Exploring the Finite-sample Behavior of Residual Diagnostics for Linear Mixed-effects Models

Yicheng Shen sheny2@carleton.edu
Department of Mathematics and Statistics, Carleton College
and
Yucheng Yang yangy2@carleton.edu [*]
Department of Mathematics and Statistics, Carleton College

December 19, 2021

Abstract

Understanding the residual behaviors has always been a key challenge for analysts who need to examine the validity of fitted models. In the framework of linear mixed-effects (LME) models, we employ simulation-based methods that provide an abundant set of artificial data and models with and without deficiencies to explore the finite-sample behavior of residual diagnostics. The results of our simulations point to the intertwined nature of LME model assumptions: a single or a pair of misspecifications often lead to structures in residual plots that could be flagged as problematic to other assumptions. The way in which the hierarchical data was composed, namely cluster sizes, residual variances, and longitudinal settings, are all influential to the residual diagnostics. These findings have major implications for the LME model checking procedures using residual analysis and can serve as guidelines for interpreting residual diagnostic plots and test statistics.

Keywords: linear mixed-effects models, simulation study, distributional assumptions, hierarchical linear models, residual diagnostics

[*] This research was supported by the Towsley Endowment for the Sciences.

1 Introduction

Linear mixed-effects (LME) models, which account for the interdependence of observations that arise from a hierarchical data structure, allow for the analysis of clustered data in a wide range of settings, such as agricultural and ecological experiments, longitudinal studies, and educational assessments. As with all model-based methods, the validity of model assumptions should always be carefully checked to ensure that the fitted model adequately represents the realities that exist in the data. While the existing diagnostic procedures (Singer et al. 2017) appear to work well in many situations (Schützenmeister & Piepho 2012), evidence suggests that these conventional tools are inadequate if certain structures appear in the data (Loy & Hofmann 2015). For example, the structure of residual plots under unbalanced group sizes can induce unusual patterns in residual plots when the fitted model is in fact adequate for analysis(Morrell & Brant 2000).

In this paper, we present a simulation study exploring the finite-sample behaviors of residual diagnostics for LME models. In Section 2, we discuss the backgrounds of LME model specification, model assumptions and proposed diagnostics. In Section 3, we outline our simulation study, from constructing data sets and fitting models, to extracting residuals and conducting analysis. The results of our study are presented in Section 4, where we explore diagnostic tools for detecting assumption violations under ten different scenarios along with interpretations of residual behaviors. Patterns and findings are generalized and discussed in Section 5.

The results of our study indicate that most LME model assumptions for residuals are entwined and interdependent: when one or more distributional assumption is misspecified in the data, practitioners are likely to observe patterns that suggest the breakdowns of other model assumptions in residual plots. The distribution of group sizes and the relative magnitude of the variability of error terms and random effects could also influence the frequency of violations seen in residual diagnostics: having unbalanced cluster sizes can exaggerate patterns of non-normality, or mitigate the heteroscedasticity of residual quantities, whereas the residual variance plays an important role in inducing problematic structures in residual plots. Notable exceptions are the violation of p normal assumption of random effects and the bimodality of errors, both of which have milder impacts on the validity of LME model assumptions. These results have far-reaching implications for LME model checking using residual plots and could serve as future guidelines for properly evaluating the diagnostic plots and test statistics.

2 Linear mixed-effects models

With the advancement of sophisticated data collection and processing tools, multiple-level and non-independent data structures have become the norm in many domains, particularly the biological and social sciences (Bolker 2008, Raudenbush & Bryk 2002). Neglecting to take this complex correlation into account could lead to underestimating standard errors of coefficients, overstating the significance of predictors and generating biased estimates (Gurka et al. 2011, Roback & Legler 2021). LME models allow us to appropriately model these hierarchical structures in clustered data.

2.1 Model Specification

A standard LME model for n observations nested in m groups is given by

$$\mathbf{y_i} = \mathbf{X_i}\boldsymbol{\beta} + \mathbf{Z_i}\mathbf{b_i} + \boldsymbol{\epsilon_i} \qquad \text{for i = 1, 2, 3, ... , m} \tag{1}$$

where $\mathbf{y_i}$ is a $n \times 1$ response vector; $\mathbf{X_i}$ is a $n \times p$ matrix representing p predictors; $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed-effect coefficients. $\mathbf{Z_i}\mathbf{b_i} + \epsilon_\mathrm{i}$ stands for the random components in the data structure: $\mathbf{Z_i}$ is a $n \times q$ model matrix for the q number of random effects in $\mathbf{b_i}$, and $\boldsymbol{\epsilon_i}$ is a $n \times 1$ vector of within-group measurement errors. Both random effects, $\mathbf{b_i}$, and residual errors, $\boldsymbol{\epsilon_i}$, are independent random variables, and we assume that

$$\boldsymbol{b_1, ...b_i} \overset{iid}{\sim} N_q(\mathbf{0}, \sigma^2\boldsymbol{G}) \;\text{ and }\; \boldsymbol{\epsilon_i} \overset{iid}{\sim} N_n(0, \sigma^2\boldsymbol{R_i}) \qquad \text{for i = 1, 2, 3, ... , m} \tag{2}$$

where $\mathbf{G}$ and $\mathbf{R_i}$ are q and n dimensional positive definite matrices, with elements expressed as functions of a vector of covariance parameters (Laird & Ware 1982).

LME models are featured by incorporating both fixed and random effects within a hierarchical structure (Bates & Pinheiro 1998). The fixed effects in LME models are regression coefficients that correspond to fixed quantities of interest. The random effects in LME models represent the correlation of intercepts and slopes in hierarchical data, adjusting for non-independence between observations and group-specific profiles.

To model both the mean and covariance structures, estimates must be obtained for both fixed effects and variance components. One possible approach is to maximize the likelihood function (McCulloch 1997). Due to the bias of maximum likelihood (ML) estimates for the variance components, restricted maximum likelihood (REML) has been more widely adopted (Gilmour et al. 1995). Regardless of the procedures used, reliable estimation must be based upon models that correctly depict the structure of data. Verbeke & Lesaffre (1996) illustrated that wrongly assuming normality of random effects leads to incorrect estimates of random effects coefficients. The misspecification of random effects can cause substantial bias of estimations of regression coefficients from LME models (Heagerty & Kurland 2001, Hui et al. 2021).

2.2 Model Diagnostics

Fitting correctly specified multilevel models that converge is a crucial step of formulating a solid and reliable inferences for statisticians (Gelman & Hill 2006). A broad set of model checking guidelines and techniques have been developed and implemented in statistical software over the years for inspecting the appropriateness of model assumptions, including residual normality, linearity and homoscedasticity (Pinheiro & Bates 2000, Cheng et al. 2010, Bates et al. 2015).

Residual analysis for LME models requires three types of residual quantities generated from the fitted models (Santos & Singer 2007, Haslett & Hayes 1998):

- The *marginal residuals* estimate the model's random components, $\boldsymbol{Z_i b_i} + \boldsymbol{\epsilon_i}$. They are defined by $\hat{\boldsymbol{\delta}}_{\boldsymbol{i}} = \boldsymbol{y_i} - \hat{\boldsymbol{E}}(\boldsymbol{y_i}) = \boldsymbol{Z_i}\hat{\boldsymbol{b_i}} + \hat{\boldsymbol{\epsilon}}_{\boldsymbol{i}}$. Here $\hat{\mathbf{E}}(\mathbf{y_i})$ is the best linear unbiased estimates (BLUEs) using all data.

- The *conditional residuals* are the residual deviations that estimate the model's error term, $\boldsymbol{\epsilon}_i$. They are defined by $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{i}} = \boldsymbol{y_i} - \hat{\boldsymbol{E}}(\boldsymbol{y_i}|\boldsymbol{b_i}) = \boldsymbol{y_i} - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}_{\boldsymbol{i}} - \boldsymbol{Z_i}\hat{\boldsymbol{b_i}}$.

- The *predicted random effects residuals*, $\mathbf{Z_i}\hat{\mathbf{b}}_{\mathbf{i}}$, estimate the model's random effects, $\mathbf{Z_i b_i}$. They are the best linear unbiased predictors (BLUPs) given by $\hat{\boldsymbol{E}}(\boldsymbol{y_i}|\boldsymbol{b_i}) - \hat{\boldsymbol{E}}(\boldsymbol{y_i})$.

In our simulation study, we standardized residuals for model diagnostics. We employed Cholesky residuals as standardized marginal residuals (Houseman et al. 2004). The original marginal covariance matrix and its Cholesky decomposition as $L(y)$, which is defined as $V(y)^{-1} = L(y)L(y)^\top$. Then the Cholesky residual is defined as $\delta_i^* = L(y)^\top \delta_i$. Santos & Singer (2007) suggested standardizing the raw conditional residuals, $\hat{e}_i$, by its $\hat{\sigma}$. The standardized conditional residuals are thus given by $\hat{e_i}^* = \frac{\hat{e_i}}{\hat{\sigma}\sqrt{\hat{p}_{kk}}}$, where the elements $\hat{p}_{kk}$ is the functions of the joint leverage of the fixed and random effects.

Snijders & Berkhof (2008) discussed the diagnostics for two-level hierarchical linear mixed-effects models, highlighting the important role of residual analysis and deriving various properties of the residuals at both the individual and cluster levels. They suggest that the main use of the marginal residuals is to investigate the specification of within-cluster linearity and homoscedasticity. The presence of outliers and potential effects of omitted variables can also be studied using the marginal residuals. The conditional and random effects residuals, on the other hand, can be plotted to check their normality assumptions.

Singer et al. (2017) further generalized the usage of visual diagnostic tools: plotting standardized marginal and conditional residuals versus the explanatory variables or fitted values is the standard approach for LME model diagnostics of linearity, normality, independence, and heteroscedasticity of errors. The normality assumptions of random effects are measured by the Mahalanobis Distance (MD) (Mahalanobis 1936, McLachlan 1999): if random effects follow multivariate normal distribution, their MD should follow a chi-squared distribution with q degrees of freedom, where q denotes the number of random effects.

However, residual methods that work well in linear regression with i.i.d errors have been less successful in LME model diagnostics because the empirical distribution of residuals does not necessarily converge to the true distribution of the errors (Jiang 1998). In practice, conventional residual plots and tests for model validation perform poorly in finite sample situations (Loy et al. 2017). The assessment of the normality assumption by interpreting conventional quantile–quantile (QQ) plots suffers from inflated Type I error rates (Loy & Hofmann 2015). Schützenmeister & Piepho (2012) argued that another problem for QQ plots is the difficulty of assessing whether the curvature in the plotted residuals is indicative of a departure from normality or whether there are possible outliers. Moreover, if a cluster has exactly one observation, the plot of any estimated random effect against any other estimated random effect will fall on a straight line (Morrell & Brant 2000), requiring researchers to take extra cautions when evaluating heteroscedasticity in the set of repeated-measures or longitudinal data.

3 Simulation Study

To address the inadequacy of the scope of existing literature on this topic, we propose a simulation-based approach to study residual diagnostic of LME models under finite samples.

