Multilevel Models Project

Working paper

9-Dec-98

Causal Inferences from repeated measures data

by

Harvey Goldstein

Institute of Education

The multivariate repeated measures model

Suppose we have several measurements made at each occasion for a sample of individuals. These might be, say, blood pressures, anthropometry, ventricular mass, diet etc. We are interested in asking questions about the causal priority or sequence, for example whether a change in blood pressure is associated with a later alteration in ventricular mass, rather than the reverse.

Traditionally such questions have been approached through path type models with measurements at successive (fixed) occasions and relationships across occasions studied to examine the strength of the relationships (path coefficients) in various time directions. Thus, if a prior measurement of x_1 has a stronger relationship with a later measurement of x_2 than a prior measurement of x_2 has with x_1 we might conclude that x_1 is causally prior to x_2. In reality the situation would seem to be more complicated than this simple model since the time lag itself presumably enters into any interpretation and this itself may be a function of age etc. Furthermore, we are restricted with such models to fixed common occasions, whereas in many longitudinal studies the occasion times vary. For such data structures repeated measures models are more appropriate.

With repeated measures data we can also study the relationship across time, and have the potential for modelling the strength (correlation) between measurements in continuous time. It is worth pointing out that with these models we do not have the scaling problems (identified by Goldstein, 1979 and Plewis, 1985) since we can have physically different measurements of the 'same' construct (such as educational achievement) and so long as we can describe the change over time (perhaps by prestandardising) we are then only concerned with correlation structures rather than in making ibferences about absolute change. Of course, these correlations may depend on the precise form of standardisation and this would need to be investigated. In the case where we are relating measurements across occasions we also need to consider the interpretation of possibly different measurements at each occasion.

There is a special case where the repeated measures approach approximates the fixed occasion model but also provides more flexibility. Consider the repeated measurement of students over time, where different measurements are used for different age ranges. If there is just one measurement on each student within each age range then we can formulate the repeated measures model as a multivariate model, fitting a term to each age range measurement, with some possibly missing. In the fixed part of the model we would adjust for any age relationship *within* each age range. Correlations estimated from this model would then be age adjusted correlations between the measurements. The age ranges may overlap and if we have more than one occasion for some individuals where the same measurement is used, this is easily accomodated using a further level in the model, below the individual

subject level. We note that this model avoids the need to consider 'vertical equating' procedures designed to produce a common scale for all measurements, which in general has considerable drawbacks (Goldstein and Wood, 1988).

What is required is a procedure for cross-correlating a pair of measures with lags and estimating separate parameters according to whether x_1 or x_2 comes later in the time sequence and to be able to model the correlation as a function of the time difference. If we can do this then we are in a position to judge the relative strengths of the relationships as described above. In the next section we show how such correlations can be derived from a multivariate repeated measures model and then introduce an alternative formulation.

The multivariate repeated measures model

To illustrate the procedure consider the following 3-level model for two measurements

$$y_{ijk} = \delta \sum_l \beta_{lk}^{(1)} x_{ljk}^{(1)} + (1-\delta) \sum_l \beta_{lk}^{(2)} x_{ljk}^{(2)} + \delta e_{jk}^{(1)} + (1-\delta) e_{jk}^{(2)} \qquad (1)$$

$$\delta = 1 \quad if \quad measurement\ 1, \quad otherwise\ 0.$$

where *i* referes to the measurement, *j* refers to occasion and *k* refers to individual subject. The terms under the summations include both fixed and random terms, the latter for example including polynomial terms in age with random coefficients. Level 1 has no variation and we need to specify the covariance structure at level 2 which is the between-occasion level.

The level 3 variation is that across individual subjects. If we regard the level 2 (within subject) variation as composed of measurement error and essentially random flutuations then we may use this level 3 variation as the basis for inference.

To illustrate the procedure we use data taken from a longitudinal study of aging (Stini, 1990). We have the following between-subject covariance matrix for subjects between the ages of 60 and 99 years, measured on up to 12 occasions and with age centered at 70 years.

