{2j} x_{2,ij} + R_{ij}, \quad i = 1, \cdots, n, \quad j = 1, \cdots, 4 \tag{13}$$

and

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{1,ij} + R_{ij}, \quad i = 1, \cdots, n, \quad j = 1, \cdots, 4 \tag{14}$$

are considered as competing models and the relative frequency of the L_2 boosting, proposed boosting algorithm and the stepwise method for model (13) (RF_{M1}) and model (14) (RF_{M2}) and other mentioned criteria in the previous simulations are presented in Table 5.

TABLE 5. Comparison of Stepwise, L_2 Boosting, and Proposed Boosting under model (12) with a near zero coefficient.

Method	n	RF_{M1}	RF_{M2}	R^2	R^2_{adj}	AIC	MSE
Stepwise	50	0.5562	0.0662	0.4416	0.4330	143.0206	0.9094
	100	0.5275	0.0708	0.4302	0.4259	285.3776	0.9550
	200	0.5295	0.0680	0.4218	0.4197	568.9626	0.9758
	300	0.5010	0.0698	0.4048	0.4035	852.7473	0.9833
	500	0.5125	0.0813	0.4133	0.4125	1420.548	0.9903
	1000	0.4972	0.0670	0.4079	0.4075	2839.894	0.9955
L_2 Boost	50	0.0200	0.7800	0.9200	0.9000	1.9238	0.0040
	100	0.0150	0.8300	0.9280	0.9050	2.1074	0.0030
	200	0.0100	0.8700	0.9320	0.9085	2.9138	0.0021
	300	0.0060	0.9000	0.9430	0.9150	5.2341	0.0018
	500	0.0030	0.9250	0.9490	0.9200	10.3823	0.0010
	1000	0.0020	0.9400	0.9550	0.9300	30.2161	0.0007
Proposed Boost	50	0.0022	0.8495	0.9406	0.9101	0.3847	0.0028
	100	0.0017	0.8937	0.9411	0.9110	0.2860	0.0009
	200	0.0002	0.9272	0.9412	0.9120	0.7088	0.0012
	300	0.0002	0.9425	0.9597	0.9297	3.8138	0.0004
	500	0.0000	0.9550	0.9609	0.9309	3.6059	0.0002
	1000	0.0000	0.9627	0.9605	0.9405	56.7524	0.0001

The results show that when one of the regression coefficients is close to zero, the proposed boosting algorithm reliably identifies the reduced model by assigning very small selection frequency to model M_1 and consistently selecting model M_2

with high RF_{M_2}, which approaches one as the sample size increases. L_2 boosting demonstrates better performance than stepwise but remains weaker than the proposed method, producing lower RF_{M_2} values, moderately larger MSE, and higher AIC across all sample sizes. In contrast, stepwise fails to detect the near-zero coefficient and continues to select the full model with $RF_{M_1} \approx 0.5$, showing no improvement with increasing sample size. Overall, the proposed boosting algorithm exhibits the highest sensitivity to small coefficients, the strongest model selection stability, and the best predictive performance among the three competing methods.

4.4. **Variable selection methodology, Part 3.** In this part, we consider the model

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{1,ij} + \beta_{7j}x_{7,ij} + R_{ij}, \quad i = 1, \cdots, n, \quad j = 1, \cdots, 4 \tag{15}$$

where the residuals, R_{ij}, are independent and identically distributed as the standard normal distribution $N(0, 1)$,

$$\boldsymbol{\Gamma} = \begin{pmatrix} \gamma_{00} \\ \gamma_{10} \\ \gamma_{70} \end{pmatrix} = \begin{pmatrix} 2 \\ 0.9 \\ 0.8 \end{pmatrix}, \ \boldsymbol{U}_j = \begin{pmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \end{pmatrix} \sim N_2\left(\mathbf{0}, \boldsymbol{\Sigma}\right), \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.6 & 0.8 \\ 0.6 & 1 & 0.4 \\ 0.8 & 0.4 & 1 \end{pmatrix}$$

and $x_{7,ij} = 0.9x_{2,ij} + 0.8x_{3,ij}$. The random intercept and slope models

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{1,ij} + \beta_{7j}x_{7,ij} + R_{ij}, \quad i = 1, \cdots, n, \quad j = 1, \cdots, 4 \tag{16}$$

and

$$y_{ij} = \beta_{0j} + B_{1j}x_{1,ij} + \beta_{2j}x_{2,ij} + \beta_{3j}x_{3,ij} + R_{ij}, \quad i = 1, \cdots, n, \quad j = 1, \cdots, 4 \tag{17}$$

are considered as competing models. The relative frequencies of the L_2 boosting, proposed boosting algorithm and the stepwise method in selecting the model (16) (RF_{M3}) and the model (17) (RF_{M4}) and other mentioned criteria are presented in Table 6. The results clearly show that the proposed boosting algorithm consistently outperforms both stepwise and L_2 boosting across all sample sizes. It achieves the highest relative frequency in selecting the true model (M_3) and offers notably improved predictive accuracy, as reflected by substantially lower AIC and MSE values. While L_2 boosting performs better than stepwise, particularly in reducing prediction error, it still falls short of the proposed method, which demonstrates superior stability and accuracy in variable selection. Overall, the findings confirm that the proposed boosting framework is the most effective and reliable approach among the three competing methods.

5. Real data analysis

In this section, the life expectancy dataset is used to variable selection based on the stepwise method, L^2 boosting algorithm and proposed boosting algorithm.

TABLE 6. Variable selection results for Models (15) using Stepwise, L_2 Boosting, and the Proposed Boosting algorithm.

