VARIABLE IMPORTANCE FOR FIXED EFFECTS IN LINEAR MIXED MODEL

Yongzhe Wang, Lingbo Ye, Zifan Yu
Department of Biostatistics
University of Washington
Seattle, WA
{yzw1996, ylb10998, zfredyu}@uw.edu

ABSTRACT

Many scientific applications are of interest to evaluate the relative shares of influence of variables in a given model through the change in prediction values or metrics, namely to explore the variable importance for covariates. Researchers have proposed different approaches to investigate variable importance for cross-sectional data with parametric and non-parametric models already. However, this topic is less brought up in the context of longitudinal data. To tackle the problem, we introduced a variable importance measurement (VIM), invented by Lindeman, Merenda, and Gold, for fixed effects in the linear mixed effect model. To cooperate with the nature of cluster effects in longitudinal data, we used marginal and conditional R^2 to obtain the variable importance, which offered two interpretations of the VIM through the improvement of R^2 from the subject level and the population level. Meanwhile, it was robust for assessing contributions to fixed effects under the presence of multicollinearity. Throughout simulations, we showed that our proposed VIM for covariates matched the true rank of covariates in data generation process for simulating longitudinal data.

1 Introduction

Many scientific applications are of interest to evaluate the relative shares of influence of variables in a given model through the change in prediction values or metrics, namely to explore the variable importance for covariates [1, 2, 3]. Various approaches have been proposed to understand this problem in cross-sectional data during the past few decades [4, 5, 6, 7, 8] and the majority of these methods were focused on or had the aid of parametric models (e.g. linear regression). However, it was less mentioned in the literature on longitudinal data analysis. During the last few decades, the linear mixed model has been widely used in longitudinal and cluster data analysis within lots of biomedical and clinical studies [9, 10]. Different approaches such as regression coefficients with or without standardization, their corresponding p-values in t-test, and p-values in F-test or variance component test with the stepwise procedure have been widely applied to investigate intrinsic variable importance in linear regression. Since the linear mixed model is composed of a segment of fixed effects in a parametric form, variable importance measurement in linear regression is still applicable in a mixed model setting. However, variable importance measure based on ranking estimated standardized coefficients and their associated p-values in t-test is suffered from a problem of inaccuracy caused by multicollinearity, which theoretically brings up wider confidence intervals [11]. When multicollinearity occurs in a model, the moment matrix (e.g. X^TX or X^TWX) may not be inverted anymore and it leads to a very small

determinant in the moment matrix which brings up inflation in the estimated variance of coefficients [12, 13, 14, 15]. Since the estimated coefficients have been affected by multicollinearity, variable importance measure from standardized coefficients is inapplicable as well. In general, variable importance under multicollinearity has a problem of intrinsic ambiguity in the sense that variables "share" importance. These problems also exist in the context of the linear mixed model and several techniques have been introduced to diagnose the illness of multicollinearity [16]. One way to deal with multicollinearity and provide variable importance for covariates in a linear model is to consider the relative percentage contribution of each covariate, namely, considering the average difference in R^2 for each covariate X_i over all possible combinations of a set of covariates [5, 17]. This is the LMG measure of variable importance, first proposed by Lindeman et al. [5, 18] and expanaded by Genizi and Grömping [6, 19].

$$\phi_i = \frac{1}{p} \sum_{u \subseteq -\{i\}} \frac{1}{\binom{p-1}{|u|}} \left[R^2_{u+i} - R^2_u \right], \tag{1}$$

where p is the number of covariates, u is a subset that includes covariates from $\{X_1, X_2, ..., X_p\}$ excluded X_i, and $|u|$ is the cardinality of set u [18]. Since it provides a weighted average for each subset of covariates and sums up all improvements in R^2 over all combinations, even though multicollinearity appears among covariates, the contributions of correlated covariates will be average out. Our goal of the paper is to extend this method to a linear mixed model, which offers two advantages – providing variable importance under multicollinearity and the interpretation of covariates' importance with relative average percentage contributions regarding R^2.In (1), we can find that the coefficient of determination R^2 plays a vital role in LMG variable importance but there is no consensus for the definition of R^2 in the linear mixed model and different version of R^2 in the linear mixed model have been proposed with their corresponding merit and problem [20, 21, 22, 23]. In our paper, we followed two types of residual-based R^2 (marginal case considering variable importance in the population and conditional case considering variable importance n the sample of people in the dataset) proposed by Xu [21], due to their heuristic structure and similar interpretation as they are in linear regression, and provide more discussion on R^2 in the linear mixed model in later sections.