Table 1. Between-individual covariance matrix for boys 11 - 16 years

	Log (wt.)	BMI	Age (logwt.)	Age (BMI)
Log (wt.)	0.028			
BMI	0.0052	0.0069		
Age (logwt.)	-0.00016	0.0	0.000041	
Age (BMI)	0.0	0.000019	0.0000070	0.000012

If we denote this matrix by Ω then the correlation between any function of the random coefficients at this level can be written as

$$R^{-0.5}(X\Omega X^T)R^{-0.5}$$
$$R = diag(X\Omega X^T) \qquad (1)$$

where the rows of X define the function of the random coefficients. Thus to calculate the correlation between log weight at age 69 and bone mineral index at age 71 we form

$$X = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

which gives a value of 0.36. The corresponding correlation between BMI at 69 years and log weight at 71 years is 0.37. Extending the time intervals we find that the correlation between log weight at 65 years and BMI at 75 years is 0.33 whereas that between BMI at 65 years and log weight at 75 years is 0.37.

We have fitted a very simple model here to describe the between subject variation and there appears to be little difference between the correlations. We now fit a more complex mode with random quadratic coefficients at the subject level. The results are in table 2.

Table 2. Between-individual covariance matrix for boys 11 - 16 years - correlations off-diagonal

	Log (wt.)	BMI	Age (logwt.)	Age (BMI)	Age^2 (logwt)	Age^2 (BMI)
Log (wt.)	0.032					
BMI	0.39	0.0080				
Age (logwt.)	0.11	-0.07	0.000086			
Age (BMI)	0.13	0.07	-0.05	0.000037		
Age^2 (logwt)	-0.82	-0.42	-0.44	0.56	2.1×10^{-7}	
Age^2 (BMI)	-0.00009	-1.95	-1.45	0.13	0.87	1.8×10^{-8}

We note that two correlations are less than -1 as a result of sampling fluctuations and indicates that further elaboration of the model would be useful. With this model we find the correlations at 69 and 71 years are respectively 0.403 and 0.376 and at 65 and 75 years are respectively 0.474 and 0.290. Thus, we obtain greater differentiation and the direction from log weight to BMI is associated with a higher correlation

The estimated correlations will depend upon the complexity of the model and we note that they are derived as functions of the parameterisation we have adopted rather than being estimated directly. In the next section we look at ways of directly estimating these correlations as model parameters, which should provide estimates less sensitive to the precise form of model.

Non linear estimation of cross-variate correlations

At the subject level we specify, for each variate separately, a random coefficient structure, for example with the age and age squared coeffs varying across subjects. We then specify the cross-variate covariance structure as follows by defining the following covariance functions

$$\begin{aligned} &\mathrm{cov}(e_{1j_1k(t)}e_{2j_2k(t-s)}) = \sigma_{12}g_1(s) \\ &\mathrm{cov}(e_{2j_2k(t)}e_{1j_1k(t-s)}) = \sigma_{12}g_2(s) \\ &s \geq 0,\ \ g_1(0) = g_2(0) = 1 \end{aligned} \qquad (2)$$

We can implement this in MLn by specifying the functions g_1, g_2 as random design vectors for the covariance term where g_1 is zero when $x_1 > x_2$ and vice versa for g_2. Several choices for these functions are possible, for example

$$\begin{aligned} &g(s) = \exp(-\alpha s) \\ &g(s) = \exp[-(\alpha_0 + \alpha_1 s^{\gamma})] \quad s > 0 \\ &\qquad 0 \qquad\qquad\qquad\qquad\ \ s = 0 \end{aligned} \qquad (2)$$

In both cases in (2) when *s*=0 the common covariance is $2\sigma_{12}$. Note that in the discrete occasion case with equal intervals the first function is a first order autoregressive process and we are then interested in a comparison of the autoregressive correlations. We can also have different functions for each correlation.

These functions are similar to those used by Goldstein et al (1994) in modelling multilevel time series, but here the correlation structure is at the highest rather than the lowest level of the hierarchy. The model is fitted using the linearisation procedures described by Goldstein (1995, Appendix 5.1).

In general, to make inferences about causality, we would wish to investigate the coefficients of the fitted functions g_1, g_2, and plotting these functions against *s*. We can elaborate the model by allowing these functions to depend on further explanatory variables, notably age itself, and also allowing the variances to change with age etc.

In practice we would expect the second function in (2) with $\gamma = 1$ to describe the structure with sufficient flexibility. Further work on fitting such functions is in progress.

Reference

Stini, W. A. (1990). Osteoporosis: etiologies, prevention and treatment. *yearbook of physical anthropology* **33**: 151-194.