Method	n	RF_{M3}	RF_{M4}	R^2	R^2_{adj}	AIC	MSE
Stepwise	50	0.6575	0.1007	0.8731	0.8560	144.5306	0.9096
	100	0.6232	0.1073	0.8707	0.8621	286.8709	0.9554
	200	0.6172	0.1013	0.8701	0.8657	570.4839	0.9761
	300	0.6022	0.1057	0.8684	0.8656	854.2920	0.9838
	500	0.6085	0.1085	0.8691	0.8674	1422.021	0.9905
	1000	0.5997	0.1128	0.8686	0.8677	2841.428	0.9956
L_2 Boost	50	0.7020	0.2105	0.9028	0.9004	61.4203	0.4021
	100	0.7102	0.2181	0.9054	0.9041	120.3094	0.3884
	200	0.7155	0.2210	0.9071	0.9065	242.1180	0.3712
	300	0.7224	0.2150	0.9090	0.9079	354.2291	0.3605
	500	0.7180	0.2193	0.9082	0.9076	602.4428	0.3528
	1000	0.7251	0.2124	0.9104	0.9099	1192.330	0.3455
Proposed Boost	50	0.7495	0.2342	0.9252	0.9249	17.7122	0.2442
	100	0.7422	0.2542	0.9256	0.9252	73.8998	0.2554
	200	0.7422	0.2567	0.9258	0.9256	147.3572	0.2569
	300	0.7640	0.2375	0.9326	0.9322	202.2468	0.2365
	500	0.7422	0.2577	0.9282	0.9278	366.7582	0.2572
	1000	0.7502	0.2497	0.9310	0.9307	710.0968	0.2498

5.1. **Analysis of data based on the classical multiple regression.** We consider the life expectancy dataset. This dataset consists of information such as the country name, the status of the country (developed or developing), factors affecting life expectancy such as income, mortality rates, vaccination factors, economic factors, social factors and other health related factors for 183 different countries. A summary of the information related to this dataset is provided in Table 7. This dataset is available on the kaggle website.

In the dataset, the life expectancy is considered as dependent variable (y) and the variables such as adult mortality, infant deaths, alcohol, expenditure percentage, Hepatitis B, measles, BMI, under-five deaths, polio, Total expenditure, Diphtheria, HIV/AIDS, GDP, population, thinness 1-19 years and thinness 5-9 years, income composition of resources and schooling are considered as independent variables $X_1, X_2, \cdots, X_{18}$ respectively.

The correlation coefficient and the p-value of the correlation test between the dependent variable and the independent variables are reported in Table 8. It shows that there is a correlation (-0.6589) between the Life Expectancy (y) and Adult Mortality (X_2). Also, all p-values of correlation test between the dependent variable and the independent variables (except X_{14}) are less than 0.05, it shows a significant correlation between the dependent variable and the independent variables. Also, the correlation values between the independent variables are given in Table 9. It shows that there is correclation 0.939 between variable X_{15} and X_{16}.

TABLE 7. A summary of life expectancy dataset.

Number	Variables	Icon displayed	nature
1	$Life\ expectancy$	y	numeric
2	$Country$		$There\ are\ 183\ countries$
3	$Status$		$Developed - Developing$
4	$Adult\ Mortality$	X_1	numeric
5	$infant\ deaths$	X_2	numeric
6	$Alcohol$	X_3	numeric
7	$percentage\ expenditure$	X_4	numeric
8	$HepatitisB$	X_5	numeric
9	$Measles$	X_6	numeric
10	BMI	X_7	numeric
11	$under - five\ deaths$	X_8	numeric
12	$Polio$	X_9	numeric
13	$Total\ expenditure$	X_{10}	numeric
14	$Diphtheria$	X_{11}	numeric
15	$HIV/AIDS$	X_{12}	numeric
16	GDP	X_{13}	numeric
17	$Population$	X_{14}	numeric
18	$thinness\ 1 - 19\ years$	X_{15}	numeric
19	$thinness\ 5 - 9\ years$	X_{16}	numeric
20	$Income\ composition\ of\ resources$	X_{17}	numeric
21	$Schooling$	X_{18}	numeric

TABLE 8. The correlation coefficient and p-value of the correlation test between dependent and independent variables.

$variables$	ρ	$p - value$
(y, X_1)	-0.6589	0
(y, X_2)	-0.1872	$1.629776e^{-24}$
(y, X_3)	0.3641	$1.603314e^{-92}$
(y, X_4)	0.3718	$1.060857e^{-96}$
(y, X_5)	0.2362	$2.059512e^{-38}$
(y, X_6)	-0.1500	$3.344361e^{-16}$
(y, X_7)	0.5450	$3.083808e^{-226}$
(y, X_8)	-0.2122	$3.460444e^{-31}$
(y, X_9)	0.4477	$2.259156e^{-144}$
(y, X_{10})	0.1610	$1.812727e^{-18}$
(y, X_{11})	0.4605	$1.131846e^{-153}$
(y, X_{12})	-0.5343	$6.919814e^{-216}$
(y, X_{13})	0.4177	$4.634611e^{-124}$
(y, X_{14})	-0.0272	$1.404545e^{-01}$
(y, X_{15})	-0.4323	$1.114056e^{-133}$
(y, X_{16})	-0.4272	$2.970185e^{-130}$
(y, X_{17})	0.5862	$6.103773e^{-270}$
(y, X_{18})	0.5699	$7.486033e^{-252}$

