STATISTICS IN MEDICINE, VOL. 16, 2349–2380 (1997)

TUTORIAL IN BIOSTATISTICS

USING THE GENERAL LINEAR MIXED MODEL TO ANALYSE UNBALANCED REPEATED MEASURES AND LONGITUDINAL DATA

AVITAL CNAAN[1]*, NAN M. LAIRD[2] AND PETER SLASOR[2]

[1]*Division of Biostatistics, Department of Pediatrics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, U.S.A.*

[2]*Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, U.S.A.*

SUMMARY

The general linear mixed model provides a useful approach for analysing a wide variety of data structures which practising statisticians often encounter. Two such data structures which can be problematic to analyse are unbalanced repeated measures data and longitudinal data. Owing to recent advances in methods and software, the mixed model analysis is now readily available to data analysts. The model is similar in many respects to ordinary multiple regression, but because it allows correlation between the observations, it requires additional work to specify models and to assess goodness-of-fit. The extra complexity involved is compensated for by the additional flexibility it provides in model fitting. The purpose of this tutorial is to provide readers with a sufficient introduction to the theory to understand the method and a more extensive discussion of model fitting and checking in order to provide guidelines for its use. We provide two detailed case studies, one a clinical trial with repeated measures and dropouts, and one an epidemiological survey with longitudinal follow-up. © 1997 by John Wiley & Sons, Ltd.

Statist. Med., **16**, 2349–2380 (1997)
No. of Figures: 6 No. of Tables: 8 No. of References: 34

1. INTRODUCTION

This tutorial deals with the use of the general linear mixed model for the regression analysis of correlated data. The correlation arises because subjects may contribute multiple responses to the data set. The model assumes a continuous outcome variable which is linearly related to a set of explanatory variables; it expands on the ordinary linear regression model by allowing one to incorporate lack of independence between observations and to model more than one error term. The types of data that can be analysed using the general linear mixed model include longitudinal

* Correspondence to: Avital Cnaan, Division of Biostatistics, Department of Pediatrics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, U.S.A.

Contract grant sponsor: National Institutes of Health
Contract grant number: GM 29745, RR00240-31

CCC 0277–6715/97/202349–32$17.50

data, repeated measures data (including cross-over studies), growth and dose–response curve data, clustered (or nested) data, multivariate data and correlated data. Despite this wide applicability, the results of such analyses are only now beginning to appear in the research literature; we will describe briefly several recent applications. The model's limited use is due in part to the fact that commercial software has been limited in the past, and in part to a lack of familiarity with the model by the general statistical and medical community. The goal of this tutorial is to introduce the model and its analysis to applied statisticians, to provide some guidelines on model fitting, and to illustrate its use with two data sets. Two recent books which cover aspects of the model and its analysis in more detail are those of Longford[1] and Diggle *et al.*[2]

We use the term longitudinal data to imply that each subject is measured repeatedly on the same outcome at several points in time. The main interest is usually in characterizing the way the outcome changes over time, and the predictors of that change. A typical example of a longitudinal study is the Tucson Epidemiological Study of Airways Obstructive Disease, in which the effects of smoking onset and cessation on changes in pulmonary function were assessed.[3] This study followed 288 subjects for 17 years. Using a random-effects model, the authors showed that the largest beneficial effect related to quitting smoking was in younger subjects and the effect decreased linearly with age at quitting. The advantage of this study over many studies in this area is that the modelling enabled the authors to account for both the age at onset of smoking and at quitting, in addition to other important covariates, and thus give a more comprehensive picture than previous studies which approached only one aspect or another of the problem.

A widely used and general term is repeated measures data, which refers to data on subjects measured repeatedly either under different conditions, or at different times, or both. An example is a study of stress and immune response in mothers of very low birth weight infants and mothers of normal weight infants.[4] The two groups of mothers, comparable in sociodemographic variables, were measured for stress and various immune function markers at delivery of the infant, one, two and four months post-delivery. The data for each outcome were analysed separately using the balanced and complete repeated measures data approach with an unstructured covariance matrix for the four measurements. The mothers of the very low birth weight infants had increased anxiety and decreased lymphocyte proliferation as well as decreased percentages of some immunologic cell subsets. Longitudinal models of these markers as either a linear or a quadratic function of time showed that resolution of the immunosuppression of pregnancy was substantially faster in mothers of very low birth weight infants than in mothers of normal weight infants, although neither group reached normal levels by four months. The advantage of using longitudinal models in this study was that the time of actual measurement of the immune markers was used, which often deviated for these mothers from the original preset schedule. This enabled the authors to obtain a well-fitting function of time, and model the resolution of immune markers' levels.

In growth and dose–response curve data, the subjects are ordinarily measured repeatedly at a common set of ages or doses. In a study conducted in Pittsburg, the purpose was to identify the effect of prenatal alcohol exposure on growth.[5] It has previously been established that prenatal alcohol exposure is associated with smaller birth size. However, it has not been clear whether there is subsequent catch-up growth, and cross-sectional studies have not been able to satisfactorily answer this question. This study followed the infants from birth to three years, and, using a growth curve model, was able to ascertain that there was no long-term catch-up growth; the smaller size observed at birth is maintained. The model used for analysis was a general unbalanced repeated measures model with a fully parameterized covariance matrix.

Clustered data arise commonly in surveys and observational studies of populations which have a natural hierarchical structure, such as individuals clustered in households and patients clustered in hospitals or other service delivery systems. Here interest may centre on which characteristics of the individual or of the cluster, or both, affect the outcome. An example of a data set where mixed-effects models were used in a cluster setting involves a study for predicting microbial interactions in the vaginal ecosystem.[6] The data for the study consisted of bacteria concentrations from *in vivo* samples; samples could be obtained from a subject once or multiple times during a study. The explanatory variables were concentrations of various specific bacteria, timing in menstrual cycle, and flow stage. The general approach for inclusion of variables was a backward elimination process. The mixed-effects modelling enabled the researchers to use all the data available, account for repeats within subjects, and model total aerobic bacteria, total anaerobic bacteria and mean pH values, as three different mixed-effects models, with different fixed and random effects.

Multivariate data refers to the case where the same subject is measured on more than one outcome variable. In this setting, one is looking at the effects of covariates on several different outcomes simultaneously. Using the general linear mixed model analysis allows one more flexibility than using traditional multivariate regression analysis because it permits one to specify different sets of predictors for each response, it permits shared parameters for different outcomes when covariates are the same, and it allows subjects who are missing some outcomes to be included in the analysis. A study with multivariate outcomes looked at individual pain trends in children following bone marrow transplantation.[7] Children aged 6–16 years and their parents were both asked to evaluate the child's pain on a standardized analogue scale daily for 20 days. Empirical Bayes' methodology was used to model pain as a quadratic function of number of days post-transplant. Separate models were used for describing children's and parent's pain ratings and curves were compared informally. Using a single analysis with a multivariate response of the children and parents would have allowed a direct statistical comparison of the pain curves. The analysis did show that the empirical Bayes' approach to the modelling gave a better fit than ordinary least squares (OLS) modelling separately for each child, which was possible in this example due to the large number of observations per child. In general, parents reported higher pain levels than the children, and there was some variability in reporting within the children by age group.

Correlated data is a generic term which includes all of the foregoing types of data as special cases. The only restriction is that the correlation matrix of the entire data matrix must be block diagonal, thus the model does not accommodate time series data, in which a non-zero correlation is assumed at least between every two consecutive observations. In all of these cases the observations are clustered, or grouped within subject. Thus we can structure the data by first identifying the subject or cluster, and then the repeated observation on the subject or the cluster. In describing this hierarchical structure, one might encounter in the literature the terms level one and level two, stage one and stage two, observation level and subject level, or subject level and cluster level; the latter two are especially confusing since sometimes subject is level one and sometimes it is level two, depending upon the application. Rather than exclusively using a single terminology in this paper, we will use the terminology most appropriate to the context and indicate in parenthesis what is meant, if necessary. In the absence of a particular context we will generally use subject and observation to denote the two levels of the hierarchy.

A wide variety of names are also used in the statistical literature to describe versions of the same model, reflecting the diversity of its use in many fields. These names include: mixed linear

model;[8] two-stage random effects model;[9] multilevel linear model;[10] hierarchical linear model;[11] empirical Bayes' model,[12] and random regression coefficients.[13] The hierarchical structure for the data leads naturally to the terms hierarchical linear model or multilevel linear model. Strictly speaking, both of these terms are used to describe models which may include more than two levels. Examples of more than two levels are commonly encountered in the educational literature where we have students grouped in classes and classes grouped in schools. Since many of the biomedical applications deal with only two levels, we will not consider the higher level models in this paper.

One can broadly characterize many versions of the models as methods for handling the between- and within-subject variability in the data. There are basically two ways which are commonly used to model these two types of variability. The approach that seems most natural to many investigators is to develop the model in two stages. At the first stage, we specify a separate linear regression for the observations on each subject, as a function of covariates made on the observations, for example, time of the observation. There will be as many of these regressions as we have subjects; each regression must use the same set of predictor variables, so that each regression has the same vector of coefficients, but the regression coefficients themselves are allowed to vary over subjects. At the second stage, the regression coefficients modelled at stage one are the random outcome variables, hence the terms random coefficient regressions and random effects models. The term mixed effects comes from combining the two regressions into a single equation with two types of regression coefficients. The two-stage approach is attractive to many investigators, since the subjects are explicitly treated as the unit of analysis at stage two, and one models directly the two types of regressions: the within at stage one and the between at stage two.

However, the approach can be limiting in the way models are developed for both the mean and the variance-covariance structure. The mixed model approach is to simply write out a single regression model for each observation. The model has the usual linear regression predictor for the mean response, but has two types of random error terms: between-subject errors and within-subject errors. All of the observations on the same subject will have the same between-subject errors; their within-subject errors will differ, and can be correlated within a subject. Both within- and between-subject errors are assumed independent from subject to subject, thus observations on different subjects are independent.

In Section 2 we introduce two case studies. In Section 3 we will introduce notation, and show how these two approaches to formulating the model are related. We then address estimation and testing issues and offer suggestions about strategies for model specification and checking. In Section 4 we discuss statistical software and illustrate the techniques applied to the case studies in Section 5. Source code for the software programs is provided in the Appendix.

2. DESCRIPTION OF CASE STUDIES

2.1. Clinical trial in patients with Schizophrenia

As an example to motivate this tutorial, we consider a clinical trial of a medication for the treatment of schizophrenia during an acute phase of illness.[14] This clinical trial was a double-blinded study with randomization among four treatments: three doses (low, medium and high) of an experimental drug and a control drug with known antipsychotic effects as well as known side-effects. Initial studies prior to this double-blinded study suggested that the experimental drug

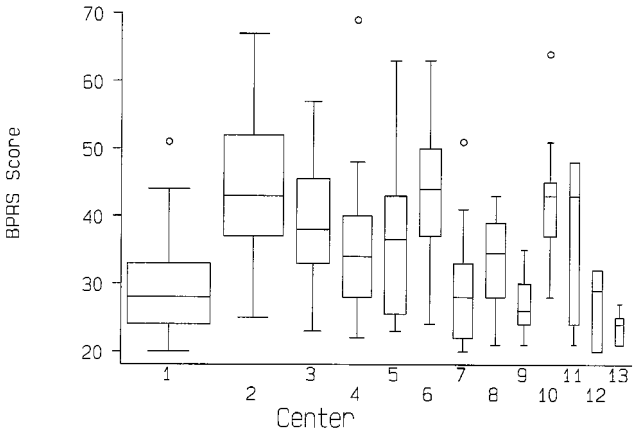

Figure 1. Distribution of BPRS by centre for the schizophrenia trial. In this box and whisker plot the box gives the interquartile range and the median. The whiskers are three halved the interquartile range rolled back to where there are data. Values beyond these are plotted as outliers. The width of the box is proportional to the number of subjects observed at baseline.

had equivalent antipsychotic activity, with lesser side-effects. The primary objectives of this study was the determination of a dose–response relationship for efficacy, tolerability and safely, and the comparison to the control drug. The study was conducted at 13 clinical centres, and a total of 245 patients were enrolled. The primary efficacy parameter was the Brief Psychiatric Rating Scale (BPRS).[15] This scale measures the extent of a total of 18 observed behaviours, reported behaviours, moods and feelings, and rates each one on a seven-point scale, with a higher score reflecting a worse evaluation. The total BPRS score is the sum of the scores on the 18 items. Study entry criteria included, among others, a score of at least 20 on initial evaluation. The distribution of BPRS scores at baseline was different in the different centres, as is shown in Figure 1. The width of the boxes is proportional to the number of patients contributing to the box, showing the different sample sizes in the different centres. Analysis of variance on the baseline BPRS shows a significant difference between centres ($p < 0{\cdot}0001$). In fact, 36 per cent of the baseline variability in BPRS scores is explained by centre. There were no differences in baseline BPRS scores between treatment groups. Because of the differences at baseline between the centres and because the randomization was done by blocks within centre, all analyses will include centre in the model.

