ANALYZING LONGITUDINAL DATA AND USE OF THE GENERALIZED LINEAR MODEL IN HEALTH AND SOCIAL SCIENCES

Jaume Arnau[1,2], Roser Bono[1,2], Rebecca Bendayan[3] and Maria J. Blanca[3]

[1] Department of Behavioural Sciences Methodology (Faculty of Psychology), University of Barcelona
[2] Institute for Brain, Cognition, and Behavior (IR3C), University of Barcelona
[3] Department of Psychobiology and Behavioural Sciences Methodology (Faculty of Psychology) at the University of Malaga

Address for correspondence:

Roser Bono

Department of Behavioural Sciences Methodology

Faculty of Psychology, University of Barcelona

Passeig de la Vall d'Hebron, 171

08035 Barcelona (Spain)

Email: rbono@ub.edu

Tel. +34 93 312 50 80 * Fax +34 93 402 13 59

ANALYZING LONGITUDINAL DATA AND USE OF THE GENERALIZED LINEAR MODEL IN HEALTH AND SOCIAL SCIENCES

Abstract

In the health and social sciences, longitudinal data have often been analyzed without taking into account the dependence between observations of the same subject. Furthermore, consideration is rarely given to the fact that longitudinal data may come from a non-normal distribution. In addition to describing the aims and types of longitudinal designs this paper presents three approaches based on generalized estimating equations that do take into account the lack of independence in data, as well as the type of distribution. These approaches are the marginal model (population-average model), the random effects model (subject-specific model), and the transition model (Markov model or auto-correlation model). Finally, these models are applied to empirical data by means of specific procedures included in SAS, namely GENMOD, MIXED, and GLIMMIX.

Keywords: generalized linear model, longitudinal data, marginal model, random effects model, transition model.

1. Introduction

Longitudinal designs are used to study processes of change that are directly associated with the passing of time. There are two reasons why, in recent years, longitudinal studies have become widely used in applied contexts. The first concerns the development of advanced analytic techniques, while the second is that current software packages have much greater potential in terms of analysis and simulation. This combination of improved statistical modelling and more powerful computational programs has led to considerable interest in longitudinal designs, especially in those areas where the study of processes is particularly relevant, such as the social, psychological, educational, psychotherapeutic, and epidemiological contexts.

This paper describes current versions of the statistical models that take into account the metric and non-metric nature of longitudinal data. As such, it offers a systematic account of the most up-to-date analytic approaches to a class of data that are usually correlated. The correlation among within-subject observations is the main problem faced by longitudinal research, and it poses a considerable challenge in terms of developing more powerful and flexible models. Indeed, the correlation among observations of the same subject must be taken into account both in the design and when analyzing data. Here we focus on the examination and analysis of data from longitudinal designs not only because of the considerable plasticity and increasing popularity of this approach, but also due to the variety of analytic techniques that can be used to infer hypotheses. One of the currently most popular of these techniques is based on the hierarchical or multilevel longitudinal model, which is considered to be a good option for

the analysis of repeated measures data (Bryk & Raudenbush, 1992; Goldstein, 2011; Raudenbush, 1988).

Our aim in this paper is to provide a general framework for analyzing longitudinal designs, and this is why the various models are presented within the context of the generalized linear model. Specifically, the goal is to compare three approaches based on the generalized linear model which have been developed specifically for use with non-normal data, a form of data that is increasingly common within the field of applied longitudinal research. The generalized linear model was chosen precisely because it can be used with non-normal data, whether quantitative or discrete.

2. Generalized linear model

The generalized linear model (GLM) is designed to analyze non-normally distributed data in the context of regression. The GLM also covers a wide range of data distributions within the exponential family.

In the GLM a single equation, initially formulated by Nelder and Wedderburn (1972), combines the systematic and random components (predictors and measurement variable, respectively) by means of a link function. If we take as our starting point the linear regression model then the response variable, $\boldsymbol{y}$, is normally distributed with mean $\boldsymbol{\mu}$ and constant variance, such that

$$\mathrm{E}(\boldsymbol{y}) = \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} \tag{1}$$

$$\mathrm{E}(\boldsymbol{y} - \boldsymbol{\mu}) = \sigma^2 \tag{2}$$

The general or classical linear regression model (Equation 1) assumes that the y_i observations are normal and independent with standard deviation σ, such that the parameter vector $\boldsymbol{\beta}$ is estimated by means of the least squares criterion $(\mathbf{y} - \boldsymbol{\mu})'(\mathbf{y} - \boldsymbol{\mu})$. Under this assumption both y_i and μ_i are determined by a large number of values. However, if $y = 0$ or $y = 1$, as is the case when responses are categorized according to the absence or presence of a given characteristic, then $0 < \mu_i < 1$. In this case the general linear model is no longer appropriate due to the restrictions that must be imposed on $\boldsymbol{\beta}$ so that μ_i falls within the possible range of values. This is why generalized models include a linear link function for $\boldsymbol{\mu}$, $g(\boldsymbol{\mu})$. This function, $g(\cdot)$, transforms μ_i into a scale on which the values are not subject to such drastic restrictions. Thus, for example, one could use $g(\mu) = \log(\mu)$ if $\mu_i > 0$ or $g(\mu) = \log[\mu / (1-\mu)]$ if $0 < \mu_i < 1$.