TABLE 9. Correlation values between the independent variables.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}
X_1	1	0.07	-0.18	-0.24	-0.13	0.03	-0.39	0.09	-0.28	-0.10	-0.28	0.52	-0.28	0.00	0.28	0.29	-0.39	-0.37
X_2		1	-0.10	-0.08	-0.13	0.50	-0.22	0.99	-0.16	-0.09	-0.16	0.02	-0.10	0.55	0.46	0.46	-0.13	-0.17
X_3			1	0.35	-0.04	-0.04	0.30	-0.10	0.22	0.39	0.21	-0.03	0.30	-0.02	-0.38	-0.37	0.27	0.31
X_4				1	-0.11	-0.05	0.23	-0.08	0.14	0.20	0.14	-0.09	0.90	-0.01	-0.24	-0.24	0.36	0.35
X_5					1	-0.12	0.18	-0.15	0.41	0.00	0.493	-0.12	-0.04	-0.03	-0.07	-0.07	0.20	0.19
X_6						1	-0.17	0.50	-0.12	-0.07	-0.132	0.03	-0.06	0.23	0.22	0.22	-0.14	-0.14
X_7							1	-0.23	0.29	0.20	0.29	-0.23	0.27	-0.06	-0.49	-0.50	0.43	0.45
X_8								1	-0.18	-0.10	-0.18	0.037	-0.10	0.53	0.46	0.46	-0.15	-0.18
X_9									1	0.14	0.69	-0.15	0.19	-0.04	-0.19	-0.19	0.34	0.35
X_{10}										1	0.15	0.02	0.10	-0.04	-0.21	-0.21	0.09	0.14
X_{11}											1	-0.16	0.18	-0.02	-0.20	-0.19	0.37	0.37
X_{12}												1	-0.11	-0.01	0.20	0.20	-0.19	-0.16
X_{13}													1	-0.02	-0.25	-0.26	0.42	0.40
X_{14}														1	0.23	0.23	0.01	-0.00
X_{15}															1	0.93	-0.29	-0.30
X_{16}																1	-0.28	-0.28
X_{17}																	1	0.86
X_{18}																	1	

In passing, we consider the multiple linear regression model

$$y_i = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_k x_{k,i} + \varepsilon_i, \quad i = 1, 2, \ldots, N,$$

where $N = n \times m = 183 \times 16 = 2928$, ε_i's are independent and identically distributed as a normal distribution, $N(\mu, \sigma^2)$. The results of variable selection based on the stepwise (forward, backward and both) method and L_2 boosting algorithm are given in Table 10. It shows that the fitted model based on the L_2 boosting algorithm has the lowest AIC_c and MSE values and the highest R^2 and R^2_{adj} values. Although 12 variables are selected based on the backward and both methods, 18 variables are selected based on the forward method. The variables X_4, X_5, X_6, X_{10}, X_{14} and X_{16} are not significant. These variables removed from the model and the results of the reduced model are presented in Table 11. So, we can write

$$\begin{aligned} Y = {} & 59.67 - 0.020X_1 + 0.125X_2 + 0.289X_3 + 0.057X_7 - 0.093X_8 + 0.0316X_9 \\ & + 0.033X_{11} - 0.501X_{12} + 7.560e^{-05}X_{13} - 0.148X_{15} + 4.924X_{17} + 0.204X_{18} + \varepsilon. \end{aligned} \tag{18}$$

The values of R^2, R^2_{adj}, AIC_c, MSE for model (18) are given in Table 10.

TABLE 10. Variable selection using the boosting algorithm and the stepwise method.

	n	*Shapiro*	R^2	R^2_{adj}	AIC_c	MSE	$N.V$
Step(both)	2928	$5.6223e^{-50}$	0.7253	0.7241	17953.99	26.67	13
Step(backward)	2928	$5.6223e^{-50}$	0.7253	0.7241	17953.99	26.67	13
Step(forward)	2928	$5.9295e^{-50}$	0.7256	0.7239	17960.87	26.64	18
$L2$ Boost	2928	$8.5864e^{-28}$	0.9918	0.9918	7646.325	0.7870	18

- $N.V$ is the number of selected variable.

The results of variable selection based on the L_2 boosting algorithm are presented in Table 12. It shows that the variables X_{12} , X_{17} and X_1 have the highest values of the variables relative influence.

5.2. **Analysis of data based on the random intercept and slope models.** We consider the regresion model

$$Y_{ij} = \beta_{0j} + \beta_{1j} x_{1,ij} + \ldots + \beta_{kj} x_{k,ij} + \varepsilon_{ij}, \quad i = 1, \ldots, 16, \quad j = 1, \ldots, 183$$

with $k = 18$ independet variable and selecte variables based on the stepwise method and L_2 boosting algorithm, where ε_{ij}'s are independent and distributed as normal distribution $N(\mu_j, \sigma_j^2)$. The results are given in Table 13.

TABLE 11. Results of variable selection using the stepwise (both) method.

	Coefficients	Estimate Std. Error	*t value*	$Pr(>\|t\|)$
(Intercept)	59.670	$5.583e^{-01}$	106.881	$< 2e^{-16}$
X_1	-0.020	$1.004e^{-03}$	-20.018	$< 2e^{-16}$
X_2	0.125	$1.046e^{-02}$	11.955	$< 2e^{-16}$
X_3	0.289	$2.711e^{-02}$	10.685	$< 2e^{-16}$
X_7	0.057	$6.054e^{-03}$	9.440	$< 2e^{-16}$
X_8	-0.093	$7.712e^{-03}$	-12.129	$< 2e^{-16}$
X_9	0.031	$5.613e^{-03}$	5.647	$1.79e^{-08}$
X_{11}	0.033	$5.620e^{-03}$	6.011	$2.07e^{-09}$
X_{12}	-0.501	$2.239e^{-02}$	-22.406	$< 2e^{-16}$
X_{13}	$7.560e^{-05}$	$8.173e^{-06}$	9.250	$< 2e^{-16}$
X_{15}	-0.148	$2.901e^{-02}$	-5.108	$3.46e^{-07}$
X_{17}	4.924	$8.003e^{-01}$	6.153	$8.67e^{-10}$
X_{18}	0.204	$4.709e^{-02}$	4.334	$1.51e^{-05}$

TABLE 12. Results of variable selection using the L_2 boosting algorithm method.