3.1 Baseline Data Generation

In this section, we introduce how we simulated data from a model with two fixed effects and two random effects. Fitting these data with properly specified models serves as the baseline for residual performance for two-level LMEs.

$$Y_{i,j} = \underbrace{\beta_0 + \beta_1 X_{i,j} + \beta_2 Z_{i,j}}_{\text{Fixed Effects}} + \underbrace{u_i + v_i X_{i,j}}_{\text{Random Effects}} + \underbrace{\epsilon_{i,j}}_{\text{Error}} \quad \text{for } i \in \{1:N\} \ \& \ j \in \{1:n\} \tag{3}$$

where $\epsilon_{i,j} \overset{iid}{\sim} N(0, \sigma^2)$ and $\begin{bmatrix} u_i \\ v_i \end{bmatrix} \sim \text{MVNorm}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho_{1,2}\sigma_1\sigma_2 \\ \rho_{1,2}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \right)$

To account for the influences of unbalanced cluster sizes, coefficients and variability of fixed effects, random effects and errors, we considered the following settings when simulating data.

- ***Generate Clusters***: We have three kind of cluster distribution settings:
  - Same size group setting: 25 observations per group, 50 groups;
  - Balanced group setting: 20-30 observations per group, 50 groups;
  - Unbalanced group setting: 2-50 observations per group, 50 groups.
- ***Simulate Predictors***: The two predictors, denoted as $X_{i,j}$ and $Z_{i,j}$, were drawn from independent $N(0, 1)$ distributions. The intercept, β_0, was set as 5 and both slopes, β_1 and β_2, were set to 1.
- ***Simulate Random Effects***: We included two random effects, a random intercept and a random slope for $X_{i,j}$. The random intercept and slope are assumed to be normally distributed and correlated, with mean zero and correlation coefficient $\rho = 0.5$. For high variance random effects, we set $\sigma_1 = \sigma_2 = 5$; for low variance random effects, both standard deviations were set to 1.
- ***Simulate Error Terms***: For large error variance, we set $\epsilon_{i,j} \overset{iid}{\sim} N(0, 5^2)$, and for small error, we set $\epsilon_{i,j} \overset{iid}{\sim} N(0, 1^2)$.

All simulations were programmed in `R` 4.0.3 (R Core Team 2021) and the resulting data were analysed with `lme4` package (version 1.1-27.1) for fitting LME models (Bates et al. 2015). `HLMdiag` (version 0.5.0) was used to extract residuals from the simulated models (Loy & Hofmann 2014).

3.2 Violating Distributional Assumptions

Our design incorporates nine different misspecification scenarios and one baseline scenario. The baseline's simulated data follows the above data generation mechanism without any deliberate assumption violations. Within each scenario, we have several settings based on all possible combinations of group sizes, two variance component setups and possible violations of model assumption(s) of interests. An example design matrix for non-normality of residuals is shown in Table 1 below.

Table 1: Design Setting Example

Variance Setting	Balance Setting	Misspecification
High Variance Error	Same-sized Groups	Non_normality
High Variance Random Effects	Same-sized Groups	Non_normality
High Variance Error	Balanced Groups	Non_normality
High Variance Random Effects	Balanced Groups	Non_normality
High Variance Error	Unbalanced Groups	Non_normality
High Variance Random Effects	Unbalanced Groups	Non_normality

There are 138 settings designed in total for this study, with 1000 artificial data sets and fitted models simulated for every setting. To assess the performance of residual plots with model misspecifications, we modified the data sets in each scenario (See the exact layout of the entire design matrix in Table 3 of the Appendix).

Scenario 1. Non-Normality: The error term is drawn from a skew-normal or a bimodal distribution using the `rpearson` function from the `PearsonDS` R package (v1.2; (Becker et al. 2021)) . Our study examines three levels of skewness of errors in contrast with the normal distribution, which has a skewness parameter of zero.

When errors are slightly skewed, the skewness parameter is 0.8 and kurtosis of distribution is 4. For moderately skewed errors, the skewness parameter is 1.5 and kurtosis is 6. For extreme skewness, the skewness parameter is 3 and kurtosis is 11. In the bimodal setting, the residual variance is drawn from a mixture of two normal distributions: 60% from $N(-1, 1)$ and 40% from $N(1, 1)$.

Scenario 2. Non-constant Variance: The error term is drawn from distributions where the variance is a function of $X_{i,j}$, as given by: $\sigma^2_{Hetero} = \sigma^2_{baseline} + \lambda(X_{i,j} - \min(X_{i,j}))$ where λ, the heteroscedasticity factor, is set to 2, 4, or 8. Note that $\lambda = 0$ yields the assumed normal distribution and a larger heteroscedasticity factor of 8 ensures that the nonconstant variance pattern of residuals is recognizably extreme (Schielzeth et al. 2020).

Scenario 3. Non-linearity: The data generation formula is changed to $Y_{i,j} = \beta_0 + \beta_1 X_{i,j} + \beta_2 Z^2_{i,j} + u_i + v_i X_i + \epsilon_{i,j}$.

Scenario 4. Omitting a Fixed Effect: We fit a reduced LME model $Z_{i,j}$: $Y_{i,j} = \beta_0 + \beta_1 X_{i,j} + u_i + v_i X_i + \epsilon_{i,j}$, but the data are still generated from equation (3).

Scenario 5. Non-normal Random Effects: To violate the multivariate normal assumption of random effects, we draw two random effect components from skewed bivariate distributions where either the random intercept or the random slope is skewed while keeping the other random effect normally distributed. For simplicity, the non-normal component follows the Pearson distribution similar to scenario 1 and is always set as moderately skewed (skewness = 1.5, kurtosis = 6).

Scenarios 6. 7. 8. (Combined Scenarios): Considering the inherent complexity of hierarchical data in which more than one assumptions on the residual quantities could be violated, we designed scenarios in which homoscedasticity and normality of conditional residuals, homoscedasticity of conditional residuals and linearity of marginal residuals, or normality of conditional residuals and linearity of marginal residuals are violated simultaneously. The process of misspecification in these simulations follows the same procedures as in Scenario 1 to 3.

Scenario 9. Special Cases: The intricacy of real-life data means there are always other ways that the fitted models may fail to accurately capture key characteristics existing in data. In this scenario, we consider three types of model misspecification in addition to typical assumptions of residuals.

- ***Only random intercept***: We also include settings where there is only a random intercept component in the random effects. In this case, we keep the random intercept component normally distributed. When fitting the LME model, we misspecify the random effects structure in still having both the random intercept and random slope.

- ***Time variable***: To simulate longitudinal data, we construct one more fixed effect, $T_{i,j}$ as a numerical sequence, to mimic the time index variable for repeated observations taken within each subject. When fitting the LME model, we omit this time index variable.

- ***Autocorrelated errors*** : Errors are drawn from a first order autoregressive process to mimic the correlated errors that often appear in longitudinal studies:

$$\epsilon_t = \phi\epsilon_{t-1} + e_t \quad \text{and} \quad e_t \overset{iid}{\sim} N(0, \sigma_e) \tag{4}$$

where the autocorrelation coefficient $\phi = 0.4$ and the standard deviation of error $e_t = 1.5$.

Due to the technical difficulties in acquiring sufficient subjects' responses and maintaining balanced group sizes in most longitudinal studies, the simulated data sets in the latter two settings have three types of dropout rates to account for these fluctuations: one with 50 groups and 25 observations per group to simulated large-scale, long-term longitudinal studies with no dropout; one with 20 groups and from 8 to 10 observations per group to simulate relatively small and balanced longitudinal studies with occasional dropouts; and the last one with 20 groups and 2 to 10 observations per group to simulate relatively small and unbalanced longitudinal studies with multiple dropouts.

3.3 Diagnosing Model Violations

Our objective is to explore whether using visualizations of standardized conditional residuals, Cholesky marginal residuals, and Mahalanobis Distance would still be valid diagnostics tools when fitting LME models. However, conducting visual diagnostics of more than $138 \times 1000 = 138,000$ plots is not reasonable for human observers, so we used conventional diagnostics tests as feasible substitutes for having to manually analyze residual plots. If the tests perform as expected, then residual plots can be easily interpreted as in the regression setting.

After purposefully violating model assumptions, we fit LME models using REML.

Normality test: We adopt the Shapiro–Wilk (Shapiro) test for testing normality of the error term $\epsilon_{i,j}$ (conditional residuals). In cases when there is only one random intercept in the data sets but we misspecify the random effects structure as having both random intercept and random slope components, we also use the Shapiro test on the Mahalanobis Distance for normality of the random effects.

Homoscedasticity test: The Breusch–Pagan test (BP) is often used to assess heteroscedasticity in the residuals of a linear regression model (Breusch & Pagan 1979). We use the BP test to check homoscedasticity of the error terms. The null hypothesis is that the variances of the error term do not depend on the independent variables $X_{i,j}$ and therefore are homoscedastic; if the test statistic has a p-value below the significance level of 0.05, we reject the null hypothesis and conclude that the error term is heteroscedastic.

Assessing the distribution of the random effects: Following the recommendations from Singer et al. (2017), we use the Kolmogorov-Smirnov (KS) test as a goodness-of-fit test on the Mahalanobis QQ-plot to assess how well the distribution of the Mahalanobis Distance agrees with the corresponding χ_2^2 distribution in cases where there are random intercept and random slope, χ_1^2 distribution in cases where there is only random intercept.

Linearity test: After extracting the Cholesky marginal residuals, we fit linear and quadratic models to the marginal residuals against the fitted. We then use Analysis of Variance (ANOVA) (Bates et al. 1992) to conduct a nested F-test on these models to determine whether there is any discernible curvature in the error terms.

4 Evaluating Residual Plots

Understanding the behaviors of residual plots is the ultimate goal for our simulation study of LME model diagnostics. It is important to note that different aspects of the specification are often entwined, and this issue is particularly prominent in assessing the fit of LME models (Snijders & Berkhof 2008). For example, it is possible that deviations from random effects normality are caused by a misspecification of the fixed effects structure (**?**) rather than the residual distribution itself. We conduct all four diagnostics tests in every scenario to provide a thorough examination of the residuals.

Baseline setting

We first explore the performance of residual diagnostics from properly specified LME models. The results for these baseline models are shown in Table 2. The Shapiro and ANOVA tests achieve the nominal type I error rate in these settings despite changing variances and group sizes. The baseline BP tests show slightly inflated type I error rate (9.5% - 11.8%). For normality of the estimated random effects, we also see slighly higher rejection rates (7.1% - 11.3%) if the error terms are more variable than the random effects, even when models are correctly specified.

Table 2: Test Results from Residual Diagnostics of Properly Specified LME Models

Variance	Balance	Shapiro	BP.Test	KS.Test	ANOVA
High Variance Error	Same-sized Groups	0.045	0.100	0.104	0.046
High Variance RE	Same-sized Groups	0.049	0.101	0.004	0.059
High Variance Error	Balanced Groups	0.053	0.098	0.085	0.050
High Variance RE	Balanced Groups	0.051	0.118	0.003	0.056
High Variance Error	Unbalanced Groups	0.059	0.102	0.079	0.041
High Variance RE	Unbalanced Groups	0.049	0.095	0.011	0.053

Non-normal Error terms

In scenario where we simulate non-normal errors, we would like the Shapiro test to successfully pick up the non-normality of the errors while the BP, KS and ANOVA tests maintain their nominal Type I error rates, as this would allow straight-forward interpretation of the residual plots.