In sum, the generalization of the classical linear model consists in assuming that $E(\boldsymbol{y})$ is not identical to the linear combination $\mathbf{X}\boldsymbol{\beta}$, since the relationship is mediated by a function that takes into account the nature of the data ($\boldsymbol{y}$). Technically, the GLM has three components:

a) A *random component* or data vector $\boldsymbol{y}$ (response variable) formed by independent observations derived from a distribution of the exponential family or similar, with a canonical parameter β that determines the form of the response. This implies being able to reformulate the distributions of the exponential family in canonical form, which is possible for most exponential distributions. Note, however, that some distributions of this family, such as the log-normal distribution, cannot be written in canonical form. It is assumed, therefore, that the observation data, $\boldsymbol{y}$, follow an independent normal distribution with mean $\boldsymbol{\mu}$ and constant variance σ^2.

b) A *systematic component* referring to the model's predictor variables (covariates **X**) or explanatory part. This component refers not only to the variables that must be taken into account but also to the way in which they should be introduced into the equation. The set of covariates determines a linear predictor $\boldsymbol{\eta}$, expressed by

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} \tag{3}$$

c) The *link function* enables the distribution parameters to be related to the model's predictors. Thus, for example, the function *g* links the systematic component with the parameter of the mean μ

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} \tag{4}$$

where $g(\cdot)$ is the *link function* that defines the linear relationship between the mean of $\boldsymbol{y}$ and the predictors. Its inverse is known as the *response function.*

$$\mathrm{E}(\mathbf{y}) = \boldsymbol{\mu} = g^{-1}(\mathbf{X}\boldsymbol{\beta}) \tag{5}$$

The function $\pi = g^{-1}(\mathbf{X}\boldsymbol{\beta})$ used in the expression of the mean (Equation 5) is referred to in the GLM as the canonical link function, and it enables better estimators of the model parameters to be obtained.

Note that the majority of distributions encountered in social and psychological research, regardless of whether they refer to continuous or discrete data, have a probability density function that belongs to the exponential family. Their mathematical expression may take either a normal or canonical form. The general expression of the distribution from the exponential family is as follows:

$$f(y,\theta) = \exp\{a(y)\, b(\theta) + c(y) + d(\theta)\} \quad (6)$$

where a, b, c, and d are known functions that have the same form for all y. If there are parameters other than θ they are considered as nuisance parameters that form part of the functions a, b, c, and d, and they are treated, therefore, as if they were known. If $a(y) = y$ the distribution is said to have a canonical form, and in this case $b(\theta)$ is referred to as the *natural* or *canonical parameter*. Note, however, that the mathematical expression of the common exponential distribution is not usually described in terms of this general form (Equation 6), since the parameter θ is replaced by $b(\theta)$, where $\eta = b(\theta)$. As a result, Equation 6 is redefined as follows:

$$f(y,\eta) = \exp\{a(y).\eta + c(y) + d(\eta)\} \quad (7)$$

where η is a function of θ. Obviously, the function $d(\cdot)$ is not the same as in the general expression (Equation 6). If y is normally distributed with mean θ and variance σ^2, and $\theta = \eta$, then one is dealing with the ordinary linear model with normal errors. In other words, the GLM uses a special subclass of the natural exponential family, where $b(\theta) = \eta$ and $a(y) = y$. This natural form can also be written in terms of the mean μ rather than θ by means of a simple transformation $\mu = g(\theta) = E(y,\theta)$ or reparameterization of the value of the mean. In the GLM the probability distribution is reparameterized such that the distribution is a function of unknown parameters based on known data.

2.1. Fit of the model

If we apply a link function to the response mean the resulting model must be fitted by means of the maximum likelihood (ML) method. Having defined the likelihood function of the model the next step consists in determining which parameters make the data most likely. This is done by log-transforming the likelihood function to convert it into an additive rather than the multiplicative scale.

The estimation of GLM parameters uses the Newton-Raphson algorithm, which resolves the log-likelihood function. The log transformation converts a multiplicative model into a linear one, thereby facilitating parameter estimation and the use of the ML estimation algorithms. The log-likelihood of the exponential model is expressed as:

$$l(\theta, y, \phi) = \Sigma\{(y\theta - b(\theta))/\phi - c(y, \phi)\} \tag{8}$$

where ϕ is a dispersion constant for all y_i (McCullagh & Nelder, 1989) and where the corresponding deviance function is defined as:

$$2\Sigma\{l(y, y) - l(y, \mu)\} \tag{9}$$

The deviance gives us an idea of the variability in the data. Hence, a measure of the variability explained by the model can be obtained by comparing the null deviance with the residual deviance. Therefore, the deviance represents the amount of variability in the response variable that is not explained by the model:

$$D' = (\boldsymbol{y}, \boldsymbol{\mu}) = 2(l(\boldsymbol{y}, \boldsymbol{y}) - l(\boldsymbol{y}, \boldsymbol{\mu})) \quad (10)$$

where $l(\boldsymbol{y}, \boldsymbol{\mu})$ is the *log-likelihood* function defined in terms of the predicted mean $\boldsymbol{\mu}$ and the response vector $\boldsymbol{y}$. As for *D*, this is the total value of the discrepancy of the GLM. The deviance statistic for an observation also reflects its contribution to the overall goodness of fit of the model. There are two useful statistics for evaluating the goodness of fit of the GLM: the *scaled deviance* and the Pearson chi-squared. For a fixed value of the dispersion parameter, ϕ, the scaled deviance is defined to be twice the difference between the maximum achievable log-likelihood and the log-likelihood at the maximum likelihood estimates of the regression parameters. The scaled version of these two statistics can be used as an approximate index of the goodness of fit of the model. Note, however, that when the value of the dispersion parameter is unknown its estimation can be used as a good approximation to both the scaled deviance and the Pearson chi-squared statistic.

3. Generalized estimating equations: generalized linear models and maximum quasi-likelihood estimation

It has already been pointed out that the correlation among repeated measures of the same subject is the main challenge faced when analyzing longitudinal data. In addition, many of the data encountered in social and behavioral research take the form of binary, frequency, or categorical responses, and hence it is necessary to use models designed for the analysis of discrete data. Note that it cannot be assumed that this kind of data fulfill the assumption of multivariate normality.