Variables	X_{12}	X_{17}	X_1	X_8	X_{11}	X_3	X_{16}	X_{18}	X_7
Rel.inf	39.65	27.71	21.38	2.11	1.75	1.52	1.37	1.006	0.91
Variables	X_9	X_{15}	X_{10}	X_5	X_{14}	X_{13}	X_2	X_4	X_6
Rel.inf	0.71	0.36	0.32	0.25	0.25	0.22	0.19	0.13	0.09

TABLE 13. Variable selection using the stepwise method and L_2 boosting algorithm.

	R^2	R^2_{adj}	AIC_c	MSE
Step(both)	0.9244	0.8656	442.7846	0.5495
Step(backward)	0.9243	0.8656	442.7846	0.5499
Step(forward)	0.9375	0.7459	448.0395	0.4690
L_2Boost	0.8304	0.8080	33.4955	0.5800

It shows that although the estimated model based on the stepwise (forward) method has the lowest MSE value and the estimated model based on the stepwise (both) method has the highest R^2_{adj} value, the difference between these values is not significant for the models estimated by the stepwise method and the L_2 boosting algorithm. Also, the model estimated using the L_2 boosting

algorithm has the lowest AIC_c value. Note that 9 variables are selected by the stepwise method, whereas 3 variables are selected by the boosting algorithm. The number of selected variables, AIC_c, R^2, R^2_{adj}, MSE and p-value of shapiro test of estimated model based on the stepwise (both) method and L_2 boosting algorithm for each country are given in Figures 1 and 2 respectively. They show that the performance of L_2 boosting algorithm is better. Also, for each country the residuals follow normal distribution, the p-values of shapiro test is greater than 0.05.

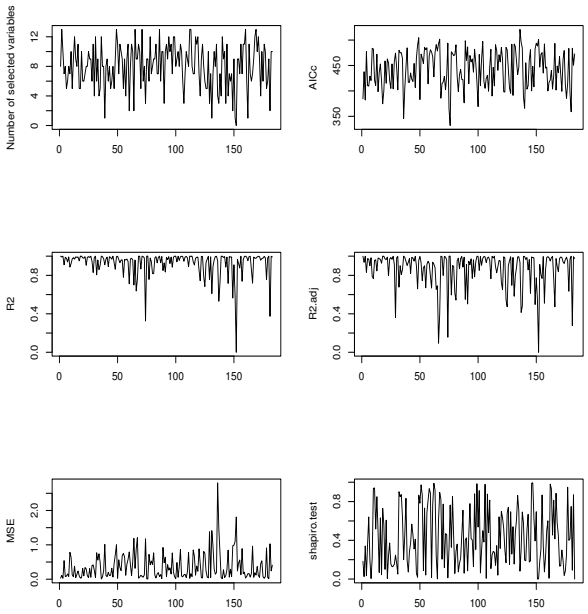

FIGURE 1. The number of selected variables, AIC_c, R^2, R^2_{adj}, MSE and p-value of shapiro test of estimated model based on the stepwise (both) method for each country.

Now, we consider the random intercept and slope model

$$\mathbf{Y}_i = \mathbf{X}_i\mathbf{\Gamma} + \mathbf{Z}_i\mathbf{U} + \mathbf{R}_i,$$

where

$$\mathbf{Y}_i = (\,\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m\,)^T, \mathbf{\Gamma} = (\,\mathbf{\Gamma_1}, \mathbf{\Gamma_2}, \ldots, \mathbf{\Gamma_m}\,)^T, \mathbf{\Gamma_j} = \begin{pmatrix} \gamma_{00} \\ \gamma_{10} \\ \vdots \\ \gamma_{k0} \end{pmatrix},$$

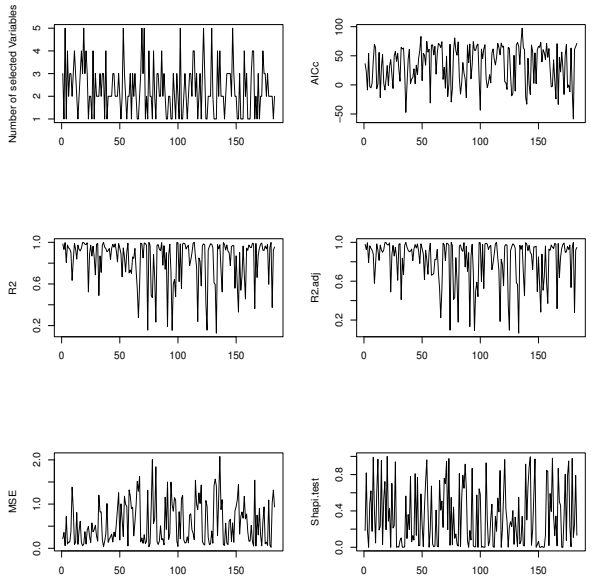

FIGURE 2. The number of selected variables, $AICc$, R^2, R^2_{adj}, MSE and p-value of shapiro test of estimated model based on the L^2 boosting algorithm for each country.

