Manuscript

Sergio Olmos Pardo

22/3/2018

Abstract

This project presents an assessment of LMM and GEE in the context of cluster-correlated data analysis. Simulation studies will be performed to asses the performance of these two methods for different hierarchical structures (see Hallgren 2013; Bruyndonckx, Hens, and Aerts 2018; Li and McKeague 2013). More specifically, the stability of parameter estimates, confidence interval coverage and F test performance will be evaluated for different number of clusters and different sample sizes in each cluster. The impact of ignoring the correlation structure within clusters could also be evaluated.

Keywords: cluster data, linear mixed models, generalized estimating equations

Contents

Introduction

Context and Motivation of the Project

General Description

When data have a clustered/grouped structure, measurements on units within a cluster are in general more similar than measurements on units in different clusters. Statistical models for this type of data must account for the intra-cluster correlation at each level, otherwise inferences may be misleading.

There are several methods for analyzing multilevel data, which require different sets of assumptions and their adequacy differs depending on the structure of the data. The two most commonly used regression methods are linear mixed models (LMM) and generalized estimating equations (GEE). There are others, such as conjugate generalized linear mixed models (Lee, Green, and Ryan 2017).

This project presents an assessment of LMM and GEE in the context of cluster-correlated data analysis. Simulation studies will be performed to asses the performance of these two methods for different hierarchical structures (see Hallgren 2013; Bruyndonckx, Hens, and Aerts 2018; Li and McKeague 2013). More specifically, the stability of parameter estimates, confidence interval coverage and F test performance will be evaluated for different number of clusters and different sample sizes in each cluster. The impact of ignoring the correlation structure within clusters could also be evaluated.

Furthermore, real data sets will be analyzed following the guidelines suggested by the simulation study. An R Markdown template will be created and used to report the analyses of these data sets. The functions used in these analyses for graphical displays in the exploratory and modeling phases will be provided. The data, report template and code will be provided as an R package.

Motivation

The fields where this type of data arise are numerous, from the social sciences to the natural sciences. Examples of clustered data are multi-center clinical trials where measurements on patients are nested within clinics, toxicity studies where repeated measurements are obtained from a single individual at different times or agricultural experiments with complex experimental designs. Knowing what method works best with a given data set is not straightforward.

A review of the literature suggests that the two most common methods for analyzing hierarchical data are LMM and GEE. Each method has advantages and drawbacks. Linear mixed models require additional assumptions beyond those of classical regression, which can be difficult to verify (Alan E. Hubbard 2010). On the other hand, GEE can be problematic in small sample settings and in unbalanced designs (Rogers and Stoner 2015). The present project aims to provide some guidelines for researchers trying to analyze multilevel data.

Objectives

General Objectives

1. Simulation-based evaluation of the linear mixed model and generalized estimating equation approach under different hierarchical structures in the data.
2. Analysis of two or three real data sets with a hierarchical structure from different fields in science with the methods previously assessed.
3. Creation of an R package with all the data, R Markdown templates and code used.

Approach and Methods

The project will extensively use the R programming environment to accomplish the proposed objectives. R is a free and open source programming language specifically designed for statistical computation and graphics (Hornik 2017), with hundreds of packages that perform specific tasks. In the context of cluster-correlated data, the packages `nlme` and `lme4` provide two frameworks to fit linear mixed models, while the `geepack` package implements the generalized estimating equations approach.

Moreover, the R programming environment provides a workflow for reproducible research. As opposed to point-and-click statistical software, in R it is easier for independent researchers to reproduce the analysis since it weaves the principle of reproducibility throughout the entire project. In particular, with self-contained R packages it is straightforward to share and reproduce the analysis performed, by attaching the data and source code into a single file that can be loaded within the R environment.

Since a large amount of code will be written, version control will be used. Git and Github will help keep track of changes in the project and correct potential bugs in the code (see Bryan 2017).

Regarding the main topic of the project, a more theoretical approach in the assessment of LMM and GEE could be adopted, but seems to be out of reach given the time constraint and the expected scope of the project. Instead, a simulation approach will be used, taking advantage of the R programming environment.

For the simulation study the performance of the fitted models are assessed using the following performance characteristics as described in Bruyndonckx, Hens, and Aerts (2018):

- Relative difference between the mean of the parameter estimates and the true parameters.
- Relative difference between the mean estimated standard error and the empirical standard error, where the estimated standard error represents the variability within simulations and the empirical standard error represents the variability between simulations.

- Coverage of the confidence interval, calculated as the percentage of times the true parameter falls within the estimated 95% Wald confidence interval.
- Stability of the F test by comparing the number of times the null hypothesis was rejected under different hierarchical structures.

Background

Cluster-correlated data arise

Alan E. Hubbard, Nancy L. Fleischer, Jennifer Ahern. 2010. "To Gee or Not to Gee: Comparing Population Average and Mixed Models for Estimating the Associations Between Neighborhood Risk Factors and Health." *Epidemiology* 21: 467–74. doi:10.1097/EDE.0b013e3181caeb90.

Bruyndonckx, Robin, Niel Hens, and Marc Aerts. 2018. "Simulation-based Evaluation of the Linear-mixed Model in the Presence of an Increasing Proportion of Singletons." *Biometrical Journal* 60 (1): 49–65. doi:10.1002/bimj.201700025.

Bryan, Jennifer. 2017. "Excuse Me, Do You Have a Moment to Talk About Version Control?" *PeerJ Preprints*. https://doi.org/10.7287/peerj.preprints.3159v2.

Hallgren, Kevin A. 2013. "Conducting Simulation Studies in the R Programming Environment." *Tutorials in Quantitative Methods for Psychology* 9 (2): 43–60. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4110976/.

Hornik, Kurt. 2017. "R FAQ." https://CRAN.R-project.org/doc/FAQ/R-FAQ.html.

Lee, J. Y. L., P. J. Green, and L. M. Ryan. 2017. "Conjugate generalized linear mixed models for clustered data." *ArXiv E-Prints*, September.

Li, Zhigang, and Ian W McKeague. 2013. "Power and Sample Size Calculations for Generalized Estimating Equations via Local Asymptotics." *Statistica Sinica* 23 (1): 231–50. doi:10.5705/ss.2011.081.

Rogers, Paul, and Julie Stoner. 2015. "Modification of the Sandwich Estimator in Generalized Estimating Equations with Correlated Binary Outcomes in Rare Event and Small Sample Settings." *American Journal of Applied Mathematics and Statistics* 3 (6): 243–51. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4793734/.