Liang and Zeger (1986) and Zeger and Liang (1986) proposed a unified method for analyzing longitudinal data that incorporates the most important aspects of multivariate analysis. Their approach is based on generalized estimating equations (GEEs) and examines the dependence between the response variable and a set of explanatory covariates. The GEE method, which constitutes an integration of the GLM and quasi-likelihood (QL) methods, can be used with both normally and non-normally distributed response variables (Davis, 1991; Park, Shin, & Park, 1998; Shoukri & Edge, 1996). When the assumption of multivariate normality is fulfilled the GEE method is comparable to the maximum likelihood (ML) procedure, although the ML criterion provides more efficient estimates (Schwartz & Stone, 1998). The GEE method only requires specification of the mean, the variance, and the working correlation matrix of the repeated measures vector for a given subject. By means of this procedure, one can obtain efficient and consistent estimates of the regression parameters, even when the correlation matrix is misspecified.

Depending on the aim of the analysis, and taking into account the correlation among the repeated measures of the same subject, Diggle, Liang, and Zeger (1994) and Zeger and Liang (1986, 1992) propose three approaches to the GEE method: the *marginal model* (*population-average model*), the *random effects model* (*subject-specific model*), and the *transition model* (*Markov model* or *auto-correlation model*).

QL estimation is a regression procedure that, in contrast to the maximum likelihood criterion, requires few assumptions about the distribution of the dependent variable. It is therefore applicable to a wide variety of data. With longitudinal data it is assumed that the observations of different subjects are independent and that those from the same subject are correlated. When the data are normally distributed the dependence

between observations of subjects can be analyzed by means of multivariate techniques. If, however, the data follow other distributions, as in the case of binary longitudinal data, other approaches such as GEEs may be applied. In this approach the marginal distribution of observations is specified together with the working correlation matrix for observations of the same subject. As a result, the GEE method proposed by Liang and Zeger (1986) and Zeger and Liang (1986), which is based on the QL approach, provides consistent estimations of the parameters and their corresponding variances, under what can be considered as weak assumptions regarding the correlation among within-subject observations. Consequently, this kind of GEE method can be used to calculate the values of $\boldsymbol{\beta}$. Although the GEE method developed by Liang and Zeger (1986) was initially applied to the analysis of covariance structure models, it can also be extended to linear mixed models (LMM).

3.1. Marginal model

The marginal model analyzes the relationship between the response variable and the covariates without taking into account the between-subject heterogeneity (Zeger, Liang, & Albert, 1988). This model is an extension of the GLM with correlated observations, and it estimates the parameters by means of the QL criterion (Liang & Zeger, 1986). The QL criterion, associated with the marginal model, enables the equations for parameter estimation to be derived. Given that, in this model, the coefficients have a population interpretation rather than an individual one, the model is also referred to by Zeger *et al*. (1988) as the *population-average model*.

By way of an example, let us suppose, following Howes and Matheson (1992), that we wish to study the development of competent play with peers, and that we will examine this from infancy through preschool, taking measurements every six months over a period of three years. If we are solely interested in studying the effect of age on children's average social behaviors during play across the established age intervals, then our study is centered on average population values and their dependence on a series of covariates. We are not studying individual development, but rather the mean development of a sample of observed children. The marginal model, which is centered on the population average, estimates the effect of covariates on the marginal expectation of the response variable. According to this model the regression coefficients associated with the covariates must be specified separately to the correlation structure of the within-subject observations. The main interest here is estimating the fixed parameters, since the parameters that define the covariance matrix are considered as nuisance parameters and are required to calculate the accuracy of the fixed parameter estimates (Burton, Gurrin, & Sly, 1998; Omar, Wright, Turner, & Thompson, 1999).

The marginal model for longitudinal data consists of:

1) The mean or marginal expectation of the response at time *t* for subject *i* and is given by

$$\mathrm{E}(y_{it}) = \mu_{it} \tag{11}$$

where the response y_{it} is a random variable at time *t* for subject *i*. The outcome y_{it} depends on a set of explanatory variables, $\mathbf{X}_{it}$, by means of the function

$$g(\mu_{it}) = \mathbf{X}_{it}\boldsymbol{\beta} \tag{12}$$

According to the GLM approach the mean is related to a set of covariates through a known link function, *g*, for example, the logit function for binary responses or the log function for amounts or frequencies. In Equation 12, $\boldsymbol{\beta}$ is a vector *p* x 1 of unknown parameters and it represents the way in which the average population response is a function of the covariates. According to Zeger *et al*. (1988) the advantage of these models is that, for a given covariate, the average population response can be estimated without assumptions regarding the heterogeneity among individual parameters.

2) The marginal variance is a function of the marginal mean, in other words, the variance of y_{it} is a function of $\boldsymbol{\mu}_{it}$, as follows:

$$\text{var}(y_{it}) = g(\mu_{it})\cdot\phi = \sigma_{it}^{2} \tag{13}$$

where *g* is a function of the known variance. It is assumed that the marginal distribution of y_{it} follows a GLM, such that $g(\mu_{it})$ is completely determined by the assumption of the exponential family. Similarly, ϕ is a scale or dispersion parameter that has to be estimated. For binomial and Poisson distributions this scale parameter is fixed at 1.

