Quality & Quantity **34:** 323–330, 2000.
© 2000 *Kluwer Academic Publishers. Printed in the Netherlands.*

Note

Discrete Response Multilevel Models for Repeated Measures: An Application to Voting Intentions Data

MARIA FERRAO BARBOSA[1] and HARVEY GOLDSTEIN[2]
[1]*Pontificia Universidade Catolica do Rio de Janeiro;* [2]*Institute of Education, University of London*

Abstract. Repeated measures data can be modelled as a two-level model where occasions (level one units) are grouped by individuals (level two units). Goldstein et al. (1994) proposed a multilevel time series model when the response variable follows a Normal distribution and the measurements are taken with unequal time intervals. This paper extends the methodology to discrete response variables. The models are applied to British Election Study data consisting of repeated measures of voting intention.

Key words: discrete response, longitudinal data, multilevel model, repeated measures, time series, underdispersion, voting intentions.

1. Introduction

Repeated measures data can be modelled as a two-level structure where measurement occasions are level one units and individual subjects are level two units. Consider a data set consisting of repeated measurements of the heights of a random sample of children. Thus, for linear growth we can write a simple model as

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + e_{ij}. \tag{1}$$

This model assumes that height (Y) is linearly related to age (X) with each subject having their own intercept and slope so that, assuming Normality, we have

$$\begin{pmatrix} \beta_{0j} \\ \beta_{1j} \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \Omega\right), \quad \Omega = \begin{pmatrix} \sigma^2_{u0} & \\ \sigma_{u01} & \sigma^2_{u1} \end{pmatrix}, \quad e_{ij} \sim N(0, \sigma^2_e).$$

There is no restriction on the number or spacing of ages, so that we can fit a single model to subjects who may have one or several measurements. We can clearly extend Equation (1) to include further explanatory variables, measured either at the occasion level, such as time of year or state of health, or at the subject level such as birthweight or gender.

For measurements such as growth the specification of the level 2 variation serves to model a separate curve, typically a polynomial, for each individual. We can think of each curve as a smooth summary of growth with small random departures at each measurement occasion. If, however, measurements on an individual are obtained very close together in time, consecutive measurements will have a similar departure from that individual's underlying growth curve. This implies that the level 1 residuals will be positively correlated; there will be 'autocorrelation' between them. Examples occur in other areas, such as economics, where successive measurements on each unit, for example an enterprise or economic system, exhibit an autocorrelation structure and where the parameters of the separate time series will vary across units at level 2.

A detailed discussion of multilevel time series models is given by Goldstein et al. (1994) who provide maximum likelihood estimates for multivariate Normal models. They discuss both the discrete time case, where the measurements are made at the same set of equal intervals for all level 2 units, and the continuous time case where the time intervals can vary. We are concerned here with the more general continuous time model and its extension to discrete responses.

To simplify the presentation, we shall drop the level 1 and 2 subscripts and write a general model for the level 1 residual covariance structure as follows

$$\text{cov}(e_t e_{t-s}) = \sigma_e^2 f(s). \tag{2}$$

This states that the covariance between two measurements s units in time apart, depends on the level 1 variance (σ_e^2, which in a more general model could also be a function of age and other covariates) and a function involving the time difference. The latter function is conveniently described by a negative exponential together with the common assumption that with increasing time difference the covariance will tend to a fixed value, $\alpha\sigma_e^2$. We have

$$f(s) = \alpha + \exp(-h(\beta, z, s)), \tag{3}$$

where β is a vector of parameters for further explanatory variables z. The choice of h should be parsimonious and depend on the context. Goldstein (1995) presents a table with several possible functions and we explore some of these below.