- As seen in Figure 1-1, non-normality of the conditional residuals is accurately detected in all simulations with skew-normal errors, indicating that the skewness of conditional residuals can be discerned from residual diagnostics. Interestingly, compared with the baseline results, the non-normality of the estimated conditional residuals is not apparent when errors have larger variances and follow bimodal distributions. The connection between bimodality and variability of the errors could be further explained by the phenomenon that when these two distributions are more diffuse due to high variances, the resulting mixture distribution would appear normal.

- More severe skewness in the conditional residuals leads to higher false positive rates of heteroscedasticity, as shown in Figure 1-2, especially if the error terms possess larger variances than the random effects. Compared with the baseline, heteroscedasticity is not more frequently detected among settings with errors in bimodal distribution (baseline: 9.5% - 11.8%; bimodal: 7.6% - 11.5%), regardless of variance components and group sizes.

- Using KS tests, we rarely detect false non-normality when the random effects have lower variances in Figure 1-3 (0.2% - 1.2%), but this error rates increases when the errors have higher variances (Type I error rates increases to 7.1% - 11.3%).

- Meanwhile, ANOVA tests achieve nominal Type I error rates in Figure 1-4 (4.1% - 5.9%), meaning that no unusual quadratic structure is detected and the Cholesky marginal residuals behave as we expected.

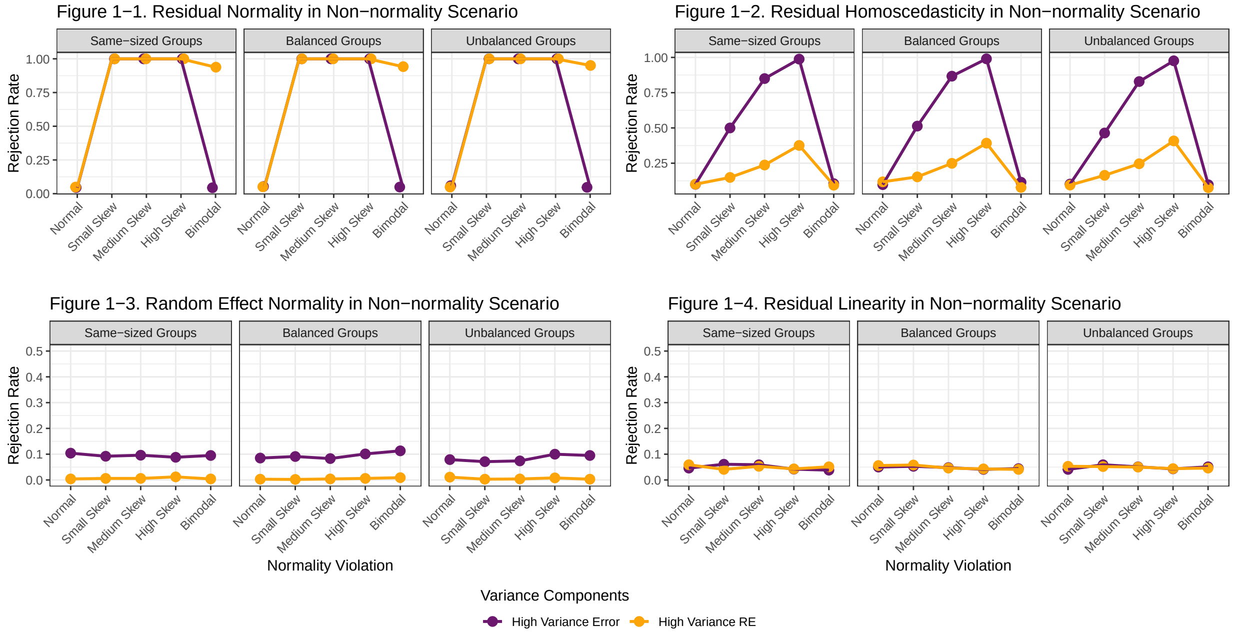


Figure 1. The behaviors of residual quantities when the normality assumption of error terms was deliberately violated. 1-1 presents the rejection rate of Shapiro tests: Non-normality of conditional residuals is well detected, except bimodality with high variance errors. BP tests in 1-2 suggest that the increasing skewness of errors is linked to more severe false positive rates of heteroscedasticity. 1-3 and 1-4 show that normality of random effects and linearity of errors are not strongly deviated from the baseline statistics.

Non-Constant Variance

In scenario where we simulate heteroscedastic errors, we would like the BP test to successfully pick up the heteroscedasticity of the error terms while the Shapiro, KS and ANOVA tests maintain their nominal Type I error rates.

- When error terms are heteroscedastic and random effects are given high variances, we observe higher rejection rates of Shapiro tests for the conditional residuals in Figure 2-1.

As the heteroscedasticity factor gets higher, the false rejection rates increase. With more unbalanced group sizes, we also see more frequent erroneous detection of non-normality (15.6% - 21%) than with balanced groups (11.1% - 17.4%).

- BP tests on the conditional residuals (Figure 2-2) detect heteroscedasticity more often with higher heteroscedasticity factors. When the level of heteroscedasticity is relatively low ($\lambda = 2$ or 4), the heteroscedasticity of errors is detected more frequently if the random effects have larger variances (20.4% - 27.7% with high variance random effects and 11% - 23.2% with high variance errors). In the extreme heteroscedasticity settings ($\lambda = 8$), the variances of the error terms and the random effects don't appear to affect the powers of BP tests.

- With error terms having higher variances, the false rejection rates of KS tests in Figure 2-3 increase when λ becomes higher (22.6% - 26.7% if $\lambda = 8$), and remain low with high variance random effects (0.3% - 1.1%). This suggests that analysts are likely to identify non-normal random effects from looking at the Mahalanobis distances when errors possess strong heteroscedasticity and large variability.

- The ANOVA tests show that the estimated Cholesky marginal residuals behave as expected, with the false rejection rate fluctuating around 5% in Figure 2-4.

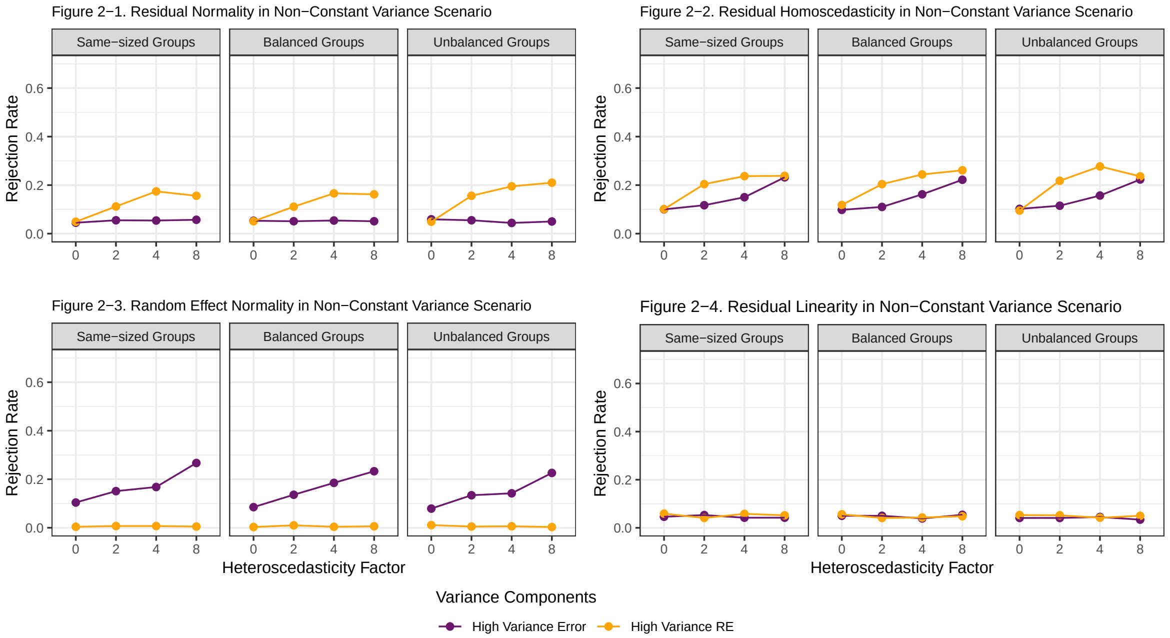


Figure 2. The behaviors of residual quantities when homoscedasticity assumption of error terms was deliberately violated. 2-1 shows that conditional residuals are more likely to exhibit non-normality when error variances are low. BP tests in 2-2 suggest more severe heteroscedasticity with high variance random effects. The normality of random effects is more likely to break in 2-3 as errors possess higher variability. 2-4 shows that linearity of errors are not strongly deviated from the baseline.

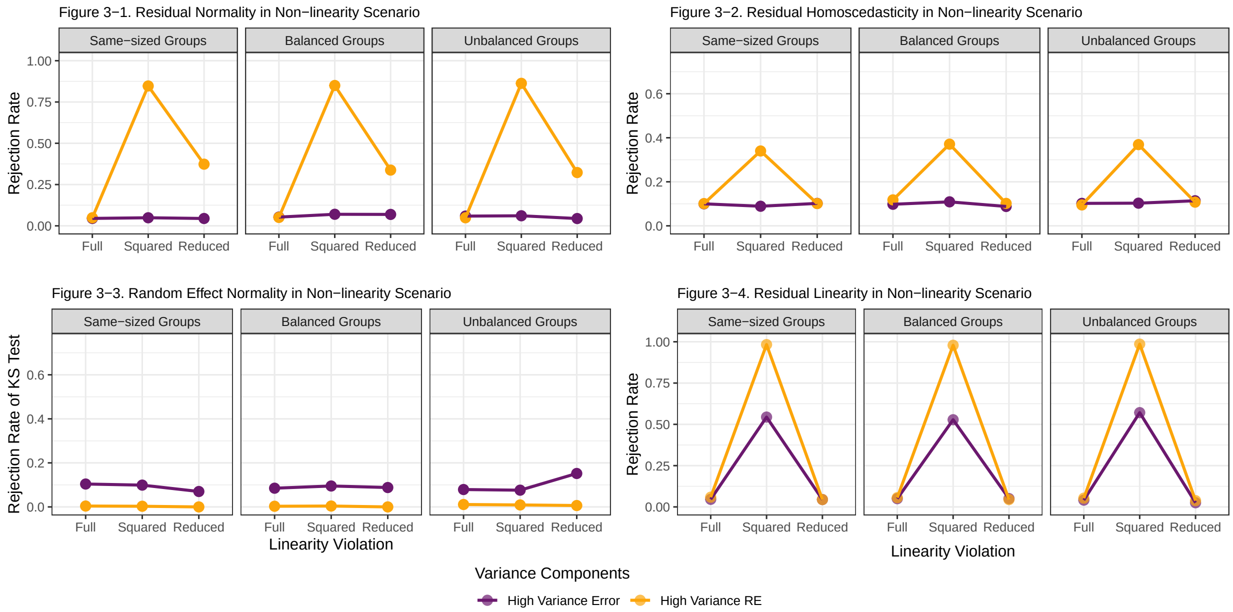


Figure 3. The behaviors of residual quantities when the fixed effects were squared or omitted. From 3-1 and 3-2, the misspecification of fixed effects would intensify non-normality and heteroscedasticity of low variance errors. The normality of random effects is more likely to break in 3-3 as errors possess higher variability and cluster sizes are unbalanced. 3-4 shows squaring one fixed effect leads to more quadratic distribution of errors.