3) In order to take into account the within-subject dependence in y_{it} (i.e., the correlations among observations taken from a given subject) it is necessary to specify the working correlation matrix, $\mathbf{R}_i(\boldsymbol{\alpha})$. This matrix depends on an unknown parameter vector, $\boldsymbol{\alpha}$, which is the same for all subjects. In accordance with the QL criterion the working correlation matrix is defined by

$$\mathbf{V}_i = \mathbf{A}_i^{1/2}\,\mathbf{R}_i(\boldsymbol{\alpha})\mathbf{A}_i^{1/2}\,/\phi \qquad (14)$$

where $\mathbf{A}_i$ is a diagonal matrix n_i x n_i, for each subject i, with $g(\mu_{it})$ as t element of the diagonal ($\mathbf{A}_i$ = diag[$g(\mu_{i1,}\,\mu_{i2},...,\,\mu_{it}\,)$]), and $\mathbf{R}_i(\boldsymbol{\alpha})$ is the working correlation matrix n_i x n_i for each subject i, with parameter $\boldsymbol{\alpha}$. Since $\mathbf{R}_i(\boldsymbol{\alpha})$ is not expected to be correctly specified the GEE method provides consistent estimates even when $\mathbf{R}_i(\boldsymbol{\alpha})$ is not the correct correlation matrix. With independent observation, $\mathbf{V}_i = \mathbf{A}_i \cdot \phi$ and $\mathbf{R}_0 = \mathbf{I}$. Given that a correlation is also expected among the repeated measures of a given subject, $\mathbf{R}_i(\boldsymbol{\alpha})$ is defined as a function of a vector s x 1 of unknown parameters $\boldsymbol{\alpha}$. Finally, ϕ is an overdispersion parameter whose square root is called the scale parameter and which is estimated from the data.

The GEE method for estimating the parameters $\boldsymbol{\beta}$, as proposed by Liang and Zeger (1986) and Zeger and Liang (1986), extends the concept of QL to correlated observations and has the following function:

$$\mathrm{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathrm{D}_i^{'}\mathrm{V}_i^{-1}\mathrm{S}_i \;= \mathbf{0} \qquad (15)$$

where $\mathbf{D}_i$ is the derivative matrix with $\partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}$ elements, $\mathbf{V}_i$ is the covariance matrix as specified in Equation 14, and $\mathbf{S}_i = \mathbf{Y}_i - \boldsymbol{\mu}_i$. Given that the repeated measures of a given subject are expected to be correlated, $\mathbf{R}_i(\boldsymbol{\alpha})$ is a working correlation matrix that depends on the vector of unknown parameters α.

The GEE of $\boldsymbol{\beta}$ is given by

$$U(\boldsymbol{\beta}) = \Sigma \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1}(\boldsymbol{\alpha})(\mathbf{Y}_i - \boldsymbol{\mu}_i) = 0 \qquad (16)$$

where $\mathbf{V}_i(\boldsymbol{\alpha}) = \mathbf{A}_i^{½}\mathbf{R}_i(\boldsymbol{\alpha})\mathbf{A}^{½}$. The marginal regression coefficients, $\boldsymbol{\beta}$, are interpreted as coefficients of cross-sectional regression, such that marginal models with correlated data are naturally analogous to the GLM with independent data (Zeger *et al*., 1988).

3.2. Random effects model

If a researcher is more interested in the individual response than the population value it is preferable to use the *random effects* (or *subject-specific*) *model*. This model assumes that the subject-specific effects follow a parametric distribution in the population. As in the case of linear random effects models it is assumed that the response is a linear function of the explanatory variables, with coefficients that vary from one subject to another. This variability reflects the heterogeneity attributable to unmeasured factors. The study by Howes and Matheson (1992) is a good example of the classical linear regression of children's development, where the coefficients represent the measures of initial social play and the rate of peer play development. Obviously, children exhibit different ~~weights~~ social behaviors and they develop at different rates, due, for example, to environmental factors that are difficult to quantify. A random effects model is a reasonable description when the set of coefficients for a population of children can be considered as a sample from a distribution. Given the current coefficients for a child the linear random effects model also assumes that the repeated observations of this individual are independent. The correlation among repeated observations arises because it is not possible to observe

the underlying growth curve, that is, the true regression coefficients, since we only have imperfect measures of each child's social behavior during play.

The *generalized linear mixed model* (GLMM), one of whose applications is the subject-specific model, assumes that the regression coefficients vary among subjects according to a normal distribution. The use of GLMMs is required when the data are binomial and non-Gaussian (Hand & Crowder, 1996). The GLMM is specified as follows:

1) Let y_{it} be a random outcome variable, $\mathbf{X}_{ij}$ a vector p x 1 of fixed covariates at time t for subject i, $\mathbf{Z}_{it}$ a vector q x 1 of covariates associated with the random effects $\boldsymbol{\gamma}_i$, and $\mu_{it} = \mathrm{E}(y_{it}|\boldsymbol{\gamma}_i)$. The GLMM model then assumes that the responses of subject i satisfy

$$\mu_{it} = f(y_{it}|\boldsymbol{\gamma}_i) = \mathbf{X}'_{it}\boldsymbol{\beta} + \mathbf{Z}'_{it}\boldsymbol{\gamma}_i \qquad (17.1)$$

$$\mathrm{var}(y_{it}|\boldsymbol{\gamma}_i) = g(\mu_{it})\cdot\phi \qquad (17.2)$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_i$ are vectors of fixed and random effects parameters, respectively, as in the LMM. It is assumed that the random effects are independent and come from an F distribution.

2) The marginal moments, $\boldsymbol{\mu}_i$ and $\mathbf{V}_i$, are calculated from the conditional moments and the F distribution of the random effects. Thus, given the conditional moments (equations 17.1 and 17.2) and the F distribution for the random effects the expected marginal value, μ_{it}, is

$$\mu_{it} = E(y_{it}) = E[E(y_{it}|\boldsymbol{\gamma_i})] \quad (18.1)$$

and the marginal covariance matrix is given by

$$\mathbf{V}_i = \text{cov}[E(\mathbf{Y}_i|\boldsymbol{\gamma_i}] + E[\text{cov}(\mathbf{Y}_i|\boldsymbol{\gamma_i})] \quad (18.2)$$

3) When $\boldsymbol{\mu}_i$ and $\mathbf{V}_i$ have been calculated for each subject, $\boldsymbol{\beta}$ is estimated by means of Equation 17.1. The idea underlying the random effects model is that there is natural heterogeneity among individuals as regards their regression coefficients, and that this heterogeneity can be represented by a probability distribution.

