Yaxu Zhuang

October 20, 2021

Purpose of Requesting the REST Study Data from the PCRC Repository

This dataset will be used for two purposes: 1) to explore novel statistical methods for clustering longitudinal data; 2) to satisfy the requirements for a longitudinal data analysis course in the Colorado School of Public Health, BIOS 6643.

**Title: “Clustering Methods for Linear Mixed Models”**

Traditional methods of multi-level linear mixed model often include a “organization” as highest level such as schools, medical centers, or hospitals. Although some people in these clusters may share similar unmeasured characteristics (e.g. socioeconomic status, distance to care, food insecurity/access), there may be other features that would make them similar to groups other than their organization or institution.

Here, I describe “Clustering LMM” a new method to generate a highest level for a multi-level linear mixed model, by reassigning people into different clusters based on demographic and socioeconomic variables available in the data. Below, I describe the potential benefits of this method.

First, a “representative” subject can be used several times and be assigned to different groups. This method of clustering, also called “soft clustering” could reduce residual error in the estimate of the treatment effect by reducing the size of the random variation term(s).

Secondly, clustering methods are flexible on which/how many variables to generate clusters. For example, if investigators want to classify subjects by socioeconomic status, then several representative economic variables could be used to generate clusters, such as income, education level, and insurance status. Alternatively, investigators can group subjects into different risk groups (for example, using performance status, functional status, comorbidities, etc). Furthermore, clusters can be generated by vital statistics, such as blood pressure, heart rate and other variables associated with patient’s health status.

Finally, clustering methods can capture the possible relationships among patients at different “organizations”, and reorganizing clusters based on measured features, rather than somewhat arbitrary groups could result in efficiency gains when estimating effects of interventions.

In this project, I will apply both traditional multi levels LMM methods and “Clustering LMM” methods into this longitudinal data set. The models fitted by both methods will be compared by AIC score and standard error of the estimates.

There are three major hypotheses for this study:

* “Clustering LMM” methods have better performances (lower AIC score and smaller standard errors for estimates) than traditional LMM methods.
* The performance of “Clustering LMM” method depends on which/how many variables used to generate clusters and also number of clusters generated (how many groups to divide the entire cohort).
* Fuzzy clustering method will outperform K-mean clustering, because one representative subject is allowed to be in several different groups. There will be more subjects in fuzzy clustering LMM than K-means LMM. A more reliable estimate will be achieved by Fuzzy clustering LMM because of more subjects (data points) in each group. Thus, fuzzy clustering LMM might have lower standard error of estimates.

R 4.0.1 is used for data cleaning and generating clusters. Package “factoextra” and “cluster” are used for clustering analysis. Package “psych”, “data.table”, “CIDAtools”, “knitr” and “pander” are used for data cleaning and generating tables and figures. A processed data set with an extra variable “clusters” will then be import to SAS 9.4 for modeling. Proc mixed procedure is used to fit LMMs.