2. Discrete Response Multilevel Models

There are many situations where the response variable is not Normally distributed, for example where the response is a proportion or count. For such generalised linear models we can write a 2-level generalisation (Goldstein, 1995) as

$$\pi_{ij} = f(X_{ij}\beta_j),$$

where π_{ij} is the expected value of the response for the ijth level 1 unit and f is a nonlinear function of the 'linear predictor' $X_{ij}\beta_j$. Note that we can have random coefficients at level 2. The model is completed by specifying a distribution for the *observed* response $y_{ij}|\pi_{ij}$. Where the response is a proportion this is typically taken to be binomial and where the response is a count taken to be Poisson. It remains for us to specify the nonlinear 'link' function f. In the remainder of this paper we shall be concerned with responses which are binary or proportions, the most common link function for which is the logit so that we write

$$\pi_{ij} = \{1 + \exp(-[\beta_0 + \beta_1 x_{1ij} + u_{0j}])\}^{-1}. \tag{4}$$

The observed responses y_{ij} are proportions with the standard assumption that they are binomially distributed, namely

$$y_{ij} \sim \text{Bin}(\pi_{ij}, n_{ij}), \tag{5}$$

where n_{ij} is the denominator for the proportion and

$$\text{var}(y_{ij}|\pi_{ij}) = \pi_{ij}(1 - \pi_{ij})/n_{ij}. \tag{6}$$

Following Goldstein (1995) we fit this into a standard multilevel framework by writing

$$y_{ij} = \pi_{ij} + e_{ij} z_{ij}, \quad z_{ij} = \sqrt{\pi_{ij}(1 - \pi_{ij})/n_{ij}}, \sigma_e^2 = 1, \tag{7}$$

and we use the mean and variance properties as specified in Equations (4) and (6) to produce quasilikelihood estimates of the model parameters (McCullagh and Nelder, 1989). To carry this out we need to linearise Equation (4) and this leads to marginal (MQL) or predictive (penalised) (PQL) estimates. Details can be found in Goldstein and Rasbash (1996).

3. Discrete Multilevel Models for Repeated Measures Longitudinal Data

Since repeated measures data have a 2-level structure it may seem that we can just apply a 2-level model which is the discrete response analogue of the continuous response model (1) (Diggle et al., 1994). For some kinds of data this will be reasonable, but in other cases the assumption in Equations (4) and (5) that we have level 1 binomially distributed responses which are independent, conditionally on the covariates and random effects, is untenable. We shall be considering the particular case of repeated measurements of voting intentions where a proportion of the population have the same response at each occasion; their probabilities are therefore either zero or one and this implies that the linear predictor in Equation (4) is infinite. Other examples will occur in repeated measures of attitudes, disease states etc.

Yang et al. (1998) discuss this issue in detail for the case where there is a small number of fixed occasions. Their data consist of responses to the question 'do you vote Conservative?' in each of 3 years, 1983, 1986, 1987 using the British Election Study data (Heath et al., 1985). They set up a multivariate binary response model where each occasion (year) is treated as a variate with binomial variation between individuals at each occasion and covariances across occasions (variates) which are estimated from the data. Thus the repetition at level 1 (indicated by t) is nested within individuals (indicated by i), while individuals are nested within constituency j. Let z_t be the vector of indicator variables for $t = 1, 2, 3$ or 1983, 1986 and 1987 respectively,

$$\left.\begin{aligned} z_{1ij} &= 1 \ \text{ if } \ t = 1983 \\ z_{2ij} &= 1 \ \text{ if } \ t = 1986 \\ z_{3ij} &= 1 \ \text{ if } \ t = 1987 \end{aligned}\right\} \text{and 0 otherwise.}$$

Since year is level 1 the notation reflects this with t being the index for the first subscript. The term s_{tij} denotes the measurement of time (1, 2, 3) as a continuous variable. For simplicity we ignore covariates and write a variance components model, fitting only an intercept in the fixed part of the model for the probability of a positive response π_{tij}

$$\log it(\pi_{ij}) = \sum_{t=1}^{3} \beta_{0,t} z_{tij} + \sum_{t=1}^{3} (v_{tj} + u_{tj}) z_{tij}, \qquad v_{tj} \sim N(0, \Omega_v),$$

$$\Omega_v = \begin{pmatrix} \sigma_{v1}^2 & & \\ \sigma_{v12} & \sigma_{v2}^2 & \\ \sigma_{v13} & \sigma_{v23} & \sigma_{v3}^2 \end{pmatrix},$$

$$\Omega_u = \begin{pmatrix} \pi_{1ij}(1-\pi_{1ij}) & & \\ \sqrt{\pi_{1ij}(1-\pi_{1ij})\pi_{2ij}(1-\pi_{2ij})}\sigma_{u12} & \pi_{2ij}(1-\pi_{2ij}) & \\ \sqrt{\pi_{1ij}(1-\pi_{1ij})\pi_{3ij}(1-\pi_{3ij})}\sigma_{u13} & \sqrt{\pi_{2ij}(1-\pi_{2ij})\pi_{3ij}(1-\pi_{3ij})}\sigma_{u23} & \pi_{3ij}(1-\pi_{3ij}) \end{pmatrix}. \tag{8}$$

The v_{tj} are the constituency level random effects. At the subject level the variance terms in Ω_u reflect the binomial assumption and the $\sigma_{ut_1t_2}$ terms represent the 'point biserial' correlations between occasions which are to be estimated.