The basic difference between the subject-specific model and the population-average model concerns the objectives. The former is mainly interested in subjects, whereas the latter focuses on the mean response. In addition, the subject-specific model explains the covariance among repeated measures, whereas the population-average model only describes this covariance. Although subject-specific models are preferable when one wishes to determine individual responses and profiles, their use is limited due to the minimal information available about each subject. As Zeger *et al*. (1988) point out, longitudinal studies often involve only a few observations per subject, and this makes it difficult to estimate the regression coefficients separately.

3.3. Transition model

Transition models analyze the effect of covariates on the transitional patterns of continuous, binary, and categorical responses across time. These are conditional models,

since the explanatory variables and the previous responses act as predictors of the current response. In transition models the correlation among the observed data is explained by the action of past values on the current outcome. Hence, the main characteristic of transition models is that they include past observed values as additional predictor variables. Note that the extent to which y_{it} depends on $\mathbf{X}_{it}$ remains a basic objective. However, given that the observations are serial and probably dependent, Zeger and Qaqish (1988) propose Markov models in which the actual expected response depends not only on the associated covariates but also on past responses.

These models also use a QL approach with Gaussian and non-Gaussian time series data. In general terms the transition model can be defined by means of the following expression:

$$y_{it} = \mathbf{X'}_{it}\boldsymbol{\beta} + \varepsilon_{it} \tag{19}$$

where

$$\varepsilon_{it} = \alpha\varepsilon_{it-1} + u_{it} \tag{20}$$

and where α = exp(-ϕ) and the u_{it} are mutually independent random variables that follow a normal distribution of mean 0 and variance $\sigma_u^2 = \sigma^2(1 - \alpha^2)$. By substituting Equation 20 into Equation 19, we obtain the conditional distribution of y_{it}, given the preceding response, y_{it-1}, as follows:

$$y_{it}|y_{it-1} \sim N\{\mathbf{X'}_{it}\boldsymbol{\beta} + \alpha(y_{it-1} - \mathbf{X'}_{it-1}\boldsymbol{\beta}), \sigma_u^2\} \tag{21}$$

Equation 21 considers both the explanatory variables and the previous responses as explicit predictors of the current outcome. Using this expression the transition model can easily be defined. Thus, with metric data a linear regression model with autoregressive errors takes the form

$$y_{it} = \mathbf{X'}_{it}\boldsymbol{\beta} + \sum_{r=1}^{q} \alpha_r (y_{it-r} - \mathbf{X'}_{it-r}\boldsymbol{\beta}) + u_{it} \qquad (22)$$

According to Equation 22 the current observation y_{it} is a linear function of $\mathbf{X}_{it}$ or of the explanatory covariates and the prior observations, $r = 1,..., q$, and the $\boldsymbol{u}_{it}$ are the mutually independent random variables. From here the transition model can be reformulated by means of the binary or categorical responses across time. For binary data the formula is reduced to

$$\text{logit}\ \ pr(y_{it} = 1 |\ y_{it-1},\ y_{it-2,...,}\ y_{it-q}\ ) = \mathbf{X'}_{it-q}\ \boldsymbol{\beta} + \Sigma\alpha y_{it-q} \qquad (23)$$

where the $\mathbf{X}_{it}$ are time-dependent subject-specific variables and q is the order of Markov dependence. The regression coefficients can be interpreted, in this case, as the effect of the variables on the probability of a binary event adjusting for the past history of the process.

4. Analysis of data from an empirical example using SAS

In this section we present a more extensive analysis of data in order to illustrate the analysis of GEEs using SAS version 9.4. Specifically, we examine a set of data taken from the Millennium Cohort Study (MCS). This study followed the lives of around 19,000 children in the UK since 2000-2001. The present study used the first four sweeps of the MCS, the first being when the children were around 9 months old, the second when they were around 3 years old, the third when they were 5 years old, and the fourth when they were around 7 years old. More information about the MCS data sets can be found in Hansen (2012).

Although the MCS is a longitudinal study with a large sample, the aim of the illustrative study being considered here requires the researcher to use subsamples, which might be small. Note that some of the values of the variables of interest may have been changed in order to make the example clearer and easier to follow for the reader.

The aim of the illustrative study was to examine alcohol consumption in two groups of mothers. More specifically, a group of 60 mothers who were diagnosed with depression at the first measurement point, when their child was 9 months old, was compared with a control group of 96 mothers. The variable frequency of current alcohol consumption was regrouped into five main categories: (1) five times or more per week; (2) three or four times per week; (3) one or two times per week; (4) one or two times per month; and (5) never.

The aim of the study was to examine whether there were systematic between-individual differences in within-individual change in the mothers' alcohol consumption over time as a consequence of depression.

The data file was organized according to the longitudinal format in SAS, and for the transition models a new variable was generated with the lag data of each subject

(Table 1).

The syntax for organizing the data, shown in Table 1, was developed following Singer and Willett (2003). In the original format, each subject has a row of data containing the values of the outcome variable on each of four occasions (*alcohol9m*, *alcohol3y*, *alcohol5y*, and *alcohol7y*). Each record also contains an identifying variable, *id*, and the variable *depression*. In the longitudinal format, the data set contains two variables identical to those in the original format (*id* and *depression*) and two new variables (*time* and *alcohol*). *Time* identifies the measurement occasion to which the record refers, and *alcohol* records the individual's score at that measurement occasion. The longitudinal data set for this example has a total of 624 records, four for each of the 156 mothers.

[INSERT TABLE 1 ABOUT HERE]

4.1. Model specification and analysis

In this example we would expect the alcohol consumption values at the successive time points to be correlated, and also that they would be affected by a set of covariates such as time and depression.