The layout of this paper is as follows. Section 2 is to provide notation and definition for linear mixed model, discuss R^2 in the linear mixed model, and propose our extension of LMG type of variable importance measure in the linear mixed model. Section 3 is used to describe our schema for simulation study and Section 4 is for the results from simulation study. Section 5 and section 6 will be the discussion and the conclusion.

2 Variable Importance in the Linear Mixed Model

2.1 Linear Mixed Model

We consider the linear mixed model [9] as given a subject i for all $i \in \{1, 2, ..., m\}$,

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \tag{2}$$

where m is the number of subjects, $\mathbf{y}_i \in \mathbb{R}^{n_i}$ is a column vector for a response variable, $\mathbf{X}_i$ is a $n_i \times (p+1)$ fixed-effects design matrix, $\mathbf{Z}_i$ is a $n_i \times q$ random-effects design matrix , $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ is a column vector for fixed-effects coefficients, $\mathbf{b}_i \in \mathbb{R}^q$ is a column vector for random-effects coefficients, and $\boldsymbol{\epsilon}_i \in \mathbb{R}^{n_i}$ is a column vector for within-subject errors. We also assume the normality for random effects and within-subject errors such that

$$\mathbf{b}_i \sim \mathcal{N}_q(\mathbf{0}, \mathbf{D}_i) \quad \text{and} \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \mathbf{R}_i),$$

where $\mathbf{D}_i$ is a $q \times q$ covariance matrix for the random effects and $\mathbf{R}_i$ is a $n_i \times n_i$ covariance matrix for within-subject errors. Without loss of generality, we assume the conditional independence for error terms $\mathbf{R}_i = \sigma^2_\epsilon \mathbf{I}_{n_i}$ where $\mathbf{I}_{n_i}$ is an identity matrix with dimension n_i as well.

We also can stack subject-level vectors and matrices and express the linear mixed model as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}, \tag{3}$$

where

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_m \end{bmatrix}_{m\times 1}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_m \end{bmatrix}_{m\times(p+1)}, \quad \mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{Z}_m \end{bmatrix}_{m\times m},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}_{(p+1)\times 1}, \quad \mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_m \end{bmatrix}_{m\times 1}, \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_m \end{bmatrix}_{m\times 1}.$$

In addition, the covariance matrix for $\mathbf{b}$ is $\mathbf{D} = \text{diag}(\mathbf{D}_1, \mathbf{D}_2, ..., \mathbf{D}_m)$ and the covariance matrix for $\boldsymbol{\epsilon}$ is $\mathbf{R} = \text{diag}(\mathbf{R}_1, \mathbf{R}_2, ..., \mathbf{R}_m)$. Consequentialy, we can obtain the marginal variance of $\mathbf{Y}$ as $\text{Var}(\mathbf{Y}) = \mathbf{V} = \mathbf{Z}\mathbf{D}\mathbf{Z}^T + \mathbf{R}$.

The parameters estimation in the linear mixed model (2) and (3) traditionally followed the maximum likelihood estimation (MLE) or restricted maximum likelihood estimation (REML) [24, 25]. The empirical Bayes estimate with EM algorithm was also feasible and equivalent for parameter estimate [9]. However, we also noticed that if REML was applied within the context of linear regression, the R^2 in linear regression will be adjusted R^2 [26] so we used MLE for parameter estimate in our simulation study. The partition of R^2 is held under the initial definition of R^2 instead of adjusted R^2.

2.2 Proposed $\mathbf{R^2}$ in Linear Mixed Model

There is a long history that using the coefficient of determination (R^2) as a metric to evaluate the performance of a regression model for cross-sectional data. However, there is no concurrence on the definition of R^2 for the linear mixed

effect model. This is partially because people less used it as a prediction model and different proposed definitions of R^2 share some practical and theoretical problems [20]. In ordinary linear regression, previous literature has proposed some criteria for assessing R^2 [27, 28] and it can be summarized as a dimensionless, unit-free, and comparable measurement of goodness-of-fit without correction of the degree of freedom and limitation of any specific model-fitting process. It provides interpretation as explaining the variation of the response variable ($\mathbf{Y}$) through explanatory variables ($\mathbf{X}$). In general, the definition of R^2 characterizes as the total percentage of variation in response variable minus the ratio of two sum of squared residuals in a fitted model and in a null model

$$R^2 = 1 - \frac{RSS}{RSS_0} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \hat{y}_i^{(0)})^2}, \tag{4}$$

