(Entry for *Encyclopaedia of Biostatistics*, Armitage, P., & Colton,T. (Eds.), 1998, Wiley.)

Random coefficient repeated measures models

by

Harvey Goldstein

Institute of Education

London, WC1H 0AL

h.goldstein@ioe.ac.uk

Introduction

This section is concerned with modelling data where measurements of one or more attributes are repeated on the same set of individuals over time. Typical applications are to the modelling of anthropometric growth of children or animals. The model specification will be developed for the case where a single continuous measurement is made on several occasions for a sample. This will then be extended to consider the case of multiple measurements at each time point and mention will be made of extensions to latent variable models and to discrete response data.

To begin with we look at the simple, restricted, data structure where there are a fixed number of measurement occasions and each individual has a measurement at each occasion.

Multivariate models

Consider the data matrix of responses

Individual	**Occasion 1**	**Occasion 2**	**Occasion 3**	**Occasion 4**
1	y_{11}	y_{21}	y_{31}	y_{41}
2	y_{12}	y_{22}	y_{32}	y_{42}
3	y_{13}	y_{23}	y_{33}	y_{43}

The first subscript refers to occasion and the second to individual. We assume multivariate normality and so for the response vector we have initially

$$Y \sim N(\mu, \Sigma) \qquad (1)$$

This constitutes a null model and in general we will wish to include further variables, notably age or time. Suppose we wish to express the response, say a weight measurement, as a linear function of time (*t)* measured at each occasion. We may then write

$$y_{ij} = \beta_{0j} + \beta_{1j} t_{ij} + \varepsilon_{ij} \qquad (2)$$

where we allow the intercept and average growth rate to vary across individuals. Suppose further that

$$\begin{pmatrix}\beta_{0j}\\ \beta_{1j}\end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix},\begin{pmatrix}\sigma^2_{\beta_0} & \\ \sigma_{\beta_{01}} & \sigma^2_{\beta_1}\end{pmatrix}\right] \quad \varepsilon_{ij} \sim N(0,\sigma^2_\varepsilon) \tag{3}$$

We have replaced the general mean and covariance structure given by (1) by the specific structure given by (3). Thus, for example, the goodness of fit of (3) can be judged and the model elaborated with suitable explanatory variables. Grizzle and Allen [9] provide details of estimation and test procedures.

This multivariate model cannot deal satisfactorily with the typical situation where the spacing and number of measurement occasions is variable and has generally been superceded, except in one or two special cases such as that of latent growth models mentioned below. We now develop an alternative approach to fitting models such as (2), based upon a *multilevel* model.

The 2-level repeated measures model

Model (2) and the associated covariance structure (3) as they are written make no particular assumptions about the number or spacing of measurement occasions and in fact constitute a special case of a 2-level model (see entry on multilevel models). Level 1 units are the measurement occasions and level 2 units are individuals. All the usual procedures for estimation and inference in such models are therefore available, including cases of multivariate responses, nonlinear models etc. We can additionally

consider individuals as nested within further hierarchies, say animal litters or schools for students and cross classifications may also occur.

A consequence of (2) is that measurements made on the same individual are correlated, through the sharing of common intercept and slope parameters, and it is this dependency that leads to the inadequacy of simple estimation procedures, for example based upon ordinary least squares. Furthermore, interest will usually lie just as much in the covariance matrix estimates as in the average growth parameters and we may also wish to form posterior mean estimates of the individual growth parameters (β_{0j}, β_{1j}) and we shall illustrate below how these can be used for efficient prediction. As in the general multilevel model case, we may have a Bayesian formulation for the model with prior distributions upon the parameters (see for example Best et al [1]).

In the following sections we shall consider in more detail nonlinear models, multivariate response models with more than one response at each occasion and complex structures for the level 1 residuals. For a detailed exposition of further aspects of these topics and some alternative approaches as well as a discussion of issues related to informatively missing data and transition type models the reader should consult Diggle et al [3]. In particular these authors consider the so-called 'population average' model where interest centres on the estimation of the fixed or average component of (2). This often allows simplified estimation procedures to be used with no requirement for the separate estimation of the random components. This may be appropriate in certain circumstances, but since it ignores the specific nature of repeated measurements data is not considered further here.