Nonlinearity & Omitted fixed effects

In this scenario where we simulate one squared fixed effects term or omit one fixed effects term, we would like the ANOVA test to pick up the nonlinearity of the Cholesky residuals while the Shapiro test, BP test, and KS test maintain their nominal Type I error rates.

- Fitting a model with only linear terms has great influence on the conditional residual plots only when random effects are given higher variability. In this case, there is also a high chance that the estimated errors will not appear to be normally distributed (84.7% - 86.3%), and a moderate chance that they will appear to be homoscedastic (34% - 37.1%). With errors possessing higher variances, the significance tests no longer recognize the non-normality or heteroscedasticity more often than in the baseline situation (Figure 3-1 & 3-2).

- The spikes shown in Figure 3-4 for the ANOVA test rejection rates are expected since one fixed effects term was squared. They are evidence that misspecifying fixed effects have direct impacts on the behavior of marginal residuals. This impacts, however, differ by the variability of residual quantities: the rejection rates are between 97.9% and 98.5% when errors have larger variances than the random effects and between 52.8% and 57.1% in the opposite situation. This again illustrates the importance of taking into consideration of the ratio of residuals variance when assessing model inadequacy by looking at residual plots.

- Omitting a fixed effect shows no strong repercussions on assessing the distributional assumptions of residual quantities. Despite missing a term, the fitted models generally yield the same rejection rates as in the baseline situation. There are two exceptions as shown in Figure 3-1 and 3-3. There are high false positive rates of the conditional residuals' normality assumption (32.3% - 37.3%) when we have high variance random effects. The other is the elevated rejection rate of random effects normality assumption (15.8%) with high variance random effects and unbalanced clusters.

Non-normal Random Effects

In the scenario where we simulate skewed bivariate normal random effects, we would like the KS test to successfully pick up the non-normality of the random effects while the Shapiro, BP and ANOVA tests maintain their nominal Type I error rates. Overall, we discover that using KS test on MD is not a reliable diagnostics tool for misspecified LME:

- Regardless of whether the skewed component is the random intercept or the random slope, the power of the KS test is higher when the errors possess larger variances (8.7% - 11.2%) as opposed to cases when the random effects are given higher variances (2.3% - 3.6%). Overall, the highest power of KS tests is only 11.2% across all settings, and in cases when the random effects are given larger variances we see the lowest detection power (well below 5%). This implies that when errors have higher variances, the random effects are more likely to appear non-normal, no matter whether there is true model misspecification or not, calling into question of the validity of using Mahalanobis QQ-plots to diagnose random effects.
- The other test results, indicate that having skewed bivariate normal random effects does not heavily impact the normal and constant-variance properties of the estimated conditional residuals, nor the normality of the Cholesky marginal residuals.
- We also simulate two-level data with only a random intercept component while misspecifying the random effects structure as also having a random slope component when fitting the LME model. Since there is no underlying multivariate normal distribution but just the normally distributed random intercept, we employ the Shapiro test on the random effects to detect any potential non-normal behavior.

In this case, we see that normality assumption of error terms and the Cholesky residuals generally holds well. For testing the homogeneity assumption of the error terms, the Type I error rate of the BP test fluctuates around 10%.

To see if the residual diagnostics can pick up the fact that the random effects do not follow multivariate normality distribution, we use the KS test and observed that the power of the KS test vary between 33.6% and 44.5%. To see if we can diagnose that the random effects follows univariate normality distribution, we use Shapiro test on the MD that gives false rejection rates of 99.8% to 100% (Figure D-3). Overall, we notice that using residual diagnostics to test the multivariate normality assumption of the random effects are not reliable.

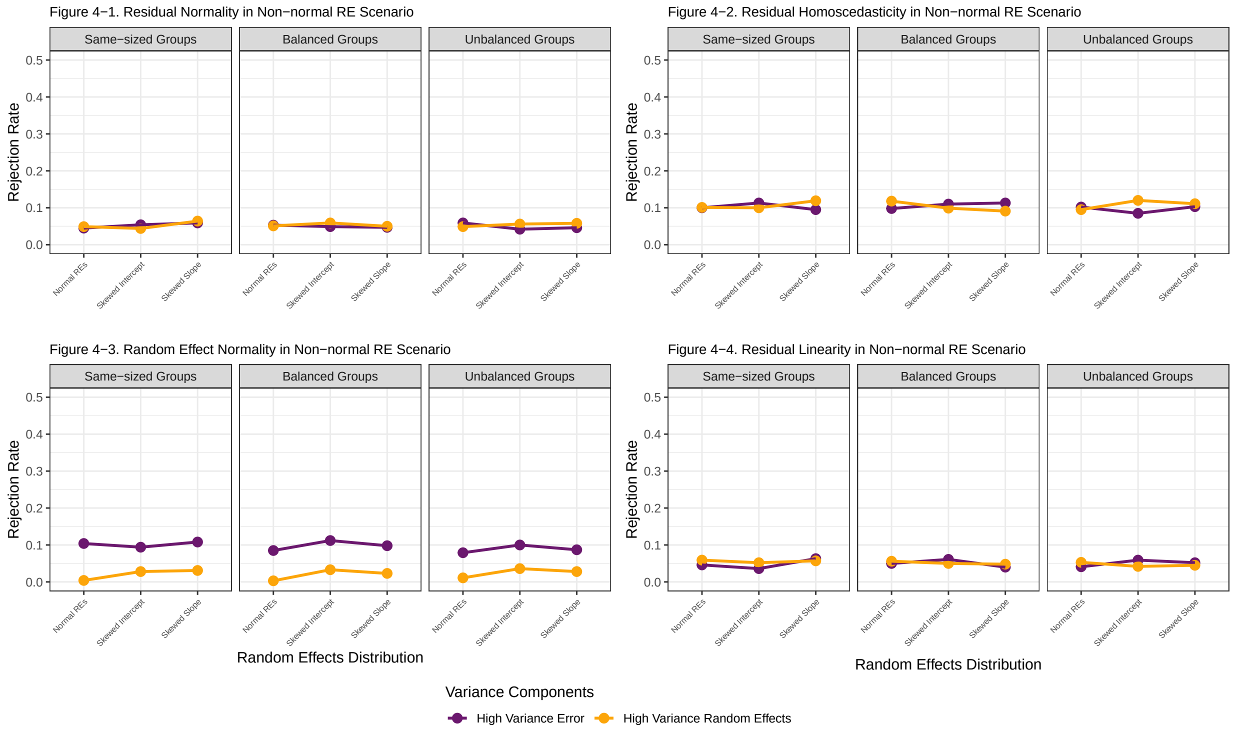


Figure 4. The behaviors of residual quantities when the normality assumption of random effects was violated. The influences of having skewed random effects distributions are genearly weak in inducing other assumption violations as shown in 4-1, 4-2 and 4-4. Higher chance of non-normal random effects continued to be associated with high variance errors in 4-3.

Heteroscedastic & Moderately skewed error

In the scenario where we simulate heteroscedastic errors with $\lambda = 2$, 4, and 8 and also set the errors to be moderately skewed, we would like the BP test to successfully pick up the non-constant variance property and Shapiro test to pick up the non-normality of the estimated conditional residuals while the KS and ANOVA tests maintain their nominal Type I error rates.

- The non-normal behavior of the estimated conditional residuals would be detected accurately using the Shapiro test regardless of the changing variances and group sizes.
- The heteroscedastic pattern of the estimated conditional residuals would be detected more accurately (50.3% - 65.3%) when errors have larger variances, as against when the random effects have larger variances (19.4% - 31.6%). As the heteroscedasticity factor, λ, gets larger, the power of BP tests does not necessarily get stronger (Appendix Figure A-2), but rather stays in the range of 50% - 65% (high errors variance) and 25% - 32% (low errors variance). Furthermore, holding λ constant, as cluster sizes of data become more balanced, the estimated conditional residuals appear to be less heteroscedastic.
- With heteroscedastic and skewed errors, MD tends to appear non-normal when the errors have larger variances. As λ gets bigger, the false rejection rate of KS test tends to

get higher in cases when errors have a larger variance (from 51.7% - 80.2%). Comparably, When the errors have smaller variances, the MD seems to be behaving too "normally" with the rejection rate of the KS test being between 0.1% and 0.6%, sometimes below the nominal Type I error rate.

- The ANOVA test results show that the estimated Cholesky marginal residuals behave as we would expect, with the false rejection rate fluctuating around 5%.

Heteroscedastic error & Squaring fixed effects

In the scenario where we simulate heteroscedastic errors with $\lambda = 2$, 4, and 8 and squared one of the fixed effects, we would like the BP test to successfully pick up the non-constant variance property of the estimated conditional residuals and ANOVA test to pick up the nonlinearity of the Cholesky residuals while the Shapiro and KS tests maintain their nominal Type I error rates.

- When the errors have smaller variances, we could see that the false rejection rates of the Shapiro test (16.9% - 21.7%) are much higher than when the errors have comparably higher variances (4.8% - 6.4%). (See Appendix Figure B-1)
- The heteroscedastic pattern of estimated conditional residuals are detected most accurately when errors have smaller variances and λ is at its lowest value. When λ increases to 4, whether the errors or the random effects have larger variances, we see that powers of BP tests are roughly the same (25% - 30%). When $\lambda = 8$, the power increases especially in cases when errors have higher variances. Overall, as λ increases from 2 to 8, we see the overall powers of BP tests on estimated conditional residuals will get higher. (See Appendix Figure B-2)
- The false rejection rates of KS tests are much higher (18.7% - 50.4%) when the errors have larger variances than when the errors had lower variances (0.1% - 1.1%). (See Appendix Figure B-3)
- Checking the Cholesky marginal residuals using the ANOVA test, its power decreases as λ gets higher. Regardless of whether the cluster sizes are balanced or not, when the random effects have larger variances, we see higher powers of ANOVA tests (44.3% - 78.6%); when errors have higher variances, the powers of ANOVA tests are lower (31.1% - 45.7%). (See Appendix Figure B-4)

Squaring fixed effect & Non-normal error

In the scenario where we simulate non-normal error terms (moderate skewed normal and bimodality) and induced nonlinearity, we would like the Shapiro test to successfully pick up the non-normality of the estimated conditional residuals and the ANOVA test to pick up the nonlinearity of the Cholesky residuals while BP and KS tests maintain their nominal Type I error rates.

- Overall, our Shapiro tests pick up the moderate skewness of the errors better (77.2% - 100%) than the bimodal distribution (average 4.8% - 12.3%, one outlier rejection rate of 92.1%). As the cluster sizes get more unbalanced, the estimated conditional residuals appear more normal (Appendix Figure C-1).

- When distribution of the error term is moderately skewed, the false rejection rate of BP tests using conditional residuals is much higher (23.6% - 81.5%) than when the errors are bimodal (8.4% - 18.1%).

- Overall, the false rejection rates of KS tests are not stable and are likely to be higher (7.9% - 71.5%) when the errors have larger variances compared than when the random effects have higher variances (0.1% - 0.9%).

