**Project 2: Linear mixed model with non-normal distribution of random components**

Wenru Zhou

**Introduction**

Mixed model is very popular in dealing with correlated data. It provides a representation of among-individual variation by including random effects. Normality of random effects is an important assumption of mixed model, however, this assumption does not always hold in realistic situations, especially sometimes the important features of among-individual variation is unknown[1]. We investigated the influence of non-normal random components on the inference of fixed effects assuming the random components are normally distributed in linear mixed model. Our goal is to check the performance of REML estimator of fixed effects including the estimation, standard deviation, and standard variance, as well as type I and type II error.

**The Model Notation**

Suppose, the response for subject i at time j. is continuous. Then the linear mixed effect can be expressed as:

is a vector representing the coefficient of fixed effects. is a vector representing the fixed effects. is a vector representing the coefficient of random effects. is a vector representing the random effects. Usually we have the following assumption: , in which G is covariance matrix for random effects. , in which R is covariance matrix for errors.

Therefore, we know that , where is a matrix, and . The combination of restricted maximum likelihood method and numerical method (usually Newton Raphson) is common to make inference of fixed effects and random effects.

However, the random effect is not always normally distributed. Some extreme measurements might happen on some subjects due to obscure reasons, which leads to a skewed distribution of random intercept. The effect of the continuous variable might have extremely strong or weak effect on some of the subjects to let their individual slope very different from others. In addition, subjects might also have mixture normal distribution because they are actually from different groups or belong to different classifications, which we might never know or be interested in.

**Simulation**

Our goal is to evaluate how the maximum likelihood estimators of fixed effect will behave when random components have skewed distribution and mixture distribution. We would like to evaluate the estimated coefficients, bias, and the corresponding standard error, standard deviation estimators for fixed effects using random intercept model as well as random intercept and slope model. Type I and II error of the linear mixed model would also be summarized with significant level 0.05.

In simulation datasets, subjects are equally allocated into a treatment group and a control group. Several repeated measurements are made on each subject. We analyzed these data with linear mixed models with random intercept, random intercept and slopes, and with group, time and group×time as fixed effects as follows:

1. Random intercept model:

(2) Random intercept and slope model:

is the outcome for subject i at time j. Time

, , , : fixed intercept and slope

: Random intercept. : Random slope for time

We set three settings of subject sizes i: 20, 200, 1000 and two settings of even spaced time points j: 5 time points and 10 time points. We set zero and non-zero coefficients. The non-zero coefficients of fixed effects are set to be 2, 0.5, 1, 0.25, which representing intercept, time, and group and interaction. The time is treated as continuous. The group is a dummy variable, which has two levels.

We set two kinds of distributions rather than normal distribution for random components: a lognormal distribution, which represents a skewed situation, and symmetric mixture distribution, which is a combination of two different normal distribution. All the variance of random intercept are set to be 3, and all the variance of random time are set to be 2. The correlation between the random intercept and the random slope is set to be 0. The error is normally distributed with mean equals 0 and variance equals 3.

The mean and variance of log normal were generated from by the following transformation:

The mean and variance of mixture normal distribution were generated from distribution and by the following transformation:

, , where and represent contribution that each normal distribution gives to the final distribution. The distribution of each random intercept and slop can be find in figure 1

The replicates of datasets in each scenario, which is the simulation number, is set to be 500. To better compare the results, the random part and error are simulated for each simulation. But within each simulation, we did not change the seed. This details of simulation settings can be found in table 1.

Non-convergence is a common problem in mixed model. We used Nelder–Mead, quasi-Newton and conjugate-gradient algorithms to let models converge, which has the best performance in convergence.

**Results**

The simulation results for N=20, 200, 1000 are summarized in Table 2.1, 2.2 and 2.3.

As the increase of the number of size of the subject, the standard error, bias, and type II error of coefficients decrease. When the number of subjects equals to 20, the bias is at the level of 0.01. When the number increases to 200, many of the bias is at the level of 0.001. When the number increases to 1000, many biases has been limited to 10E-4, and the estimated coefficients are close to the simulation setting. This indicated that, as the number observations increases, the REML estimator is unbiased in this study, even though the random effect is not normally distributed.

The type II error also decrease fast as the size of subjects increases. When the number of subjects changes from 20 to 1000, the type II error decreases from around 0.8 to almost 0. We can conclude that when the sample size is small, it is likely to miss the significance. The only exception is the intercept. Even when the number of subjects=20, the type II error for intercept is around 0.09. This happens because the overall mean of the model apparently deviates from zero, and detect a larger difference required smaller sample size while holding type I/II error constant.

Especially, the type II error of interaction between time and group is higher than the type II error of other effects when we added random intercept and slope parts, which means that we are more likely to assert that the interaction is not significant, while it is actually significant. This indicated that when we added a noise to the interaction, the difference of the time slopes between the two groups becomes smaller for detection compared to the difference between main effects and zero.