Nonlinear and generalised linear models

Most attempts to fit nonlinear models to repeated measurements have fitted separate curves to each individual's set of measurements and then combined these to describe the between-individual variation. A major problem with this approach is that it requires many measurements on each. Also, while nonlinear curves have been used successfully to describe change, for example in pharmacokinetic studies, in other areas, such as growth they can also impose inflexible relationships among growth events which are not empirically supported (Goldstein [5]).

Bock [2] describes a maximum likelihood analysis of a human growth model using the superimposition of three logistic functions. Lindstrom and Bates [13] describe an approximate estimation procedure for nonlinear models and Goldstein [7] gives an example using the so called Jenss-Bayley [10] curve for children aged 5 to 10 years. Davidian and Giltinan [4] give a detailed discussion of different approaches to the fitting of nonlinear models to repeated measures data.

Where the response is discrete, for example binary or ordered as in the case of recording developmental stages over time, then a generalised linear model will be appropriate. Consider the following example where each individual (j) is measured at several times (t) and their nutritional state (y) at occasion (i) is classified as adequate (1) or inadequate (0). A standard model would be written as

$$\begin{aligned} &\log it\{\pi_{ij}\} = a_j + b_j t_{ij} \\ &\pi_{ij} = pr(y_{ij} = 1) \\ &y_{ij} \sim Bin(\pi_{ij}, 1) \end{aligned} \qquad (4)$$

This expresses the logit of the probability of having an adequate nutritional state as a linear function of time. Such a model might be appropriate, for example, in evaluating a nutritional intervention programme and further covariates for group membership, age, etc. can readily be introduced. We can also try alternative link functions and study the possibility of further random coefficients. For responses such as counts we would typically use a log link with a Poisson or related distributional assumption.

For many longitudinal data we are effectively measuring the *cumulative* probability of a response over time. Thus, when studying the onset of menarche the probability of occurrence is an increasing function of time and successive observations will consist of a string of zeros (non-occurrences) followed by a string of ones (occurrences). More generally, we will have an ordered sequence of stages through which all individuals pass and (4) will be modified to reflect this. One such *proportional hazards* model can be written as

$$\gamma_{ij}^{(s)} = \{1 - \exp(-\exp[\beta_0 + \beta_1 t_s])\}, \qquad \beta_1 > 0 \tag{5}$$

where *s* indexes the stages and the cumulative probability is

$$\gamma_{ij}^{(s)} = \sum_{h=1}^{s} \pi_{ij}^{(h)}$$

We can add further covariates and random coefficients as before.

An extension of both (4) and (5) is to the multivariate case where multiple responses are measured on each individual at each time point, with possibly missing responses at some occasions and where some responses are discrete and some continuous. A

discussion of such models and their estimation is given by Goldstein [7, Chapter 7] and the multivariate model with continuous only responses is discussed in the next section.

In these models so far we have made the basic assumption that the level 1 errors are independent. We shall deal with violations of this assumption for continuous responses below, but there are also many cases for discrete responses where this assumption is untenable and this gives rise to particular difficulties. As an example consider a repeated survey of attitudes to abortion where we wish to study the characteristics of individual and group changes over time. For a large proportion, perhaps the majority, of the population there will be no change in their attitudes; thus the probability that they will agree with a 'pro-abortion' statement will be very close to one or zero. A model such as (4) would generally require such individuals to have extremely large positive or negative random effects since it is unlikely that we would have covariates which could discriminate precisely among such individuals. This then poses severe distributional problems for parametric models.

An obvious way to avoid this difficulty is to consider the vector of, say binary, responses for each individual as a multivariate vector where the distribution at each occasion is binomial and the between-occasion covariances are estimated from the data. Lipsitz et al [14] study such models with examples. While this approach is satisfactory for a number of fixed occasions, even with missing data, and while it can also be extended to other than binary responses, it is unable directly to handle the general case of arbitrary occasions. To do this requires an extension of the serial correlation models discussed below, but that is beyond the scope of this article.

Multivariate continuous responses

Where several responses are recorded on individuals at each occasion we will generally wish to model the average time relationship for each response and the covariance matrix among the responses as a function of time. This is readily done by considering the multivariate response structure as a further, lowest, level in the data hierarchy with measurements nested within occasions within individuals (see entry on multilevel models).