- Using the Cholesky marginal residuals, the ANOVA test has higher power when the cluster sizes are more balanced, as well as in cases when the error term is moderately skewed. Holding cluster sizes and skewness type constant, the power of the ANOVA tests are much higher in cases when the random effects have higher variances (26.5% - 98.2%) than when errors have higher variances (7.9% - 53.2%). (Appendix Figure C-4)

Longitudinal settings

In this scenario, we are interested in exploring the effects of model structure misspecifications on residual plots. We considered omitting a time index variable in longitudinal data set; and misspecifing AR(1) errors as i.i.d errors. When fitting LME models, we use the standard model mentioned in section 3.1 equation (3).

Time variable case vs auto-correlated error

- The overall rejection rates of BP tests are in the range of 9.4% to 15%, with higher rejection rates when the longitudinal characteristics appears in the auto-correlated error term (9.4% - 14.6%) compared to the scenario when it appears in fixed-effect components as the time index variable (8.5% - 12.7%).

Shared results

- Holding all other factors constant, when random effects have larger variances, we see that the false rejection rates of using BP tests on testing the estimated conditional residuals' normality are slightly higher than when errors have higher variances than random effects.

- Similar to previous scenarios, using KS tests on MD is not a stable diagnostics tool (rejection rates range from 0.3% to 69.5%) (Appendix Figure D-3); especially when the simulated data sets are designed to have unbalanced clusters, we tend to see the non-normal behaviors of random effects being exaggerated (rejection rate up to 69.5%).

- With the estimated conditional residuals, Shapiro tests generally achieve the nominal Type I error rate with occasionally higher rejection rates when the data set has unbalanced group sizes and small sample sizes (see small, unbalanced setting in Section 3). The ANOVA tests on the Cholesky marginal residuals also maintain the nominal Type I error rate.

Special case

- When the random effects have larger variances, the group sizes are balanced with a total of 50 groups in the data set, and the longitudinal characteristics appears as the time index variable, we see one "outlier" Shapiro rejection rate of 100% compared to the average 5% rejection rate in other longitudinal settings (Appendix Figure D-1). After investigating the visuals of the standardized conditional residuals, we determine that omitting the time index variable when we have many groups and a large number of observations per group does skew the distribution of the standardized conditional residuals to the extent that using the Shapiro test is not advised.

5 Discussion

Our simulation study offers valuable insight into how residual diagnostics can be used and interpreted for LME models.

Single Misspecification: Our simulation results confirm and strengthen the intertwined nature of residual behaviors: When only one model assumption is misspecified, typically more than one model assumption appears to be violated based on diagnostic plots. For example, models with skewed normal error terms often appear to be non-normal random effects and heteroscedastic errors. If the homoscedasticity assumption is violated alone by having a high heteroscedasticity factor, analysts will be likely to observe non-normality of both conditional residuals and random effects along with the expected non-constant variance pattern in residual plots. Bimodality is a noteworthy exception: plots do not often show non-normality of estimated errors or other inadequacy when true bimodal errors have high variance.

The violation of the multivariate normal assumption of random effects alone also does not result in strong deviations from any other model assumptions. This finding in part is in agreement with Schielzeth et al. (2020) and supports the robustness of LME models in the case of random effects misspecification.

Whether the longitudinal characteristic appears in the fixed effects as a time index variable or in the errors as auto-correlated errors, the distribution and variance patterns of the estimated standardized conditional residuals and the Cholesky marginal residuals are not very different from the correctly specified LME models. The one exception is the MD, especially when the errors have larger variances, the MD tend to show heavy deviations from the multivariate normality assumption. In cases when we see errors with lower variance, the deviations are much less discernible.

Combined Misspecifications: When a pair of assumptions are violated, analysts will tend to observe more issues in residual plots. The two truly violated assumptions will be detected in their corresponding diagnostic plots, along with some concerning issues with other assumptions. For example, normality of random effects may be falsely detected when non-normality and heteroscedasticity of errors occur. The combination of non-linearity and non-constant variance misspecifications in particular cause all four distributional assumptions to be flagged.

Residual Variability: The variability of errors and random effects can lead to drastically different behaviors of residual plots. We notice that normality of random effects are more likely to be identified as problematic when error terms were given higher variance than random effects. On the other hand, violating the linearity assumption has more severe consequences if random effect variances are larger. When there is a single misspecification, normality, linearity and homoscedasticity assumptions of errors are more likely to be violated in residual plots if their variances are smaller. The interesting exception is heteroscedasticity induced by non-normality of errors, which is more alarming if errors have larger variances. The impacts of the relative magnitude of the residual variances are similar in combined scenarios. Heteroscedasticity continues to be more frequently detected where random effects are less variable than errors. The normality assumption of errors, if not directly violated, tend to be misdiagnosed more often with high variance random effects.

Influences of Group Sizes: The distribution of cluster sizes plays a more significant role in the combined scenarios. Simulation results suggest that the normality and homoscedasticity assumptions on the error terms are more likely to appear to be violated with unbalanced clusters than evenly distributed ones. When linearity and normality assumptions on the errors are violated, only unevenly distributed clusters have their random effects flagged as non-normal, and the heteroscedasticity and non-linearity of residuals become less discernible as group sizes become more unbalanced.

Overall, our study revealed that employing a single diagnostic plot on LME models' residuals is not sufficient to accurately assess the validity of model assumptions. One remedy is to employ the lineup protocol that will allows us to simultaneously assess several model assumptions for a thorough investigation (Loy et al. 2017). The intertwined nature of these model specifications require practitioners to examine not only a full set of diagnostic tools available for each assumption, but also the ways in which the hierarchical structure of fitted models was composed. Other important characteristics of the interested data sets, for example cluster sizes, relative magnitude of residual variances, omitted fixed effects and longitudinal sampling, should also be checked to avoid exaggerating or overlooking any violations to certain model assumptions. Failure to take these factors into consideration will result in misinterpretation of diagnostics plots and misjudgment of model assumptions, which can lead to biased model estimates.

There are still questions left unaddressed beyond the scope of our study. The validity of using least squares residuals for the diagnostic purposes are still yet to be explored. Researchers could also look into residual plots under more complex combinations of LME model misspecifications, like three of more violations at the same time. In comparison to the classic maximum likelihood approach there have also been some alternative algorithms and methodologies developed for LME model estimation, such as the Marquardt algorithm (Proust & Jacqmin-Gadda 2005) and the robust estimation method (Koller 2016). We recommend more future research in the direction of correctly evaluating expected diagnostic protocols for these approaches.

6 Acknowledgement

We are deeply indebted to Professor Adam Loy from Carleton College who kindly advised us throughout this research program.

7 Appendices

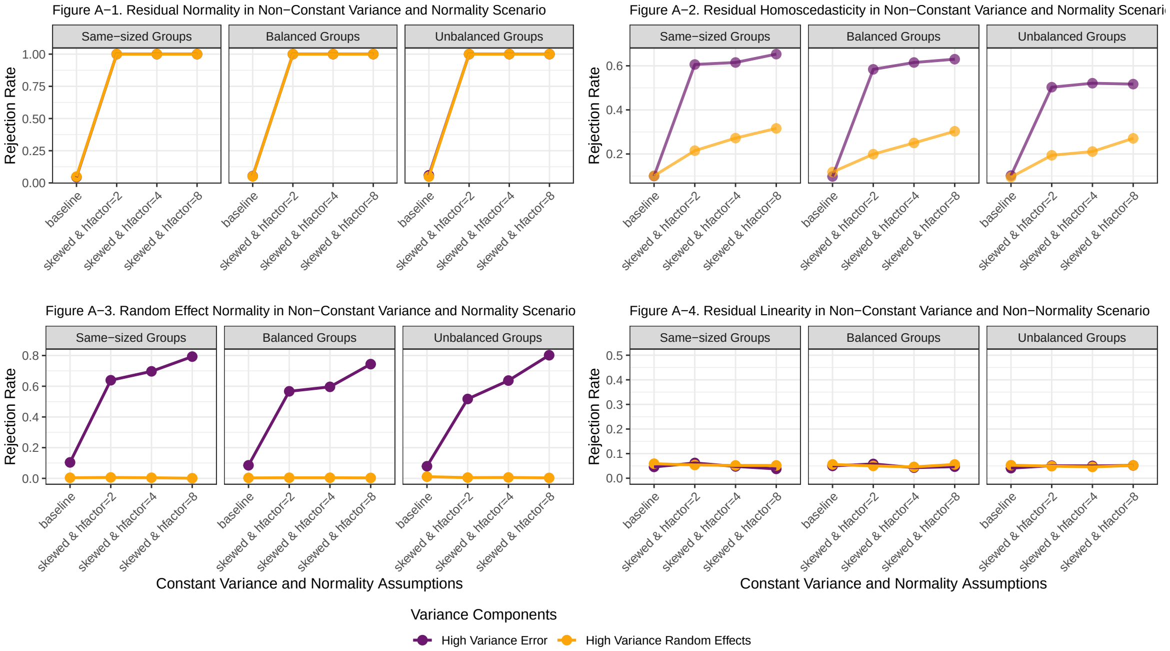


Figure A−1. Residual Normality in Non−Constant Variance and Normality Scenario

Figure A−2. Residual Homoscedasticity in Non−Constant Variance and Normality Scenari

Figure A−3. Random Effect Normality in Non−Constant Variance and Normality Scenario

Figure A−4. Residual Linearity in Non−Constant Variance and Non−Normality Scenario

Appendix A. The behaviors of residual quantities in Non-Constant Variance and Non-Normality Scenario. A-1 suggests that non-normality of moderately skewed errors would be well captured by residual plots. Heteroscedasticity in A-2 is more likely to be flagged with high variance errors. A-3 shows deviations of random effects normality induced by heteroscedastic and non-normal errors.

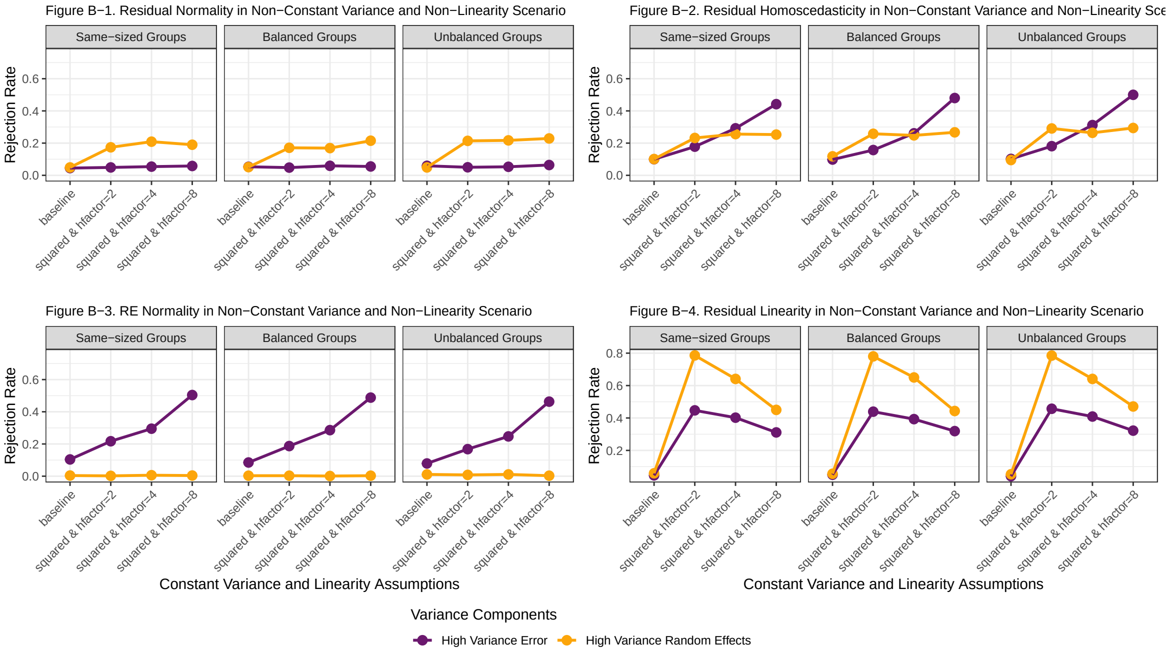


Figure B−1. Residual Normality in Non−Constant Variance and Non−Linearity Scenario