Patients were evaluated at baseline and after one, two, three, four and six weeks of treatment. Patients were admitted to the hospital for the first four weeks of treatment, and discharged as the clinical condition permitted for the final two weeks. Since each patient had a different baseline value (range 20–69), the focus of this analysis is on the rate of change during the six weeks, rather than on a target desirable BPRS value. Of the patients, 134 (55%) completed the study, and 11 additional patients had a week six evaluation, even though they were technically considered non-completers. Figure 2 gives the box plots of BPRS values for each of the four treatments at each of the observation points. The width of the box plots gives a feel for the dropout rate.

The primary reason for discontinuation was a perceived lack of effectiveness of the treatment by the physician; there were also several withdrawals due to side-effects. Table I gives the number

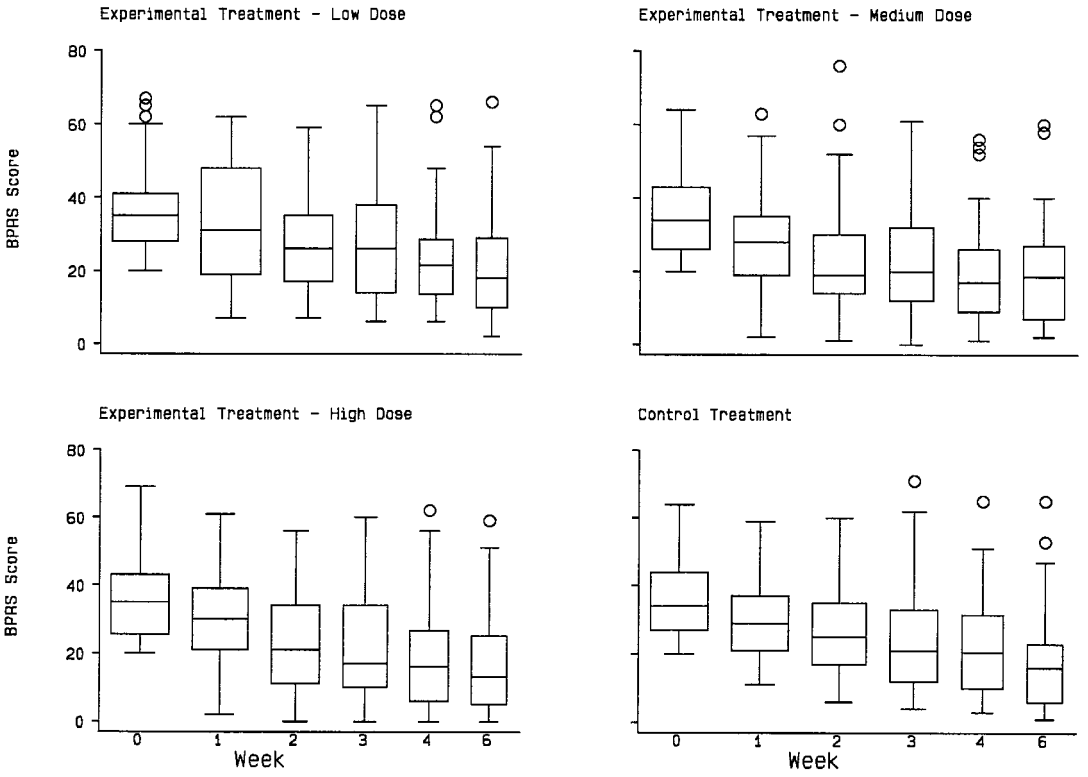

Figure 2. Distribution of BPRS by treatment group at each measurement occasion for the schizophrenia trial

Table I. Schizophrenia case study: number of patients measured at each time point

Week	Exp-low	Exp-medium	Exp-high	Control	Total
0	61	61	60	63	245
1	55	57	57	60	229
2	45	53	50	57	205
3	38	46	46	51	181
4	32	43	40	44	159
6	29	42	36	38	145

of patients at each observation point on each treatment. Table II gives the reasons for discontinuation by treatment. Table I shows that the largest dropout occurred on the experimental low-dose treatment, and Table II shows that the primary reason for this loss was perceived lack of therapeutic effectiveness. However, the tables also show that there was substantial dropout due to side-effects in the control treatment. When a discontinuation or a dropout occurred, every attempt was made to conduct a final observation at that time, often between two scheduled

Table II. Schizophrenia case study: reasons for discontinuation from clinical trial

Reason	Exp-low	Exp-medium	Exp-high	Control	Total
Completed*	27	40	33	34	134
Adverse experience	2	1	2	12	17
Lack of effect	21	7	13	11	52
Other reasons	11	13	12	6	42

* Several patients discontinued between weeks 4 and 6 and had a final observation, which was recorded as a week 6 observation, with the appropriate non-completer coding

observation points. For purposes of this analysis, the final observation was assumed to be at the next observation point, since the exact times were not available to us. Thus, for a patient who discontinued after the week 4 observation, their final observation was recorded at the week 6 observation point. Therefore, there are more patients with week 6 data (Table I) than there are completers (Table II).

In order to further understand the dropout, Cox regression models were applied to time until discontinuation. In these models, discontinuation was considered a failure event, and subjects who completed the study were considered censored at six weeks. High BPRS scores either at baseline, or later, or large differences (increases) in BPRS scores, were all associated with a shorter time to discontinuation. However, discontinuation was not related to treatment directly (in the presence of BPRS score), nor, in general, to centre. Results were similar either when the failure event was defined as discontinuation of any kind or when it was defined only as discontinuation due to lack of therapeutic response. The hazard ratio for all causes of discontinuation based on a difference of ten points on the BPRS scale was estimated as 2·4; for discontinuation due to lack of effect, the corresponding hazard ratio is 3·9.

2.2. Epidemiologic study of pulmonary function development in children

The Six Cities Study of Air Pollution and Health was designed to characterize pulmonary function growth between the ages of six and eighteen and the factors that affect growth.[16] A cohort of 13,379 children born in or after 1967 was enrolled in six communities: Watertown, MA; Kingston and Hariman, TN; a section of St. Louis, MO; Steubenville, OH; Portage, WI; and Topeka, KN. Most children were enrolled in the first or second grade and participants were seen annually until high school graduation or less to follow-up. At each examination, spirometry was performed and FEV_1, the volume of air exhaled by force in the first second, was measured. For the purpose of the analyses demonstrated here, a subset of 300 girls from Topeka, KN, with a total of 1994 measurements is presented. The data included their FEV_1, age and height. Baseline mean age was 8.29 years (SD 1·38) and mean height was 1·29 metres (SD 0·095). The number of measurements (observations) for each girl ranged from 1 to 12, and the total follow-up duration for each girl ranged from zero to 11 years.

In our analysis, we will replicate the modelling approach suggested by Hopper *et al.*[17] in a similar Australian study. In that study, the natural logarithm of the FEV_1 was found to be well predicted by both current height and age. despite the strong correlation between the two. He also found that an autocorrelation structure represented the correlation in the data better than a random effects structure.

3. THE DEVELOPMENT OF THE GENERAL LINEAR MIXED MODEL

3.1. Design considerations

We will let $i = 1, \ldots, N$ index the subjects in the study and $\boldsymbol{y}_i$ denote the vector of observations made on the ith subject. Since subjects are allowed to have unequal numbers of observations, the vectors $\boldsymbol{y}_i$ may have different dimensions. In designed experiments, the number of observations is usually fixed to be the same for all subjects, and inequalities arise as a result of missing data. In the schizophrenia trial (Section 2.1), the length of $\boldsymbol{y}_i$ varies from one to six depending upon if and when the patient dropped out. In clustered data applications, where the second level might be a hospital, and observations are made on different patients within the hospital, the dimension of $\boldsymbol{y}_i$ can vary arbitrarily from hospital to hospital. We will denote the dimension of $\boldsymbol{y}_i$ by n_i, so that the total number of observations that we have in the data set are $\sum_{i=1}^{N} n_i$.

Each subject also has a vector of covariates; in the schizophrenia trial (Section 2.1), the subject level covariates are treatment group and centre. In addition, each observation typically has its own covariates. The subject level covariates remain constant for all the repeated observations on a subject, whereas observation level covariates may vary as the repeated observations are made. In the schizophrenia trial, time and functions of time such as time squared etc., are the main observation level covariates. With clustered data where observations on different subjects are the repeated measures, the observation level covariates can be numerous, for example, age, sex, severity score, etc. In cross-over trials where one subject receives several treatments in different periods, both period and treatment are observation level covariates.

Many longitudinal studies where an individual is the subject typically have time as the only observation level covariate. The FEV_1 data example (Section 2.2) we present is an exception, where two important predictors, age and height, are characteristics of the observation. Sometimes a variable can appear as a covariate at both levels. For example, Wechsler *et al.* used a two-stage cluster survey to study drinking behaviour in college students.[18] First a random sample of colleges was drawn, then a random sample of students were surveyed in each college appearing in the stage one sample. In the analysis, both the sex of the individual student (observation level) and the per cent of students in the college who were a specific sex (cluster level) were used as covariates. Another example arises in longitudinal studies where subjects are measured at different initial ages, and both current age and initial age can appear as predictors in the model.[19] In this case, the initial age is a subject level covariate and current age is a covariate at the observation level.

This distinction between observational level and subject level covariates is important, since this typically drives model formulation and selection, and determines to some extent the amount of information available in the data to estimate the effects of different variables. The repeated observations on a subject are usually assumed to be positively correlated; this implies that the effects of within-subject (or observation level) variables will typically be estimated with greater precision than the effects of between-subject variables. This is because the former are estimated using within-subject contrasts, and the variance of a contrast is reduced when the observations are positively correlated. This is the basic principle of a split-plot design, where the whole plot (or between-subject) factor is estimated with less precision than the subplot (or within-subject) factor and the interaction of the two factors.[20] Likewise, the cross-over design has potential for greater efficiency than the 'parallel' design because treatment contrasts are constructed using positively correlated responses.[21] In the context of longitudinal studies, the implication is that time trends, and the effects of subject level variables on time trends, will be the effects best estimated.[22]

In classically designed experiments with balanced and complete data, covariates (or factors) typically vary either within subjects (observation level) or between subjects (subject level) but not both. For example, if each schizophrenia patient (Section 2.1) were observed on all six occasions, time would vary within but not between patients, since each patient would have the same vector of values for time. Treatment and centre vary between but not within subjects. This clean separation of between- and within-subject variables is a hallmark of designed experiments and has the advantage of leading to orthogonal designs and a simplified repeated measures analysis. With observational studies, unbalanced designs and/or missing data, it is rarely possible to achieve this clear separation of between- and within-subject variables. For example, in the FEV_1 data set (Section 2.2), it is not logistically possible to design the study so that each subject is measured at the same set of ages and heights. If a covariate varies both within and between subjects, it may be important to specify a model which reflects this, as in the examples previously cited.[18,19]

3.2. Model development

We will begin in this section by formulating a two-stage or two-level model and then show its relationship to the general linear mixed model. Let $\boldsymbol{Z}_i$ denote the $n_i \times q$ matrix of observation level covariates for subject i. Typically the first column is a vector of ones for the intercept, but the rest of the columns must be variables that vary within the subject (observation level covariates). It is not necessary that the values vary within every subject, that is, $\boldsymbol{Z}_i$ may be less than full rank for some individuals, but by definition, $\boldsymbol{Z}_i$ contains the observation level covariates and should consist of variables which model within-subject variation. For example, in the schizophrenia trial, we might posit that the response over time for each subject follows a quadratic curve. In this case $\boldsymbol{Z}_i$ would be $n_i \times 3$; the first column would be all ones, the second would be the time of the observation, and the third would be time squared.

