ARTICLE IN PRESS

Research in Developmental Disabilities xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Research in Developmental Disabilities

journal homepage: www.elsevier.com/locate/redevdis

Modelling correlated data: Multilevel models and generalized estimating equations and their use with data from research in developmental disabilities

Dimitrios Vagenas[a,*], Vasiliki Totsika[b,c]

[a] Institute of Health and Biomedical Innovation, Queensland University of Technology, Australia
[b] Centre for Educational Development, Appraisal, and Research (CEDAR) and Centre for Education Studies (CES), University of Warwick, UK
[c] Centre for Developmental Psychiatry and Psychology, Department of Psychiatry, Monash University, Australia

ARTICLE INFO

Number of reviews completed is 3
The paper is part of a special issue on Clustered/Longitudinal.

Keywords:
Clustered
Longitudinal
Multilevel
Generalized Estimating Equations
Repeated
Developmental disability

ABSTRACT

Background: The use of Multilevel Models (MLM) and Generalized Estimating Equations (GEE) for analysing clustered data in the field of intellectual and developmental disability (IDD) research is still limited.
Method: We present some important features of MLMs and GEEs: main function, assumptions, model specification and estimators, sample size and power. We provide an overview of the ways MLMs and GEEs have been used in IDD research.
Results: While MLMs and GEEs are both appropriate for longitudinal and/or clustered data, they differ in the assumptions they impose on the data, and the inferences made. Estimators in MLMs require appropriate model specification, while GEEs are more resilient to misspecification at the expense of model complexity. Studies on sample size seem to suggest that Level 1 coefficients are robust to small samples/clusters, with any higher-level coefficients less so. MLMs have been used more frequently than GEEs in IDD research, especially for fitting developmental trajectories.
Conclusions: Clustered data from research in the IDD field can be analysed flexibly using MLMs and GEEs. These models would be more widely used if journals required the inclusion of technical specification detail, simulation studies examined power for IDD study characteristics, and researchers developed core skills during basic studies.

What this paper adds?

Research data cease to be independent when a super-ordinate structure or repeated measurements create correlation amongst individual data points. Ignoring this correlation in model specification leads to a bias in standard errors that is proportionate to the magnitude and direction of the correlation. Multilevel models (MLM) and Generalized Estimating Equations (GEEs) model the data taking into account this correlation. The two approaches differ in the way they handle this correlation, and selecting between the two relies on the research aims and study characteristics. The paper discusses some of the core features of MLMs and GEEs for researchers who are considering how to analyse their longitudinal or clustered data. We review how IDD researchers have used the models so far, in the hope that other researchers will consider using them. We believe their use would be more widespread if researchers were taught these models as part of their studies, if journals required researchers to include more technical details on how MLM or GEE models were fitted, and if further research focused on examining how powerful these models could be for IDD studies that often rely on modest sample sizes, few clusters or large cluster to participant ratios.

[*] Corresponding author.
E-mail address: dimitrios.vagenas@qut.edu.au (D. Vagenas).

https://doi.org/10.1016/j.ridd.2018.04.010
Received 29 September 2017; Received in revised form 19 March 2018; Accepted 12 April 2018

1. Introduction

Research in the field of intellectual and developmental disabilities (IDD) often generates longitudinal and/or correlated data. One source of correlation comes from clustering such as data from mothers and children who share the same household; data from parents (couples) of children with IDD. Another often encountered source of correlation is repeated measurements obtained from a group of participants over time.

Researchers have three options when they analyse correlated data (e.g. longitudinal data on the same individuals, or data clustered within a hyper-ordinate structure): (1) ignore the correlation, (2) bypass it by withholding one part of the data, or (3) deal with the correlation using appropriate analytic techniques. The first two approaches are not really efficient or appropriate since they (1) either result in inappropriate inferences or (2) do not make full use of the data. Statistical expertise to deal with clustered and/or longitudinal data is required at an advanced level, one which often exceeds the training researchers in the IDD field may have. In addition, IDD research presents some unique challenges: (a) low prevalence of condition examined (typically resulting in small sample size), (b) often a high number of super-ordinate clusters (for example, a relatively small number of children within a large number of genetic syndromes). In addition, where research is applied and the focus is on informing educational or social policy, there is often an interest in drawing conclusions about the population with a particular need, and not about the way individuals' diagnostic labels/clinical services/educational settings are clustered. In other words, the clustering is not always part of the substantive research question. This often results in a higher than expected frequency of the first two options, i.e., either ignoring the clustering or bypassing it by not using part of the data.

With the present paper, we would like to encourage IDD researchers to use appropriate modelling techniques when having correlated data. We focus on two analytic techniques: Multilevel Models (MLM; also known as Mixed Models proposed by Laird & Ware, 1982) and Generalized Estimating Equations (GEE); proposed by Zeger and Liang (1986). Using non- technical terms, we will discuss: (i) why it is important to account for the correlation and (ii) what MLMs and GEEs do and how they could be used in research in IDD drawing on examples from published research in the field.

1.1. Generalized Linear Models (GLM)

The standard ANOVA and regression models are part of a bigger family called the Generalized Linear Models (GLMs). GLMs assume that one variable – referred to as the outcome or dependent variable – is explained by or depends on some other variables called the explanatory or independent variables. The outcome and explanatory variables are assumed to be related ("linked") with a function called the "link function". One of the purposes of a GLM is to estimate a unique combination of the explanatory variables which explain as much variation of the outcome variable as possible. This is done by estimating different weights called "regression coefficients". Up to this point the whole process is a mathematical process referred to as optimization, as its aim is to minimize the unexplained variance of the outcome (i.e. the amount of variation not accounted for by the explanatory variables). This is usually done by using the so called "least squares" optimization method. Additionally, we want to know how certain we are about these estimated coefficients given that their estimation was based on a sample rather than a population. This is achieved by estimating confidence intervals and the associated *p* values (i.e., hypothesis testing). The usual test checks if the coefficients are significantly different from zero; for this we rely on distributional assumptions, and the issue now becomes a statistical one.

1.2. Assumptions of GLMs

A standard GLM assumes: (i) a distributional assumption (i.e. the distribution of residuals of the regression has a particular shape), needed for estimating confidence intervals (ii) a link function which connects the outcome with the explanatory variables (iii) constant variance, so that the inferences we make are valid for all the range of the dependent and independent variables and (iv) independence of the individual measurements. This last assumption is violated when analysing longitudinal and clustered data in standard ANOVAs or regressions. We demonstrate the effect of this violation with an example in 1.2.4.

The distributional assumption is used for mathematical calculations and approximations during the optimization procedure. Importantly, it is also used for making inferences and especially for estimating the confidence intervals. The confidence intervals are used to determine if an estimated coefficient is different from zero (hence statistically significant) or not.

1.2.1. Linear relationship

In every GLM we assume a mathematical relationship between the dependent variables and the independent variables. In a linear regression, we assume that the link function takes the form $f(y) = y$, and thus in most cases it is not stated. On the other hand, in Poisson regression for example, we assume the log() link function and thus $f(y) = \log(y)$ and the Poisson regression can be written as $\log(y) = a + bx + e$. The equivalent form of the normal linear regression is $Y = a + bx + e$.

1.2.2. Homoscedasticity

Homoscedasticity is a term of Greek origin and means "equal variance". This is more of a practical assumption since it allows us to use the same standard error for the range of the values we have available. Otherwise, we would have to estimate a separate standard error for each value.