Let us begin by applying the marginal model (Model 1) to the data. The syntax corresponding to the GEEs is shown in Table 2. *PROC GENMOD* in SAS is used as it enables this kind of model to be fitted.

With respect to the distribution followed by these data, it should be taken into account that they refer to a categorical variable, and it is therefore possible to fit a

multinomial distribution with cumulative logit link. The cumulative logit model is the most popular model for ordered categorical data (for details on the cumulative logit model, see McCullagh & Nelder, 1989).

Table 2 displays the parameter estimates obtained with the GEE model. The intercept terms correspond to the four cumulative logits defined on the categories of alcohol consumption. Thus, *Intercept1* is the intercept for the first cumulative logit, $log\left(\frac{p_1}{1-p_1}\right)$, *Intercept2* is the intercept for the second cumulative logit, $log\left(\frac{p_1+p_2}{1-(p_1+p_2)}\right)$, and so forth. Note the strong effect of *time* (p=0.022). By contrast, the variable *depression* shows no significant differences (p=0.913). Neither was the *time*depression* interaction significant (p=0.067).

[INSERT TABLE 2 ABOUT HERE]

The second model analyzed, the random effects model, can be resolved analytically by means of the LMM with *PROC MIXED* or the GLMM with *PROC GLIMMIX*, depending on whether the distribution is normal or another member of the exponential family. In this example, following Stroup (2012), *PROC GLIMMIX* was used as the data are not normally distributed. Random effect models for longitudinal data are regression models in which the regression coefficients are allowed to vary across subjects. Therefore, it is necessary to specify the random variables and the fixed variables, depending on whether they come from a random sample or are held constant across different subjects. For example, in the analysis of the effect of depression on alcohol consumption among our sample of 156 subjects, depression is taken as a fixed

effect, and we are interested in comparing the mean alcohol consumption across the two levels of depression. The unique subject identification (*id*) for each of the 156 subjects would be treated as a random factor. Given that the aim of the study was to analyze the variables *time* and *depression*, these variables must be considered as fixed effects.

The corresponding syntax is shown in Table 3. As regards the random effects, the *random* statement in Model 2 enables us to estimate the between-subjects variance or the degree to which subjects vary around the intercept and the within-subject residual variance.

The estimates of the fixed for Model 2 are shown in Table 3. Note that the parameter estimates and their significance are similar to those of Model 1.

[INSERT TABLE 3 ABOUT HERE]

To conclude, let us analyze the transition model, which represents the true nature of longitudinal, or change over time. To this end, at the level of the data structure, it is helpful to create a new variable based on the lag of the values of each subject, such that the influence of each observation on the next can be ascertained (Table 1). Based on the transition structure, in Table 4 we specify an initial model with *PROC GENMOD* (Model 3) and another model with *PROC GLIMMIX* (Model 4). In all these models the model statement includes the new variable that takes into account the effect of the preceding observation on the current one (*lag_1*).

Note that when the transition models include the lag variable (*lag_1*) this variable is statistically significant under both approaches, *PROC GENMOD* (Model 3) and *PROC MIXED* (Model 4). The variable “depression” and its interaction with the lag variable

(*lag_1*) are also significant in the transition models analyzed. As expected, the increase in alcohol consumption is greater among mothers diagnosed with depression than in the control group, when taking into account the immediately preceding level of consumption.

[INSERT TABLE 4 ABOUT HERE]

In the above example, different SAS procedures have been applied depending on the kind of model. Thus, *PROC GENMOD* is suitable with the marginal model, while with random effects or transition models one may use *PROC GENMOD* or *PROC GLIMMIX*.

5. Discussion

One alternative to classical techniques for modelling longitudinal data is the LMM, which assumes a multivariate normal distribution of observations. One of the concepts on which this model is based is that the mean profile of the observations of a given subject is formed by an average population term and a subject-specific term. This is reinforced by the form of the covariance matrix, which comprises components of within-subject and between-subject variance.

When the data do not follow a Gaussian distribution it is possible to apply the estimation method based on GEEs. The GEE method is a regression procedure within the context of the GLM and it applies the QL criterion to estimate the regression parameters. The GEE method was initially described in relation to two non-dynamic models,

depending on whether the focus of interest was on the average population parameters (marginal model) or subject-specific parameters (random effects model). Thus, the main distinction between these two models concerns whether the regression coefficients describe an average change in response with respect to the covariates or an individual change. In addition, the marginal model describes the covariance among repeated measures of a given subject, whereas the random effects model focuses on the source of this covariance. According to Zeger and Liang (1992) there are three advantages to the GEE approach. Firstly, the regression coefficients are almost as efficient as ML estimators. Secondly, even if the covariance structure for correlated repeated measures is misspecified, the estimates for the regression coefficients remain consistent as long as the sample size is large enough. And thirdly, the statistical inference of the regression coefficients is not influenced by the covariance matrix, provided that one uses a robust estimate of the covariance matrix of the regression coefficient estimators, as suggested by Liang and Zeger (1986). A further point is that use of the GEE method means that we can apply the transition model, which is the most suitable in longitudinal studies.

To conclude, note that the models examined (marginal, random effects, and transition) are extensions of the GLM and they can be applied to continuous or discrete longitudinal data. Marginal models are recommended in population studies, such as epidemiological research, where the difference in the mean response between groups is more important than the change in the response of a given subject. Random effects models are used when the individual response is of greater interest than the population response, as would be the case, for example, in growth curve studies. However, the effectiveness of random effects models is limited by the minimal information available per subject. Many longitudinal studies involve only a few observations of each subject,

and it is not possible to estimate the regression coefficients separately. A further point to note is that time series models with linear regression methods have been widely studied with Gaussian data, but very little attention has been paid to non-Gaussian data. Hence the importance of the QL approach for regression with time series data (Zeger & Liang, 1992).

