Cluster-Correlated Data

Sergio Olmos Pardo

3/4/2018

Clustered data are characterized as data that can be classified into a number of distinct groups or clusters within a particular study. Such data implies a hierarchical or multilevel structure, where the term "level" refers to the position of a unit of observation within a given hierarchy or group structure. By convention, the lowest level of the hierarchy is referred to as level 1. The outcome measure is always assessed at the lowest level, while the explanatory variables can be measured at any level of the hierarchy. Consequently, clustered data provide opportunities to explore, in greater depth, the interrelationships among variables at any level.

Clustering usually implies that measurements on units within a cluster are more similar than measurements on units in different clusters, implying that observations within a cluster are not independent and therefore standard data analysis methods may not be apropriate. Linear models (LM) and generalized linear models (GLMs) assume that the observations are conditionally independent given the specified fixed effects. Thus applying these methods to cluster-correlated data could lead to incorrect inferences (Fitzmaurice 2005; Ananth, Platt, and Savitz 2005).

Several regression methods that account for the intra-cluster correlation have been developed. These methods can be divided in three general approaches:

1. GLMs with fixed effects to account for clustering.
2. Multilevel Models
3. Generalized estimating equations

Clustered data can arise due to a naturally occurring hierarchy in the target population and/or a consequence of study design.

The group structure does not necessarily need to be nested; that is, level 1 units nested within level 2 units, nested within level 3 units and so on. Non-nested data can arise when units are characterized by overlapping attributes.

Ananth, Cande V., Robert W. Platt, and David A. Savitz. 2005. "Regression Models for Clustered Binary Responses: Implications of Ignoring the Intracluster Correlation in an Analysis of Perinatal Mortality in Twin Gestations." *Annals of Epidemiology* 15 (4). Elsevier: 293–301. doi:10.1016/j.annepidem.2004.08.007.

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.