where i is the index for subjects, $\hat{y}$ is the prediction from a fitted model, and $\hat{y}^{(0)}$ is a value from a specific null model. In linear regression, the null model is always described as

$$H_0 : y_i = \beta_0^{(0)} + \epsilon_i^0, \tag{5}$$

where it is just a model with intercept term. Hence, the $RSS_0 = \sum_i (y_i - \bar{y})^2$ is the marginal empirical variance of all subjects. Since (5) only includes the intercept from fixed effects in the linear mixed model, it tends to measure the portion of reduction in residual variation explained by a set of fixed effects. Compared with R^2 measured deviance from the population-level mean in (5), it is of interest to consider the null model that includes subject-level information and corresponding R^2 can explain a portion of the reduction in residual variation explained by the subject's response, so the null model can be treated as

$$H_0 : y_i = \beta_0 + b_{i0} + \epsilon_i, \tag{6}$$

where the null model only consists of random intercept and global intercept without any other covariates. Various ways of calculating R^2 in linear mixed model have been proposed [20, 21, 22, 29]. In general, they can be categorized into two types, likelihood-based and residual-based R^2. The likelihood-based R^2 was first introduced by Kent as a method to explain randomness in linear, generalized linear, and proportional hazard models [30, 31] but it suffered from a problem of decreased or even negative R^2 with the introduction of additional covariates [20]. Comparatively, the residual-based R^2 is more similar to the original definition in linear model with heuristic and intuitive interpretation. Following the definition of residual-based R^2 proposed by Xu [21], we use the marginal (corresponding to (5)) and conditional (corresponding to (6)) definition of R^2 to obtain the LMG variable importance, providing for people to choose different approaches to interpret variable based on their scientific questions.

2.3 LMG Variable Importance Measure in Linear Mixed Model

The LMG measure of variable importance, first proposed by Lindeman et al. [5] and expanaded by Genizi and Grömping [6, 19], can be expressed as

$$\phi_i = \frac{1}{p} \sum_{u \subseteq -\{i\}} \frac{1}{\binom{p-1}{|u|}} \left[R^2_{u+i} - R^2_u \right],$$

where p is the number of covariates, u is a subset that includes covariates from $\{X_1, X_2, ..., X_p\}$ but excludes X_i, $|u|$ is the cardinality of set u, and R^2_{u+i} and R^2_u are adpated from the marginal or conditional R^2 for fixed effects in Section 2.2. When the fixed effect X_i is excluded from model, its associated random effect, if there is one, will be excluded as well. We also normalized ϕ_i for each variable over all $i \in \{1, 2, ..., p\}$ to provide percentage version of LMG variable importance measure in linear mixed model

$$\Phi_i = \frac{\phi_i}{\sum_{j=1}^{p} \phi_j}. \tag{7}$$

The LMG variable importance measure considers offering weights for subsets of covariates (u), obtaining the improvements in R^2 at the moment that the i-th covariate plug into the regression with a subset of covariates from u, and averages out at all increments in R^2 for X_i over all subsets involved X_i. In the context of the linear mixed model, covariates occur in the fixed effects so this variable importance measure is designed to provide a rank for a set of fixed effects. As Cheng et al. mentioned [32], similar to the linear model, longitudinal data was affected by multicollinearity problem as well, which decreased the power of inference from t-test in estimated fixed effects. However, compared with methods affected by multicollinearity, the LMG variable importance can measure the variable importance for correlated fixed effects, since for two given correlated covariates X_i and X_j, the LMG method will go through all possible subsets of covariates that involve X_i or X_j separately. We can find that the LMG variable importance measurement provides a weighted average for each subset of covariates and sums up all improvements in R^2 over all combinations, though multicollinearity appears among covariates, the contributions of correlated covariates will be average out. Hence, the contribution of X_i or X_j in explaining the variation of the response variable is evaluated independently. Meanwhile, the averaged improvements in R^2 for a fixed effect represent an averaged contribution of it to the explained the variation of the response from populational level or person's level.

Besides the original definition of LMG variable importance, Owen [18] pointed out that the expression of ϕ_i coincided with the Shapley value in game theory [33, 34]. In cooperative game theory, when players purchase a service and create costs related to the purchase, the Shapley value was the only way fulfilled four compelling properties that allocated cost among the players. As Owen described [18], these four properties were

- Efficiency: $\sum_{i=1}^{p} \phi_i = val(1:p)$;
- Symmetry: if $val(u \cup \{i\}) = val(u \cup \{j\})$ for all $u \subseteq 1:p - \{i, j\}$, then $\phi_i = \phi_j$;
- Dummy: if $val(u \cup \{i\}) = val(u)$ for all $u \subseteq 1:p$, then $\phi_i = 0$;
- Linearity: if val_j and val'_j have Shapley value ϕ_j and ϕ'_j respectively, then $val_j + val'_j$ have Shapley value $\phi_j + \phi'_j$ for all $j \in 1:p$;

where val is the value function, often chosen R^2 or predicted value $\hat{\mathbf{Y}}$, for Shapley value. These properties were preserved in linear regression with a value function as R^2.