In this paper we have described how the GEE method can be used to study the retroactive effects that are typical of time series or longitudinal data. This is especially relevant to those cases in which the interdependence of data is an important aspect to take into account. The GEE method is also suitable for studies involving categorical data, and in those where subjects' responses are influenced by prior responses (Zeger & Liang, 1991). The three models described in this article have been analyzed empirically using procedures incorporated within SAS, although this does not mean that other software packages could not also be used.

Acknowledgements

1. This research was supported by grant PSI2012-32662 from the Spanish Ministry of Economy and Competitiveness.

2. The authors are grateful to the Centre for Longitudinal Studies (CLS), Institute of Education, University of London, for the use of these data, and also to the UK Data Archive and Economic and Social Data Service (ESDS) for making them available. However, neither CLS nor ESDS bear any responsibility for the analysis or interpretation of these data.

References

Brown, H., & Prescott, R.: Applied mixed models in medicine (Statistics in practice) (3rd ed.). Chichester: Wiley (2015)

Bryk, A. S., & Raudenbush, S. W.: Hierarchical linear models for social and behavioural research: Applications and data analysis methods. Newbury Park, CA: Sage Publications (1992)

Burton, P., Gurrin, L., & Sly, P.: Extending the simple linear regression model to account for correlated responses. Statistics in Medicine, 17, 1261-1291 (1998). doi: 10.1002/0470023724.ch1a

Davis, C. S.: Semi-parametric and non-parametric methods for the analysis of repeated measurements, with applications to clinical trials. Statistics in Medicine, 10, 1959-1980 (1991). doi: 10.1002/sim.4780101210

Diggle, P. J., Liang, K. -Y., & Zeger, A. L.: Analysis of longitudinal data. Oxford: Oxford University Press (1994)

Goldstein, H.: Multilevel statistical models (4th ed.). Chichester: Wiley (2011)

Hand, D., & Crowder, M.: (1996). Practical longitudinal data analysis. London, Chapman & Hall (1995)

Hansen, K.: Millennium Cohort Study: First, second, third and fourth surveys. A guide to the datasets (7th ed.). London, UK: Centre for Longitudinal Studies (2012)

Liang, K. -Y., & Zeger, S. L.: Longitudinal data analysis using generalized linear models. Biometrika, 73, 13-22 (1986). doi: 10.1093/biomet/73.1.13

McCullagh, P., & Nelder, J. A.: Generalized linear models (2nd ed.). London: Chapman & Hall (1989)

Nelder, J. A., & Wedderburn, R. W. M.: Generalized linear models. Journal of the Royal Statistical Society, A135, 370-384 (1972)

Omar, R. Z., Wright, E. M., Turner, R. M., & Thompson, S.G.: Analysing repeated measurements data: a practical comparison of methods. Statistics in Medicine, 18, 1587-1603 (1999). doi: 10.1002/(SICI)1097-0258(19990715)18:13<1587::AID-SIM141>3.0.CO;2-Z

Park, T., Shin, D. W., & Park, C. G.: A generalized estimating equations approach for testing ordered group effects with repeated measurements. Biometrics, 54, 1645-1653 (1998). doi: 10.2307/2533689

Raudenbush, S. W.: Educational applications of hierarchical linear models: A review. Journal of Educational Statistics, 13, 85-116 (1988)

Schwartz, J. E., & Stone, A. A.: Strategies for analyzing ecological momentary assessment data. Health Psychology, 17, 6-16 (1998). doi: 10.1037/0278-6133.17.1.6

Shoukri, M. M., & Edge, V. L.: Statistical methods for health sciences. Boca Raton, FL: CRC Press (1996)

Singer, J. D., & Willett, J. B.: Applied longitudinal data analysis: Modeling change and event occurrence. New York, NY: Oxford University Press (2003)

Stroup, W. W.: Generalized linear mixed models: modern concepts, methods and applications. Boca Raton, FL: Chapman and Hall/CRC (2012)

University of London Institute of Education. Centre of Longitudinal Studies. Millennium Cohort Study: First survey, 2001-2003 [Computer file] (9th ed.). Colchester, Essex, UK: UK Data Archive [distributor], April 2012. doi: 10.5255/UKDA-SN-4683-1

University of London Institute of Education. Centre of Longitudinal Studies. Millennium Cohort Study: Second survey, 2003-2005 [Computer file] (6th ed.). Colchester, Essex, UK: UK Data Archive [distributor], April 2012. doi: 10.5255/UKDA-SN-5350-1

University of London Institute of Education. Centre of Longitudinal Studies. Millennium Cohort Study: Third survey, 2006 [Computer file] (4th ed.). Colchester, Essex, UK: UK Data Archive [distributor], April 2012. doi: 10.5255/UKDA-SN-5795-1

University of London Institute of Education. Centre of Longitudinal Studies. Millennium Cohort Study: Fourth survey, 2008 [Computer file] (1st ed.). Colchester, Essex, UK: UK Data Archive [distributor], April 2012. doi: 10.5255/UKDA-SN-6411-1

Zeger S. L., & Liang K. -Y.: Longitudinal data analysis for discrete and continuous outcomes. Biometrics, 42, 121-130 (1986)

Zeger, S. L., & Liang, K. -Y.: Feedback models for discrete and continuous time series. Statistica Sinica, 1, 51-64 (1991)

Zeger S. L., & Liang, K. -Y.: An overview of methods for the analysis of longitudinal data. Statistics in Medicine, 11, 1825-1839 (1992). doi: 10.1002/sim.4780111406

Zeger S. L., Liang, K. -Y., & Albert, P. S.: Models for longitudinal data: a generalized estimating equation approach. Biometrics, 44, 1049-1060 (1988)

Zeger S. L., & Qaqish B.: Markov regression models for time series: a quasi-likelihood approach. Biometrics, 44, 1019-1031 (1988)