The type I error did not decrease much with the increase in the number of subjects. When the number of subjects=20, it is around 0.05~0.07. When the number of subjects=1000, it is around 0.03~0.05. It is reasonable because the common way to minimize type I error is to set a more conservative significance level, in our case, can be set as 0.01.

As time increase, for random intercept model both random intercept and slope model, the standard deviation and error of time decrease. This indicated that with more time points, the estimator of time becomes more precious. The bias of time also decreased from 0.01 level to 0.001 level (N=20 and N=200) as time points increase from 5 to 10. When N=1000, because when the number of replicates is very large, even with only 5 points we can have a smaller bias at 0.01 level. Therefore, based on this study, the estimation of coefficients of time is not influenced by non-normal random effects especially when the number of observations is large.

As for random intercept model, the coefficient of time and interaction did not change between two different distributions. This happens because only a random intercept is added to the model. The random intercepts are different by subjects, such that related to treatment groups, while a random intercept was not related to time effect at all. The interaction between time and treatment also did not change, which means that the interaction of time and group is not affected even though we added a random intercept. The standard deviation and error of time and interaction are lower than intercept and treatment. Those did not happen in random intercept and time model, where we added a random time slope for each subject.

In each scenario, between the coefficients equal to zero and not equal to zero situation, for each coefficient, their standard deviation, standard error, and bias are the same. This happens because we keep the random components and error the same for those two situations, so the only thing that changes is the setting of coefficients.

**Future work**

This study evaluated the performance of mixed model for inference of fixed effect when random components are skewed or mixture distribution when the outcome is continuous and normally distributed. It is worthy to investigate other types of outcomes (binomial, Poisson or gamma distribution, etc.) and extended linear mixed model to generalized linear mixed model. In this study, the lognormal is not very skewed, it would be interesting to test a more skewed distribution, for example, exponential distribution, or asymmetric mixture distribution to evaluate the performance of REML estimator of fixed coefficients. In addition, we want to repeat this study on more subjects, and then plot bias and coefficients versus the number of subjects to check the efficiency and consistency of the REML estimator. We only calculated the standard error of coefficients to check efficiency, further, the standard error of bias is also needed to check consistency.

**Reference**

[1] **Daowen Zhang\* and Marie Davidian,** Linear Mixed Models with Flexible Distributions of Random Effects for Longitudinal Data

**Reproducible research information**

Link to Code

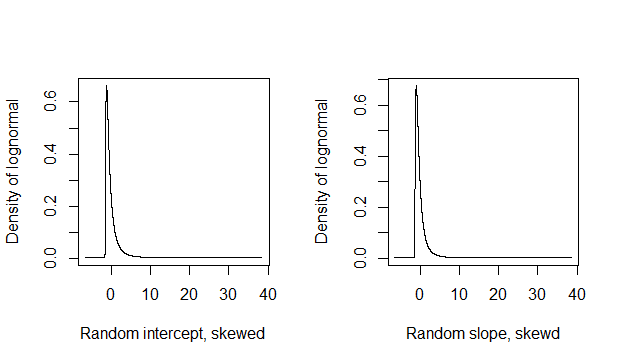
**https://github.com/BIOS6624-UCD/bios6624-zhwr7125/tree/master/Project2**

**Tables and Figures**

**Table1: Simulation settings**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Simulation settings (simulation number =500) | | | | |
| **Items** | **Values** | | | |
| Subject size | 20 | 200 | | 1000 |
| time points | 0~4, 0~9 | | | |
| Sample size | 100, 200 | 1000, 2000 | | 5000, 10000 |
| Intercept | 2 | | 0 | |
| Time | 0.5 | | 0 | |
| Treatment | 1 | | 0 | |
| Treatment\*time | 0.25 | | 0 | |
| Error | N(0.3) | | | |
| Random components | Skewd:  ~    Skewd:  ~,  Mixture:  ~  Mixture:  ~,, | | | |

\* i represents each subjects

****

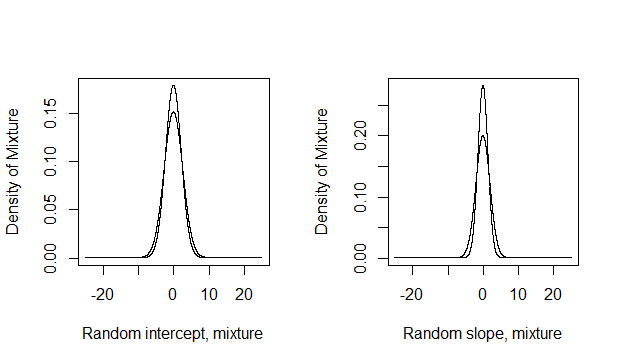
****

Figure 1: Different distribution of random intercept and slope

