# Comparisons of R Package Outputs of GLS, LME, GEE functions based on Simulations

## UBC SPPH 501 Term Project

Chloe You
xinyuan.you@stat.ubc.ca

# Background

- Generalized least squares (GLS), linear mixed effect models (LME), and generalized estimating equations (GEE) are commonly used methods for fitting data with correlation
- However, the performance of these methods can vary depending on factors such as the sample size, model specifications, and algorithm implementation.
- Many packages in R: `nlme` , `lme4` , `geepack`, `glmmTMB` . etc

# Refresher I

- In the case of linear models, the GLS estimator of the coefficients can be considered a special case of GEE
- In both LME and GEE, we assume that the errors are correlated and that the covariance structure can be modeled using some set of unknown parameters.
- Both methods estimate these unknown parameters and use them to improve the estimation of the fixed effects.
- However, the difference lies in how they handle the random effects.

# Refresher II

- In GLS/GEE, we assume that the errors are correlated, but we do not explicitly model any random effects.
- Instead, we estimate the covariance matrix of the errors and use it to calculate weights for each observation.
- Different specifications of the covariance matrix of the error:
  - **Compound symmetry:** Equal variances and covariances between errors. Diagonal matrix with common value for off-diagonal elements.
  - Autoregressive: Covariance between errors decreases as time between them increases. Specific structure based on distance between time points.
  - Exchangeable: Equal variances and covariances between errors, but covariances may differ from compound symmetry. Symmetric matrix with common value for off-diagonal elements.
  - Unstructured: No restrictions on variances or covariances between errors. Full matrix with all elements being different.

# Refresher III

- In LME, we model both fixed and random effects.
- Fixed effects are similar to those in GLS, but the random effects capture the between-subject variation that is not explained by the fixed effects

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + b_i + \epsilon_{ij}$$

where $y_{ij}$ is the response variable for the $j$th observation in the $i$th group, $\beta_0$ and $\beta_1$ are the fixed intercept and slope, $x_{ij}$ is the predictor variable for the $j$th observation in the $i$th group, $b_i$ is the random intercept for the $i$th group, and $\epsilon_{ij}$ is the error term for the $j$th observation in the $i$th group.

# Objective

- Compare the outputs of R packages that implement GLS, LME, and GEE based on simulated scenarios.
- Simulation 1: generate data via a GLS model
- Simulation 2: generate data via a LME model
- Investigate the performance of these methods and their R implementations to recover the true coefficients and error terms, and examine the heterogeneity captured in the different models.
- Aim to provide insights into the strengths and limitations of these methods and inform the selection of appropriate methods for practical applications.

# Simulation 1

Fit a GLS on the CD4 data and extract the coefficients to generate data with n = 10, 50, 100. We will then try to fit GLS, GEE, and LME models from varying packages to get the coefficients and error terms.

$$Y_{ij} = \beta_0 + \beta_1 Time_{ij} + \epsilon_{ij}, \epsilon \sim N(0, \Sigma)$$

$$\Sigma = \sigma^2(\rho J + (1-\rho)I)$$

$Y_{ij}$ represents the CD4 count for the $i$th individual at time $j$ , and $Time_{ij}$ represents the time (in weeks) for the $i$th individual at time $j$. $\epsilon$

$\sigma^2$ is the variance of the errors, $\rho$ is the correlation coefficient between any two measurements taken at the same time for the same individual, $J$ is a matrix of ones, $I$ is an identity matrix. This is the compound symmetry structure that corresponds to a constant correlation.

# Simulation 2

Use coefficients estimated via a random intercept linear mixed effect model. Similarly, we will generate data with sample sizes n = 10, 50, 100, and fit GLS, GEE, LME models to recover the coefficients.

$$Y_{ij} = \beta_{0i} + \beta_1 Time_{ij} + \epsilon_{ij}$$

where $\beta_{0i}$ represents the random intercept for the $i$th individual and $\beta_1$ represents the fixed effect of time. $\epsilon_{ij}$ represents the residual error for the $i$th individual at time $j$, assumed to be normally distributed with mean 0 and constant variance $\sigma^2$. We assume that $\beta_{0i} \sim N(0, \tau^2)$, and that the random intercepts for different individuals are independent of each other.