3 Data Generation for Simulation Study

We conducted a simulation study to estimate the LMG variable importance measurement for a linear mixed model and assess its performance through the comparison with p-values from t-test for estimated coefficients of fixed effects with and without multicollinearity appeared in fixed effects. In the case without multicollinearity, we generated a cluster of continuous outcomes with $p = 6$ covariates and random intercepts were always assumed in the model. We considered there were $m = 100, 200, 400$ subjects and each subject had $n = 5, 7, 9$ repeated measures. Since simulated covariates for each subject i ($i = 1, 2, ..., m$) was a cluster denoted as $\mathbf{X}_i = [\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{i6}]$ where $\mathbf{x}_{ik}$ was a column vector for a covarite k ($k = 1, 2, ..., p$), we assumed each element in $\mathbf{x}_{ik}$ was from Unif(0,1) and corresponding random effect matrix was $\mathbf{Z}_i$. Due to continuous covariates from the same distribution, we assumed $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6] = [1/2, 1/2, 1/2, 1/4, 1/8, 1/16, 1/32]$ which was under a decreasing order and represented for a true importance of covariates and two cases of random effects, a random intercept $b_{i0} \sim N(0, \theta)$ where $\theta = 1$ and a random intercept and a random slope for $\mathbf{x}_{i1}$ where $\mathbf{b}_i = [b_{i0}, b_{i1}] \sim \mathcal{N}_2(\mathbf{0}, \theta\mathbf{I}_2)$, were assumed for our simulation. Hence, the value of true response variable is highly determined by the top two covariates. We assumed an error vector for each cluster i as $\boldsymbol{\epsilon}_i \sim \mathcal{N}_n(\mathbf{0}, \mathbf{R}_i)$ where $\mathbf{R}_i = \sigma^2\mathbf{I}_n$ and $\sigma^2 = 1$. Hence, the continuous outcome $\mathbf{y}_i$ for each cluster was simulated from a multinormal distribution $\mathcal{N}_n(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i)$, where elements in the diagonal and the off-diagonal of $\boldsymbol{\Sigma}_i$ were $\sigma^2 + \theta$ and θ respectively (e.g. compound symmetry). In the case with multicollinearity, the simulation schema was similar to the above and except that we only assumed random intercepts existed. We assumed that two of covariates (X_1 and X_2) were highly associated with correlation $\rho = 0.99$ and $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6] = [1/2, 1/2, 1/2, 1/4, 1/4, 1/8, 1/8]$ where correlated covariates shared the same true coefficient. Each element in $\mathbf{x}_{ik}$ was generated from a normal distribution $N(0, 1)$.

4 Result

For each combination of m and n, we first fitted the response variable using all possible combinations of covariates (i.e. 2^p combinations including a null model and full models) under the model (3). We then obtained the marginal and conditional R^2 defined in Section 2.2 corresponding to (5) and (6) for each linear mixed model and calculated the LMG variable importance for each covariate. The p-value from t-test for each covariate in the fitted model with all covariates was recorded as well.

In the simulation setting, the true rank of variable importance was based on the magnitude of true $\boldsymbol{\beta}$. Table 1 provides mean and standard deviation of the estimated LMG variable importance for each covariate over 100 simulations under different combinations of the number of subjects and the number of repeated measures. The proposed variable importance measurement for the linear mixed model can capture the true rank of variable importance regardless of the size of the dataset. For the top two significant covariates (X_1 and X_2), there were only slight differences in the estimated LMG variable importance between using marginal R^2 and conditional R^2 proposed by Xu [21] and they contributed a large portion of weighted average improvements in R^2 compared with other covariates. The same simulation study

Table 1: 100 times simulation results for LMG variable importance from the linear mixed model with random intercepts using MLE

(a) Mean and standard deviation of LMG variable importance for each covariate over 100 times simulation