Figure B−2. Residual Homoscedasticity in Non−Constant Variance and Non−Linearity Sce

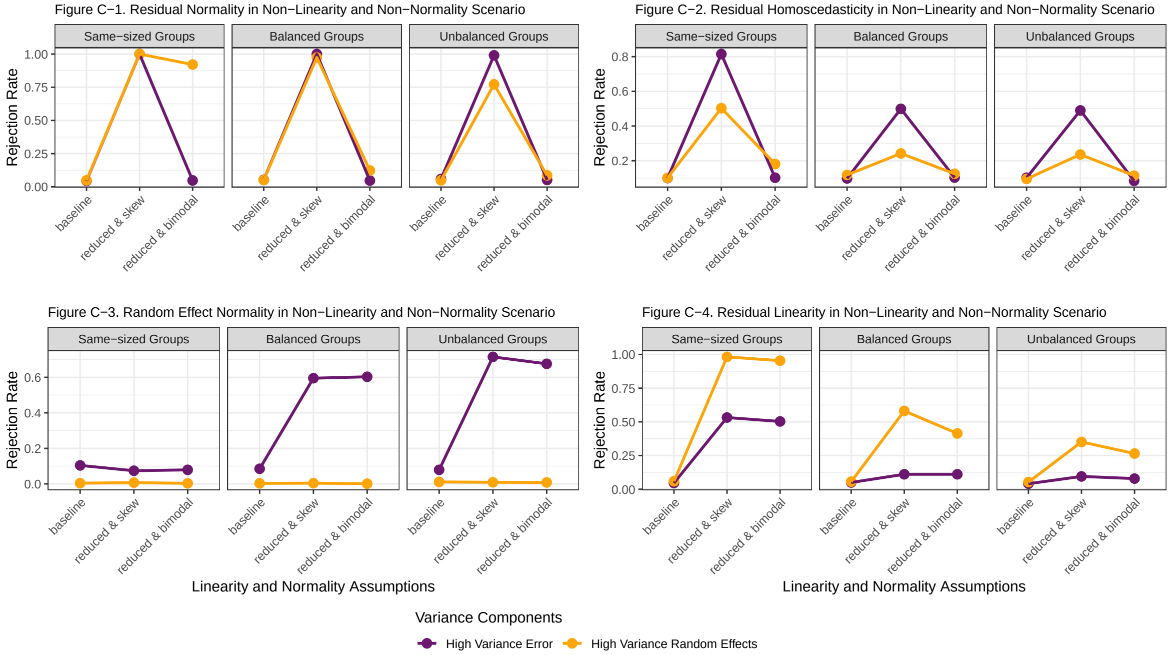
Figure B−3. RE Normality in Non−Constant Variance and Non−Linearity Scenario

Figure B−4. Residual Linearity in Non−Constant Variance and Non−Linearity Scenario

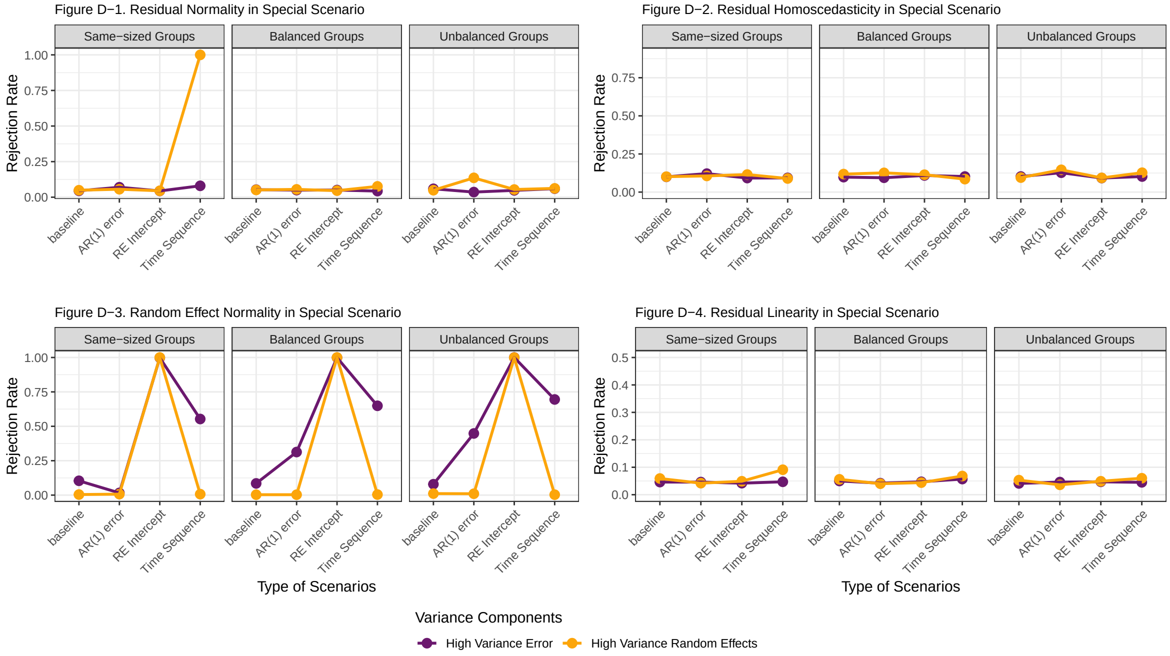
Appendix B. The behaviors of residual quantities in Non-Constant Variance and Non-Linearity Scenario. Both B-2 and B-4 show direct consequences of non-constant variance and non-linearity. B-1 presents deviations of error term normality caused by misspecification of constant Variance and linearity assumptions. Deviations of random effects normality are apparent only when errors have higher variability in B-3.