Parenthetically, when modelling time trends, it is often desirable to use orthogonal polynomials for numerical stability or interpretation. In doing so, it is important to use a single set of orthogonal polynomials for all subjects, even though different subjects may be observed at different times. In the schizophrenia data set we could use the protocol design, that is observations at weeks 0, 1, 2, 3, 4, 6 to calculate our orthogonal polynomials. Then $\boldsymbol{Z}_i$ would be orthogonal only for the subjects observed at all six occasions, but the meaning of the coefficients remains the same for each subject. The first-stage (or level one, observation level) regression model is given by

$$y_{ij} = \boldsymbol{z}_{ij}\beta_i + e_{ij}, \qquad i = 1, \ldots, N \tag{1}$$

where $\boldsymbol{\beta}_i$ represents the $q \times 1$ vector of regression coefficients for the ith subject, $\boldsymbol{z}_{ij}$ is the jth row of $\boldsymbol{Z}_i$, and the e_{ij} are zero mean error terms. The $\boldsymbol{\beta}_i$ represent inherent characteristics of the subjects, for example, parameters of a subject's 'true' growth curve, and the e_{ij} can be thought of as sampling error, or random perturbations. The e_{ij} are typically taken to be independently and identically distributed with variance σ^2. In cases where the observations have a clear ordering or structure, some investigators alternatively assume that correlation among the e_{ij} is non-zero, and varies in a systematic way. For example, with equally spaced points in time, one might assume simple autoregressive structure (AR1) for $\mathrm{var}(\boldsymbol{e}_i)$, where $\boldsymbol{e}_i^{\mathrm{T}} = (\boldsymbol{e}_{i1}, \ldots, e_{in_i})$. When the $\boldsymbol{e}_i$ can be thought of as measurement or sampling error the assumption of independence is natural. We let $\boldsymbol{R}_i = \mathrm{var}(\boldsymbol{e}_i)$. In the second-stage regression (or level two, subject level), the $\boldsymbol{\beta}_i$ are regarded as

N independent q-dimensional random vectors, hence the term random regression coefficients. Their mean depends upon subject level characteristics. Let $\boldsymbol{a}_{ki}$ denote the vector of subject level characteristics which affect the mean of the kth coefficient. To model the level two regression, we assume the $\boldsymbol{\beta}_i$ are independently distributed with

$$E(\boldsymbol{\beta}_i) = \boldsymbol{A}_i\boldsymbol{\alpha} \tag{2}$$

where

$$\boldsymbol{A}_i = \begin{bmatrix} \boldsymbol{a}_{1i}^{\mathrm{T}} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{a}_{2i}^{\mathrm{T}} & & \\ \vdots & & & \\ \mathbf{0} & . & \cdots & \boldsymbol{a}_{qi}^{\mathrm{T}} \end{bmatrix}$$

$$\boldsymbol{\alpha}^{\mathrm{T}} = (\boldsymbol{\alpha}_1^{\mathrm{T}}, \ldots, \boldsymbol{\alpha}_q^{\mathrm{T}}).$$

and the $\boldsymbol{\alpha}_k$'s are regression parameter vectors, of varying length depending upon the number of subject level covariates (in $\boldsymbol{a}_{ik}$) which affect the mean of the kth coefficient. We further assume

$$\mathrm{var}\,(\boldsymbol{\beta}_i) = \boldsymbol{D}.$$

The diagonal elements of $\boldsymbol{D}$ (a $q \times q$ matrix) tell us how much the individual regression coefficients, the $\boldsymbol{\beta}_i$, vary from subject to subject, after adjusting for the covariates in $\boldsymbol{A}_i$. Thus the $\boldsymbol{D}$ matrix models the between-subject variance, while the σ^2 from level one models the within-subject (observation level) variance.

Because we have specified a linear model for the mean of β_i, it is convenient to write β_i as

$$\boldsymbol{\beta}_i = \boldsymbol{A}_i\boldsymbol{\alpha} + \boldsymbol{b}_i$$

where $\boldsymbol{A}_i\boldsymbol{\alpha}$ are as defined above, and the $\boldsymbol{b}_i$ are independent and identically distributed with zero mean and variance $\boldsymbol{D}$. Here each $\boldsymbol{b}_i$ can be regarded as the ith subject's random deviation from the mean ($\boldsymbol{A}_i\boldsymbol{\alpha}$). Rewriting equation (1) in vector and matrix notation we have

$$\boldsymbol{y}_i = \boldsymbol{Z}_i\boldsymbol{\beta}_i + \boldsymbol{e}_i$$

which in turns implies

$$\boldsymbol{y}_i = \boldsymbol{Z}_i\boldsymbol{A}_i\boldsymbol{\alpha} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i. \tag{3}$$

As a result

$$E(\boldsymbol{y}_i) = \boldsymbol{Z}_i\boldsymbol{A}_i\boldsymbol{\alpha}$$

and

$$\mathrm{var}\,(\boldsymbol{y}_i) = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^{\mathrm{T}} + \boldsymbol{R}_i.$$

Notice that when we combine the two regressions into a single model for the response as in (3), we have two types of regression parameters, the $\boldsymbol{\alpha}$, and the $\boldsymbol{b}_i$'s (since there are N subjects there are N different $\boldsymbol{b}_i$'s). The $\boldsymbol{\alpha}$ parameter is often termed the fixed effect, in contrast to the $\boldsymbol{b}_i$'s, which are called random effects or random coefficients. The $\boldsymbol{b}_i$'s can also be viewed as error terms sine they are random variables with zero mean.

The form of the design matrix for the fixed effects in model (3) implies that any observation level variables must be specified as random effects if they are to be included in the model. This can be an unattractive modelling strategy in situations where there are several observation level variables or they are all indicator variables, as in cross-over designs. One way to avoid this requirement is to force components of $\boldsymbol{D}$ to be zero. Alternatively, the model is made much more flexible by replacing $\boldsymbol{Z}_i\boldsymbol{A}_i$ by an arbitrary design matrix $\boldsymbol{X}_i$, and writing

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i. \tag{4}$$

Equation (4) defines what we call the general linear mixed model. We use the term general linear mixed model to emphasize that $\boldsymbol{X}_i$, $\boldsymbol{Z}_i$ and $\boldsymbol{R}_i$ can be quite general; other versions of the model place implicit constraints on one or all of these matrices. The model has two types of error terms: the $\boldsymbol{b}_i$ are the random subject effects and the $\boldsymbol{e}_i$ are observation level error terms. In the general linear mixed model we assume only that the $\boldsymbol{b}_i$ and $\boldsymbol{e}_i$ are independently distributed with zero mean and variances $\boldsymbol{D}$ and $\boldsymbol{R}_i$, respectively, implying that

$$E(\boldsymbol{y}_i) = \boldsymbol{X}_i\boldsymbol{\alpha}$$

and

$$\mathrm{var}(\boldsymbol{y}_i) = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^{\mathrm{T}} + \boldsymbol{R}_i.$$

3.3. The general linear mixed model: estimation and testing

Under the assumption that the $\boldsymbol{b}_i$ and $\boldsymbol{e}_i$ are independently distributed as multivariate normal, estimation of the parameters by maximum likelihood (ML) is straightforward.[23] The variance-covariance parameters can also be estimated using Restricted Maximum Likelihood (REML). As the name implies, the REML estimates maximize the likelihood of the error contrasts, rather than the full data, and are often preferred over ML estimates since they yield well known method-of-moment estimators in balanced cases which have closed form solutions.

The maximum likelihood estimate of $\boldsymbol{\alpha}$ is given by

$$\hat{\boldsymbol{\alpha}} = (\sum X_i^{\mathrm{T}}\hat{\boldsymbol{V}}_i^{-1}\boldsymbol{X}_i)\sum\boldsymbol{X}_i^{\mathrm{T}}\hat{\boldsymbol{V}}_i^{-1}\boldsymbol{y}_i \tag{5}$$

where

$$\boldsymbol{V}_i = \mathrm{var}(\boldsymbol{y}_i) = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^{\mathrm{T}} + \boldsymbol{R}_i$$

and $\hat{\boldsymbol{V}}_i$ is $\boldsymbol{V}_i$ with $\boldsymbol{D}$ and $\boldsymbol{R}_i$ replaced by their ML estimators. In practice, REML estimators are often used to estimate $\boldsymbol{V}_i$. The estimator of $\boldsymbol{\alpha}$ defined in equation (5) is also called the Generalized Least Squares (GLS) estimator because it is inverse variance weighted least squares, where each subject's estimated variance, $\hat{\boldsymbol{V}}_i$, determines their weight in the estimation of $\boldsymbol{\alpha}$. The variance of $\hat{\boldsymbol{\alpha}}$ is usually estimate by

$$\mathrm{var}(\hat{\boldsymbol{\alpha}}) = (\sum\boldsymbol{X}_i^{\mathrm{T}}\hat{\boldsymbol{V}}_i^{-1}\boldsymbol{X}_i)^{-1}. \tag{6}$$

The GLS estimator (and hence the ML estimator) of $\boldsymbol{\alpha}$ has good optimality properties that hold without the assumption of normality for the error terms ($\boldsymbol{e}_i$ and $\boldsymbol{b}_i$). It is consistent, asymptotically normal, and fully efficient if $\boldsymbol{V}_i$ correctly specified the $\mathrm{var}(\boldsymbol{y}_i)$. If we have misspecified the variance, so that $\mathrm{var}(\boldsymbol{y}_i) \neq \boldsymbol{V}_i$, $\hat{\boldsymbol{\alpha}}$ is still consistent and asymptotically normal, but not fully efficient, and

equation (6) is not a valid estimate of var $(\hat{\boldsymbol{\alpha}})$. Liang and Zeger[24] suggest using an alternative estimator of var$(\hat{\boldsymbol{\alpha}})$ which is valid when var$(\boldsymbol{y}_i) \neq \boldsymbol{V}_i$

$$\text{var}(\hat{\boldsymbol{\alpha}}) = (\sum \boldsymbol{X}_i^{\mathrm{T}} \hat{\boldsymbol{V}}_i^{-1} \boldsymbol{X}_i)^{-1} \sum \boldsymbol{X}_i^{\mathrm{T}} \hat{\boldsymbol{V}}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\alpha}})(\boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\alpha}})^{\mathrm{T}} \hat{\boldsymbol{V}}_i^{-1} \boldsymbol{X}_i (\sum \boldsymbol{X}_i^{\mathrm{T}} \hat{\boldsymbol{V}}_i^{-1} \boldsymbol{X}_i)^{-1}.$$

Because this estimate of var$(\hat{\boldsymbol{\alpha}})$ is not currently available via commercial software, we do not consider it further. Tests for covariance structure are considered in Section 3.5.

Three approaches are used to test hypotheses of the form $H_0: \boldsymbol{C}\boldsymbol{\alpha} = \boldsymbol{0}$. Likelihood ratio tests can be used with large samples, providing one uses ML, rather than REML, for model fitting. Standard Wald tests are widely available; these are also asymptotically χ^2. Approximate F-tests for class variables can be carried out by dividing the Wald test by the numerator degrees-of-freedom and approximating the denominator degrees-of-freedom. All of these tests are large sample, and more research on small sample adjustments is needed.[25] As with ordinary univariate regression, normality of the errors is not required for asymptotic normality of the estimate of $\boldsymbol{\alpha}$, but highly skewed error distributions can lead to invalid tests and confidence intervals, thus it is wise to consider transformations to promote normality, if appropriate.

Sometimes it is useful to have estimates of the individual random effects, the $\boldsymbol{b}_i$'s, or equivalently, the $\boldsymbol{\beta}_i$'s. The random effects are estimated using 'shrinkage' or empirical Bayes estimators which represent a compromise between estimates based only on an individual subject's data, and estimates based only on the population mean. These estimates can be used to construct individual growth curves for each subject, as we illustrate with the FEV_1 data (Section 2.2). For example, Tsiatis *et al.*[26] used estimates of individual trajectories of CD4 counts to study the relationship between fall in CD4 count and disease progression in AIDS. Subjects with substantial data will have estimates close to their individual least squares curves, whereas subjects with sparse data will have estimates close to the population mean. These shrinkage estimators are also known as best linear unbiased predictors (BLUP).[27]

3.4. Model selection

Selecting a model means specifying a design matrix for each subject, $\boldsymbol{X}_i$, the random effects (or $\boldsymbol{Z}_i$), and the variance-covariance structure for $\boldsymbol{e}_i$, $\boldsymbol{R}_i$. In many cases, the mean parameters, or $\boldsymbol{\alpha}$, are those of most interest and the random effects and correlation structure can be viewed as nuisance quantities. Even in this latter case it is important to model var$(\boldsymbol{y}_i)$ carefully, since it affects both the efficiency of $\hat{\boldsymbol{\alpha}}$, and the validity of the estimate of var$(\hat{\boldsymbol{\alpha}})$.

Because $E(\boldsymbol{y}_i) = \boldsymbol{X}_i \boldsymbol{\alpha}$, the mean parameters are determined entirely by the design matrix $\boldsymbol{X}_i$; it can include any arbitrary combination of subject level and observation level covariates that we desire. The $\boldsymbol{Z}_i$ matrix specifies the design for the random effects. It must contain only observation level covariates, apart from an intercept. In growth curve studies it typically contains the design on time or age. Whereas $\boldsymbol{X}_i$ determines our specification for the mean, the $\boldsymbol{Z}_i$ matrix, or specification of random effects, only affects variance. To see this, note that it is possible to specify no random effects so that $\boldsymbol{Z}_i = \boldsymbol{0}$. Here we still have $E(\boldsymbol{y}_i) = \boldsymbol{X}_i \boldsymbol{\alpha}$ but now var$(\boldsymbol{y}_i) = \boldsymbol{R}_i$. Thus, in many cases the use of random effects can be viewed simply as a device for modelling correlation structure.

