SPPH 501 Term Project Proposal

Comparisons of R Package Outputs of Generalized Least Squares, Mixed Effect Models, and Generalized Estimating Equation based on Simulations

Chloe You
xinyuan.you@stat.ubc.ca

Background

Generalized least squares (GLS), generalized estimating equations (GEE), and linear mixed effects (LME) models are commonly used methods for estimating associations in clustered or correlated data to make unbiased statistical inferences. Fixed effects are population parameters that are constant across individuals, while random effects are individual-specific parameters that vary randomly across units and are modeled as random variables to account for residual heterogeneity not explained by fixed effects. LME models allow for both fixed and random effects, while GEE models allow only for fixed effects. In the case of linear models, the GLS estimator of the coefficients can be considered a special case of GEE (Fitzmaurice et al., 2012).

There are several software packages that can fit GLS, GEE, and LME models, including SAS, R, WinBUGS, and SPSS. R is the most popular software for these models because it is open-source and has a vast repository of packages. For example, the `geepack` package is popular for fitting GEEs and allows for a variety of correlation structures and link functions. Other commonly used packages for fitting GEEs include `gee` and `geeglm`. Popular R packages for fitting GLS models include `gls`, `geeglm`, `nlme`, and `lme4`. Some packages support fitting more than one of the three approaches and are listed multiple times. Common packages for fitting mixed-effect models include `lme4` and `nlme`.

Objective

Although R is popular for these models, the implementation of the functions varies across packages and the interpretation of the results also varies based on the model and package used. This project aims to compare popular GLS, GEE, and LME packages in R when applied to simulated data. The goal is to recover the true coefficients and error terms and to model the heterogeneity captured in the different models. The project should help to gain a better understanding of the different models and packages in R.

Anticipated Data

Two simulation settings will be created: one that estimates a random intercept mixed-effect model and one with a random intercept and random slope mixed-effect model. Log-transformed CD4 counts data from an AIDS clinical trial (Henry et al., 1998) on 1309 patients will be used to fit a linear mixed-effect model via `lme4` with random intercepts:

$$y_{ij} = \beta_{0i} + \beta_1 t_{ij} + \epsilon_{ij}, \quad i = 1, \dots, M, \quad j = 1, \dots, n_i \tag{1}$$

Where M is the number of individuals in the study, y_{ij} is the observed CD4 count for time observation j for individual i, t_{ij} is the week of observation, and ϵ_{ij} are independent $\mathcal{N}(0, \sigma_\epsilon^2)$. The random intercept can be expressed as

$$\beta_{0i} = \beta_0 + b_{0i}, \quad b_{0i} \sim \mathcal{N}(0, \sigma_0^2)$$

Next, a linear mixed-effect model via `lme4` with a form that allows the rate of CD4 depletion to vary across individuals will be fit:

$$y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + \epsilon_{ij}, \quad i = 1, \ldots, M, \quad j = 1, \ldots, n_i \tag{2}$$

Where y_{ij} is the observed CD4 count for time observation j for individual i, t_{ij} is the week of observation, and ϵ_{ij} are independent $\mathcal{N}(0, \sigma_\epsilon^2)$. The random intercept and random slope also can be expressed as

$$\beta_{0i} = \beta_0 + b_{0i}, \quad b_{0i} \sim \mathcal{N}(0, \sigma_0^2)$$

$$\beta_{1i} = \beta_1 + b_{1i}, \quad b_{1i} \sim \mathcal{N}(0, \sigma_1^2)$$

Based on the estimated $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}_0^2$, $\hat{\sigma}_1^2$, $\hat{\sigma}_\epsilon^2$ extracted from the `lme4` fits, simulated data will be generated by Equation 1 and Equation 2 separately.

Anticipated Methods

We fit 2 models, LME and GLS to the data generated by Equation 1 and compare the estimated coefficients and standard errors to the oracle. Likewise, we perform the same procedure to the data generated by Equation 2.

For data generated by Equation 1, we compare the following models:

1. Random intercept mixed-effect model

   $$y_{ij} = \beta_{0i} + \beta_1 t_{ij} + \epsilon_{ij}, \quad \beta_{0i} = \beta_0 + b_{0i}, \quad b_{0i} \sim \mathcal{N}(0, \sigma_0^2), \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

   We will fit it using the ‘lme4‘ and ‘nlme‘ package in R.

2. GLS

   $$y_{ij} = \beta_0 + \beta_1 t_{ij} + \epsilon_{ij}$$

   We will fit it using ‘gls‘, ‘geeglm‘, ‘nlme‘, ‘lme4‘ in R.

For data generated by Equation 2, we compare the following models:

1. Random intercepts and random slopes for t_{ij}

   $$y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + \epsilon_{ij}$$

   We will fit it using the ‘lme4‘ and ‘nlme‘ package in R.

2. GLS

   $$y_{ij} = \beta_0 + \beta_1 t_{ij} + \epsilon_{ij}$$

   We will fit it using ‘gls‘, ‘geeglm‘, ‘nlme‘, ‘lme4‘ in R.

Brief Reasons for the Anticipated Methods

We aim to compare the fixed-effect term estimates of both the Generalized Least Squares (GLS) and the Linear Mixed Effects (LME) methods, to assess their ability to recover the true coefficients and to determine the extent of their differences from each other and from the oracle. Additionally, we want to examine the heterogeneity captured in both models. While the LME model accounts for more sources of heterogeneity, the GLS model only models a single source. As we are aware of the true heterogeneity, we aim to evaluate the similarity of the values obtained from both models.

References

- Wang, Wei, and Michael O. Harhay. "A comparative study of R functions for clustered data analysis." Trials 22 (2021): 1-8.
- Hubbard, Alan E., et al. "To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health." Epidemiology (2010): 467-474.
- Fitzmaurice, Garrett M., Nan M. Laird, and James H. Ware. Applied longitudinal analysis. Vol. 998. John Wiley & Sons, 2012.
- Henry, Keith, et al. "A randomized, controlled, double-blind study comparing the survival benefit of four different reverse transcriptase inhibitor therapies (three-drug, two-drug, and alternating drug) for the treatment of advanced AIDS." JAIDS Journal of Acquired Immune Deficiency Syndromes 19.4 (1998): 339-349.