Thesis proposal

PEC 0

Sergio Olmos Pardo

24/2/2018

Methods for Analyzing Cluster-Correlated Data

Keywords: `clustered-correlated data`, `multilevel/hierarchical models`, `regression`, `linear mixed models`, `general estimating equations`.

Topic

Clustered data presents a hierarchical/multilevel structure where observations are grouped or nested within different clusters. Clustering can be due to a naturally occuring hierarchy in the target population or a consequence of study design (Fitzmaurice 2005). The fields where this type of data arise are numerous. Examples of clustered data are multi-center clinical trials where measurements on patients are nested whithin clinics or toxicity studies where repeated measurements are obtained from a single individual at different times.

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. Statistcal 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 adecuacy differs depending on the structure of the data. The two most commonly used regression methods are linear mixed models and generalized estimating equations. There are others, such as conjugate generalized linear mixed models (Jarod Y.L. Lee 2017).

Several statistical packages for the implementation of these methods exist. The R packages `nlme`, `lme4` and `geepack` provide a powerful and easy-to-use framework to work with these methods.

Due to the wide applicability of these statistical models in the context of clustered data, a review of the different methods is in order.

Line of Research

Knowing what method works best with a given dataset is not straightforward. We will focus our analysis in the two most commonly used statistical models for clustered data, namely, linear mixed models and generalized estimating equations. We can explore how the asymptotic properties of the different statistical models depend on the structure of the data: number of clusters and sample size in each cluster (Li and McKeague 2013). Also, we can study the effects of ignoring the correlation structure within a cluster on our results. Simulation studies can provide some guidelines in the model specification process.

Furthermore, better graphical displays could be helpful in this context. This includes exploratory data analysis of clustered data as well as results of the modeling process.

Finally, a few example datasets from real studies can be used to illustrate all of the above.

Objectives

- Review the different statistical methods that can be used to model clustered data.
- Study the asymptotic properties of linear mixed models and the generalized estimating equations approach under different number of clusters and different sample sizes in each cluster through simulation.
- Study the effects of ignoring the correlation structure within clusters on results.
- Give some guidelines and recomendations for the model specification process.
- Better graphical displays in R for the exploration and modeling of clustered data.
- Create an R Markdown template for reporting results of clustered data analyses.
- Analyze a few examples of real clustered data.

References

Fitzmaurice, Garret. 2005. "Overview of Methods for Analyzing Cluster-Correlated Data." Boston: Harvard School of Public Health; https://catalyst.harvard.edu/docs/biostatsseminar/Fitzmaurice_BSP-Workshop-Slides.pdf.

Jarod Y.L. Lee, Louise M. Ryan, Peter J. Green. 2017. "Conjugate Generalized Linear Mixed Models for Clustered Data." http://arxiv.org/abs/arXiv:1709.06288.

Li, Zhigang, and Ian W. McKeague. 2013. "Power and Sample Size Calculations for Generalized Estimating Equations via Local Asymptotics." http://www3.stat.sinica.edu.tw/statistica/j23n1/j23n111/j23n111.html.