There are several advantages to considering the joint modelling of several responses. The ability to estimate their covariance matrix as a function of time allows one to study the distribution of any function of the responses with respect to time. For example, when studying issues of prior determination it may be useful to see whether the correlation between two variables a given time apart is greater when one is the prior variable rather than the other. Likewise, it provides a general prediction procedure for one measurement, conditional on any set of observed prior measurements. We illustrate this with an example concerned with the prediction of adult height given a series of height measurements taken during a period of childhood growth. In this case, one of our response measurements, adult height, is made at the level of the individual and the others are made at the occasion level.

Consider the following extension to (2)

$$
\begin{aligned}
y_{ij} &= \beta_{0j} + \beta_{1j} t_{ij} + \beta_2 t_{ij}^2 + \beta_3 t_{ij}^3 + \varepsilon_{ij} \\
y_j &= \sum_{\mathrm{k}} \gamma_{jk} x_{jk} + \alpha_j , \\
\begin{pmatrix} \alpha_j \\ \beta_{0j} \\ \beta_{1j} \end{pmatrix} &\sim N\left[\begin{pmatrix} \alpha \\ \beta_0 \\ \beta_1 \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & & \\ \sigma_{\alpha\beta_0} & \sigma_{\beta_0}^2 & \\ \sigma_{\alpha\beta_1} & \sigma_{\beta_{01}} & \sigma_{\beta_1}^2 \end{pmatrix} \right]
\end{aligned}
\qquad (6)
$$

where the adult measurement (y_j) is allowed to depend on further covariates, and we may also wish to incorporate covariates into the growth period component of the model. The key feature of this model is that we have a joint covariance matrix for the adult height component and the growth curve parameters, all of which vary at the level of the individual. Given the model parameters and any set of growth measurements for an individual, say $Y_j^* = (y_{1j}^*, y_{2j}^*, ... y_{pj}^*)$ we can estimate $E(y_j | Y_j^*)$ together with an estimate of its standard error etc. Details are given in Goldstein [6].

A further development of the multivariate growth model is the so called ‘latent growth model’. In essence this considers each of the sets of random coefficients β_{0j}, β_{1j} etc. as a latent variable or factor. Each observed response is thus a linear function of factor scores where the coefficients are for example polynomials in time, or more generally may be estimated from the data. One restriction of such models is that they require the same set of discrete occasions for all measurements and thus lose the flexibility of the continuous time formulation. A full discussion can be found in Muthen [15].

Serial correlation models

For some kinds of repeated measurements the structure implied by (2) or (4) is inadequate. For example, daily measurements of animal weights over a long period will not usually fluctuate completely randomly about a long term smooth trend for each animal, the departure from such a trend on any one day being more like the departures on neighbouring days than on days further distant. In a study of human growth, Goldstein et al [8] found that residuals from measurements of height made on adolescent boys had a noticeable serial correlation when made less than three months apart. To incorporate such possibilities we can extend (2) by adding the following covariance condition for two level 1 residuals *s* time units apart, where time is continuous

$$\mathrm{cov}(\varepsilon_t, \varepsilon_{t-s}) = \sigma_\varepsilon^2 \exp(-g(s,z)) \qquad (7)$$

Here *g* is a positive function and may depend on covariates, *z*, which may be measured at the individual or occasion level, thus, for example, allowing the exponential decay rate implied by (7) to vary with time.

One possible simple choice, which is the continuous time analogue of a first order autoregressive series is $g = \alpha s$ and other possibilities are discussed by Goldstein et al [8] and Diggle et al [3, Chapter 5]. An alternative approach is via State-Space modelling which leads to similar although not generally identical models (Jones [11]).

We give an example of a model with a simple correlation structure, together with the estimation of a seasonal effect for a set of 3-monthly height measurements made on a sample of 26 boys aged between 11 and 14 years. Full details are given by Goldstein et al [8]. A fourth degree polynomial is fitted for the average growth curve with a cosine

term representing seasonal growth. The first three coefficients are random at level 2 and the serial covariance structure is given by $g = \alpha s$ fitted at level 1.

Table 1. Height in cm as a fourth degree polynomial on age, measured about 13.0 years. Standard errors in brackets; correlations in brackets for covariance terms. Serial correlation structure fitted for level 1 residuals.