Table 1. SAS syntax for transforming data

Data structure	Syntax

Longitudinal	data alcohol;
	input id depression alcohol9m alcohol3y alcohol5y alcohol7y;
	alcohol=alcohol9m; time=1; output;
	alcohol=alcohol3y; time=2; output;
	alcohol=alcohol5y; time=3; output;
	alcohol=alcohol7y; time=4; output;
	drop alcohol9m alcohol3y alcohol5y alcohol7y;

Lag data for each subject (transition model)	data alcohol; set alcohol; lag_1=lag(alcohol); if time=1 then lag_1=.; run;

Note. *data* is the data set for the analysis; *input* statement defines the variables to be read in each line of data; *alcohol=alcohol9m* creates new variable called alcohol using the original variable of alcohol consumption among mothers when their child was 9 months old (*alcohol9m*); *time=1* creates new variable called time and sets the value for the first time measure as 1; *output* statement defines output to the new converted univariate dataset (the same commands apply to the next three command lines for different time measures: when the children were 3, 5, and 7 years old); *drop* statement specifies the names of the variables to omit from the output data set; *set alcohol* statement modifies an existing SAS data set; *lag_1=lag(alcohol)* creates the variable lag_1 and lag function returns the value of the previous observation; *if time=1 then lag_1=*. assign "." to the first response of each subject.

Table 2. Syntax and analysis of the marginal model (Model 1)

```
proc genmod;
class id;
model alcohol = time depression time*depression
                / dist=multinomial link=cumlogit;
repeated subject=id /corr=ind;
run;
```

Analysis of GEE parameter estimates

Parameter	Estimate	SE	*Z*	*p*
Intercept1	-2.831	0.630	-4.49	<0.001
Intercept2	-1.887	0.596	-3.16	0.002
Intercept3	-0.670	0.587	-1.14	0.254
Intercept4	0.084	0.582	0.14	0.885
Time	0.370	0.161	2.29	0.022
Depression	-0.038	0.347	-0.11	0.913
Time*Depression	-0.166	0.091	-1.83	0.067

Note. *proc genmod* calls the *PROC GENMOD* in SAS; *class* statement defines the classification variables or grouping variables; *id* is the subject identifier or subject variable; *model* statement specifies an equation in which the dependent variable is to the left of the equals sign and the effects or predictor variables are to the right (by default, the intercept is included on the right); the multinomial distribution is fitted by *dist=multinomial* and *link=cumlogit*; *repeated* command indicates there are repeated data for each subject; the option *subject=id* refers to the individual subjects specified in the input data set by the variable id (this variable must be listed in the class statement); *corr=ind* indicates that the working correlation type is independent (note that *corr=ind* is the only correlation structure allowed whenever *dist=multinomial* is specified).

Table 3. Syntax and analysis of random effects model (model 2)

```
proc glimmix;
class id;
model alcohol = time depression time*depression
                / dist=multinomial link=cumlogit solution;
random intercept / type =un subject=id;
run;
```

Estimates of the fixed effects

Parameter	Estimate	SE	*t*	*p*
Intercept1	-5.004	1.053	-4.75	<0.001
Intercept2	-3.294	1.035	-3.18	0.002
Intercept3	-0.989	1.024	-0.97	0.336
Intercept4	0.595	1.023	0.58	0.562
Time	0.622	0.253	2.46	0.014
Depression	-0.162	0.612	-0.26	0.792
Time*Depression	-0.283	0.153	-1.85	0.065

Note. *proc glimmix* calls the *PROC GLIMMIX* in SAS; *class* specifies the categorical variable *id*; *model* specifies the fixed effects; the multinomial distribution is fitted by *dist=multinomial* and *link=cumlogit*; the option *solution* requests the parameter estimates and their corresponding standard errors; *random* specifies the random effects; *random intercept/subject=id* indicates that each subject has its own intercept; *type=un* next to the *random* command specifies the structure of the between-individual covariance matrix as non-structured.

Table 4. Syntax and analysis of transition models

Models	Syntax

Model 3

```
proc genmod;
class id;
model alcohol = time lag_1 depression time*depression
                lag_1*depression / dist = multinomial link = cumlogit;
repeated subject = id /corr = ind
run;
```

Model 4

```
proc glimmix;
class id;
model alcohol = time lag_1 depression time*depression
                lag_1*depression / dist = multinomial link = cumlogit
                solution;
random intercept / type=un subject=id;
run;
```

Model 3

Analysis of GEE parameter estimates

Parameter	Estimate	SE	*Z*	*p*
Intercept1	-1.505	1.471	-1.02	0.306
Intercept2	0.022	1.503	0.01	0.988
Intercept3	2.140	1.525	1.40	0.160
Intercept4	3.430	1.540	2.23	0.026
Time	0.614	0.356	1.73	0.084
Lag_1	-0.973	0.298	-3.26	0.001
Depression	1.823	0.902	2.02	0.043
Time*Depression	-0.280	0.214	-1.30	0.192
Lag_1*Depression	-0.378	0.167	-2.26	0.024

Model 4

Estimates of the fixed effects

Parameter	Estimate	SE	*t*	*p*
Intercept1	-1.505	1.491	-1.01	0.314
Intercept2	0.022	1.489	0.02	0.988
Intercept3	2.140	1.498	1.43	0.155
Intercept4	3.431	1.503	2.28	0.024
Time	0.614	0.381	1.61	0.108
Lag_1	-0.973	0.257	-3.78	<0.001
Depression	1.823	0.908	2.01	0.045
Time*Depression	-0.280	0.231	-1.21	0.227
Lag_1*Depression	-0.378	0.155	-2.46	0.014

Note. The commands of *PROC GENMOD* are explained in Table 2, while those of *PROC GLIMMIX* are described in Table 4. The variable *lag_1* (lag data) is added to the *model* statement.