# Fit Models

```
# gls using nlme package
gls <- gls(log_CD4 ~ Week,
           correlation = corCompSymm(form = ~ 1 | ID))
# LME using nlme package
lme_nlme <- nlme::lme(log_CD4 ~ Week,
                      random = ~1|ID)
# LME using lme4 package
lme_lme4 <- lme4::lmer(log_CD4 ~ Week + (1|ID))
# GEE using geepack package
gee_geepack <- geeglm(log_CD4 ~ Week, id = ID)
# generalized linear mixed model using glmmTMB package
glmm_glmmTMB <- glmmTMB(log_CD4 ~ Week + (1|ID))
```

# Results: Simulation 1

Table 1: Parameter estimates on GLS-generated data

| Sample Size | Model | FE intercept | FE Week | RE sd | Error sd |
|---|---|---|---|---|---|
| n=10 | GLS (nlme) | 2.65 | 0.21 | | 1.00 |
| | LME (nlme) | 2.65 | 0.21 | 0.44 | 0.90 |
| | LME (lme4) | 2.65 | 0.21 | 0.44 | 0.90 |
| | GEE (geepack) | 2.65 | 0.21 | | 0.96 |
| | GLMM (glmmTMB) | 2.65 | 0.21 | 0.40 | 0.87 |
| n=50 | GLS (nlme) | 3.27 | -0.13 | | 0.94 |
| | LME (nlme) | 3.27 | -0.13 | 0.08 | 0.94 |
| | LME (lme4) | 3.27 | -0.13 | 0.08 | 0.94 |
| | GEE (geepack) | 3.27 | -0.13 | | 0.93 |
| | GLMM (glmmTMB) | 3.27 | -0.13 | 0.06 | 0.93 |
| n=100 | GLS (nlme) | 2.97 | 0.02 | | 0.93 |
| | LME (nlme) | 2.97 | 0.02 | 0.08 | 0.92 |
| | LME (lme4) | 2.97 | 0.02 | 0.08 | 0.92 |
| | GEE (geepack) | 2.97 | 0.02 | | 0.92 |
| | GLMM (glmmTMB) | 2.97 | 0.02 | 0.07 | 0.92 |

True parameters are $\sigma = 0.938, \beta_0 = 3.06, \beta_1 = -0.008$.

# Results: Simulation 2

Table 2: Parameter estimates on LME-generated data

| Sample Size | Model | FE intercept | FE Week | RE sd | Error sd |
|---|---|---|---|---|---|
| n=10 | GLS (nlme) | 3.19 | -0.07 | | 0.63 |
| | LME (nlme) | 3.19 | -0.07 | 0.41 | 0.48 |
| | LME (lme4) | 3.19 | -0.07 | 0.41 | 0.48 |
| | GEE (geepack) | 3.19 | -0.07 | | 0.60 |
| | GLMM (glmmTMB) | 3.19 | -0.07 | 0.38 | 0.47 |
| n=50 | GLS (nlme) | 3.25 | -0.01 | | 0.89 |
| | LME (nlme) | 3.25 | -0.01 | 0.68 | 0.58 |
| | LME (lme4) | 3.25 | -0.01 | 0.68 | 0.58 |
| | GEE (geepack) | 3.25 | -0.01 | | 0.88 |
| | GLMM (glmmTMB) | 3.25 | -0.01 | 0.67 | 0.57 |
| n=100 | GLS (nlme) | 2.99 | 0.03 | | 0.86 |
| | LME (nlme) | 2.99 | 0.03 | 0.68 | 0.54 |
| | LME (lme4) | 2.99 | 0.03 | 0.68 | 0.54 |
| | GEE (geepack) | 2.99 | 0.03 | | 0.86 |
| | GLMM (glmmTMB) | 2.99 | 0.03 | 0.67 | 0.53 |