Table 3: Shapiro, BP, KS and ANOVA Test Results of all 138 Settings

Setting	Variance	Balance	Normality	Linearity	H.factor	RE	FE	Hetero_lin	Hetero_norm	Lin_norm	Special	Shapiro	BP.Test	KS.Test	ANOVA
1	High Variance Error	Same-sized Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.0450000	0.1000000	0.1040000	0.0460000
2	High Variance RE	Same-sized Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.0490000	0.1010000	0.0040000	0.0590000
3	High Variance Error	Balanced Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.0530000	0.0980000	0.0850000	0.0500000
4	High Variance RE	Balanced Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.0510000	0.1180000	0.0030000	0.0560000
5	High Variance Error	Unbalanced Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.0590000	0.1020000	0.0790000	0.0410000
6	High Variance RE	Unbalanced Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.0490000	0.0950000	0.0110000	0.0530000
7	High Variance Error	Same-sized Groups	skewness_3	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	1.0000000	0.9880000	0.0880000	0.0420000
8	High Variance RE	Same-sized Groups	skewness_3	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	1.0000000	0.3760000	0.0120000	0.0430000
9	High Variance Error	Balanced Groups	skewness_3	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	1.0000000	0.9900000	0.1010000	0.0410000
10	High Variance RE	Balanced Groups	skewness_3	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	1.0000000	0.3920000	0.0060000	0.0430000
11	High Variance Error	Unbalanced Groups	skewness_3	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	1.0000000	0.9760000	0.1000000	0.0430000
12	High Variance RE	Unbalanced Groups	skewness_3	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	1.0000000	0.4080000	0.0080000	0.0440000
13	High Variance Error	Same-sized Groups	skewness_1.5	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	1.0000000	0.8500000	0.0960000	0.0590000
14	High Variance RE	Same-sized Groups	skewness_1.5	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	1.0000000	0.2370000	0.0060000	0.0530000
15	High Variance Error	Balanced Groups	skewness_1.5	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	1.0000000	0.8670000	0.0830000	0.0480000
16	High Variance RE	Balanced Groups	skewness_1.5	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	1.0000000	0.2490000	0.0040000	0.0460000
17	High Variance Error	Unbalanced Groups	skewness_1.5	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	1.0000000	0.8290000	0.0740000	0.0510000
18	High Variance RE	Unbalanced Groups	skewness_1.5	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	1.0000000	0.2460000	0.0040000	0.0500000
19	High Variance Error	Same-sized Groups	skewness_0.8	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	1.0000000	0.5000000	0.0920000	0.0610000
20	High Variance RE	Same-sized Groups	skewness_0.8	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	1.0000000	0.1490000	0.0060000	0.0410000
21	High Variance Error	Balanced Groups	skewness_0.8	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	1.0000000	0.5130000	0.0910000	0.0530000
22	High Variance RE	Balanced Groups	skewness_0.8	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	1.0000000	0.1530000	0.0020000	0.0580000
23	High Variance Error	Unbalanced Groups	skewness_0.8	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	1.0000000	0.4640000	0.0710000	0.0590000
24	High Variance RE	Unbalanced Groups	skewness_0.8	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	1.0000000	0.1640000	0.0030000	0.0520000
25	High Variance Error	Same-sized Groups	bimodal	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.0440000	0.1050000	0.0950000	0.0380000
26	High Variance RE	Same-sized Groups	bimodal	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.9380000	0.0940000	0.0040000	0.0510000
27	High Variance Error	Balanced Groups	bimodal	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.0490000	0.1150000	0.1130000	0.0440000
28	High Variance RE	Balanced Groups	bimodal	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.9420000	0.0780000	0.0090000	0.0410000
29	High Variance Error	Unbalanced Groups	bimodal	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.0470000	0.0970000	0.0950000	0.0510000
30	High Variance RE	Unbalanced Groups	bimodal	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.9510000	0.0760000	0.0030000	0.0460000
31	High Variance Error	Same-sized Groups	norm	linear	2	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.0550000	0.1170000	0.1510000	0.0530000
32	High Variance RE	Same-sized Groups	norm	linear	2	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.1120000	0.2040000	0.0070000	0.0410000
33	High Variance Error	Balanced Groups	norm	linear	2	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.0510000	0.1100000	0.1360000	0.0500000
34	High Variance RE	Balanced Groups	norm	linear	2	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.1110000	0.2040000	0.0100000	0.0410000
35	High Variance Error	Unbalanced Groups	norm	linear	2	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.0550000	0.1150000	0.1340000	0.0410000
36	High Variance RE	Unbalanced Groups	norm	linear	2	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.1560000	0.2180000	0.0050000	0.0520000
37	High Variance Error	Same-sized Groups	norm	linear	4	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.0540000	0.1500000	0.1680000	0.0420000
38	High Variance RE	Same-sized Groups	norm	linear	4	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.1740000	0.2370000	0.0070000	0.0580000
39	High Variance Error	Balanced Groups	norm	linear	4	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.0540000	0.1620000	0.1850000	0.0390000
40	High Variance RE	Balanced Groups	norm	linear	4	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.1660000	0.2440000	0.0040000	0.0430000
41	High Variance Error	Unbalanced Groups	norm	linear	4	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.0440000	0.1570000	0.1420000	0.0450000
42	High Variance RE	Unbalanced Groups	norm	linear	4	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.1950000	0.2770000	0.0060000	0.0420000
43	High Variance Error	Same-sized Groups	norm	linear	8	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.0570000	0.2320000	0.2670000	0.0420000
44	High Variance RE	Same-sized Groups	norm	linear	8	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.1560000	0.2380000	0.0050000	0.0520000
45	High Variance Error	Balanced Groups	norm	linear	8	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.0510000	0.2220000	0.2330000	0.0540000
46	High Variance RE	Balanced Groups	norm	linear	8	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.1620000	0.2610000	0.0060000	0.0480000
47	High Variance Error	Unbalanced Groups	norm	linear	8	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.0500000	0.2230000	0.2260000	0.0340000
48	High Variance RE	Unbalanced Groups	norm	linear	8	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.2100000	0.2360000	0.0030000	0.0500000
49	High Variance Error	Same-sized Groups	norm	sq	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.0490000	0.0890000	0.0990000	0.5440000
50	High Variance RE	Same-sized Groups	norm	sq	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.8470000	0.3400000	0.0030000	0.9830000
51	High Variance Error	Balanced Groups	norm	sq	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.0700000	0.1090000	0.0950000	0.5280000
52	High Variance RE	Balanced Groups	norm	sq	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.8500000	0.3710000	0.0040000	0.9790000
53	High Variance Error	Unbalanced Groups	norm	sq	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.0610000	0.1030000	0.0760000	0.5710000
54	High Variance RE	Unbalanced Groups	norm	sq	0	norm_re	full	linear_homo	0_skew	linear_norm	standard	0.8630000	0.3690000	0.0090000	0.9860000
55	High Variance Error	Same-sized Groups	norm	linear	0	mildly_skewed_re_intercept	full	linear_homo	0_skew	linear_norm	standard	0.0540000	0.1130000	0.0940000	0.0360000
56	High Variance RE	Same-sized Groups	norm	linear	0	mildly_skewed_re_intercept	full	linear_homo	0_skew	linear_norm	standard	0.0440000	0.1000000	0.0280000	0.0520000
57	High Variance Error	Balanced Groups	norm	linear	0	mildly_skewed_re_intercept	full	linear_homo	0_skew	linear_norm	standard	0.0490000	0.1100000	0.1120000	0.0610000
58	High Variance RE	Balanced Groups	norm	linear	0	mildly_skewed_re_intercept	full	linear_homo	0_skew	linear_norm	standard	0.0590000	0.0990000	0.0330000	0.0500000
59	High Variance Error	Unbalanced Groups	norm	linear	0	mildly_skewed_re_intercept	full	linear_homo	0_skew	linear_norm	standard	0.0420000	0.0850000	0.1000000	0.0590000
60	High Variance RE	Unbalanced Groups	norm	linear	0	mildly_skewed_re_intercept	full	linear_homo	0_skew	linear_norm	standard	0.0560000	0.1200000	0.0360000	0.0420000
61	High Variance Error	Same-sized Groups	norm	linear	0	mildly_skewed_re_slope	full	linear_homo	0_skew	linear_norm	standard	0.0590000	0.0950000	0.1080000	0.0630000
62	High Variance RE	Same-sized Groups	norm	linear	0	mildly_skewed_re_slope	full	linear_homo	0_skew	linear_norm	standard	0.0640000	0.1190000	0.0310000	0.0570000
63	High Variance Error	Balanced Groups	norm	linear	0	mildly_skewed_re_slope	full	linear_homo	0_skew	linear_norm	standard	0.0470000	0.1130000	0.0980000	0.0400000
64	High Variance RE	Balanced Groups	norm	linear	0	mildly_skewed_re_slope	full	linear_homo	0_skew	linear_norm	standard	0.0500000	0.0910000	0.0230000	0.0480000
65	High Variance Error	Unbalanced Groups	norm	linear	0	mildly_skewed_re_slope	full	linear_homo	0_skew	linear_norm	standard	0.0460000	0.1030000	0.0870000	0.0520000
66	High Variance RE	Unbalanced Groups	norm	linear	0	mildly_skewed_re_slope	full	linear_homo	0_skew	linear_norm	standard	0.0580000	0.1110000	0.0280000	0.0450000
67	High Variance Error	Same-sized Groups	norm	linear	0	norm_re	full	linear_homo	2_skew	linear_norm	standard	1.0000000	0.6060000	0.6390000	0.0620000
68	High Variance RE	Same-sized Groups	norm	linear	0	norm_re	full	linear_homo	2_skew	linear_norm	standard	1.0000000	0.2150000	0.0060000	0.0540000
69	High Variance Error	Balanced Groups	norm	linear	0	norm_re	full	linear_homo	2_skew	linear_norm	standard	1.0000000	0.5840000	0.5670000	0.0580000
70	High Variance RE	Balanced Groups	norm	linear	0	norm_re	full	linear_homo	2_skew	linear_norm	standard	1.0000000	0.1990000	0.0040000	0.0500000
71	High Variance Error	Unbalanced Groups	norm	linear	0	norm_re	full	linear_homo	2_skew	linear_norm	standard	1.0000000	0.5030000	0.5170000	0.0510000
72	High Variance RE	Unbalanced Groups	norm	linear	0	norm_re	full	linear_homo	2_skew	linear_norm	standard	1.0000000	0.1940000	0.0050000	0.0490000
73	High Variance Error	Same-sized Groups	norm	linear	0	norm_re	full	linear_homo	4_skew	linear_norm	standard	1.0000000	0.6150000	0.6970000	0.0480000
74	High Variance RE	Same-sized Groups	norm	linear	0	norm_re	full	linear_homo	4_skew	linear_norm	standard	1.0000000	0.2720000	0.0040000	0.0520000
75	High Variance Error	Balanced Groups	norm	linear	0	norm_re	full	linear_homo	4_skew	linear_norm	standard	1.0000000	0.6150000	0.5960000	0.0430000
76	High Variance RE	Balanced Groups	norm	linear	0	norm_re	full	linear_homo	4_skew	linear_norm	standard	1.0000000	0.2500000	0.0040000	0.0460000
77	High Variance Error	Unbalanced Groups	norm	linear	0	norm_re	full	linear_homo	4_skew	linear_norm	standard	1.0000000	0.5210000	0.6370000	0.0500000
78	High Variance RE	Unbalanced Groups	norm	linear	0	norm_re	full	linear_homo	4_skew	linear_norm	standard	1.0000000	0.2110000	0.0060000	0.0460000
79	High Variance Error	Same-sized Groups	norm	linear	0	norm_re	full	linear_homo	8_skew	linear_norm	standard	1.0000000	0.6530000	0.7930000	0.0380000
80	High Variance RE	Same-sized Groups	norm	linear	0	norm_re	full	linear_homo	8_skew	linear_norm	standard	1.0000000	0.3160000	0.0010000	0.0520000
81	High Variance Error	Balanced Groups	norm	linear	0	norm_re	full	linear_homo	8_skew	linear_norm	standard	1.0000000	0.6300000	0.7440000	0.0470000
82	High Variance RE	Balanced Groups	norm	linear	0	norm_re	full	linear_homo	8_skew	linear_norm	standard	1.0000000	0.3030000	0.0030000	0.0560000
83	High Variance Error	Unbalanced Groups	norm	linear	0	norm_re	full	linear_homo	8_skew	linear_norm	standard	1.0000000	0.5170000	0.8020000	0.0520000
84	High Variance RE	Unbalanced Groups	norm	linear	0	norm_re	full	linear_homo	8_skew	linear_norm	standard	1.0000000	0.2710000	0.0030000	0.0520000
85	High Variance Error	Same-sized Groups	norm	linear	0	norm_re	full	sq_2	0_skew	linear_norm	standard	0.0490000	0.1780000	0.2170000	0.4470000
86	High Variance RE	Same-sized Groups	norm	linear	0	norm_re	full	sq_2	0_skew	linear_norm	standard	0.1740000	0.2320000	0.0020000	0.7860000
87	High Variance Error	Balanced Groups	norm	linear	0	norm_re	full	sq_2	0_skew	linear_norm	standard	0.0480000	0.1570000	0.1870000	0.4390000
88	High Variance RE	Balanced Groups	norm	linear	0	norm_re	full	sq_2	0_skew	linear_norm	standard	0.1710000	0.2580000	0.0030000	0.7800000
89	High Variance Error	Unbalanced Groups	norm	linear	0	norm_re	full	sq_2	0_skew	linear_norm	standard	0.0500000	0.1810000	0.1680000	0.4570000
90	High Variance RE	Unbalanced Groups	norm	linear	0	norm_re	full	sq_2	0_skew	linear_norm	standard	0.2140000	0.2910000	0.0080000	0.7850000
91	High Variance Error	Same-sized Groups	norm	linear	0	norm_re	full	sq_4	0_skew	linear_norm	standard	0.0540000	0.2920000	0.2950000	0.4020000
92	High Variance RE	Same-sized Groups	norm	linear	0	norm_re	full	sq_4	0_skew	linear_norm	standard	0.2090000	0.2560000	0.0060000	0.6410000
93	High Variance Error	Balanced Groups	norm	linear	0	norm_re	full	sq_4	0_skew	linear_norm	standard	0.0590000	0.2610000	0.2860000	0.3930000
94	High Variance RE	Balanced Groups	norm	linear	0	norm_re	full	sq_4	0_skew	linear_norm	standard	0.1690000	0.2480000	0.0010000	0.6500000
95	High Variance Error	Unbalanced Groups	norm	linear	0	norm_re	full	sq_4	0_skew	linear_norm	standard	0.0530000	0.3120000	0.2470000	0.4090000
96	High Variance RE	Unbalanced Groups	norm	linear	0	norm_re	full	sq_4	0_skew	linear_norm	standard	0.2170000	0.2640000	0.0110000	0.6410000
97	High Variance Error	Same-sized Groups	norm	linear	0	norm_re	full	sq_8	0_skew	linear_norm	standard	0.0580000	0.4420000	0.5040000	0.3110000
98	High Variance RE	Same-sized Groups	norm	linear	0	norm_re	full	sq_8	0_skew	linear_norm	standard	0.1900000	0.2530000	0.0040000	0.4500000
99	High Variance Error	Balanced Groups	norm	linear	0	norm_re	full	sq_8	0_skew	linear_norm	standard	0.0550000	0.4800000	0.4880000	0.3190000
100	High Variance RE	Balanced Groups	norm	linear	0	norm_re	full	sq_8	0_skew	linear_norm	standard	0.2150000	0.2670000	0.0030000	0.4430000
101	High Variance Error	Unbalanced Groups	norm	linear	0	norm_re	full	sq_8	0_skew	linear_norm	standard	0.0640000	0.5000000	0.4630000	0.3220000
102	High Variance RE	Unbalanced Groups	norm	linear	0	norm_re	full	sq_8	0_skew	linear_norm	standard	0.2290000	0.2940000	0.0030000	0.4710000
103	High Variance Error	Same-sized Groups	norm	linear	0	norm_re	reduced	linear_homo	0_skew	linear_norm	standard	0.0445860	0.1019108	0.0700637	0.0445860
104	High Variance RE	Same-sized Groups	norm	linear	0	norm_re	reduced	linear_homo	0_skew	linear_norm	standard	0.3734177	0.1012658	0.0000000	0.0445860
105	High Variance Error	Balanced Groups	norm	linear	0	norm_re	reduced	linear_homo	0_skew	linear_norm	standard	0.0691824	0.0880503	0.0880503	0.0503145
106	High Variance RE	Balanced Groups	norm	linear	0	norm_re	reduced	linear_homo	0_skew	linear_norm	standard	0.3375796	0.1019108	0.0000000	0.0445860
107	High Variance Error	Unbalanced Groups	norm	linear	0	norm_re	reduced	linear_homo	0_skew	linear_norm	standard	0.0443038	0.1139240	0.1518987	0.0253165
108	High Variance RE	Unbalanced Groups	norm	linear	0	norm_re	reduced	linear_homo	0_skew	linear_norm	standard	0.3227848	0.1075949	0.0063291	0.0379747
109	High Variance Error	Same-sized Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	reduced_skew	standard	1.0000000	0.8150000	0.0740000	0.5320000
110	High Variance RE	Same-sized Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	reduced_skew	standard	1.0000000	0.5030000	0.0070000	0.9820000
111	High Variance Error	Balanced Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	reduced_skew	standard	1.0000000	0.4990000	0.5950000	0.1110000
112	High Variance RE	Balanced Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	reduced_skew	standard	0.9740000	0.2420000	0.0040000	0.5810000
113	High Variance Error	Unbalanced Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	reduced_skew	standard	0.9900000	0.4900000	0.7150000	0.0950000
114	High Variance RE	Unbalanced Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	reduced_skew	standard	0.7720000	0.2360000	0.0090000	0.3510000
115	High Variance Error	Same-sized Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	reduced_bimodal	standard	0.0480000	0.1020000	0.0790000	0.5030000
116	High Variance RE	Same-sized Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	reduced_bimodal	standard	0.9210000	0.1810000	0.0030000	0.9540000
117	High Variance Error	Balanced Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	reduced_bimodal	standard	0.0470000	0.1030000	0.6030000	0.1110000
118	High Variance RE	Balanced Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	reduced_bimodal	standard	0.1230000	0.1250000	0.0010000	0.4140000
119	High Variance Error	Unbalanced Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	reduced_bimodal	standard	0.0520000	0.0840000	0.6760000	0.0790000
120	High Variance RE	Unbalanced Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	reduced_bimodal	standard	0.0860000	0.1140000	0.0080000	0.2650000
121	High Variance Error	Same-sized Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	ar_error	0.0700000	0.1210000	0.0160000	0.0460000
122	High Variance RE	Same-sized Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	ar_error	0.0560000	0.1060000	0.0070000	0.0420000
123	High Variance Error	Balanced Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	ar_error	0.0500000	0.0940000	0.3130000	0.0420000
124	High Variance RE	Balanced Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	ar_error	0.0550000	0.1260000	0.0030000	0.0400000
125	High Variance Error	Unbalanced Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	ar_error	0.0360000	0.1270000	0.4480000	0.0460000
126	High Variance RE	Unbalanced Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	ar_error	0.1360000	0.1460000	0.0100000	0.0360000
127	High Variance Error	Same-sized Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	re_int	0.0440000	0.0920000	0.9980000	0.0420000
128	High Variance RE	Same-sized Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	re_int	0.0450000	0.1150000	1.0000000	0.0490000
129	High Variance Error	Balanced Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	re_int	0.0500000	0.1090000	0.9990000	0.0470000
130	High Variance RE	Balanced Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	re_int	0.0460000	0.1140000	1.0000000	0.0440000
131	High Variance Error	Unbalanced Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	re_int	0.0480000	0.0920000	1.0000000	0.0470000
132	High Variance RE	Unbalanced Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	re_int	0.0540000	0.0930000	0.9990000	0.0490000
133	High Variance Error	Same-sized Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	time_seq	0.0800000	0.0930000	0.5530000	0.0470000
134	High Variance RE	Same-sized Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	time_seq	1.0000000	0.0890000	0.0070000	0.0910000
135	High Variance Error	Balanced Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	time_seq	0.0430000	0.1020000	0.6490000	0.0570000
136	High Variance RE	Balanced Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	time_seq	0.0760000	0.0850000	0.0040000	0.0680000
137	High Variance Error	Unbalanced Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	time_seq	0.0600000	0.1020000	0.6950000	0.0450000
138	High Variance RE	Unbalanced Groups	norm	linear	0	norm_re	full	linear_homo	0_skew	linear_norm	time_seq	0.0620000	0.1270000	0.0030000	0.0600000