3.4.1. Modelling the mean

Although it is technically possible to specify $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ independently, in the case where we do choose to include random effects, there are some guidelines one should use in choosing models. In

general, the columns of $\boldsymbol{Z}_i$ should be a subset of the columns of $\boldsymbol{X}_i$ to avoid the problem of specifying as 'error terms' effects which do not have zero mean. To return to our schizophrenia example (Section 2.1), we could assume that a quadratic curve is needed to characterize the mean response for each subject, implying an initial large effect, followed by a levelling off of the effect. We would not want to include a quadratic effect in $\boldsymbol{Z}_i$ and omit it in the mean ($\boldsymbol{X}_i$), since this model would imply that each subject had a quadratic curve, but that the population curve was linear.

Alternatively, omitting observation level effects from $\boldsymbol{Z}_i$ that are contained in $\boldsymbol{X}_i$ can lead to serious underestimation of the standard errors of $\hat{\boldsymbol{\alpha}}$. Consider the case where we fit linear time trends for the mean, but assume that only the intercepts are random. In this case, $\boldsymbol{Z}_i$ is a column of ones for each subject, and, if $\boldsymbol{R}_i$ is taken to be $\sigma^2\boldsymbol{I}$, then the variance of $\boldsymbol{y}_i$ has the compound symmetry assumption. This assumption will cause the estimated standard errors of estimated slopes to be too small if it fails to hold. Including the linear term as a random effect corrects this problem.[28] Assuming $\boldsymbol{R}_i$ has a richer structure is an alternative solution.

One advantage of the two-stage approach to modelling is that if $\boldsymbol{X}_i = \boldsymbol{A}_i\boldsymbol{Z}_i$, where $\boldsymbol{A}_i$ contains all the subject level covariates and $\boldsymbol{Z}_i$ contains all of the observation level covariates, then the $\boldsymbol{Z}_i$ are necessarily a subset of $\boldsymbol{X}_i$ and any observation level variables modelled in $\boldsymbol{X}_i$ are also contained in $\boldsymbol{Z}_i$. Recall that $\boldsymbol{A}_i$ contains the vectors of covariates, $\boldsymbol{a}_{ki}$, for the regression of the kth element of $\boldsymbol{\beta}_i$ on subject level variables. Each of the components in $\boldsymbol{\beta}_i$ may have a different set of predictor covariates. In most settings there will be considerable overlap in the covariate set, and we may want to assume $\boldsymbol{a}_{ki} = \boldsymbol{a}_i$ for all k.

To fix ideas, consider the schizophrenia study (Section 2.1) where we fit a quadratic response function in time, so that each $\boldsymbol{Z}_i$ is $n_i \times 3$ with an intercept, a linear and a quadratic term. The subject level characteristics are centre and treatment group. The regression model we choose for $\boldsymbol{A}_i$ will depend upon how we have parameterized time. Suppose we do not use orthogonal polynomials so that the intercept estimates baseline (or prerandomization) means. We would certainly want to include centre as a predictor of baseline means since we know that mean BPRS scores are different at baseline. This may be due to differences in centre populations or to differences in interpreting the BPRS rating system. We could have centre as a predictor of the slope and quadratic terms as well to account for possible systematic differences in time trends from centre to centre. However, if data are sparse at the centre level, we might fit a more parsimonious model, omitting centre as a predictor of the quadratic or linear and quadratic terms. Since the study is randomized, treatment is formally not a predictor of intercept, although it is preferable to include it in the model, to avoid possible model misspecification. Thus we would include both centre and treatment in $\boldsymbol{a}_{1i}$. If we want to test for treatment effects, we would omit treatment as a predictor for both the linear and quadratic terms, and compare the results to models fit including treatment as a predictor of the linear or the linear and quadratic terms. If we use orthogonal polynomials, then treatment effects would also have to be tested by looking at the effect of treatment on the intercept, which now estimates a mean level over all times. Since the groups were designed to be equivalent at baseline, testing for a treatment difference in a mean level which includes baseline is not very compelling. Notice that the 'main' effects are all specified by the regression of the intercept on $\boldsymbol{a}_{1i}$ and the remaining portions of $\boldsymbol{A}_i$ specify interactions of subject level and observation level variables.

When the observation level variables are class variables, as in cross-over and many other repeated measures designs, using random effects beyond a random intercept is not natural, hence the more general form of $\boldsymbol{X}_i$ is useful. In a cross-over study, one choice of $\boldsymbol{X}_i$ might be a column for the intercept, $T-1$ columns for the T treatments and $P-1$ columns for the P periods (or

occasions), all of which are observation level covariates. However, $\boldsymbol{Z}_i$ can still include only a random intercept. In many longitudinal studies with a small number of observations per subject it may also be attractive to use class variables to model the mean response over time. This allows one to model the data without making any assumptions about time trend, rather than assuming a parametric model for how the mean response varies as a function of time. For example in the schizophrenia study, there are six possible observation times. One could use three columns of $\boldsymbol{X}_i$ for intercept, linear and quadratic terms. Alternatively, one could use five indicator variables to uniquely define the observation times.

Another issue that often arises, especially in the context of longitudinal clinical trials, is how to handle the response made at baseline. It can be included as part of the response vector, or one can use the baseline measurement as a subject level covariate in $\boldsymbol{X}_i$, or one can subtract the baseline response from each subsequent response and model the vector of changes. In the schizophrenia study (Section 2.1), the first approach implies that each subject has six observations. The time trend (linear or quadratic) is fit from baseline to week 6. The second approach implies that each subject has five observations and the intercept in the model is part of the treatment effect. The coefficient of the baseline covariate captures the changes in response from baseline to week 1, while the time trend coefficients capture changes from week 1 to week 6. In both approaches, because there are large differences in the response at baseline between centres, 12 indicator variables for the 13 centres are required in $\boldsymbol{X}_i$ for centre effects. However, in the first approach, because the baseline is part of the response, one might need additional 12 or 24 variables for centre by time linear or linear and quadratic interactions. In the second approach, because the baseline in not part of the response, and the time trend coefficients model only the changes from week 1 to week 6, there is a potential for simpler models without centre by time interactions. The third approach of using changes from baseline rather than baseline as a covariate is appealing to clinicians, but is generally less statistically efficient.

3.4.2. Modelling the Variance

The variance of $\boldsymbol{y}_i$, $\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^{\mathrm{T}} + \boldsymbol{R}_i$, is determined by how we model $\boldsymbol{R}_i$ and the random effects. One commonly used strategy is to set $\boldsymbol{R}_i = \sigma^2\boldsymbol{I}$, and use random effects to give additional structure to $\boldsymbol{V}_i$. For example, in the schizophrenia study, setting $\boldsymbol{R}_i = \sigma^2\boldsymbol{I}$ implies independence between observations within the subject. If $\boldsymbol{Z}_i$ is the $n_i \times 3$ vector with intercept, linear and quadratic time trends, then $\boldsymbol{D}$ is a 3×3 matrix which provides the additional structure to $\boldsymbol{V}_i$. This generally works well when we are fitting parametric curves in longitudinal studies; it is especially useful when subjects are observed at a large number of irregular times. A limiting feature is that if each subject is observed on only a few occasions, lack of identifiability and convergence problems can occur even for only two or three random effects. This strategy is also limiting when we are using indicator variables to model the time course, because it implies estimation of $n \times (n+1)/2$ parameters in the matrix, which may be a large number and result in poor estimation.

At the other extreme, we may assume no random effects ($\boldsymbol{Z}_i = \boldsymbol{0}$), and put additional structure on $\boldsymbol{R}_i$. If each subject has observations at the same set of T occasions, and T is modest relative to N, we may choose to let $\boldsymbol{R}$ be an arbitrary $T \times T$ correlation matrix. Because $\boldsymbol{R}$ can be totally arbitrary (positive definite) $\boldsymbol{R}$ can be called an 'unstructured matrix'. If a subject is missing an observation, $\boldsymbol{R}_i$ is $\boldsymbol{R}$ with the rows and columns corresponding to the missing observations removed. In the schizophrenia study, since there are $T = 6$ possible observation points and $N = 245$, we use this approach. This strategy is also attractive for modelling multivariate data,

where each component of $\boldsymbol{y}_i$ is a different variable, rather than a repeated measure on the same variable. For example, if we wish to model height, weight and arm circumference as a function of age and sex, the response variable has three components which are not repeats on the same variable, and an arbitrary $\boldsymbol{R}$ may be the only logical choice for a variance structure.

Diggle[29] proposed a useful model which can be viewed as an extension of the general linear mixed model. He suggests using a single subject effect (b_i is scalar and $\boldsymbol{Z}_i$ a vector of ones), setting $\text{var}(\boldsymbol{e}_i) = \sigma_e^2 \boldsymbol{I}$, and introducing a third random error vector, say $\boldsymbol{r}_i$, which has an autoregressive error structure. Specifically, he assumed $\text{var}(\boldsymbol{r}_i) = \sigma_r^2$, and that the correlation between r_{ij} and r_{ik} depends only on the distance or time between the observations. He suggests several parametric forms for the correlations which model it as a decreasing function of increasing time. This model has the attractive feature that the variance and correlation are modelled with only a few parameters, even when there are many different observation times, the correlation is highest for measurements close together in time, but does not necessarily get to zero even for measurements made far apart. An advantage of the Diggle model is that it allows one to model correlation with very irregularly spaced intervals using only one random effect. It does still assume constant variance over time. In the longitudinal data setting, this assumption is often violated. For example, in modelling growth data, values at adolescence may show much more variability than early childhood data.

Specifying $\boldsymbol{R}_i$ and the random effects involves a trade-off. If the number of repeated measures is small, say less than 5, and N is substantial, say at least 100, the most attractive strategy may be to set $\boldsymbol{Z}_i = \boldsymbol{0}$ and let $\boldsymbol{R}_i$ be arbitrary. With larger numbers of repeat observations, some modelling strategy which incorporates random effects is attractive. This is especially true for longitudinal data which is highly unbalanced, for example, observations are made at arbitrary times leading to a highly unbalanced data set such as in data registries involving patient follow-up. In this case, setting $\boldsymbol{R}_i = \sigma^2 \boldsymbol{I}$ and using random effects may be the only feasible strategy. The Diggle[29] approach is one that has been used successfully to model FEV_1 data (Section 2.2).[17]

It is not possible to include an arbitrary set of random effects and let $\boldsymbol{R}_i$ be completely general, because this results in overparameterization. Consider the following simple case. If we specify that $\boldsymbol{R}$ has compound symmetry structure and let $\boldsymbol{Z}_i$ be a vector of ones, then $\boldsymbol{V}_i$ is the sum of two matrices with identical structure (constant variance on the diagonal and same correlation at all positions off the diagonal) and they are not separately identifiable. With large numbers of repeat measures, it is possible to fit several random effects and specify some non-saturated structure for $\boldsymbol{R}_i$, such as autoregressive, but general rules for identifiability of model parameters are not available. A simple approach that seems to work well in many situations is to either load up the variance in the random effects and set $\boldsymbol{R}_i = \sigma^2 \boldsymbol{I}$, or set $\boldsymbol{Z}_i = \boldsymbol{0}$ and allow $\boldsymbol{R}_i$ to be arbitrary. Practical experience suggests that the simple compound symmetry assumption, which involves setting $\boldsymbol{R}_i = \sigma^2 \boldsymbol{I}$ and using a single random subject effect (intercept), typically fails to adequately model the variance structure. Often adding just a single additional random effect is adequate.

3.5. Model checking

Many issues of model sensitivity and checking goodness-of-fit are exactly the same as those which arise in ordinary least squares regression. The standard residual diagnostics can be helpful; it can also be useful to look at residuals stratified on time (in longitudinal studies) or by subject. Waternaux *et al.*[30] suggest methods for outlier detection based on the empirical Bayes estimates of the subject random effects. They can help identify outlying subjects, rather than outlying observations.

One issue that is different from ordinary univariate regression is that a model must be specified for var($\boldsymbol{y}_i$). Likelihood ratio tests based on the REML likelihoods can be used informally for comparing nested models for var ($\boldsymbol{y}_i$), but the asymptotic χ^2 distribution fails to hold because the null hypothesis corresponds to a boundary value. Some computer packages give goodness-of-fit measures based on Schwartz's Bayesian Criterion and the Akaike's Information Criterion. Both of these are functions of the likelihood, with a penalty for the number of covariance parameters. For both criteria, larger values imply better fitting models, and can be used to compare non-nested models. More informally, in the balanced setting, one can compare a fitted model with the empirical covariance matrix, which is obtained by fitting the unstructured variance model. This is preferable to using the 'all available pairs' estimate available from some software packages, since the latter may be biased by missing data. Diggle[29] and Laird *et al.*[19] illustrate the use of the empirical semi-variogram to check model adequacy of the correlation structure in more complex settings where models are not nested; it is not currently available with software for the general linear mixed model.