We wish to generalise this to the case of unequal time intervals where we can no longer model a multivariate (fixed occasion) structure for the between-subject variation. At each occasion we retain the assumption of binomial variation and we write the covariance as

$$\sqrt{\pi_{tij}(1-\pi_{tij})\pi_{(t+s)ij}(1-\pi_{(t+s)ij})}\, f(s), \qquad f(s) = \alpha + \exp(-h(\beta, z, s)). \tag{9}$$

The estimation procedure follows that for continuously distributed responses explained in Appendix II of Goldstein et al. (1994), operating on the linearised version of the model as described above. Macros were written in *MLwiN* (Rasbash et al., 1999) to carry out the computations.

4. An Application to Voting Measurements

The data for the following examples consist of vote/vote intention measurements on a panel of voters per constituency area as used by Yang et al. (1998) and described above. On first occasion respondents were interviewed immediately after the general election, second occasion measures were taken in the autumn of 1986, and the third immediately after the general election of 1987. There are 112 constituencies, 1633 voters, and 3434 outcomes about vote or vote intention. For the purpose of this paper the response variable is whether or not they voted or intended to vote Conservative. There are also measures of voters' fundamental values related to party policies on nuclear defence, unemployment (versus inflation), tax cuts (versus government spending) and privatisation (versus nationalisation) which were analysed by Yang et al., but we do not consider these here.

A three level model was fitted with occasions as level 1 units, voters as level 2 units and constituencies as level 3 units. The following results are second order PQL estimates (Goldstein, 1995). Initially three variance components models were fitted. The results are presented in Tables I and II. At level 1 we can fit a scale parameter for the binomial variance to estimate an under/over-dispersion parameter and as a check on the binomial assumption.

Three models presented in each table are:

(A) A standard 3-level version of Equations (4)–(5) which assumes independence across occasions;

(B) A time series model with autocorrelation function $f(s) = \exp(-\beta s)$;

(C) A time series model with autocorrelation function $f(s) = \exp[-(\beta_0 s + \beta_1 s^{-1})]$.

The natural extension to model (B) would seem to use the function $f(s) = \exp[-(\beta_0 + \beta_1 s)]$, but it was not possible to obtain convergence with this function.

Since there are only three correlation parameters to be estimated, fitting a model with three (non-dependent) parameters defining the autocorrelation function is equivalent to fitting the full multivariate model. The aim of these analyses is to see how well a time series formulation approximates the full multivariate model.

The autocorrelation function $f(s) = \exp(-\beta s)$ is equivalent to an AR(1) model for equally spaced measures. To begin with, for simplicity, we do not fit the full model for the between-constituency variation as in Equation (8), nor do we fit the covariates. Table I fits the above three models with the corresponding multivariate model fitted in Table II. An extra binomial parameter is estimated in all cases. The simple model in column A shows a large amount of under-dispersion as expected and overestimates the between-constituency and between-respondent

Table I. Voting in 1983, 1986 and 1987 with different covariance structures

Fixed	(A) Estimate (se)	(B) Estimate (se)	(C) Estimate (se)
1983	−0.55 (0.12)	−0.42 (0.08)	−0.42 (0.08)
1986	−1.57 (0.13)	−0.85 (0.09)	−0.85 (0.08)
1987	−0.36 (0.13)	−0.33 (0.99)	−0.32 (0.08)
Random			
Level 3	0.71 (0.19)	0.37 (0.08)	0.33 (0.08)
Level 2	6.99 (0.37)	–	–
Extra-binomial	0.37 (0.01)	0.99 (0.03)	0.97 (0.03)
β	–	0.35 (0.02)	–
β_0	–	–	0.12 (0.01)
β_1	–	–	0.42 (0.03)
Correlations			
$s = 1$	–	0.70	0.58
$s = 2$	–	0.50	0.64
$s = 3$	–	0.35	0.61

variation. The second column with the autocorrelation function $f(s) = \exp(-\beta s)$ demonstrates too rapid a decay and the model in column C provides the best fit and corresponds closely to the estimates obtained from the full multivariate model for the between-year correlations. The extra binomial parameters for B, C and the multivariate model are close to 1.