Appendix C. The behaviors of residual quantities in Non-Linearity and Non-Normality Scenario. The skewness of errors is again more well captured than bimodality with low variance errors in C-1. C-2 and C-4 show the rates of heteroscedasticity and non-linearity alleviate as clusters become more unevenly distributed. Highly unbalanced cluster sizes, however, are associated with more severe non-normality of low variance random effects induced by non-linearity and non-normality of errors in C-3.



Appendix D. Behaviors of residual quantities when misspecifying autocorrelated errors, time variable and only random intercept. From D-1 and D-4, missing the time variable of longitudinal settings can induce problems of error term normality and linearity when random effects variances are higher. In D-2, higher chance of heteroscedasticity is detected with misspecified autocorrelated errors. Violations to random effect normality occur more frequently with unbalanced cluster sizes in D-3.

References

Bates, D., Chambers, J. & Hastie, T. (1992), *Statistical models in S*.

Bates, D. M. & Pinheiro, J. C. (1998), 'Computational methods for multilevel modelling', *University of Wisconsin, Madison, WI* pp. 1–29.

Bates, D., Machler, M., Bolker, B. & Walker, S. (2015), 'Fitting linear mixed-effects models using lme4', *Journal of Statistical Software* **67**(1).

Becker, M., Klößner, S. & Becker, M. M. (2021), 'Package 'pearsonds'', *Aust. NZJ Stat* **50**(2), 199–205.

Bolker, B. M. (2008), *Ecological models and data in R*, Princeton University Press.

Breusch, T. S. & Pagan, A. R. (1979), 'A simple test for heteroscedasticity and random coefficient variation', *Econometrica: Journal of the econometric society* pp. 1287–1294.

Cheng, J., Edwards, L. J., Maldonado-Molina, M. M., Komro, K. A. & Muller, K. E. (2010), 'Real longitudinal data analysis for real people: building a good enough mixed model', *Statistics in medicine* **29**(4), 504–520.

Gelman, A. & Hill, J. (2006), *Model checking and comparison*, Analytical Methods for Social Research, Cambridge University Press.

Gilmour, A. R., Thompson, R. & Cullis, B. R. (1995), 'Average information reml: an efficient algorithm for variance parameter estimation in linear mixed models', *Biometrics* pp. 1440–1450.

Gurka, M. J., Edwards, L. J. & Muller, K. E. (2011), 'Avoiding bias in mixed model inference for fixed effects', *Statistics in medicine* **30**(22), 2696–2707.

Haslett, J. & Hayes, K. (1998), 'Residuals for the linear model with general covariance structure', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60**(1), 201–215.

Heagerty, P. J. & Kurland, B. F. (2001), 'Misspecified maximum likelihood estimates and generalised linear mixed models', *Biometrika* **88**(4), 973–985.

Houseman, E. A., Ryan, L. M. & Coull, B. A. (2004), 'Cholesky residuals for assessing normal errors in a linear model with correlated outcomes', *Journal of the American Statistical Association* **99**(466), 383–394.

Hui, F. K. C., Müller, S. & Welsh, A. H. (2021), ‘Random effects misspecification can have severe consequences for random effects inference in linear mixed models’, *International Statistical Review* **89**(1), 186–206.

Jiang, J. (1998), ‘Asymptotic properties of the empirical blup and blue in mixed linear models’, *Statistica Sinica* pp. 861–885.

Koller, M. (2016), ‘robustlmm: an r package for robust estimation of linear mixed-effects models’, *Journal of statistical software* **75**(1), 1–24.

Laird, N. M. & Ware, J. H. (1982), ‘Random-effects models for longitudinal data’, *Biometrics* pp. 963–974.

Loy, A. & Hofmann, H. (2014), ‘Hlmdiag: A suite of diagnostics for hierarchical linear models in r’, *Journal of Statistical Software* **56**(1), 1–28.

Loy, A. & Hofmann, H. (2015), ‘Are you normal? the problem of confounded residual structures in hierarchical linear models’, *Journal of computational and graphical statistics* **24**(4), 1191–1209.

Loy, A., Hofmann, H. & Cook, D. (2017), ‘Model choice and diagnostics for linear Mixed-Effects models using statistics on street corners’, *Journal of computational and graphical statistics* **26**(3), 478–492.

Mahalanobis, P. C. (1936), On the generalized distance in statistics, National Institute of Science of India.

McCulloch, C. E. (1997), ‘Maximum likelihood algorithms for generalized linear mixed models’, *Journal of the American statistical Association* **92**(437), 162–170.

McLachlan, G. J. (1999), ‘Mahalanobis distance’, *Resonance* **4**(6), 20–26.

Morrell, C. H. & Brant, L. J. (2000), ‘Lines in random effects plots from the linear Mixed-Effects model’, *The American statistician* **54**(1), 1–4.

Pinheiro, J. C. & Bates, D. M. (2000), ‘Linear mixed-effects models: basic concepts and examples’, *Mixed-effects models in S and S-Plus* pp. 3–56.

Proust, C. & Jacqmin-Gadda, H. (2005), ‘Estimation of linear mixed models with a mixture of distribution for the random effects’, *Computer methods and programs in biomedicine* **78**(2), 165–173.

R Core Team, . (2021), 'R: A language and environment for statistical computing'.

Raudenbush, S. W. & Bryk, A. S. (2002), *Hierarchical linear models: Applications and data analysis methods*, Vol. 1, sage.

Roback, P. & Legler, J. (2021), *Beyond Multiple Linear Regression*.

Santos, J. & Singer, J. (2007), 'Residual analysis for linear mixed models', *Biometrical Journal: Journal of Mathematical Methods in Biosciences* **49**(6), 863–875.

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., Réale, D., Dochtermann, N. A., Garamszegi, L. Z. & Araya-Ajoy, Y. G. (2020), 'Robustness of linear mixed-effects models to violations of distributional assumptions', *Methods in ecology and evolution* **11**(9), 1141–1152.

Schützenmeister, A. & Piepho, H.-P. (2012), 'Residual analysis of linear mixed models using a simulation approach', *Computational statistics & data analysis* **56**(6), 1405–1416.

Singer, J. M., Rocha, F. M. M. & Nobre, J. S. (2017), 'Graphical tools for detecting departures from linear mixed model assumptions and some remedial measures', *International statistical review* **85**(2), 290–324.

Snijders, T. A. B. & Berkhof, J. (2008), Diagnostic checks for multilevel models, *in* J. d. Leeuw & E. Meijer, eds, 'Handbook of Multilevel Analysis', Springer New York, New York, NY, pp. 141–175.

Verbeke, G. & Lesaffre, E. (1996), 'A linear mixed-effects model with heterogeneity in the random-effects population', *Journal of the American Statistical Association* **91**(433), 217–221.