3.6. Missing data

From a technical point of view, it is easy to handle missing responses in the outcome since there is no requirement that all n_i are equal, or that subjects are measured at the same set of occasions. All that is required is that the design matrix and correlation structure can be specified for the vector of responses that are observed. A bigger issue is the validity of the estimates in the presence of missing data. If the missingness is unrelated to outcome, then the ML estimates are valid and fully efficient. This is probably the case with most missingness in the FEV_1 data set (Section 2.2) since data were collected in school, and missingness was related to school attendance, and likely not to child FEV_1.

The probability of missingness may often be related to covariates; in principle this causes no difficulty but in practice it implies one should take care in the process of model selection. For example, suppose children who smoke are likely to have poorer attendance records, or leave the school early. If smoking affects FEV_1, then the probability of missingness is indirectly related to outcome. In this case, smoking status should be used as a subject level covariate in the model for FEV_1 in order to avoid bias due to missing data. The parameter estimates would then model mean response conditional on smoking. If we want to estimate the marginal mean FEV_1 at each age, unconditional on smoking, we will need to take the appropriate weighted combination of smoker and non-smoker means.

If missingness is related to observed responses but not missing ones, termed Missing at Random (MAR) by Little and Rubin,[31] then the estimates will be valid and fully efficient provided that the model assumed for the data distribution is correct. Dropouts in longitudinal studies are sometimes regarded as MAR if one can argue that the likelihood of dropping out only depends upon (observed) past history, and not on future values. In the schizophrenia trial (Section 2.1), Cox regression analysis was used to show that the probability of discontinuing on protocol, and hence subsequent missingness, was strongly related to changes in BPRS. The validity of the ML analysis that we present rests on the assumption that conditional on their observed BPRS outcomes, treatment group and centre, the future BPRS values of a discontinued subject can be predicted from the assumed multivariate normal distribution. The implication of this is that the distribution of future values for a subject who drops out at time t is the same as the distribution of future values for a subject who remains in at time t, if they have the same covariates, and the same past history of outcome until and time t.

If dropout depends upon some other process which is related to outcome, then conditioning on past outcomes may not capture all of the dependence of dropout on future (unobserved) outcome, and the dropout process is informative, or nonignorable. For example, in the schizophrenia trial, discontinuing on protocol might have reflected decisions based on physician assessment of patient condition only weakly related to BPRS. The Cox regression shows that BPRS is strongly predictive of discontinuation, and is consistent with the assumption that the missingness is at random, but does not preclude a non-ignorable mechanism.

The general case of non-ignorable missingness occurs when the probability that a response is missing is directly related to the missing outcome. For example, in trials of patients undergoing chemotherapy treatment for cancer, quality-of-life assessments may be required on a quarterly schedule. Most quality-of-life forms are self-report and may require substantial effort on the part of the patient. Patients who are experiencing poor quality-of-life are likely to be far less able to complete the self-report required for response. Obtaining valid estimates of population parameters in this setting is far more complicated, since we are in a situation of having to make assumptions about the distribution of the missing outcomes which cannot be fully tested by the data.

4. SOFTWARE FOR MAXIMUM LIKELIHOOD ANALYSIS OF GENERAL LINEAR MIXED MODEL

Several statistical software packages are available for the analysis of correlated data. These include BMDP-5V, SAS, and ML3, among other. HLM and S-plus are other programs available, but these are restricted to analysis of random effects models, and are not discussed here.

Most of the software packages offer a choice between maximum likelihood and restricted maximum likelihood estimation. The optimization algorithm may be chosen as Newton–Raphson, Fisher Scoring, or the EM algorithm. The user is required to specify an equation for mean response that is linear in the fixed effects, and to specify a covariance structure. The user may select a full parametrization of the covariance structure (unstructured) or choose from among structured covariances which are more parsimonious. The covariance structure is also determined by inclusion of random effects and specification of their covariance structure.

Output generated includes a history of the optimization iterations, estimates of fixed effects, covariance parameters, and their standard errors. Estimates of user-specified contrasts and their standard errors are also printed. Graphics facilities for these software packages are currently limited. Parameter estimates, fitted values and residuals produced from a model run may be saved as data sets, and supplied to programs suited for graphics.

SAS PROC MIXED was designed for the analysis of mixed models. It provides a very large choice of covariance structures. In addition to the unstructured, random effects and autoregressive, it can fit the Diggle model with $\boldsymbol{R}_i = \sigma^2 \boldsymbol{I} + \boldsymbol{S}_i$, where $\boldsymbol{S}_i$ can have one of several autoregressive structures. In addition, covariance can be specified to depend on a grouping variable. Separate analyses can be run on subgroups of the data through use of a single BY statement. PROC MIXED conveniently provides empirical Bayes estimates (random effects and predicted values). SAS constructs approximate F-tests for class variables by dividing the Wald test by the numerator degrees-of-freedom and approximating the denominator degrees-of-freedom. All components of the SAS output, such as parameter estimates, residuals and contrast matrices, can be saved as data sets, for purposes of graphics or manipulation by other SAS procedures. The FEV_1 data set (Section 2.2) was analysed using SAS.

BMDP-5V was designed for the analysis of unbalanced repeated measures data. It also provides a large variety of options for the covariance structure, and for estimation. The data set-up is different between SAS and BMDP. In SAS PROC MIXED, every observation within a subject constitutes a line, or a record. In BMDP5V, the line or record is the entire subject, with all its observations. This makes BMDP particularly useful and easy to manipulate when studies are planned to be complete and balanced. It also allows one to very easily see the patterns of missingness. However, for epidemiological studies, which are inherently unbalanced, setting the data up in BMDP may be awkward. The schizophrenia case study (Section 2.1) was analysed using BMDP. BMDP provides a Wald test to examine overall effects of class variables, such as clinic, in the schizophrenia study. This is similar to the F-test provided by SAS.

Several models were run both in BMDP and SAS. Results were identical up to fourth decimal place for the log-likelihood, regression coefficients estimates and their standard errors, and estimates of covariance parameters. Both SAS PROC MIXED and BMDP-5V, as parts of larger packages, have the flavour of their parent packages. In both programs, in order to run a different model on the same data set, one needs to enter the editor, change the model specifications and then rerun the program. This is somewhat short of fully interactive.

ML3, which stands for Software for Three Level Analysis, was developed for applications within the fields of education and human growth.[32] Nutall *et al.*[33] used ML3 for modelling achievement scores in 96 schools in London, where the observations were on individual students and the clusters were schools. Input and manipulation of data are very similar to MINITAB, making it relatively easy to use for MINITAB users. The program is highly interactive. The user sets the model, and with a one word command sees the specifications. The user can run the model, and view, save, or print out what they want to see, and in the same session, simply change one parameter, for example, add a fixed effect, and rerun the model. In that interactive environment, ML3 is somewhat more convenient to use than either BMDP or SAS, which require an actual separate batch run, when the user wishes to change one thing in the model.

5. CASE STUDIES ANALYSIS

5.1. Schizophrenia clinical trial

5.1.1. Design considerations

The clinical trial in schizophrenia (Section 2.1) is a typical example of a study with a pre-planned repeated measures design, in which incomplete data resulted from study dropouts, necessitating the use of the mixed model, instead of the regular repeated measures approach. Because there were only six different observation points, and 245 subjects, we used an unstructured covariance matrix in the analysis. In addition, we focused on the nature of the rate of change in BPRS across time. The fact that there were only six different observation points enabled us to construct models with indicator variables for all observation points, reflecting no specific trend across time, and compare any fitted function of time (such as a linear or quadratic time trend) with the model with indicator variables to examine goodness-of-fit.

In order to stabilize polynomial effects in the time trends, we chose a location transformation by subtracting three weeks from the week number, which makes the data for the linear and quadratic effects for subjects with complete data nearly orthogonal to each other. In a model with a linear or a linear and quadratic effect using the untransformed time scale, the estimate of the

intercept reflects an estimate of BPRS score at baseline. In a model using the suggested transformed time scale, the estimate of the intercept reflects the mean BPRS at three weeks. In a linear model, the linear coefficient reflecting the linear rate of change of BPRS with time remains the same, regardless of transformation. In the quadratic model, however, the quadratic coefficient remains the same after the transformation, but the linear coefficient changes as a results of the transformation.

5.1.2. Model selection and testing

Our first approach to the analysis was to use BPRS scores from all six observations as response variables y_{ij}, i indicating subject number and j indicating the observation number ($j = 1, \ldots, 6$). We compared models with linear and quadratic time trends (observation level covariates) to the model with indicator variables for each time point. We chose not to fit trends with time higher than quadratic because those are difficult to interpret in this study. A combination of linear and quadratic effects can be interpreted as a decrease in BPRS scores linearly with time, and the decrease is larger initially, and then levels off, hence the quadratic term. Plots of the data support this interpretation. We also added first-order interactions both between subject-level covariates (for example, treatment $\times$ centre) and between subject and observation level covariates (for example, treatment $\times$ week). The model with a quadratic effect for time, as well as a quadratic effect for the interaction between time and centre provided a substantially better log-likelihood than the model with indicators at each time point and without a time by centre interaction, although at a cost of estimating 21 more parameters. A direct log-likelihood comparison is not possible, since these models are not nested.

There were several unattractive features in this quadratic model. The first is the time–centre quadratic-interaction. While it is not surprising that the centres had different baseline BPRS means (reflecting either somewhat different patient populations at the centre or somewhat different interpretations of the BPRS scoring systems), it is unexpected to see a strongly significantly different change across time between the centres, given that all centres gave the same treatments according to the same protocol. Moreover, a close examination of the within-subject correlations from the estimated model shows stronger correlations between all observation points beyond baseline than correlations with the baseline values, and a tendency of decreasing correlations with increasing time differences. Finally, a quadratic trend without the time–centre quadratic interaction did not provide a fit that was comparable to the model with no assumptions on the time trend. Observing also that 36 per cent of the baseline variability in BPRS is explained by centre alone, we chose to fit subsequent models using the baseline BPRS score as a subject level covariate, rather than a response.

Table III gives the BMDP results of models fitted to the data presented as five response values (weeks 1, 2, 3, 4, 6), with the baseline BPRS score as a subject level covariate. This reduced the number of subjects included in the model from 245 to 233, since 12 subjects had no observations beyond baseline, and thus no responses in these models. Adding the baseline BPRS score as a subject level covariate reduced substantially the variability to be fit by other terms in the model, suggesting that the approach to analysing the data as at most five responses rather than six responses was preferable. In models 1 and 2, time is entered as linear and quadratic observation level terms; model 2 also contains a subject level main effect of treatment. A comparison between models 1 and 2 shows that the addition of a treatment effect was not significant. The addition of a different linear trend in time among the treatments (observation level) was also non-significant

Table III. Models for schizophrenia study. Baseline BPRS and centre are included as covariates in all models

Model numbers	Variables in model	ML log-likelihood	Number of reg parms	Baseline BPRS coefficient	Week coefficient	Week^2 coefficient	Test effect	p-value
1	Week, week^2	−3266·33	16	0·65	−1·55	0·38		
2	Trt., week, week^2	−3263·85	19	0·64	−1·56	0·38	Trt	0·21
3	Trt., week, week^2, week × trt.	−3261·90	22	0·65	−1·50	0·38	Week × trt.	0·26
4	Trt., week indicators	−3256·29	21	0·63	–	–		
5	Trt., week, week^2, week × centre	−3257·07	31	0·64	−1·72	0·38	Week × centre	0·25
6	Trt., week, week^2, status	−3164·03	20	0·58	−2·35	0·48	status	⩽0·0001

For all models, both the linear time trend coefficient (week) and the quadratic, as well as the overall centre effect, were significant at the 0·001 level by the Wald test. Overall treatment effect was not significant in any model

(week × treatment, model 3, compared to model 2). Since the purpose of the study was to show that the experimental treatment was as efficacious as the active control, this result is not surprising. The non-significance of the treatment effect does not, of course, imply equivalence. The four treatment effects need to be tested for equivalence, which is not the focus of this tutorial, and is not considered here. Model 4, which reflects no assumption on the time trend, has all the time indicators significant. A comparison with model 2, gives a likelihood ratio test of 15·12 on two degrees of freedom, implying that a quadratic in time model does not adequately describe the time trend.