True parameters we're trying to estimate is

$\beta_0 = 3.063, \beta_1 = -0.008, \tau = 0.7626, \sigma = 0.547$

# Results: Variance from Simulation 1

Table 3: Variance estimated from GLS-generated data

| | Sample Size | Model | FE intercept | FE Week | RE sd | Error sd |
|---|---|---|---|---|---|---|
| Sample Size | Model | Random Effect | Error | Sum | | |
| | n=10 | GLS (nlme) | | 1.00 | 1.00 | |
| | | LME (nlme) | 0.19 | 0.80 | 1.00 | |
| | | LME (lme4) | 0.19 | 0.80 | 1.00 | |
| | | GEE (geepack) | | 0.93 | 0.93 | |
| | | GLMM (glmmTMB) | 0.16 | 0.76 | 0.93 | |
| | n=50 | GLS (nlme) | | 0.89 | 0.89 | |
| | | LME (nlme) | 0.01 | 0.88 | 0.89 | |
| | | LME (lme4) | 0.01 | 0.88 | 0.89 | |
| | | GEE (geepack) | | 0.87 | 0.87 | |
| | | GLMM (glmmTMB) | 0.00 | 0.87 | 0.87 | |
| | n=100 | GLS (nlme) | | 0.86 | 0.86 | |
| | | LME (nlme) | 0.01 | 0.85 | 0.86 | |
| | | LME (lme4) | 0.01 | 0.85 | 0.86 | |
| | | GEE (geepack) | | 0.85 | 0.85 | |
| | | GLMM (glmmTMB) | 0.00 | 0.85 | 0.85 | |

# Results: Variance from Simulation 2

Table 4: Variance estimated from LME-generated data

| | Sample Size | Model | FE intercept | FE Week | RE sd | Error sd |
|---|---|---|---|---|---|---|
| Sample Size | Model | Random Effect | Error | Sum | | |
| | n=10 | GLS (nlme) | | 0.40 | 0.40 | |
| | | LME (nlme) | 0.17 | 0.23 | 0.40 | |
| | | LME (lme4) | 0.17 | 0.23 | 0.40 | |
| | | GEE (geepack) | | 0.36 | 0.36 | |
| | | GLMM (glmmTMB) | 0.15 | 0.22 | 0.36 | |
| | n=50 | GLS (nlme) | | 0.79 | 0.79 | |
| | | LME (nlme) | 0.46 | 0.33 | 0.79 | |
| | | LME (lme4) | 0.46 | 0.33 | 0.79 | |
| | | GEE (geepack) | | 0.77 | 0.77 | |
| | | GLMM (glmmTMB) | 0.44 | 0.33 | 0.77 | |
| | n=100 | GLS (nlme) | | 0.74 | 0.74 | |
| | | LME (nlme) | 0.46 | 0.29 | 0.74 | |
| | | LME (lme4) | 0.46 | 0.29 | 0.74 | |
| | | GEE (geepack) | | 0.74 | 0.74 | |
| | | GLMM (glmmTMB) | 0.45 | 0.29 | 0.74 | |

# Discussion

- Estimated fixed effect was consistent across the different methods, despite slight variations in the standard deviation of the random effects for the same sample size.
- LME captures the variance through the random effect's covariance matrix , thus highlighting the importance of selecting an appropriate model and considering the data's characteristics to obtain accurate estimates for variances.
- The sum of the random effect variance and error variance was uniform across all methods, and the sum was in proximity to the true variance used to generate the data.

# Conclusion

Overall, the findings suggest that the models used in this study effectively estimated the parameters, and the selection of the appropriate model and method is crucial in accurately estimating variances. Future research could explore the performance of different packages and models for other types of data and outcomes.

Detailed analysis can be found via https://chloeyou.github.io/gls-gee-lme-comparisons/analysis/analysis.html