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.

Clustered data can arise due to a naturally occurring hierarchy in the target population and/or a consequence of study design. In the social sciences, the group structure is often given, not designed into the study. A classic example is an observational study where student performance metrics are clustered within schools and covariates can belong to the student and/or school levels. Nevertheless, only examples of clustered data in the health and biological sciences will be considered in this paper. Examples of cluster-correlated data in these fields are:

1. Multicenter clinical trials, where measurements are taken on patients nested within different clinics or centers.
2. Longitudinal data, where multiple measurements are taken over time on each individual. Here the clusters are the different individuals.
3. Cluster randomized trials, where whole clinics are randomized to an intervention, as opposed to randomization at the subject level. Here, the clusters are formed of patients within clinic.
4. Multicenter longitudinal clinical trials, where repeated measurements are taken over time, nested within subjects, nested within clinics.
5. Agricultural experiments, where treatments are assigned to different pots, then multiple plants are grown in each pot and multiple measurements are taken over time for each plant. Here, repeated measurements on plants are taken over time, nested within plants, nested within pots.

From the examples above, we can distinguish between two types of clustered data, nested and non-nested. Suppose one is interested on the effect of different treatments. In a nested structure, only one of the treatments being compared is present in each cluster. Examples 2, 3, 4 and 5 above all have a nested hierarchical structure. On the other hand, in non-nested structures at least some of the clusters contain observations from different treatment groups. Examples 1 and 4 have non-nested structures.

Analysis of clustered data

Clustering usually implies that measurements on units within a cluster are more similar than measurements on units in different clusters. Thus, observations within a cluster are not independent, while observations from different clusters are. The degree of association between the observations within clusters is usually measured as the intra-cluster correlation (ICC), given by

$$ICC = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2},$$

where σ_B^2 is the variance between clusters and σ_W^2 is the variance within clusters.

Given the nature of cluster-correlated data, it seems innaproprioate to treat cluster data as if all observations are independent. The effect of using this approach will depend on the nature of the correlation structure.

Multiple simulation studies have assessed the consequences of ignoring clustering and found dramatic effects for many cluster structures (Galbraith, Daniel, and Vissel 2010; Ananth, Platt, and Savitz 2005; Bernard Rosner, Glynn, and Lee 2006a).

There are mainly four general approaches when analyzing cluster-correlated data:

1. Reducing clusters to independent observations.
2. Classical regression.
3. Adjusting existing tests to account for clustering.
4. Modeling approaches.

Reducing data to independent observations

This approach consists in reducing the multiple observations in a cluster to a single observation by taking a suitable summary statistic. Common summary statistics are the mean and the median. In situations where the measurements within a cluster are taken over time, it also makes sense to use the difference between the basal measurement and the final measurement. Given the nature of clustered data, where observations from different clusters can be considered independent, the obtained observations can be considered independent and analyzed with standard methods.

Although this is a valid approach, it comes with some serious limitations. Reducing all the observations in a cluster to a single observation, a great amount of information can be lost, resulting in a less powerful analysis. Furthermore, complications arise if there are unequal numbers of observations per cluster or if the group structure is not nested.

Classical regression

It is possible to account for clustering by including the grouping factors in a classical regression model as indicator variables. When the number of clusters is small and the structure is not nested, this method can be a reasonable approach (Galbraith, Daniel, and Vissel 2010; Senn 1998). However, if the number of clusters is large, including an indicator variable for each cluster could be problematic. Imagine a longitudinal study with thousands of patients, this approach would require to fit a classical regression model with thousands of indicator variables which may result in collinearity problems. Moreover, this approach does not allow the correlation structure of the data to be studied.

Adjusting existing tests to account for clustering

These methods modify existing parametric and non-parametric tests to account for clustering. Modifications of the Wilcoxon rank sum test (Mann-Whitney U test) have been developed by t-test and the χ^2 test have been proposed to adjust for clustering (see Gönen, Panageas, and Larson 2001). Rank sum tests that account for clustering have been developed by Rosner and Grove (1999), Rosner, Glynn, and Lee (2003), Datta and Satten (2005). Modifications of the signed-rank tests for paired data have also been developed (Bernard Rosner, Glynn, and Lee 2006b; Datta and Satten 2008).

Although these methods can perform as well as alternative methods in some situations, they offer less flexibility than the model-based approaches that will be introduced next.

Modeling approaches

Linear mixed models and generalized estimating equations are extensively used model-based methods in the analysis of clustered data. These two approaches will be specified in detail in the next two sections. Given

the limitations of the previous methods and the flexibility of the model based approach, only LMMs and GEEs will be considered in our analysis.

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.

Datta, Somnath, and Glen A Satten. 2005. "Rank-Sum Tests for Clustered Data." *Journal of the American Statistical Association* 100 (471). Taylor & Francis: 908–15. doi:10.1198/016214504000001583.

Datta, Somnath, and Glen A. Satten. 2008. "A Signed-Rank Test for Clustered Data." *Biometrics* 64 (2): 501–7. doi:10.1111/j.1541-0420.2007.00923.x.

Galbraith, Sally, James A. Daniel, and Bryce Vissel. 2010. "A Study of Clustered Data and Approaches to Its Analysis." *Journal of Neuroscience* 30 (32). Society for Neuroscience: 10601–8. doi:10.1523/JNEUROSCI.0362-10.2010.

Gönen, Mithat, Katherine S. Panageas, and Steven M. Larson. 2001. "Statistical Issues in Analysis of Diagnostic Imaging Experiments with Multiple Observations Per Patient." *Radiology* 221 (3): 763–67. doi:10.1148/radiol.2212010280.

Rosner, B., and D. Grove. 1999. "Use of the Mann–Whitney U-test for Clustered Data." *Statistics in Medicine* 18 (11): 1387–1400. doi:10.1002/(SICI)1097-0258(19990615)18:11<1387::AID-SIM126>3.0.CO;2-V.

Rosner, Bernard, Robert J. Glynn, and Mei-Ling T. Lee. 2006a. "Extension of the Rank Sum Test for Clustered Data: Two-Group Comparisons with Group Membership Defined at the Subunit Level." *Biometrics* 62 (4): 1251–9. doi:10.1111/j.1541-0420.2006.00582.x.

———. 2006b. "The Wilcoxon Signed Rank Test for Paired Comparisons of Clustered Data." *Biometrics* 62 (1): 185–92. doi:10.1111/j.1541-0420.2005.00389.x.

Rosner, Bernard, Robert J. Glynn, and Mei-Ling Ting Lee. 2003. "Incorporation of Clustering Effects for the Wilcoxon Rank Sum Test: A Large-Sample Approach." *Biometrics* 59 (4): 1089–98. doi:10.1111/j.0006-341X.2003.00125.x.

Senn, Stephen. 1998. "Some Controversies in Planning and Analysing Multi-centre Trials." *Statistics in Medicine* 17 (15-16): 1753–65. doi:10.1002/(SICI)1097-0258(19980815/30)17:15/16<1753::AID-SIM977>3.0.CO;2-X.