Simulation Scenarios		Covariates											
		X_1		X_2		X_3		X_4		X_5		X_6	
		R^2_m	R^2_c	R^2_m	R^2_c	R^2_m	R^2_c	R^2_m	R^2_c	R^2_m	R^2_c	R^2_m	R^2_c
m=100	n=5	0.46	0.46	0.18	0.18	0.10	0.10	0.09	0.09	0.09	0.09	0.08	0.09
		(0.21)	(0.21)	(0.09)	(0.09)	(0.05)	(0.05)	(0.05)	(0.06)	(0.05)	(0.05)	(0.04)	(0.04)
	n=7	0.52	0.52	0.16	0.16	0.09	0.09	0.08	0.08	0.08	0.08	0.08	0.08
		(0.15)	(0.15)	(0.07)	(0.07)	(0.04)	(0.04)	(0.03)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)
	n=9	0.56	0.56	0.15	0.15	0.09	0.09	0.07	0.07	0.07	0.07	0.07	0.07
		(0.12)	(0.12)	(0.06)	(0.06)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
m=200	n=5	0.54	0.54	0.16	0.16	0.09	0.09	0.07	0.07	0.07	0.07	0.07	0.07
		(0.16)	(0.16)	(0.07)	(0.07)	(0.04)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	n=7	0.56	0.56	0.15	0.15	0.09	0.09	0.07	0.07	0.07	0.07	0.06	0.06
		(0.12)	(0.12)	(0.06)	(0.06)	(0.04)	(0.04)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	n=9	0.56	0.56	0.17	0.17	0.08	0.08	0.07	0.07	0.06	0.06	0.06	0.06
		(0.10)	(0.10)	(0.05)	(0.05)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
m=400	n=5	0.57	0.57	0.16	0.16	0.08	0.08	0.06	0.06	0.06	0.06	0.06	0.06
		(0.12)	(0.12)	(0.06)	(0.06)	(0.04)	(0.04)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	n=7	0.59	0.59	0.16	0.16	0.08	0.08	0.06	0.06	0.06	0.06	0.06	0.06
		(0.11)	(0.11)	(0.05)	(0.05)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	n=9	0.58	0.58	0.16	0.16	0.08	0.08	0.06	0.06	0.06	0.06	0.06	0.06
		(0.09)	(0.09)	(0.04)	(0.04)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)

* R^2_m is the marginal R^2 and R^2_c is the conditional R^2.

(b) Proportion of p-value for each estimated fixed effect over 100 simulations that each fixed effect was significant at 0.05

Simulation Scenarios		Covariates					
		X_1	X_2	X_3	X_4	X_5	X_6
m=100	n=5	1	0.68	0.15	0.12	0.05	0.04
	n=7	1	0.80	0.22	0.12	0.07	0.06
	n=9	1	0.88	0.38	0.10	0.08	0.06
m=200	n=5	1	0.93	0.37	0.12	0.08	0.10
	n=7	1	0.94	0.43	0.13	0.09	0.05
	n=9	1	1	0.65	0.17	0.06	0.03
m=400	n=5	1	0.99	0.62	0.19	0.06	0.08
	n=7	1	1	0.72	0.17	0.09	0.08
	n=9	1	1	0.86	0.30	0.11	0.05

was also performed for the linear mixed model assumed two random effects (random intercept and random slope in X_1) and the results were also aligned with Table 1 (supplementary). However, the proportion of p-value that was significant at 0.05 for covariate X_2 over 100 simulations shown in Table 1(b) was affected by the size of the dataset and a small dataset size yielded a lower proportion of significant p-value. The performance of our proposed method was reasonable in different simulation scenarios, and only when the sample size and cluster size were small, the performance of the proposed method showed larger standard deviation for estimated. The residual-based R^2 in the linear mixed model required $n_i \to \infty$ to provide accurate estimation, namely the empirical estimated distribution of random effects converges to its true distribution [31, 35]. We noticed that in the simulation study where we assumed random intercept and random slope for X_1, the estimated variable importance from models using REML resulted in negative values for unimportant covariates (supplementary). We investigated the problem and found that the R^2 proposed by Xu [21] returned negative values when the distribution assumption for random effect was inappropriate (e.g. Hessian matrix for random effects was not positive definite). In general, the performance of the proposed variable importance worked better in the linear mixed model with MLE than REML.