5.1.3. Estimation and effect of missingness

Model 5 shows the addition of a centre by week interaction (observation level covariate). The overall log-likelihood is comparable to the unrestricted model 4, although with ten more estimated parameters. The overall centre by linear week interaction was not a significant term by the Wald test, nor was it a significant contribution beyond model 2 by a likelihood ratio test. From Table III we can see that results are very similar for models 1, 2 and 3 in terms of both coefficient values and log-likelihoods. The coefficient of baseline BPRS indicates a substantial drop from baseline to week 1 in BPRS of approximately one-third, which on average means a reduction of 12 points. Correcting for the linear and quadratic effect on that first week, adds approximately five points, for a net effect of an estimated mean seven point drop during the first week of treatment. Subsequently, between weeks 1 and 6 there is a continued, slower drop which gives a total estimated additional reduction of six points by study end. Comparing model 2 to model 4 shows that the quadratic model does not provide a completely adequate description of the time trend. Table IV gives the model fitted variance-covariance matrix for model 2. Standard deviations are between 9.5 and 14 BPRS points.

Model 6 was examined in an effort to study the effect of dropout. The study protocol called for discontinuation of the study if there was a perceived lack of therapeutic effect of the study drug. The decision to discontinue a patient on those grounds was done without knowledge of treatment

Table IV. Covariance matrix for schizophrenia study. Model with baseline BPRS, treatment, centre, week, and week^2 (Correlations above the diagonal; covariances below)

	Week 1	Week 2	Week 3	Week 4	Week 6
Week 1	92·20	0·63	0·52	0·34	0·35
Week 2	66·04	120·30	0·76	0·64	0·60
Week 3	64·07	106·26	161·80	0·77	0·72
Week 4	43·67	91·75	127·41	169·22	0·85
Week 5	46·79	93·07	128·92	155·08	196·37

Table V. Covariance matrix for schizophrenia study. Model with baseline BPRS, treatment, centre, week, week^2, and patient status (correlations above the diagonal; covariances below)

	Week 1	Week 2	Week 3	Week 4	Week 6
Week 1	85·47	0·61	0·50	0·27	0·32
Week 2	55·54	95·50	0·75	0·59	0·60
Week 3	50·92	80·34	121·02	0·73	0·67
Week 4	26·14	60·01	83·07	108·13	0·77
Week 5	31·30	60·96	76·92	83·39	107·34

assignment and not based on BPRS scores. However, the final observation in the study for those cases was performed after the decision to discontinue. Thus, we added an observation-level indicator covariate of patient status. This variable was defined as 0 while the patient was on study and as 1 at the last observation if the patient was discontinued due to lack of therapeutic effect, so that the last observation reflected a status of being off-study. Model 6 provided the highest likelihood with the fewest parameters of any of the previous models.

Table V gives the variance-covariance matrix of this model, and all variances are reduced as compared to model 2 (SDs are 9–10.5 BPRS points in model 6), and this reduction becomes more substantial in later time points. Both Tables IV and V show an increase in variance with time; this is a common phenomena in longitudinal studies. One reason is that entry to the study is restricted to those patients with BPRS of 20 or greater, thus limiting the total range, as compared to later time points. In model 6, both the baseline BPRS coefficient and the linear trend coefficient showed an even greater mean reduction in BPRS than other models in Table III, which is adjusted by an increase in BPRS in those patients who were discontinued due to perceived lack of therapeutic effect. The estimated treatment effects for this model, which includes status, can be interpreted as effects of treatment, assuming that patients are maintained on therapy. In that sense it might be viewed as an explanatory analysis.[34]

To illustrate the relationship between the notation of the models in Section 3 and the models presented here, we define model 6 in the terminology of the general linear mixed model as presented in equation (4). Recall that $\boldsymbol{X}_i$ is defined by a combination of subject level and observation level covariates. The subject level covariates are: (i) a column of 1's for the intercept; (ii) the subject's baseline BPRS value; (iii) three columns of treatment indicators; (iv) 12 columns of centre indicators. The observation level covariates are: (a) a linear week effect (range -2 to $+3$; (b) a quadratic week effect (range 0 to 9); (c) a column for the status variable. The vector $\boldsymbol{y}_i$ has

Table VI. Regression coefficients and standard errors for three covariance models

Variable name	Unstructures covariance		Random effects		Autoregressive	
	Estimate	SE	Estimate	SE	Estimate	SE
Intercept	2·13	2·47	1·90	2·51	3·73	2·54
Baseline BPRS	0·64	0·067	0·65	0·067	0·58	0·069
Low dose treatment	1·93	1·04	2·04	1·06	2·69	1·10
Medium dose treatment	− 1·86	1·02	− 1·64	1·03	− 1·75	1·06
High dose treatment	0·10	1·03	− 0·13	1·05	− 0·36	1·07
Week	− 1·56	0·25	− 1·52	0·25	− 1·91	0·23
Week^2	0·38	0·084	0·37	0·085	0·36	0·090

between one and five elements, depending on the number of observations beyond baseline. We assume, as defined earlier, that $\boldsymbol{Z}_i = 0$ and that $\text{var}(\boldsymbol{y}_i)$ is arbitrary. The BMDP code associated with this model is given in Appendix I.

If we wish to use a random effects approach. we would define $\boldsymbol{Z}_i$ to have three columns including an intercept, slope and quadratic effect. The observation level covariate status, being binary, would not have an associated random effect. Examining the within-subjects model-based correlations matrices in both models 2 and 6 showed similar correlations of 0·5–0·85 between BPRS scores at different observation points, except for between week 1 and weeks 4 and 6, which were approximately 0·3. In addition, there was uniformly a decreasing correlation between observation points as the time gap was larger. This suggested exploring a random effects error structure with random intercept, linear trend and quadratic trend, that is, all observation-level covariates as recommended in Section 3, as well as exploring an autoregressive structure (AR1).

Table VI gives the estimated coefficients and SEs under all three covariance structures for model 2. Investigating the coefficients shows that all three models estimate the coefficient of the baseline value to be approximately two-thirds. The unstructured covariance matrix and the random effects model had similar linear and quadratic time trends coefficients, indicating a reduction of six points from week 1 to week 6 beyond the initial reduction in the first week. The AR1 model predicted a reduction by eight points. Part of the reason may be due to the model-estimated within-subject covariance matrices. In both the unstructured and the random effects models, the variances and covariances are not restricted, and, as Table IV shows for the unstructured covariance model, the variances increase with time. The AR1 model constrains the structure to equal variances at all observation points, and covariances depend only on the time gap. Thus, this model assumes more precision than the other models for the late time points in this study, and thus gives them relatively more weight. Because a substantial part of the dropout is due to a perceived lack of therapeutic effect, at least theoretically, the mean of week 6 BPRS values may be biased down. Thus, an estimate which puts slightly more weight on week 6 will show overall a slightly larger reduction in the BPRS score. It should be stressed that this difference between the AR1 model estimates and the two other models is primarily a result of different estimated within-subject variances and not different estimated within-subject correlations.

Figure 3 illustrates the fitted model 6 and the observed data for two subjects in the medium dose treatment group.

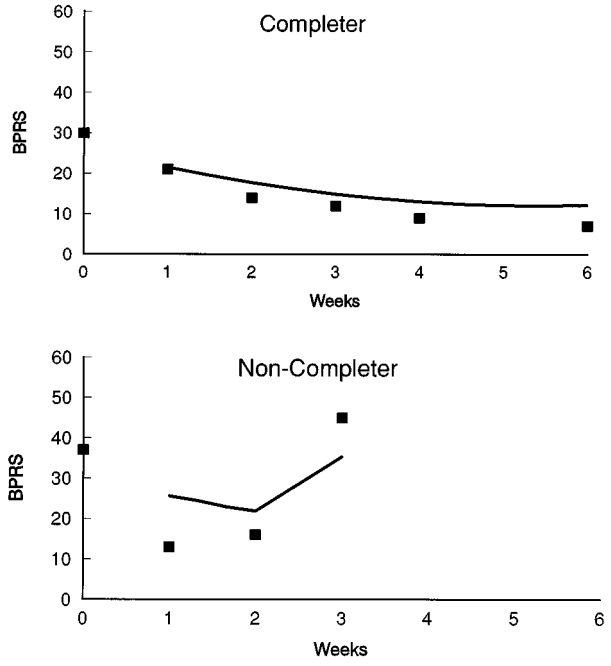

Figure 3. Fitted curves for two patients in the schizophrenia trial in the largest centre and receiving the medium dose. Model 6 is assumed

5.2. Pulmonary function development in children

5.2.1. Model selection and checking

In contrast to the schizophrenia trial, the data from this study (Section 2.2) are highly unbalanced. The study called for annual measurements, but girls were different ages at entry and there is considerable missing data. Figure 4 gives the distribution of subjects' entry age and number of measurements. Age at entry ranges from ages 6 to 14 with 81.3 per cent of girls being between ages 7 and 9 at entry. Girls entering at a late age had few measurements taken, while younger entrants had as many as 12 measurements. Dropout does occur, likely as a result of families moving between school districts.

We follow the analysis of Hopper *et al.*[17] where the natural logarithm of FEV_1 was found to depend on both age and height (observation level covariates). They found that it was best to represent the covariate age as a linear spline, with slope depending on whether or not the girl was of age less than or greater than 12.5. In their data, prior FEV_1 measurement was highly predictive of current FEV_1, so that in addition to measurement error, the covariance structure consisted of a first-order autoregressive component. They also considered a Diggle model for covariance consisting of measurement error, first-order autoregressive component, and random intercept as described in Section 3.4. The parsimonious model choice retained height and age as predictors. The first part of the spline for age was forced to have slope zero, so that age was not predictive of FEV_1 prior to age 12·5, after adjusting for height. The covariance structure employed was the

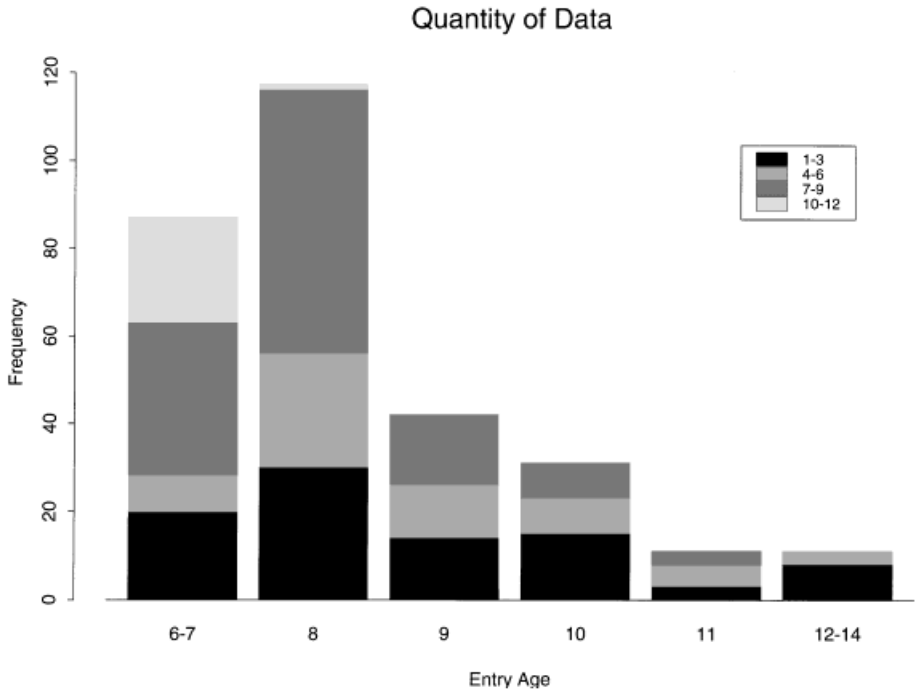

Figure 4. The distribution of age at entry for the FEV_1 data, showing also the number of measurements obtained on each subject

measurement error plus first order autoregressive. In this data set there were no subject level covariates except for intercept.

Previous analyses of our data suggested the use of log height rather than height as a predictor of FEV_1; we also had no reason to support the spline with knot at age 12·5 years. Thus our first step was to fit models with both versions of height (linear and logarithmic) and varying the knot point for the spline from 11·5 to 17·0 in six month increments. We did not assume that the slope for age prior to the knot was zero. We used the Diggle covariance structure, and then used likelihood values and residual plots to choose the best model. Residual plots revealed an observation that was clearly outlying. This observation was of a subject who only had one measurement available. We dropped this observation, so that all subsequent analysis is based on the data of 299 girls. The spline with knot at age 16 provided the highest likelihood, and the likelihood ratio test comparing this model to the simpler model with linear age, was highly significant ($p < 0{\cdot}0001$). Table VII gives the likelihood values for these models.

