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

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`. Popular R packages for fitting GLS models include `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 There will be 2 simulation settings. For the first simulation, we will 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. The second simulation will 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.

Anticipated Methods We fit multiple models, LME, GLS, and GEE from various packages to the data generated in the two simulation settings and compare the estimated coefficients and standard errors to the oracle. We will examine the bias of the estimated parameters. Particularly, we will compare the variances of the LME models to the GLS/GEE models where variance can be captured by the random effect, additional to the error term.

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.