5. Discussion

This paper extends to discrete response variables the methodology proposed by Goldstein et al. (1994) for normally distributed responses. It shows that for the data set considered the standard repeated measures assumptions are untenable and lead to biases. The bias is assumed to arise from the existence of a mixture distribution whereby some individuals have a constant response and others have responses which vary from occasion to occasion. Such a model might be characterised as a mover-stayer model with a proportion always voting Conservative and a proportion never voting Conservative. There are, however, two problems with such a characterisation.

First, the proportion of true ‘stayers’ may in fact be very small, with a proportion having a small but nonzero probability of switching in certain circumstances, for example to engage in ‘tactical voting’. Such a group could not be adequately

Table II. Basic multivariate model for 1983, 1986, and 1987

Fixed	Estimate (se)
1983	−0.39 (0.07)
1986	−0.78 (0.08)
1987	−0.31 (0.08)
Random	
Level 3	0.29 (0.07)
Level 2:	
σ^2_{u1}	0.92 (0.03)
σ^2_{u2}	0.96 (0.04)
σ^2_{u3}	0.95 (0.05)
σ^2_{u12}	0.56
σ^2_{u13}	0.63
σ^2_{u23}	0.61

The σ^2_{ut} terms are extra binomial parameters and the $\sigma^2_{ut_1t_2}$ terms are correlations estimated using these.

modelled along with the ‘movers’ unless a suitable covariate was available which correlated strongly with their propensity to change their vote. Secondly, while it is possible to carry out the estimation of a mover-stayer model, at least for a 2-level model, where there are covariates and random coefficients the estimation becomes more complicated with the model involving the sum of three non-linear components. The proposed method provides a feasible computing procedure. Results from simulations of the PQL2 estimation procedure for univariate logistic response models suggest that for moderate amounts of higher level variation and large numbers of level 1 units per level 2 unit, the estimates show little bias. In other cases the bias can be eliminated using an iterated bootstrap procedure (Goldstein and Rasbash, 1996).

The approach we have used for data which are binary or proportions can be used with minor modifications for count data and can, in principle, be extended to multinomial ordered or unordered data. Further work is planned along these lines.

We would expect data from other applications to exhibit similar properties to our voting intention responses. Longitudinal data on attitudes will often contain individuals whose attitudes do not change; the same may be found in medical studies of response to treatment and in other areas. It will therefore be useful when fitting models to such data to estimate extra-distributional parameters to check the model assumptions.

Acknowledgements

We are most grateful to Min yang and Anthony Heath for their helpful comments. This work was partly supported with a grant from the Economic and social research Council (UK) under its Analysis of Large and Complex Datasets programme. Maria F. Barbosa was sponsored by CNPq-Brazil.

References

Diggle, P., Liang, K. & Zeger, S. L.(1994). *Analysis of Longitudinal Data*. Oxford: Oxford University Press.

Goldstein, H., Healy, M. & Rasbash, J. (1994). Multilevel time series models with applications to repeated measures data. *Statistics in Medicine* 13: 1643–1655.

Goldstein, H. (1995). *Multilevel Statistical Models*, 2nd edn. London: Edward Arnold, New York: Wiley.

Goldstein, H. & Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society A* 159(3): 505–513.

Heath, A. F., Jowell, R. M. & Curtice, J. K. (1985). *How Britain Votes*. Oxford: Pergamon.

McCullagh, P. & Nelder, J. (1989). *Generalised Linear Models*, 2nd edn. London: Chapman and Hall.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Draper, D., Plewis, I., Healy, M. & Woodhouse, G. (1999). *MLwiN Users' Guide.* London: Institute of Education.

Yang M., Heath, A. & Goldstein, H. (1999). Multilevel models for repeated binary outcomes: attitudes and vote over the electoral cycle (Submitted for publication).