Analysis of Clustered Data

Statnews #88

Cornell Statistical Consulting Unit

Created December 2013. Last updated September 2020

Introduction

When faced with the analysis of clustered or multilevel data many possible options are available for linear models. In this newsletter, we will review the currently popular methods and describe some the advantages and disadvantages of each approach.

Here are several common situations where data are clustered:

- nested or multilevel data e.g. test scores of students nested within schools
- longitudinal data e.g. data on the length of caterpillars recorded daily for 2 months
- repeated measures e.g. subjects speed is recorded when repeatedly performing tasks
- various experimental designs e.g. randomized complete block design, split plot design
- or a combination of any of the above

All such data violate the assumption of independence of observations that we typically make in regression models. This is particularly influential on the variability of the model estimates, often resulting in standard errors that are smaller than they should be and thus leading to incorrect inference. We describe here the five most common approaches to account for the clustering of data, along with their different advantages and caveats.

Statistical approaches

Clustered Robust Standard Errors

Description

This method of correcting for clustering within a data set involves adjusting the standard errors of model estimates, typically inflating them. These robust standard errors (also called Huber-White, Sandwich or Empirical standard errors) take into account the level of correlation of observations within a cluster, inflating the standard errors of the estimates accordingly.

Advantages

This is a simple implementation as it often only requires asking for robust standard errors while the model estimates remain unchanged. The results are no more complex than a typical regression model, making for easy interpretation while accounting for clustering.

Caveats

The number of clusters should not be too small when using this technique, but it is likely more robust in this regard than a mixed model. When data are highly imbalanced this method can encounter problems and should be avoided, particularly in longitudinal data. This solution is rather conservative. Only two-level models can be estimated, so if there are multiple layers of clustering, this approach will be insufficient.

Generalized Estimating Equations

Description

Population average models, or marginal models, are estimated by the method of Generalized Estimating Equations (GEE). These models describe changes in the mean population response given changes in predictors, while accounting for within-cluster correlation by appropriately adjusting the variance estimates of these coefficients. The variance estimation procedure is specified by a working correlation structure for the observations within a cluster. Common choices for the correlation structure include exchangeable, unstructured, autoregressive, and independence. Choosing the most appropriate correlation structure could be based on a priori hypotheses (e.g. autoregressive based on decaying correlation with distance) but could also be determined after examining the estimated variance-covariance matrices from several different specified structures. In some instances limitations of the data may determine the most appropriate structure, e.g. if a few clusters have many observations exchangeability may be best in order to capitalize on these large sample sizes when estimating the variance-covariance matrix.

Advantages

GEE have fewer distributional assumptions than mixed models, and are robust to mis-specification of the correlation structure. Robust standard errors can be obtained for GEE estimates.

Caveats

Only two-level models can be estimated. A key assumption for valid inference is that the number of clusters is sufficiently large, though there is no specific cutoff for when this is achieved.

Fixed Effects Models

Description

Fixed effects models include an indicator (or dummy) variable for each cluster relative to a reference cluster, which amounts to a within-subjects regression model. If the number of clusters

(n) is large, this amounts to the inclusion of many (i.e. n-1) indictors variables, which are often suppressed in the model output.

Advantages

Fixed effects models have the advantage of controlling most thoroughly for unmeasured characteristics of the clusters. This approach can also best handle data when there are very few clusters present in the data.

Caveats

If there are hypotheses concerning between-subject predictors, fixed effects models are less efficient because they do not utilize any between-group information, and the main effects parameters for between-subject predictors will not be estimated. In some cases this could effectively mean throwing away a lot of information. If the number of groups is very large, you lose many degrees of freedom to the n-1 indicator variables.

Mixed Models

Description

Mixed models assume that correlation within observed data can be represented by random effects for the unit at which the data clusters, with an assumed probability distribution. The fixed effects parameters and their standard errors are typically estimated through the method of maximum likelihood (or restricted maximum likelihood). Mixed models decompose the variance in the outcome into residual and cluster-to-cluster components.

Advantages

Multiple levels of clustering (e.g. 3-level models) can be modeled using this approach using multiple random effects, while other methods cannot.

Caveats

If the number of clusters in the data is very small, the mixed model method is not reliable. Mixed models assume independence of error terms from each other and from the independent variables. They are also criticized as having untestable model assumptions, namely on the distribution of the random effects.

Generalized Least Squares

Description

The method of generalized least squares (GLS) assumes a structure to the variance- covariance matrix (e.g. independence, autoregressive), which it uses in the estimation of model parameters and standard errors. This can be done in practice by estimating the covariance structure and plugging it in as weights in a weighted least squares model.

Advantages

GLS allows for efficient estimation of both within-subject and between-subject variability. It will yield similar results to the mixed model approach, but it does not rely on the assumption of normally distributed random effects. It is a weighted between and within estimator and as such is more efficient as it uses both within and between information.

Caveats

In practice the GLS method is difficult to implement, so certain assumptions must be imposed on the variance-covariance structure resulting in a feasible generalized least squares (FGLS) estimator, which results in lower efficiency. GLS assumes that the within and between estimate of the fixed effects are the same. There are limitations on the choice covariance structure as compared to other methods.

Conclusions

The methods presented here represent the most common classes of models for analyzing clustered data currently, but constitute only a broad overview of many possibilities. Most of these methods or some variation of these, can be used not only for linear models with normally distributed response but also for many non-normally distributed responses such as logistic and poisson. These methods will not produce exactly the same results and in some cases the results might even vary wildly. The interpretation of the results will also depend on the method that you used. Deciding which method you should use will depend on your research question, your data but also on the expectations of your field of study. If you have any questions regarding implementation or interpretation of the methods presented here, please contact the CSCU Office.