Table 2 is the simulation study where multicollinearity existed. Since X_1 and X_2 were highly correlated and shared the same true β, the simulation study disclosed that two correlated covariates in the linear mixed model had indistinguishable

Table 2: 100 times simulation results for LMG variable importance from the linear mixed model with random intercepts using MLE, where linear dependence existed between X_1 and X_2

(a) Mean and standard deviation of LMG variable importance for each covariate over 100 times simulation

Simulation Scenarios		Covariates											
		X_1		X_2		X_3		X_4		X_5		X_6	
		R^2_m	R^2_c	R^2_m	R^2_c	R^2_m	R^2_c	R^2_m	R^2_c	R^2_m	R^2_c	R^2_m	R^2_c
m=100	n=5	0.38	0.38	0.38	0.38	0.06	0.06	0.07	0.07	0.05	0.05	0.05	0.05
		(0.04)	(0.04)	(0.04)	(0.04)	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
	n=7	0.38	0.38	0.38	0.38	0.07	0.07	0.07	0.07	0.05	0.05	0.05	0.05
		(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
	n=9	0.38	0.38	0.38	0.38	0.07	0.07	0.07	0.07	0.05	0.05	0.05	0.05
		(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
m=200	n=5	0.39	0.39	0.38	0.39	0.07	0.07	0.07	0.07	0.05	0.05	0.05	0.05
		(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
	n=7	0.39	0.39	0.39	0.39	0.07	0.07	0.06	0.06	0.05	0.05	0.05	0.05
		(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
	n=9	0.39	0.39	0.39	0.39	0.06	0.06	0.06	0.06	0.05	0.05	0.05	0.05
		(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
m=400	n=5	0.39	0.39	0.39	0.39	0.07	0.07	0.07	0.07	0.05	0.05	0.05	0.05
		(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
	n=7	0.39	0.39	0.39	0.39	0.06	0.06	0.06	0.06	0.05	0.05	0.05	0.05
		(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
	n=9	0.39	0.39	0.39	0.39	0.06	0.06	0.06	0.06	0.05	0.05	0.05	0.05
		(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)

* R^2_m is the marginal R^2 and R^2_c is the conditional R^2.

(b) Proportion of p-value for each estimated fixed effect over 100 simulations that each fixed effect was significant at 0.05

Simulation Scenarios		Covariates					
		X_1	X_2	X_3	X_4	X_5	X_6
m=100	n=5	0.24	0.13	0.83	0.82	0.34	0.31
	n=7	0.21	0.21	0.96	0.94	0.42	0.40
	n=9	0.21	0.35	1.00	0.97	0.62	0.63
m=200	n=5	0.30	0.31	0.99	0.98	0.57	0.55
	n=7	0.42	0.39	1	1	0.67	0.66
	n=9	0.53	0.50	1	1	0.88	0.82
m=400	n=5	0.50	0.54	1	1	0.79	0.81
	n=7	0.65	0.75	1	1	0.93	0.92
	n=9	0.82	0.74	1	1	0.98	0.99

estimated mean LMG variable importance calculated by either marginal or conditional R^2 through all simulation scenarios over 100 simulations, which provided a valid analysis to understand the importance of covariates under a high degree of multicollinearity. However, the proportion of significant p-values over 100 simulations for two covariates X_1 and X_2 with a large magnitude, affected by multicollinearity, was lower than other covariates with a small magnitude, which indicated the unstable performance of t-test on estimated coefficients in the linear mixed model.

Through simulation studies with and without multicollinearity, we noticed that the proportion of p-value that was significant at 0.05 for a covariate which was significant in reality was sensitive to the size of a dataset and large size dataset tended to have a higher proportion of p-value that was significant at 0.05 for a covariate which was significant in reality. But our proposed variable importance measurement was comparatively stable and did not vary notably along with the increment of dataset size. Besides the above simulation studies, we also checked that efficiency and linearity properties in the Shapley value didn't hold in the context of the linear mixed model under the R^2 proposed by with null model as (5) or (6). Even though the LMG variable importance for the linear mixed model didn't fulfill the properties of the Shapley value, the practical use of the LMG variable importance measure was still reasonable regarding its advantages in dealing with multicollinearity and providing an interpretation of explaining variation in response variable under the linear mixed model.