$$\mathbf{X}_i = Block\left\{\boldsymbol{X_1}, \boldsymbol{X_2}, ..., \boldsymbol{X_m}\right\}, \boldsymbol{X_j} = \begin{pmatrix} 1 & X_{1,1j} & ... & X_{k,1j} \\ 1 & X_{1,2j} & ... & X_{k,2j} \\ \vdots & \vdots & ... & \vdots \\ 1 & X_{1,nj} & ... & X_{k,nj} \end{pmatrix},$$

$\mathbf{R}_i = \left(\mathbf{R}_1, \mathbf{R}_2, ... ,\mathbf{R}_m\right)^T, n = 16, m = 183$ and $k = 18$. The results of variable selection based on the proposed boosting algorithm are given in Table 14. It shows that the estimated model based on the proposed boosting algorithm has the lowest MSE and AIC_c values and the highest R^2_{adj} values. Figure 3 confirms these results. The variables X_{15} and X_1 are selected based on the proposed boosting algorithm.

TABLE 14. Variable selection using the proposed boosting algorithm.

	R^2	R^2_{adj}	AIC_c	MSE
Proposed Boost	0.9329	0.9281	17.5489	0.4627

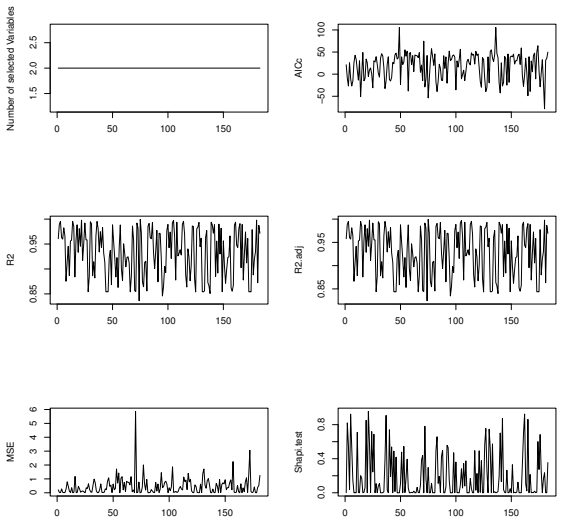

FIGURE 3. The number of selected variables, $AICc$, R^2, R^2_{adj}, MSE and p-value of shapiro test of estimated model based on the proposed boosting algorithm for each country.

6. Conclusion

In this paper, we make a number of contributions to the literature that relates to an extension of the basic multilevel regression model involving random intercept and slope model. First, we estimated the parameters of model based on the EM algorithm. Second, we proposed boosting algorithm based on the EM algorithm for variable selection in multilevel regression models. The performance of the proposed algorithm and the usual methods of variable selection were evaluated using simulated data. The simulation results show that the proposed boosting algorithm performs better in model selection for multilevel regression models than the usual variable selection methods. Also, we utilized L_2 and proposed boosting algorithm and multiple linear regression techniques in order to analyze the life expectancy dataset. Based on the results of validation methods, the built models appear to be accurate and strong. The results shows that the proposed boosting algorithm always performs better.
Future work may explore the use of Bayesian estimators for parameter estimation to assess potential gains in accuracy and robustness. Moreover, the proposed boosting algorithm can be extended by integrating Bayesian-based updates into its iterative structure, providing a promising direction for improving variable selection and prediction performance.

7. Data Availability Statement

This dataset is available on the website:
`https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who`

8. Aknowledgement

The authors are grateful to four anonymous referees and the associate editor for providing some useful comments on an earlier version of this manuscript.

9. Appendix A

The complete log- likelihood function can be expressed as

$$\begin{aligned} l(\boldsymbol{\theta};\mathbf{y}_c) &= l(\boldsymbol{\theta};\mathbf{y}|\mathbf{u}) + l(\boldsymbol{\theta};\mathbf{u}) \\ &= C - \frac{n}{2}\log \mid \boldsymbol{\Sigma}_R \mid -\frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\Gamma} - \mathbf{z}_i\mathbf{U}\right)^T \boldsymbol{\Sigma}_R^{-1}\left(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\Gamma} - \mathbf{z}_i\mathbf{U}\right) \\ &\quad - \frac{n}{2}\log \mid \boldsymbol{\Sigma}_U \mid -\frac{n}{2}\left(\mathbf{U}^T\boldsymbol{\Sigma}_U^{-1}\mathbf{U}\right), \end{aligned}$$

where C is a constant that is independent of $\boldsymbol{\Sigma}_R$, $\boldsymbol{\Sigma}_U$.
Let $\boldsymbol{\theta}^{(0)} = \left(\boldsymbol{\Gamma}^{(0)}, \boldsymbol{\Sigma}_R^{(0)}, \boldsymbol{\Sigma}_U^{(0)}\right)$ be starting values of $\boldsymbol{\theta} = (\boldsymbol{\Gamma}, \boldsymbol{\Sigma_R}, \boldsymbol{\Sigma}_U)$, so

$$\begin{aligned} Q &= E\left(l(\boldsymbol{\theta};\mathbf{y}_c \mid \mathbf{y}, \boldsymbol{\theta}^{(0)})\right) \\ &= -\frac{n}{2}\log \mid \boldsymbol{\Sigma}_R \mid -\frac{1}{2}\sum_{i=1}^{n} E\left\{\left(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\Gamma} - \mathbf{z}_i\mathbf{U}\right)^T \boldsymbol{\Sigma}_R^{-1}\left(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\Gamma} - \mathbf{z}_i\mathbf{U}\right) \mid \mathbf{y}, \boldsymbol{\theta}^{(0)}\right\} \\ &\quad - \frac{n}{2}\log \mid \boldsymbol{\Sigma}_U \mid -\frac{n}{2}\left(E\left\{\mathbf{U}^T\boldsymbol{\Sigma}_U^{-1}\mathbf{U} \mid \mathbf{y}, \boldsymbol{\theta}^{(0)}\right\}\right) + C. \end{aligned} \tag{19}$$