Plots of residuals, ordered by age, are given in Figure 5. The three graphs on the left side give residuals for models obtained with knots at ages 15, 16 and 17, respectively. The three graphs on the right are similar, except height has been transformed by taking the natural logarithm. It is difficult to discern a pattern in the residuals. A lowess curve is run through the data points of each plot. The best fitting model should have a lowess curve that is closest to a line with slope zero. Inspecting the graphs, one sees that the lowess curves are not very different. The model with knot

Table VII. Likelihood values for FEV_1 data

Age at knot	(2 log-likelihood) Covariate used	
	Height	Log height
11·5	4743·68	4734·59
12·0	4743·51	4730·25
12·5	4741·64	4720·32
13·0	4740·24	4711·09
13·5	4742·97	4708·37
14·0	4754·66	4719·16
14·5	4772·73	4739·14
15·0	4792·16	4760·68
15·5	4803·19	4773·08
16·0	4805·90	4775·90
16·5	4795·08	4764·46
17·0	4780·84	4749·42

at age 16 seems to have the flattest curve for later ages, giving further evidence that age 16 is the best choice for the knot. The plots for the models using logarithm of height are not much different from plots for models using height. We keep height (untransformed), and age as a spline with knot at age 16, in subsequent analyses.

Table VIII shows the SAS results of fitting several different models to the FEV_1 data; all have the same mean predictors (height and spline for age with knot at age 16), but the models have different covariance structures. Appendix II gives the SAS code for using the Diggle model for the covariance, and the results of this model are given in column 4 of Table VIII. The coefficient of AGE1 gives the slope prior to age 16 and the coefficient of AGE2 gives the deviation of the post-16 slope from the pre-16 slope. Age and height are significant predictors of FEV_1. Log FEV_1 is predicted to increase by 0·034 per year prior to age, 16, and 0·014 per centimetre of growth. Log FEV_1 changes with age beyond age 16 are predicted to be $-0{\cdot}006$ per year. A model was fit where the slope beyond age 16 was forced to be zero (not shown). The one degree-of-freedom likelihood ratio test was not quite significant ($P = 0{\cdot}083$). This finding differs from the Hopper result, where log FEV_1 was found to have zero slope with respect to age less than 12·5, and positive slope beyond age 12·5.

The remaining columns of Table VIII show the results of fitting the same model for the mean, but making different assumptions about the variance. Because of the highly unbalanced nature of the data, and the large number of potential observations of each subject, it is not possible to fit an unstructured covariance matrix. We fit the models used by Hopper, and also one which treats the two slopes on age as random effects. For this last model, the $\boldsymbol{Z}_i$ matrix contains an intercept, AGE1 and AGE2. Apart from the OLS fit, the estimated regression parameters are all very similar, as are their estimated standard errors. The OLS fit gives quite different estimates and standard errors, although in this case the inferences are similar. The Diggle model yields the highest likelihood, again consistent with Hopper's result.

Figure 6 gives empirical Bayes estimates for selected girls (broken curve), contrasted with ordinary least squares estimate based on the subject's data (solid curve). Notice that these curves

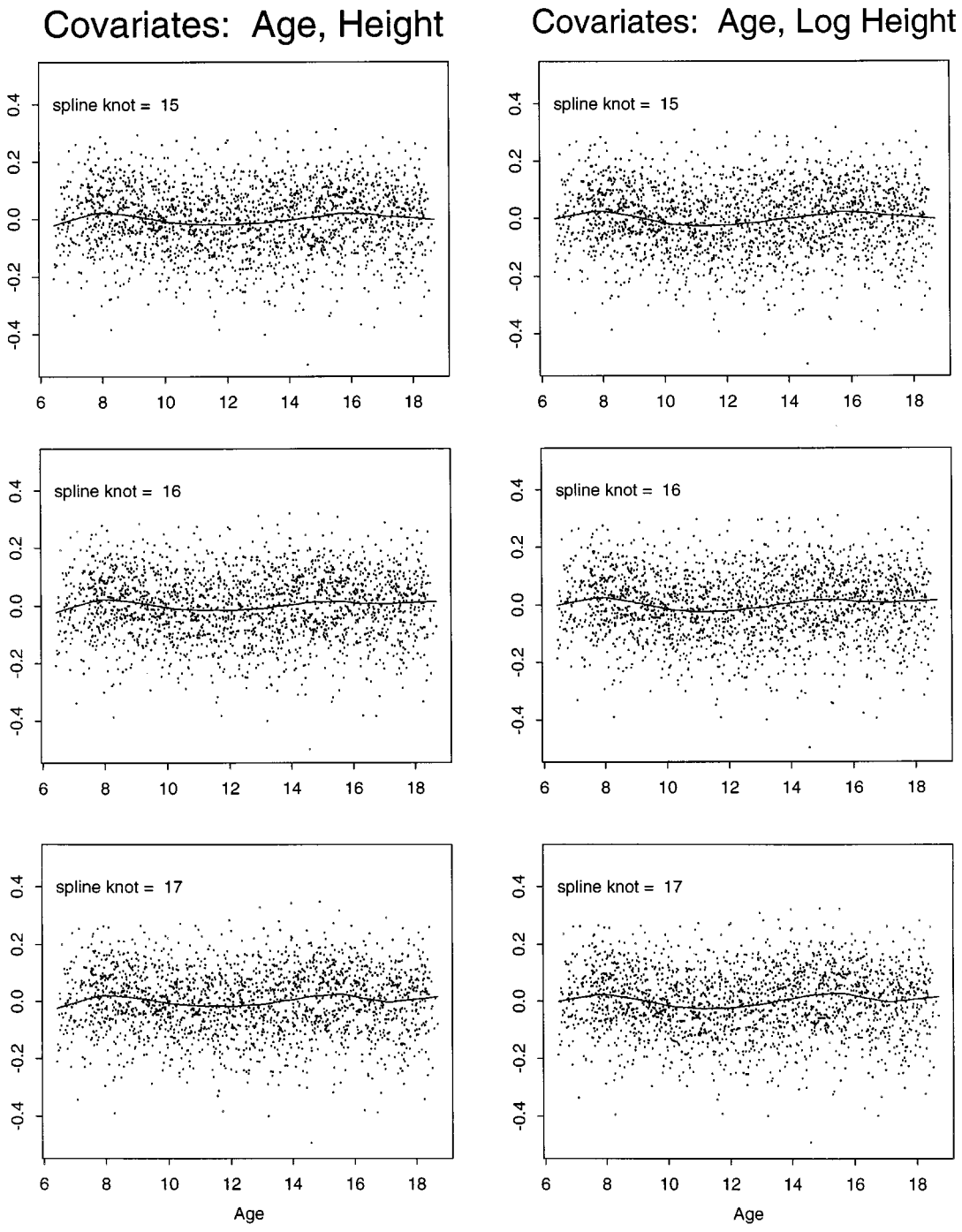

Figure 5. Residual plots for the FEV_1 data, contrasting the use of height versus log height, and using splines in age with knots at 15, 16 and 17 years of age. The lowess curve is plotted

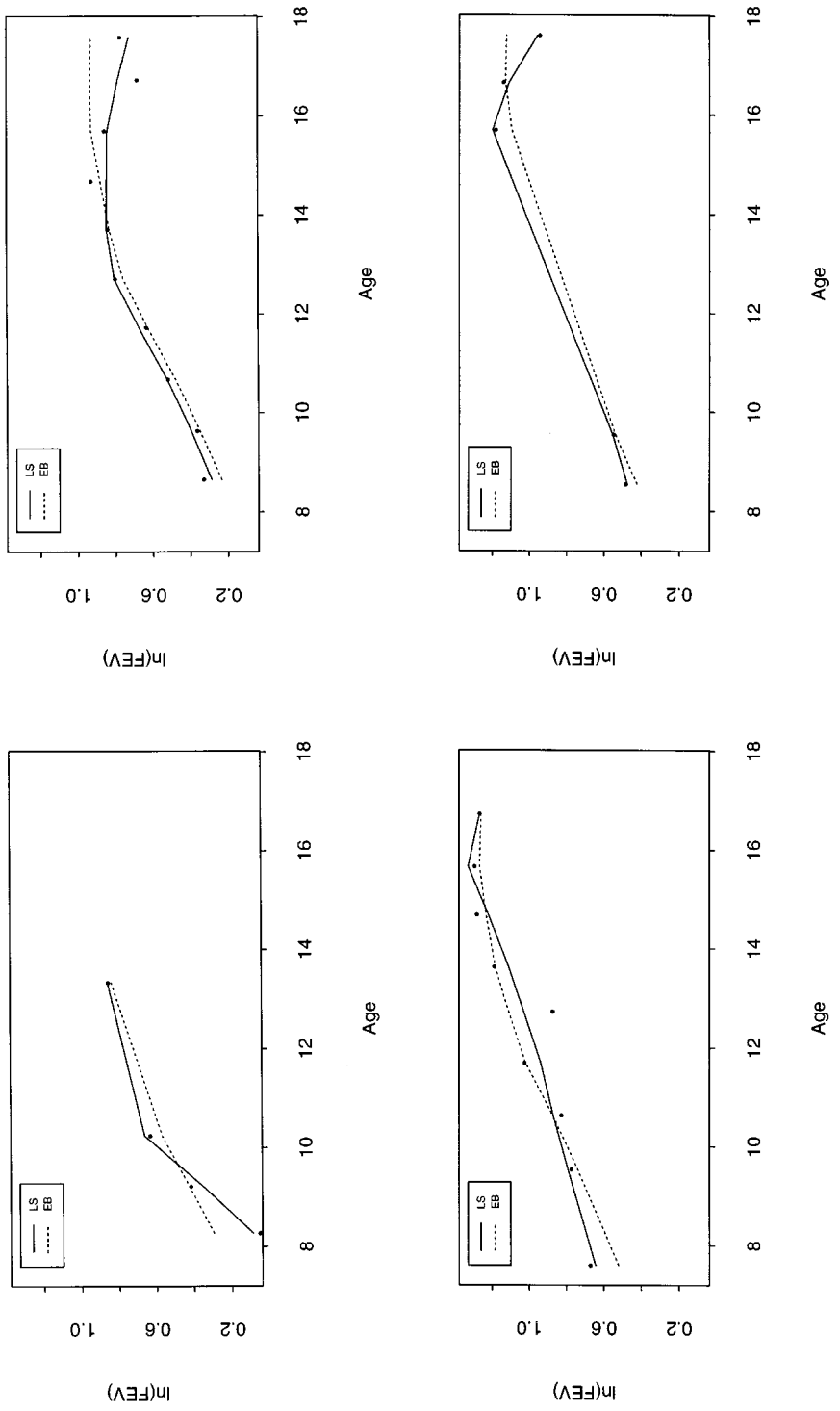

Figure 6. Empirical Bayes' and least squares curves for selected individuals. Actual data are indicated by dots

Table VIII. Estimated parameters for models fit to FEV_1 data using different co-variance structure (standard errors in parentheses)

	OLS	Compound symmetry	AR (1) plus error	Diggle	Random slopes
INT	− 1·778 (0·038)	− 1·663 (0·037)	− 1·679 (0·042)	− 1·679 (0·044)	− 1·708 (0·038)
AGE1	0·027 (0·002)	0·034 (0·002)	0·034 (0·002)	0·034 (0·002)	0·032 (0·002)
AGE2	− 0·026 (0·006)	− 0·037 (0·004)	− 0·039 (0·005)	− 0·039 (0·005)	− 0·034 (0·004)
HT (m)	1·506 (0·041)	1·380 (0·040)	1·390 (0·046)	1·389 (0·047)	1·426 (0·041)
σ_e^2	0·0129 (0·0004)	0·0039 (0·0001)	0·0023 (0·0002)	0·0017 (0·0003)	0·0033 (0·0001)
d_{00}		0·0092 (0·0008)		0·0080 (0·0009)	0·014 (0·002)
σ_r^2			0·0106 (0·0008)	0·0033 (0·0004)	
ρ			0·953 (0·007)	0·665 (0·084)	
d_{11}					0·00007 (0·00014)
d_{22}					0·00007 (0·00018)
−2log L	− 3010·58	− 4644·38	− 4791·71	− 4805·90	− 4742·03

are not simple splines in age because of the dependence of log FEV_1 on height as well. The empirical Bayes' estimates are a weighted average of population and subject parameters. For subjects with few or highly variable FEV_1 measurements, the weighing of the estimates towards the population mean is relatively strong, so that there is visible separation between the empirical Bayes' and least squares curves. The separation is also apparent in subjects who appear to have outlying data. The empirical Bayes' plots are more appealing because they are not so sensitive to outlying values at the ends of the range, and because they do not turn down, as some of the individual subject's OLS curves do.