5 Discussion

The variable importance measure in the linear mixed effect model is more complicated than linear regression. Differed from the variable selection procedure in the linear mixed model which was largely depended on penalized likelihood [36], the variable importance measure did not aim to select a subset of covariates from a set of them and rather provided a rank for covariates' contributions to explain variation in response from covariates. This procedure highly depended on how users chose quantitative metrics to define covariates' contributions. In our study, we used the coefficient of determination (R^2) to measure the contribution from a covariate. Two residual-based R^2 (marginal and conditional) we used to calculate the proposed variable importance measurement were different conceptually but we only saw a slight difference in practice. We have seen that different definitions of R^2 have been proposed for the linear mixed model and we only used the one defined by Xu [31] due to its heuristic interpretation and similarity to the original R^2 in the linear model. Therefore, one can consider using alternative definitions of R^2, collaborated with the procedure of the LMG variable importance measure that we proposed, to investigate the variable importance in the linear mixed model. The choice of R^2 statistics directly affected the performance of LMG variable importance measure. Orliean and Nakagawa [20, 23] have investigated the performance of different types of R^2 in the linear mixed model under different simulation studies and pointed out that comparatively poor performance of R^2 defined by Xu. In our simulation (supplementary), we also notified that the marginal (5) and conditional (6) R^2 defined by Xu behaved unstably and resulted in negative R^2 which indicated that a fitted linear mixed model performed even worse than the null model. We tried to understand the problem related to R^2 in the linear mixed model and noticed that in the linear regression, the $R^2 = 1 - RSS/SST$ was based on the partition of the sum of squared terms

$$SST = SS_{reg} + RSS \implies ||\mathbf{Y} - \bar{\mathbf{Y}}\mathbf{1}||^2 = ||\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\mathbf{1}||^2 + ||\hat{\mathbf{Y}} - \mathbf{Y}||^2, \tag{8}$$

where $\mathbf{Y}$, $\mathbf{1}$, and $\hat{\mathbf{Y}}$ (fitted values from linear regression) are column vector with length n and $\bar{\mathbf{Y}}$ is the mean of elements in $\mathbf{Y}$, but this partition of the sum of squared terms didn't hold in the linear mixed model due to the projection matrix $\mathbf{P}$ for $\boldsymbol{\beta}$ in the model (3)

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1} \tag{9}$$

was not symmetric and idempotent. This caused the partition of the sum of squared terms in the linear mixed model to be not unique. Hodges [37] has provided a treatment to establish a symmetric and idempotent projection matrix in the linear mixed model, but we noticed that the LMG variable importance measure would rank not only fixed effects but also random effects associated with clusters under Hodges's setting. Our goal is to better understand the intrinsic variable importance for fixed effects and therefore the variable importance for random effects is out of our scope. Since the LMG variable importance measure mainly depended on the metrics to evaluate the contribution of each fixed effect, to find a better and suitable metrics for the linear mixed model deserves to gauge in the future on its own and beyond the scope of our method.

Meanwhile, the restricted maximum likelihood estimation (REML) was the primary choice for the linear mixed model (3) due to its reduction of bias in estimating variance components compared with maximum likelihood estimation (MLE). We conducted the same simulation studies shown in Section 3 with REML (supplementary) and the result from REML was consistent with results in Section 4. We also noticed that the REML in linear regression provided adjusted R^2 for models if we used the definition of R^2 in (4) [26].

The calculation of the LMG variable importance measure was involved in building 2^p models where p was the number of covariates. Although this variable importance measure provided a conceptual method to understand covariates' average percentage contributions, the computational problems continued to be an obstacle especially when the dimension p was large. In our simulation study, we only tried a limited amount of covariates but the linear mixed model with a huge amount of covarites is worthwhile to investigate. Hence, the computation of LMG variable importance measure for linear mixed model deserves to investigate separately in future.

6 Conclusion

In our paper, we extended the LMG variable importance measure from the linear regression to the linear mixed model with a marginal and conditional coefficient of determination R^2. They showed reasonable performance in simulation studies and provided a new approach to investigating variable importance from the linear mixed model.

References

[1] Andrew J Dunning. A model for immunological correlates of protection. *Statistics in medicine*, 25(9):1485–1497, 2006.

[2] Miriam Sindelar, Ethan Stancliffe, Michaela Schwaiger-Haber, Dhanalakshmi S Anbukumar, Kayla Adkins-Travis, Charles W Goss, Jane A O'Halloran, Philip A Mudd, Wen-Chun Liu, Randy A Albrecht, et al. Longitudinal metabolomics of human plasma reveals prognostic markers of covid-19 disease severity. *Cell Reports Medicine*, 2(8):100369, 2021.

[3] Zeeshan Ahmed, Saman Zeeshan, David J Foran, Lawrence C Kleinman, Fredric E Wondisford, and XinQi Dong. Integrative clinical, genomics and metabolomics data analysis for mainstream precision medicine to investigate covid-19. *BMJ innovations*, 7(1), 2021.

[4] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[5] Richard Harold Lindeman. Introduction to bivariate and multivariate analysis. Technical report, 1980.

[6] Ulrike Grömping. Estimators of relative importance in linear regression based on variance decomposition. *The American Statistician*, 61(2):139–147, 2007.