The third term of Equation (19) is equal to

$$\begin{aligned} E\left[\mathbf{R}_i^T\boldsymbol{\Sigma}_R^{-1}\mathbf{R}_i\right] &= E\left[\left(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\Gamma} - \mathbf{z}_i\mathbf{U}\right)^T \boldsymbol{\Sigma}_R^{-1}\left(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\Gamma} - \mathbf{z}_i\mathbf{U}\right) \mid \mathbf{y}, \boldsymbol{\theta}^{(0)}\right] \\ &= E\left[tr\left\{\left(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\Gamma} - \mathbf{z}_i\mathbf{U}\right)^T \boldsymbol{\Sigma}_R^{-1}\left(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\Gamma} - \mathbf{z}_i\mathbf{U}\right) \mid \mathbf{y}, \boldsymbol{\theta}^{(0)}\right\}\right] \\ &= E\left[tr\left\{\boldsymbol{\Sigma}_R^{-1}\left(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\Gamma} - \mathbf{z}_i\mathbf{U}\right)\left(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\Gamma} - \mathbf{z}_i\mathbf{U}\right)^T \mid \mathbf{y}, \boldsymbol{\theta}^{(0)}\right\}\right] \\ &= tr\left\{E\left(\boldsymbol{\Sigma}_R^{-1}\left(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\Gamma} - \mathbf{z}_i\mathbf{U}\right)\left(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\Gamma} - \mathbf{z}_i\mathbf{U}\right)^T \mid \mathbf{y}, \boldsymbol{\theta}^{(0)}\right)\right\} \\ &= tr\left\{\boldsymbol{\Sigma}_R^{-1}E\left(\left(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\Gamma} - \mathbf{z}_i\mathbf{U}\right)\left(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\Gamma} - \mathbf{z}_i\mathbf{U}\right)^T \mid \mathbf{y}, \boldsymbol{\theta}^{(0)}\right)\right\} \\ &= tr\left\{\boldsymbol{\Sigma}_R^{-1}E\left(\left(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\Gamma} - \mathbf{z}_i\mathbf{U}\right) \mid \mathbf{y}_t, \boldsymbol{\theta}^{(0)}\right) E\left(\left(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\Gamma} - \mathbf{z}_i\mathbf{U}\right) \mid \mathbf{y}, \boldsymbol{\theta}^{(0)}\right)^T\right. \\ &\quad \left. + \boldsymbol{\Sigma}_R^{-1}Var\left(\left(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\Gamma} - \mathbf{z}_i\mathbf{U}\right) \mid \mathbf{y}, \boldsymbol{\theta}^{(0)}\right)\right\} \\ &= E\left(\left(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\Gamma} - \mathbf{z}_i\mathbf{U}\right) \mid \mathbf{y}, \boldsymbol{\theta}^{(0)}\right)^T \boldsymbol{\Sigma}_R^{-1}E\left(\left(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\Gamma} - \mathbf{z}_i\mathbf{U}\right) \mid \mathbf{y}, \boldsymbol{\theta}^{(0)}\right) \end{aligned}$$

$$+ tr\left\{\boldsymbol{\Sigma}_R^{-1} Var\left((\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\Gamma} - \mathbf{z}_i\mathbf{U}) \mid \mathbf{y}, \boldsymbol{\theta}^{(0)}\right)\right\}$$
$$= (\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\Gamma} - \mathbf{z}_i\tilde{\mathbf{u}}_1)^T \boldsymbol{\Sigma}_R^{-1} (\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\Gamma} - \mathbf{z}_i\tilde{\mathbf{u}}_1) + tr\left\{\boldsymbol{\Sigma}_R^{-1}\mathbf{z}_i\tilde{\mathbf{b}}_1\mathbf{z}_i^T\right\},$$

where tr denote trace,

$$\tilde{\mathbf{u}}_1 = E\left(\mathbf{U} \mid \mathbf{y}, \boldsymbol{\theta}^{(0)}\right) = \boldsymbol{\Sigma}_{UY}^{(0)}\boldsymbol{\Sigma}_Y^{-1(0)}(\mathbf{y} - \boldsymbol{\mu}_Y^{(0)})$$

and

$$\tilde{\mathbf{b}}_1 = Var\left(\mathbf{U} \mid \mathbf{y}, \boldsymbol{\theta}^{(0)}\right) = \boldsymbol{\Sigma}_U^{(0)} - \boldsymbol{\Sigma}_{UY}^{(0)}\boldsymbol{\Sigma}_Y^{-1(0)}\boldsymbol{\Sigma}_{YU}^{(0)}.$$