APPENDIX I: BMDP CODE FOR MIXED MODELS

```
/INPUT FILE = 'c: \bmdp\schiz.por'.
       CODE = SCHIZ.
       PORT.
/VARIABLE NAMES = TRT, CENTER, PATID, WEEKO, BPRSO, WE11, BPRS1,
       D1, WE21, BPRS2, D2, WE31, BPRS3, D3, WE41, BPRS4, D4, WE61,
       BPRS6, D6.
/TRANSFORM
       WE12 = WE11*WE11.
       WE22 = WE21*WE21.
       WE32 = WE31*WE31.
       WE42 = WE41*WE41.
       WE62 = WE61*WE61.
```

The data are read in as a record per patient. Each record contains the patient's treatment, centre, all six week numbers, associated BPRS scores and status, which is 0, except at the point of discontinuation. If the discontinuation is due to lack of effect, the relevant 'D' variable is 1. If data

are missing, for example, discontinuance after week 4, or week 2 skipped, the fields contain '·' to signify missing data.

```
/GROUP VARIABLE = TRT, CENTER.
        CODES (TRT) = 1, 2, 3, 4.
        NAMES (TRT) = 'Lo', 'Med', 'Hi', 'Cont'.
```

The categorial variables of treatment and centre are defined, so that BMDP can create the appropriate indicator variables in the model.

```
/DESIGN   DPNAME = BPRS.
          DPVAR = BPRS1, BPRS2, BPRS3, BPRS4, BPRS6.
          REPEAT = WEEK.
          LEVEL = 5.
          CVNAME = BPR0, WE1, WE2, STATUS.
          BPR0 = BPRS0, BPRS0, BPRS0, BPRS0, BPRS0.
          WE1 = WE11, WE21, WE31, WE41, WE61.
          WE2 = WE12, WE22, WE32, WE42, WE62.
          STATUS = D1, D2, D3, D4, D6.
```

The definition of dependent and independent variables is made, including the complete structure of the design matrix. BPRS is identified as the dependent variable. This paragraph identifies the fact that there is a potential for five repeats within each record, and gives a name for each variable, dependent or covariate, that can have five repeats, as well as the associated data variables which form these five repeats.

```
/STRUCTURE  TYPE = UNSTRUC.
```

This command identifies the covariance structure as unstructured, thus in this case a five by five matrix. Other options would have been: 'TYPE = CS' (compound symmetry), 'TYPE = AR(1)' (first-order autoregressive), 'TYPE = BANDED' (banded, or general autoregressive).

Structures which would require additional input would be: 'TYPE = RANDOM' (random effects), 'TYPE = ARANDOM (random effects with AR(1) residual correlation structure), 'TYPE = LINEAR' (general linear structure), 'TYPE = USER' (defined by user in FORTRAN subroutine).

If we had chosen TYPE = RANDOM and wanted to specify a random intercept, slope and quadratic, we would have added: ZVAR = INT, INT, INT, INT, INT, WE11, WE21, WE31, WE41, WE61, WE12, WE22, WE32, WE42, WE62.

```
/MODEL BPRS = 'BPR0 + TRT + CENTER + WE1 + WE2 + STATUS'.
```

Defines the linear regression model within the subject. BMDP has the option to add a /COMPUTE statement to specify number of iterations, convergence criterion, halving , and the algorithm to be used for convergence. The default is Newton–Raphson, but the EM can be

specified. There are some limitations about allowable combinations between the /STRUCTURE statement and the /COMPUTE statement.

/PRINT COVR.

There are a variety of print options, including the possibility to obtain predicted values and residuals.

/END.

APPENDIX II: SAS CODE FOR MIXED MODELS

DATA STEP

```
data fev;  set fev;  rage = round (age);  age1 = age;
age2 = max(0, age1–16);
```

Modelling logarithm of FEV_1, using covariates height and age (linear spline with knot at age 16). Diggle covariance structure.

In the data step above, the variables AGE1, and AGE2 are created as the pieces of a spline, with knot at age 16. The variable RAGE is a class (categorical) variable. This variables will be an input to the REPEATED statement of PROC MIXED. The REPEATED statement requires that one provide a class variable indicating the time point of an individual's measurement. We will use the REPEATED statement to implement the autoregressive component of the Diggle covariance structure.

PROC MIXED

```
proc mixed data = fev method = ml;  class id rage;
model lnfev = age1 age2 ht  /s p;  random int/type = un subject = id
g; repeated/local r subject = id type = sp (pow) (age1); make
'SolutionF' out = fix1; make 'Predicted' out = pred1; title 'Spline
model (Diggle Covariance)'; *next line gives alternate REPEATED
statement;
*repeated rage/local r subject = id type = AR(1);
```

METHOD = ML indicates that maximum likelihood estimation is used.

The CLASS statement declares classification variables so that SAS can create a set of dummy variables that span the categories. To fit mixed models, SAS requires a class variable to specify the subjects (see below). If repeated statement is used, an additional class variable may be necessary to define the distinct occasions of measurement (see discussion under repeated).

MODEL LNFEV = AGE1 AGE2 HT is the specification of the equation for the mean. The S option indicates that we would like estimates of the fixed effects and the P option indicates that we would like predicted values, as part of the SAS output. SAS computes predicted values based on the estimates of the fixed and random effects, so that when random effects are specified, the predicted values are empirical Bayes' estimates.

RANDOM INT specifies the random intercept component. We have only specified one random effect. If we wished to run a random slope we would write RANDOM INT AGE1. TYPE = UN specifies that the covariance matrix for the random effects in unstructured. Since we only have one random term, it does not matter which structure is specified. SUBJECT = ID provides SAS with the class variable that identifies subject (for grouping of observations within subjects) in both the RANDOM and REPEATED statements. The G option indicates that we would like an estimate of the random effects covariance printed.

REPEATED statement specifies that the covariance is modelled within subject (SUBJECT = ID); TYPE = SP(POW) (AGE1) specifies a particular covariance structure which is a continuous time generalization of the discrete time AR(1) model, it allows for a continuous representation of age. The alternative version, using REPEATED RAGE, fits a discrete time AR(1) model; for purposes of estimating the covariance parameters with this specification, the ages of measurement are first rounded to the nearest year. The LOCAL option indicates that we would like to add an additional diagonal component to the AR(1) structure. The R option requests that SAS print an estimate of the covariance block for an individual.

The first MAKE statement tells SAS to save the estimates of the fixed effects in the SAS data set named FIX1. The second MAKE statement indicates that estimates of predicted values be saved in the data set PRED1. You must specify the appropriate option (S for fixed effects, P for predicted values) in the MODEL statement, in order to have SAS save results to a data set. Most components of the computations and output of SAS can be saved to data sets. This is useful when one wishes to perform further computations or graphics.

Specification of other covariance structures is straightforward. For example, dropping the REPEATED statement above gives a specification for a random intercepts (compound symmetry) model. Dropping the RANDOM statement and the LOCAL option in the REPEATED statement, would give a first-order autoregressive structure. Dropping the RANDOM statement and specifying TYPE = UN in the REPEATED statement gives the most general unstructured covariance model.

REFERENCES

1. Longford, N. T. *Random Coefficient Models*, Clarendon Press, Oxford, 1993.
2. Diggle, P. J., Liang, K.-Y. and Zeger, S. L. *The Analysis of Longitudinal Data*, Clarendon Press, Oxford, 1993.
3. Sherrill, D. L., Holberg, C. J., Enright, P. L., Lebowitz, M. D. and Burrows, B. 'Longitudinal analysis of the effects of smoking onset and cessation on pulmonary function', *American Journal of Respiratory and Critical Care Medicine*, **149**, 591–597 (1994).
4. Gennaro, S., Fehder, W. P., Cnaan, A., York, R., Campbell, D. E., Gallaghar, P. and Douglas, S. D. 'Stress and immune response in mothers of term and preterm VLBW infants', (submitted).
5. Geva, D., Goldschmidt, L., Stoffer, D. and Day, N. L. 'A longitudinal analysis of effect of prenatal alcohol exposure on growth', *Alcoholism: Clinical and Experimental Research*, **17**, 1124–1129 (1993).
6. Ross, R. A., Lee, M.-L. T., Delaney, M. L. and Onderdonk, A. B. 'Mixed-effects models for predicting microbial interactions in the vaginal ecosystem's, *Journal of Clinical Microbiology*, **32**, 871–875 (1994).
7. Donaldson, G. W. and Monipour, C. M. 'Strengthened estimates of individual pain trends in children following bone marrow transplantation', *Pain*, **48**, 147–155 (1992).
8. McLean, R. A., Sanders, W. L. and Stroup, W. W. 'A unified approach to mixed linear models', *American Statistician*, **45**, 54–63 (1991).
9. Laird, N. M. and Ware, J. H. 'Random effects models for longitudinal data: an overview of recent results', *Biometrics*, **38**, 963–974 (1982).
10. Goldstein, H. *Multi level Models in Educational and Social Research*, Griffin, London, 1987.

11. Bryk, A. S. and Raudenbush, S. W. *Hierarchical Linear Models*: *Applications and Data Analysis Methods*, Sage Publications, Newberry Park, California, 1992.
12. Louis, T. A. General methods for analyzing repeated measures, *Statistics in Medicine*, **7**, 29–45 (1988).
13. Crowder, M. J. and Hand, D. J. *Analysis of Repeated Measures*, Chapman & Hall, London, 1990.
14. Lapierre, Y. D., Nai, N. V., Chauinard, G., Awad, A. G., Saxena, B., James, B., McClure, D. J., Bakish, D., Max, P., Manchanda, R., Beaudry, P., Bloom, D., Rotstein, E., Ancill, R., Sandor, P., Sladen-Dew, N., Durand, C., Chandrasena, R., Horn, E., Elliot, D., Das, M., Ravindra, A. and Matsos, G. 'A controlled dose-ranging study of remoxipride and haloperidol in schizophrenia—A Canadian multicentre trial', *Acta Psychiatric Scandinavica*, **82**, 72–76 (1990).
15. Overall, J. E. and Gorham, D. R. 'The Brief Psychiatric Rating Scale (BPRS): recent developments in ascertainment and scaling', *Psychopharmacology Bulletin*, **22**, 97–99 (1988).
16. Wang, X., Dockery, D. W., Wypij, D., Fay, M. E. and Ferris, B. G., Jr. 'Pulmonary function between 6 and 18 years of age', *Pediatric Pulmonology*, **15**, 75–88 (1993).
17. Hopper, J. L., Hibbert, M. E., Macaskill, G. T., Phelan, P. D. and Landau, L. I. 'Longitudinal analysis of lung function growth in healthy children and adolescents', *Journal of Applied Physiology*, **7**, 770–777 (1991).
18. Wechsler, H., Dowdall, G., Davenport, A., Moeykens, B. and Castillo, S. 'Correlates of college student binge drinking', *American Journal of Public Health*, **85**, 921–926 (1995).
19. Laird, N. M., Donnelly, C. and Ware, J. H. 'Longitudinal studies with continuous responses', *Statistical Methods in Medical Research*, **1**, 3–25 (1992).
20. Cox, D. R. *Planning of Experiments*, Wiley, New York, 1958.
21. Jones, B. and Kenward, M. G. *Design and Analysis of Crossover Trials*, Chapman Hall, London, 1989.
22. Zeger, S. L. 'Commentary', *Statistics in Medicine*, **7**, 161–168 (1988).
23. Jennrich, R. I. and Schluchter, M. D. 'Unbalanced repeated-measures models with structured covariance matrices', *Biometrics*, **42**, 805–820 (1986).
24. Liang, K. Y. and Zeger, S. 'Longitudinal data analysis using generalized linear models', *Biometrika*, **73**, 13–22 (1986).
25. Schluchter, M. D. and Elashoff, J. D. 'Small sample adjustments to tests with unbalanced repeated measures assuming general covariance structures', *Journal of Statistical Computation and Simulation*, **37,** 69–87 (1990).
26. Tsiatis, A., De Gruttola, V. and Wulfsohn, M. 'Modeling the relationship of survival to longitudinal data measured with error; applications to patients in AIDS', *Journal of the American Statistical Association*, **90**, 27–37 (1995).
27. Harville, D. A. 'Confidence intervals and sets for linear combinations of fixed and random effects', *Biometrics*, **32**, 403–407 (1976).
28. Lange, N. and Laird, N. M. 'Random effects and growth curve modeling for balanced and complete longitudinal data', *Journal of the American Statistical Association*, **84**, 241–247 (1989).
29. Diggle, P. J. 'An approach to the analysis of repeated measures', *Biometrics*, **44**, 959–971 (1988).
30. Waternaux, C., Laird, N. M. and Ware, J. H. 'Methods for analysis of longitudinal data: Blood lead concentrations and cognitive development', *Journal of the American Statistical Association*, **84**, 33–41 (1989).
31. Little, J. A. and Rubin, D. B. *Statistical Analysis with Missing Data*, Wiley, New York, 1987.
32. Prosser, R., Rasbash, J. and Goldstein, H. *ML3 Software for Three-level Analysis*, *User's Guide for V.2* Institute of Education, University of London, 1991.
33. Nutall D., Goldstein, H., Prosser, R. and Rasbash, J. 'Differential school effectiveness', *International Journal of Educational Research*, **13**, 769–776 (1989).
34. Heyting, A., Tolboom, J. T. B. M., and Essers, J. G. A. 'Statistical handling of drop-outs in longitudinal clinical trials', *Statistics in Medicine*, **11**, 2043–2061 (1992).

