Wiley StatsRef:
Statistics Reference Online

# Informative Cluster Size

By *John M. Williamson*

**Keywords:** *clustered data, nonignorable cluster size*

**Abstract:** Correlated outcomes are collected in many areas of research and occur for a variety of reasons. Valid inference is reliant upon properly accounting for the correlation among outcomes within subjects. Analysis should also consider informative cluster size, which occurs when cluster size is related to the outcome of interest and is common in studies with periodontal, surgical outcome, and reproductive toxicology data, among others. Such methods of analysis are increasingly available.

Clustered, correlated data (*see* **Clustered Data**) are common in biomedical and public health studies. Such data need to accommodate the statistical dependence among the repeated observations (*see* **Repeated Measures**) within the clusters[1]. Failure to account for this correlation among responses can lead to incorrect inferences about the regression coefficients because of incorrect variance estimation[2]. This could lead to incorrect conclusions regarding the research questions. Numerous methods have been proposed for the regression analysis of dependent outcome data. Population-averaged (marginal) and cluster-specific models are the two main classes of approaches (*see* **Marginal Models for Clustered Data**), in addition to transition and latent class models. The former class of models is often fit with generalized estimating equations (GEEs)[3,4] (*see* **Generalized Estimating Equations: Introduction**) and the latter class of models with **Generalized Linear Mixed Models** that incorporate random effects[5,6] (*see* **Fixed-, Random-, and Mixed-Effects Models**) at the cluster level. Both approaches are commonly used in the medical and life sciences.

Let $Y_{ij}$ denote the response for the $j$th observation ($j = 1, \ldots, n_i$) in the $i$th cluster ($i = 1, \ldots, N$), where $N$ is the number of clusters and $n_i$ the size of the $i$th cluster. Let $\mathbf{X_{ij}}$ be the accompanying covariate vector for the $(i, j)$th observation. Informative cluster size (also referred to as *nonignorable cluster size*) occurs when the size of the cluster is related to the risk for the outcome of interest[7]. Formally, it can be defined as any violation of the property that $E(Y_{ij}|n_i, \mathbf{X_{ij}}) = E(Y_{ij}|\mathbf{X_{ij}})$.

In reproductive toxicology studies, toxicants are frequently administered to a mother rodent and the birth-defect status of the pups in her litter are observed. The resulting correlated data clustered at the litter level may exhibit informative cluster size if litters with fewer pups, owing to more fetal resorptions, tend to have more birth defects[7]. Volume-outcome studies are often used to evaluate whether hospitals or surgeons who treat many patients for a certain disease or condition have better outcomes than those who treat fewer patients[8]. The resulting may data may exhibit informative cluster size if the treatment outcome (e.g., postoperative complication and mortality) is related to the number of patients treated. Periodontal

Centers for Disease Control and Prevention, Atlanta, GA, USA


DOI: 10.1002/9781118445112.stat07887

disease studies where disease is ascertained at multiple teeth within each subject's mouth often exhibit informative cluster size as people who are more susceptible to periodontal disease may have already lost some teeth from the disease.

There have been numerous methods proposed for the analysis of informative cluster size data. There are two types of sampling inherent when analyzing clustered data with marginal modeling[7]. The first is unit-based sampling that is implicit in the usual marginal models such as GEE, and the second is cluster-based sampling where one selects a random observation from a randomly selected cluster. For the former, larger clusters are weighted more than smaller ones. For the latter, all clusters are given equal weight regardless of size and accordingly the marginal parameter will have a cluster-based interpretation. Asymptotically the two marginal analyses will reach the same conclusion if cluster size is unrelated to the outcome of interest. However, the two marginal analyses are different for informative cluster size data.

Hoffman *et al.*[7] were among the first to describe the problem of informative cluster size data and proposed a within-cluster resampling (WCR) method that remains valid for the analysis of such data when a marginal model is of interest. Inversely weighting the GEE score equation by cluster size with an independence working correlation matrix (CWGEE) has been shown to be asymptotically equivalent to WCR[9,10]. Neuhaus and McCulloch[11] address the analysis of informative cluster size data from a cluster-specific approach through the use of generalized linear mixed models. They demonstrate that maximum likelihood methods that ignore informative cluster sizes exhibit little bias in estimating covariate effects that are uncorrelated with the random effects associated with cluster sizes. Alternatively, estimates of covariate effects may be biased if the covariate effects are associated with the random effects.

## Related Article

**Clustered Data**

## References


[1] Hu, F.B., Goldberg, J., Hedeker, D., Flay, B.R., and Pentz, M.A. (1998) Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. *Am. J. Epidemiol.* **147**, 694–703.

[2] Diggle, P.J., Heagerty, P.J., Liang, K.Y., and Zeger, S.L. (2002) *Analysis of Longitudinal Data*, 2nd edn, Oxford University Press, Oxford.

[3] Liang, K.Y. and Zeger, S.L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

[4] Zeger, S.L. and Liang, K.Y. (1986) Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121–130.

[5] Laird, N.M. and Ware, J.H. (1982) Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.

[6] McCulloch, C.E., Searle, S.R., and Neuhaus, J.M. (2008) *Generalized Linear and Mixed Models*, 2nd edn, John Wiley & Sons, Inc., New York.

[7] Hoffman, E.B., Sen, P.K., and Weinberg, C.R. (2001) Within-cluster resampling. *Biometrika* **88**, 1121–1134.

[8] Panageas, K.S., Schrag, D., Localio, A.R., Venkatraman, E.S., and Begg, C.B. (2007) Properties of analysis methods that account for clustering in volume-outcome studies when the primary predictor is cluster size. *Stat. Med.* **26**, 2017–2035.

[9] Williamson, J.M., Datta, S., and Satten, G.A. (2003) Marginal analyses of clustered data when cluster size is informative. *Biometrics* **59**, 36–42.

[10] Benhin, E., Rao, J.N.K., and Scott, A.J. (2005) Mean estimating equation approach to analysing cluster-correlated data with nonignorable cluster sizes. *Biometrika* **92**, 435–450.

[11] Neuhaus, J.M. and McCulloch, C.E. (2011) Estimation of covariate effects in generalized linear mixed models with informative cluster sizes. *Biometrika* **98**, 147–162.




*Wiley StatsRef: Statistics Reference Online*, © 2014–2016 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat07887