The last term of Equation (19) is equal to

$$\begin{aligned} E\left(\mathbf{U}^T\boldsymbol{\Sigma}_U^{-1}\mathbf{U}|\mathbf{y}, \boldsymbol{\theta}^{(0)}\right) &= tr\left\{\boldsymbol{\Sigma}_U^{-1} E\left(\mathbf{U}|\mathbf{y}, \boldsymbol{\theta}^{(0)}\right) E\left(\mathbf{U}^T|\mathbf{y}, \boldsymbol{\theta}^{(0)}\right)\right\} \\ &\quad + tr\left\{\boldsymbol{\Sigma}_U^{-1} Var\left(\mathbf{U}|\mathbf{y}, \boldsymbol{\theta}^{(0)}\right)\right\} \\ &= tr\left\{\boldsymbol{\Sigma}_U^{-1}\tilde{\mathbf{u}}_1\tilde{\mathbf{u}}_1^T + \boldsymbol{\Sigma}_U^{-1} Var\left(\mathbf{U}|\mathbf{y}, \boldsymbol{\theta}^{(0)}\right)\right\} \\ &= \tilde{\mathbf{u}}_1^T\boldsymbol{\Sigma}_U^{-1}\tilde{\mathbf{u}}_1 + tr\left\{\boldsymbol{\Sigma}_U^{-1}\left(\boldsymbol{\Sigma}_U - \boldsymbol{\Sigma}_{UY}\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\Sigma}_{YU}\right)\right\}. \end{aligned} \tag{20}$$

By substituting Equations (**??**) and (20) in Equation (19), we have

$$\begin{aligned} Q &= E\left(l(\boldsymbol{\theta}; \mathbf{y}_c)\right) \\ &= C - \frac{n}{2}\log|\boldsymbol{\Sigma}_R| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\Gamma} - \mathbf{z}_i\tilde{\mathbf{u}}_1)^T\boldsymbol{\Sigma}_R^{-1}(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\Gamma} - \mathbf{z}_i\tilde{\mathbf{u}}_1) \\ &\quad + tr\left\{\boldsymbol{\Sigma}_R^{-1}\mathbf{z}_i\tilde{\mathbf{b}}_1\mathbf{z}_i^T\right\} \\ &\quad - \frac{n}{2}\log|\boldsymbol{\Sigma}_U| - \frac{n}{2}\left(\tilde{\mathbf{u}}_1^T\boldsymbol{\Sigma}_U^{-1}\tilde{\mathbf{u}}_1 + tr\left\{\boldsymbol{\Sigma}_U^{-1}\left(\boldsymbol{\Sigma}_U - \boldsymbol{\Sigma}_{UY}\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\Sigma}_{YU}\right)\right\}\right). \end{aligned}$$

We obtain estimators by solving the estimating equations as:

$$\begin{aligned} \frac{\partial Q}{\partial \boldsymbol{\Gamma}} &= -\frac{1}{2}\frac{\partial}{\partial \boldsymbol{\Gamma}}\left(\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\Gamma} - \mathbf{z}_i\tilde{\mathbf{u}}_1)^T\boldsymbol{\Sigma}_R^{-1}(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\Gamma} - \mathbf{z}_i\tilde{\mathbf{u}}_1) + tr\left\{\boldsymbol{\Sigma}_R^{-1}\mathbf{z}_i\tilde{\mathbf{b}}_1\mathbf{z}_i^T\right\}\right) \\ &= \sum_{i=1}^{n}\mathbf{x}_i^T\boldsymbol{\Sigma}_R^{-1}(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\Gamma} - \mathbf{z}_i\tilde{\mathbf{u}}_1) = 0, \text{ so} \end{aligned}$$

$$\hat{\boldsymbol{\Gamma}} = \left(\sum_{i=1}^{n}\mathbf{x}_i^T\boldsymbol{\Sigma}_R^{-1}\mathbf{x}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{x}_i^T\boldsymbol{\Sigma}_R^{-1}(\mathbf{y}_i - \mathbf{z}_i\tilde{\mathbf{u}}_1)\right). \text{ Also,}$$

$$\begin{aligned} \frac{\partial Q}{\partial \boldsymbol{\Sigma}_R} &= \frac{\partial Q}{\partial \boldsymbol{\Sigma}_R^{-1}}\left(\frac{\partial \boldsymbol{\Sigma}_R^{-1}}{\partial \boldsymbol{\Sigma}_R}\right) \\ &= \frac{1}{2}\sum_{i=1}^{n} Vec^T\left(\boldsymbol{\Sigma}_R^T\right)\left(\boldsymbol{\Sigma}_R^{-T}\otimes\boldsymbol{\Sigma}_R^{-1}\right) \\ &\quad - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\Gamma} - \mathbf{z}_i\tilde{\mathbf{u}}_1)^T\otimes(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\Gamma} - \mathbf{z}_i\tilde{\mathbf{u}}_1) + tr\left\{\mathbf{z}_i\tilde{\mathbf{b}}_1\mathbf{z}_i^T\right\}\left(\boldsymbol{\Sigma}_R^{-T}\otimes\boldsymbol{\Sigma}_R^{-1}\right) = 0, \end{aligned}$$

where $\otimes$ denotes Kronecker product and $Vec(A)$ denotes vectorization of matrix A. So we have

$$\hat{\mathbf{\Sigma}}_R = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{y}_i - \mathbf{x}_i\mathbf{\Gamma} - \mathbf{z}_i\tilde{\mathbf{u}}_1\right)^T\left(\mathbf{y}_i - \mathbf{x}_i\mathbf{\Gamma} - \mathbf{z}_i\tilde{\mathbf{u}}_1\right) + tr\left\{\mathbf{z}_i\tilde{\mathbf{b}}_1\mathbf{z}_i^T\right\}.$$

Similarly,

$$\hat{\mathbf{\Sigma}}_U = E\left(\mathbf{U}\mathbf{U}^T \mid \mathbf{y}, \boldsymbol{\theta}^{(0)}\right).$$

References

[1] Aitkin, M. & Longford, N. (1986). Statistical modelling in school effectiveness studies (with discussion). Journal of the Royal Statistical Society, 149, 1-43. https://doi.org/10.2307/2981882