[7] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[8] Laura L Nathans, Frederick L Oswald, and Kim Nimon. Interpreting multiple linear regression: a guidebook of variable importance. *Practical assessment, research & evaluation*, 17(9):n9, 2012.

[9] Nan M Laird and James H Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.

[10] Peter J Diggle, Patrick Heagerty, Kung-Yee Liang, and Scott Zeger. *Analysis of longitudinal data*. Oxford university press, 2002.

[11] Stan Lipovetsky and Michael Conklin. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17(4):319–330, 2001.

[12] Charlotte H Mason and William D Perreault Jr. Collinearity, power, and interpretation of multiple regression analysis. *Journal of marketing research*, 28(3):268–280, 1991.

[13] Carsten F Dormann, Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R García Marquéz, Bernd Gruber, Bruno Lafourcade, Pedro J Leitão, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1):27–46, 2013.

[14] John Hannah. A geometric approach to determinants. *The American mathematical monthly*, 103(5):401–409, 1996.

[15] George AF Seber and Alan J Lee. *Linear regression analysis*. John Wiley & Sons, 2012.

[16] Sandra Sue Stinnett. *Collinearity in mixed models*. The University of North Carolina at Chapel Hill, 1993.

[17] Ulrike Grömping. Variable importance assessment in regression: linear regression versus random forest. *The American Statistician*, 63(4):308–319, 2009.

[18] Art B Owen and Clémentine Prieur. On shapley value for measuring importance of dependent inputs. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):986–1002, 2017.

[19] Abraham Genizi. Decomposition of r 2 in multiple regression with correlated regressors. *Statistica Sinica*, pages 407–420, 1993.

[20] Shinichi Nakagawa and Holger Schielzeth. A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in ecology and evolution*, 4(2):133–142, 2013.

[21] Ronghui Xu. Measuring explained variation in linear mixed effects models. *Statistics in medicine*, 22(22):3527–3541, 2003.

[22] Lloyd J Edwards, Keith E Muller, Russell D Wolfinger, Bahjat F Qaqish, and Oliver Schabenberger. An r2 statistic for fixed effects in the linear mixed model. *Statistics in medicine*, 27(29):6137–6157, 2008.

[23] Jean G Orelien and Lloyd J Edwards. Fixed-effect variable selection in linear mixed models using r2 statistics. *Computational Statistics & Data Analysis*, 52(4):1896–1907, 2008.

[24] James H Ware. Linear models for the analysis of longitudinal studies. *The American Statistician*, 39(2):95–101, 1985.

[25] David A Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American statistical association*, 72(358):320–338, 1977.

[26] Norman R Draper and Harry Smith. *Applied regression analysis*, volume 326. John Wiley & Sons, 1998.

[27] Tarald O Kvålseth. Cautionary note about r 2. *The American Statistician*, 39(4):279–285, 1985.

[28] A Colin Cameron and Frank AG Windmeijer. R-squared measures for count data regression models with applications to health-care utilization. *Journal of Business & Economic Statistics*, 14(2):209–220, 1996.

[29] Honghu Liu, Yan Zheng, and Jie Shen. Goodness-of-fit measures of r 2 for repeated measures mixed effect models. *Journal of Applied Statistics*, 35(10):1081–1092, 2008.

[30] John T Kent. Information gain and a general measure of correlation. *Biometrika*, 70(1):163–173, 1983.

[31] Ronghui Xu and John O'quigley. Ar 2 type measure of dependence for proportional hazards models. *Journal of Nonparametric Statistics*, 12(1):83–107, 1999.

[32] Jing Cheng, Lloyd J Edwards, Mildred M Maldonado-Molina, Kelli A Komro, and Keith E Muller. Real longitudinal data analysis for real people: building a good enough mixed model. *Statistics in medicine*, 29(4):504–520, 2010.

[33] L. S. Shapley. *17. A Value for n-Person Games*, pages 307–318. Princeton University Press, 2016.

[34] Eric Friedman and Herve Moulin. Three methods to share joint costs or surplus. *Journal of economic Theory*, 87(2):275–312, 1999.

[35] Jiming Jiang. Goodness-of-fit tests for mixed model diagnostics. *The Annals of Statistics*, 29(4):1137–1164, 2001.

[36] Yingying Fan and Runze Li. Variable selection in linear mixed effects models. *Annals of statistics*, 40(4):2043, 2012.

[37] James S Hodges. Some algebra and geometry for hierarchical models, applied to diagnostics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3):497–536, 1998.