Parameter	**Estimate (s.e.)**		
Fixed			
Intercept	148.9		
age	6.19 (0.35)		
age^2	2.16 (0.45)		
age^3	0.39 (0.17)		
age^4	-1.55 (0.43)		
cos (time)	-0.24 (0.07)		
Random			
level 2			
	Intercept	age	age^2
Intercept	61.5 (17.1)		
age	7.9 (0.61)	2.7 (0.7)	
age^2	1.5 (0.25)	0.9 (0.68)	0.6 (0.2)
level 1			
σ_ε^2	0.23 (0.04)		
α	6.90 (2.07)		

The serial correlation parameter value of 6.9 implies that the residual correlation three months apart is 0.19 and that six months apart is 0.04. The existence of a seasonal effect implies an average difference between Summer and Winter of about 0.5cm with no evidence of any variation between individuals.

Fitting this model, with an extra parameter to describe autocorrelation among the level 1 residuals, provides a more parsimonious model than attempting to fit, say, the cubic coefficient as random at level 2. In some cases, however, the data may be equally well explained either by such a random coefficient model with independent level 1 residuals or, alternatively, by a simpler between-individual covariance structure and a complex non-independence structure at level 1. A choice between such models will then need to be made on grounds of substantive interpretation. In the view of the present author substantive interpretations generally are best made by adopting a level 1 serial correlation structure only after fitting a suitably complex model using random coefficients alone. The use of various diagnostic tools for judging fit in multilevel models is discussed by Lewis and Langford [12].

Software

Some of the particular models described (for example the nonlinear model of Bock [2], the latent growth model or the Bayesian models) have specialised software, details of which can be found by consulting the references given. Some of the major software packages, most notably SAS, can handle many, although not all of the models and the GEE estimation procedures used by Diggle et al [3] are available in S+. The general purpose multilevel modelling package, MLn (Rasbash and Woodhouse [16]) uses both

maximum likelihood and quasilikelihood estimation and has facilities to analyse all the models described, although it can only handle the latent growth model indirectly by providing summary input for other structural equation software packages.

References

1. Best, N. G., Spiegelhalter, D. J., Thomas, A. and Brayne, C. E. G. (1996). Bayesian analysis of realistically complex models. *Journal of the Royal Statistical Society, A*, **159**: 323-42.

2. Bock, R. D., Du Toit, S. H. C. and Thissen, D. (1994). *AUXAL: Auxological analysis of longitudinal measurements of human stature.* Chicago, Scientific Software International.

3. Diggle, P. J., Liang, K.-Y. and Zeger, S. L. (1994). *Analysis of longitudinal data*. Oxford, Clarendon Press.

4. Davidian, M. And Giltinan, D. M. (1995). *Nonlinear models for repeated measurement data.*. London, Chapman and Hall.

5. Goldstein, H. (1979). *The Design and Analysis of Longitudinal Studies.* London, Academic Press.

6. Goldstein, H. (1989). *Efficient prediction models for adult height. Auxology 88; Perspectives in the Science of growth and development.* J. M. Tanner. London, Smith-Gordon: 41-48.

7. Goldstein, H. (1995). *Multilevel Statistical Models*. London, Edward Arnold: New York, Halsted Press.

8. Goldstein, H., M. J. R. Healy, and Rasbash, J. (1994). Multilevel time series models with applications to repeated measures data. *Statistics in Medicine*, **13**: 1643-55.

9. Grizzle, J. C. and D. M. Allen (1969). An analysis of growth and dose response curves. *Biometrics,* **25**: 357-61.

10. Jenss, R. M. and N. Bayley (1937). A mathematical method for studying the growth of a child. *Human Biology,* **9**: 556-63.

11. Jones, R. M. (1993). *Longitudinal data with serial correlation: a state-space approach.* London, Chapman and Hall.

12. Lewis, T. and Langford, I. (1996). Outliers in multilevel data. (Submitted for publication).

13. Lindstrom, M. J. and D. M. Bates (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics,* **46**: 673-687.

14. Lipsitz, S. R., Fitzmaurice, G. M., Sleeper, L. and Zhao, L. P. (1995). Estimation methods for the joint distribution of repeated binary observations.. *Biometrics,* **51**: 562-70.

15. Muthen, B. (1995). *Longitudinal studies of achievement growth using latent variable modelling.* Los Angeles, University of California.

16. Rasbash, J. and G. Woodhouse (1995). *MLn Command reference*. London, Institute of Education.