[2] Barr, DJ. Levy, R. Scheepers, C. & Tily, HJ. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of memory and language, 68(3), 255-278. https://doi.org/10.1016/j.jml.2012.11.001

[3] Bouwmeester, W. Twisk, JW. Kappen, TH. van Klei, WA. Moons, KG. & Vergouwe, Y. (2013). Prediction models for clustered data: comparison of a random intercept and standard regression model. BMC medical research methodology, 13, 1-10. https://doi.org/10.1186/1471-2288-13-19

[4] Bühlmann, P. (2006). Boosting for high-dimensional linear models. Annals of Statistics, 34(2), 559-583. https://doi.org/10.1214/009053606000000092

[5] Bühlmann, P. & Yu, B. (2003). Boosting with the L_2 loss: Regression and classification. Journal of the American Statistical Association, 98(462), 324-339. https://doi.org/10.1198/016214503000125

[6] Chen, LP. & Qiu, B. (2023). Analysis of length-biased and partly interval-censored survival data with mismeasured covariates. Biometrics, 79, 3929-3940. https://doi.org/10.1111/biom.13898

[7] Chen, LP. (2024a). Variable selection and estimation for misclassified binary responses and multivariate error-prone predictors. Journal of Computational and Graphical Statistics, 33, 407-420. https://doi.org/10.1080/10618600.2023.2218428

[8] Chen, LP. (2024b). Accelerated failure time models with error-prone response and nonlinear covariates. Statistics and Computing, 34, 183. https://doi.org/10.1007/s11222-024-10491-9

[9] Choi, J. (2020). Nonnegative variance component estimation for mixed-effects models. Communications for Statistical Applications and Methods, 27(5), 523-533. https://doi.org/10.29220/CSAM.2020.27.5.523

[10] Chu, J. Lee, TH. Ullah, A. & Wang, R. (2020). Boosting. Macroeconomic Forecasting in the Era of Big Data: Theory and Practice, 431-463. https://link.springer.com/book/10.1007/978-3-030-31150-6

[11] Fisher, RA. (1919). The correlation between relatives on the supposition of Mendelian inheritance. Earth and Environmental Science Transactions of the Royal Society of Edinburgh, 52(2), 399-433. https://doi.org/10.1017/S0080456800012163

[12] Friedman, JH. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232. https://doi.org/10.1214/aos/1013203451

[13] Heiling, HM., Rashid, NU. Li, Q. Peng, XL., Yeh, JJ. & Ibrahim, JG. (2024). Efficient computation of high-dimensional penalized generalized linear mixed models by latent factor modeling of the random effects. Biometrics, 80(1). https://doi.org/10.1093/biomtc/ujae016

[14] Henderson, CR. Kempthorne, O. Searle, SR. & Von Krosigk, CM. (1959). The estimation of environmental and genetic trends from records subject to culling. Biometrics, 15(2), 192-218. https://www.cabidigitallibrary.org/doi/full/10.5555/19590102115
[15] Longford, NT. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. Biometrika, 74, 817-827. https://doi.org/10.2307/2336476
[16] McLean, RA., Sanders, WL. & Stroup, WW. (1991). A unified approach to mixed linear models. The American Statistician, 45(1), 54-64. https://doi.org/abs/10.1080/00031305.1991.10475767
[17] Mei, J. Zhang, Y. & Chen, M. (2022). A flexible EM algorithm for mixture and nonlinear mixed-effects models with complex random-effects structures. Statistical Methods in Medical Research, 31(11), 2180–2198. https://doi.org/10.1177/09622802221098687
[18] Oberauer, K. (2022). The importance of random slopes in mixed models for Bayesian hypothesis testing. Psychological Science, 33(4), 648-665. 09567976211046884
[19] Raudenbush, SW. & Bryk, AS. (1986). A hierarchical model for studying school effects. Sociology of Education, 12, 241-269. https://doi.org/10.2307/2112482
[20] Robinson, GK. (1991). That BLUP is a good thing: the estimation of random effects. Statistical science, 15-32. https://doi.org/10.1214/ss/1177011926
[21] Schapire, R. (1990). The strength of weak learnability. Mach. Learn, 5, 197-227. https://doi.org/10.1007/BF00116037
[22] Schapire, R. (1999). Theoretical views of boosting and applications, in Proc. Comput. Learn. Theory, 1-10. https://doi.org/10.1007/3-540-46769-6
[23] Zakkour, A. Perret, C. & Slaoui, Y. (2023). Stochastic expectation maximization algorithm for linear mixed-effects model with interactions in the presence of incomplete data. Entropy, 25(3), 473. https://doi.org/10.3390/e25030473.

Nasrin Ghadiri Ghazvini
Orcid number: 0009-0003-7652-8098
Department of Statistics
SR.C., Islamic Azad University
Tehran, Iran
Email address: nasrin.ghadiri@gmail.com

Ramin Kazemi
Orcid number: 0000-0001-5496-7565
Department of Statistics
Imam Khomeini International University
Qazvin, Iran
Email address: r.kazemi@sci.ikiu.ac.ir

Mohammad Hassan Behzadi
Orcid number: 0000-0002-1849-5588
Department of Statistics
SR.C., Islamic Azad University
Tehran, Iran
Email address: behzadi@srbiau.ac.ir

Sedigheh Zamani Mehreyan
Orcid number: 0000-0001-9953-8203
Department of Statistics
Imam Khomeini International University
Qazvin, Iran
Email address: s.zamani@sci.ikiu.ac.ir